Towards a sustainable social media archiving strategy for Belgium

Michel, Alejandra; Pranger, Jessica; Geeraert, Friedel; Lieber, Sven; Mechant, Peter; Vlassenroot, Evelyne; Chambers, Sally ; Birkholz, Julie; Messens, Fien

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link to publication

Citation for pulished version (HARVARD):
Michel, A, Pranger, J, Geeraert, F, Lieber, S, Mechant, P, Vlassenroot, E, Chambers, S, Birkholz, J & Messens, F 2021, *Towards a sustainable social media archiving strategy for Belgium: BESOCIAL : WP1 report : an international review of Social Media Archiving initiatives*. s.n., s.l.

Towards a sustainable social media archiving strategy for Belgium

WP1 Report

An international review of Social Media Archiving initiatives

(M1-M6: December 2020)

Editors	Alejandra Michel, Jessica Pranger, Friedel Geeraert, Sven Lieber, Peter Mechant, Eveline Vlassenroot, Sally Chambers, Julie Birkholz and Fien Messens
Responsible partners	Cental (UCLouvain) CRIDS (UNamur) KBR UGhent (MICT, GhentCDH, IDLab)
Version	1.0
How to cite this?	*S. Chambers, J. Birkholz, F. Geeraert, J. Pranger, F. Messens, S. Lieber, P. Mechant, A. Michel & E. Vlassenroot, BESOCIAL: final report Work Package 1 an international review of social media archiving initiatives, April 2021.*

Table of Contents

Executive Summary

This report details the tasks completed during Work Package 1 of the BESOCIAL project; a BELSPO funded, KBR coordinated, BRAIN research project with the following partners: CENTAL from UCLouvain; CRIDS from UNamur, and the research entities of MICT, GhentCDH, and IDLab from Ghent University.

The aim of this work package was to review existing social media archiving projects and corpora in Belgium and abroad. In order to keep an overview, the Belgian part of this work package has been dealt with in a separate document. The following report outlines the existing social media archiving at an international level.

Using a threefold methodological approach, a review of the state of art of SMA (Social Media Archiving) in the context of web archiving institutions was provided, both by updating a literature review of previous work, as well as a survey and in-depth interviews to understand the practicalities and legal aspects of this type of archiving for institutions in practice. This analysis addressed following perspectives: selection, legal, technical, access, and preservation.

The scope of social media archiving initiatives vary in size, as well as the ways to preserve and provide access to them. Because it is nearly unfeasible to archive all social media, almost all initiatives resort to selective crawls using mostly Twitter as a platform. A selection is often approached organically, one should therefore document this on a transparent level in order to generate the most representative data. One step further, in the legal field, extensive consideration must also be given on how to provide and ensure access to archives and under which copyright conditions. On a technical level, APIs, in most of the initiatives studied, caused limitations in terms of information collection and it's reuse. Next to this, in the preservation section, format migration is still a nascent field.

As documented in this report, social media archiving, similar to web archiving, is occurring in a number of institutions worldwide, in an effort to document and archive records of online communication. Despite the challenges (and the still existing heterogeneity in the field) detailed in this report, there are many opportunities for learning how to accurately archive and preserve this currently underutilized information as records of our (recent) past.

The further approach of social media archiving within the BESOCIAL project, will be referred to in upcoming reports.

1. Introduction

This report, 'An international review of Social Media Archiving initiatives' aggregates the results from Work Package 1 (WP1) of the BELSPO funded, KBR coordinated BESOCIAL project. The aim of WP1 was to review existing social media archiving projects and corpora in Belgium and abroad. The report you are about to read, therefore, focuses on the existing social media archiving at an international level. A separate report entitled: *Towards a sustainable social media archiving strategy. Country report: web and social media archiving in Belgium*, documents the state of social media archiving in Belgium specifically.

In this work package, four dedicated tasks aimed to provide a concise international state-of-the art of social media archiving (SMA). *Task 1.1. (Analysis of selection and access policies)* aimed to create an overview of international best-practices for preserving and archiving social media. *Task 1.2. (Analysis of existing foreign legal frameworks)* analysed the potential foreign legal frameworks allowing SMA, inter alia, by national libraries. The analysis focussed on selected European countries but also on non-EU countries and regions that may be of interest (e.g. New-Zealand, Canada or Quebec). *Task 1.3. (Analysis of technical solutions for social media archiving and preliminary testing of tools and quality control)* surveyed existing tools, standards and techniques relevant to discover, collect and consolidate data from different social media platforms for policy-aware long-term preservation. *Task 1.4. (Analysis of preservation policies)* reviewed and synthesized the preservation policies in place for social media content by studying the policies of institutions that are archiving social media in detail. The details of these tasks are detailed here in this report.

This report is structured as follows. **Section 2** of the report is a literature review including the definitions of born-digital heritage and social media, the history of archiving social media content, and the right to information. In **Section 3**, the methodology is described for the desk research, survey, interviews and synthesis. In **Section 4**, the selection policies and practices for social media archiving are discussed. **Section 5** provides an update of the analysis of legal frameworks studied during the PROMISE research project on web archiving, including the addition of Estonia, Hungary and New Zealand. In **Section 6** the report discusses various tools to create social media archives. The description of how to access and to use social media archives is outlined in **Section 7**, including an in-depth analysis of preservation policies in **Section 8**. The findings from these various tasks are summarized in the discussion and conclusion.

2. Literature study

2.1 Social media

Coined in the nineties, it took until the mid-2000s for the phrase 'social media' to enter common parlance (Ortner, Sinner & Jadin, 2018). Although the exact meaning of the phrase is subject to ongoing discussions due to the variety of evolving stand-alone and built-in social media services, generally 'social media' refers to Information and Communication Technologies (ICT) that enable social interaction (Treem & Leonardi, 2012) that allows "the creation and exchange of user generated content" (Kaplan & Haenlein, 2010, p. 61). Social media thus encompasses interactive computer-mediated technologies that facilitate the creation or sharing of information and other forms of expression via online communities and networks (Kietzmann, Hermkens, McCarthy and Silvestre, 2011; Obar & Wildman, 2015). However, the term "social" does not account for technological features of a platform alone, its level of 'sociability' is clearly determined by the actual performances and interactions of the social platform's users (Ariel & Avidar, 2015).

Social media platforms such as Facebook, Twitter or YouTube are representative of the growing "networked information economy" (Benkler, 2006), marking a shift from an industrial information economy (content centrally produced and distributed by commercial entities) to an economy in which individuals and groups of citizens create, annotate, and distribute media de-centrally (Marwick, 2010). Social media platforms embody a key aspect of today's Internet, namely the uprise of online user participation and interaction. Websites have evolved from a collection of online static pages to continually-updated platforms that invite users not only to consume (read, listen, watch), facilitate (tag, recommend, filter) and communicate (send messages, post comments, rate, chat) but also to create (personalise, aggregate, contribute) and share (publish, upload) content.

2.2 Born-digital heritage

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

2.3 History of social media archiving

As the web evolved, web archiving evolved with it and the creation of social media platforms gave rise to SMA initiatives. One of the pioneer projects in SMA is the Occasio project, launched in 1995 that aimed to preserve political and social conversations posted between 1988 and 2002 on online discussion groups (IISH, 2020). During this period (national) libraries and archives also broadened the scope of their collections to include the web. At the National Library of New Zealand, the first Twitter archive was added to the collections in 2009 (Macnaught, 2018). The British Library started archiving social media systematically in 2010 but limited Twitter, Facebook and Youtube content was captured previous to this date, where the UK National Archives has archives of Twitter accounts dating back to 2008 in its collections (Hockx-Yu, 2014; Espley, Carpentier, Pop & Medjkoune, 2014).

In 2010, a partnership between Twitter and the Library of Congress was initiated, in order to archive public tweets published on the platform (Zellier, 2018). Since 2017 this initiative has reduced in its capacity. The change to selective collecting was prompted by the changing nature of Twitter (increased length of tweets or increasing video, images or linked content for example) and constituted an alignment with the collection policies of the Library of Congress (Library of Congress, 2017). The Bibliothèque nationale de France has archived Facebook data since the creation of its web archive in 2006, but technological changes within Facebook forced the library to stop systematically archiving it in 2010 (Le Follic & Chouleur 2018). These last two examples clearly illustrate that collection development plans are directly influenced by (technological) changes in the social media landscape.

2.4 Importance of preservation for the right to information and legal constraints

Social media archives and allow us to document the past in ways we have never previously had the ability, as well as ease to archive. They are an invaluable resource for researchers to study human behavior and history as they provide clear records of communication (Ruth & Pfeffer, 2014). The logs, social media posts and related metadata allow us to document the past in ways we have never previously had the ability, as well as ease to archive.

Social media platforms and the web in general provide an essential tool for the freedom of expression and the right to information for citizens of all ages and backgrounds. The European Court of Human Rights frequently supports this observation in its case law.[1] In such a context, the preservation of social media content and its availability for the research community and the general public are major societal challenges. Indeed, these SMA initiatives, more specifically the log files, content of social media posts and related metadata, allow people to search and access a multitude of content that can be considered as born-digital heritage and of cultural, societal, historical or scientific interest. In doing so, archiving institutions play the role of "facilitator" in the exercise of the fundamental rights conferred by Article 10 of the European Convention on Human Rights[2] and, more particularly, the right to information. This fundamental right protects both the communication of ideas, opinions and information and their reception. Furthermore, the European Court of Human Rights had the opportunity, in 2012, to consider that the constitution of archives on the Internet fell under the umbrella of Article 10 of the Convention. It was in a Times Newspapers Limited v. the United Kingdom judgment concerning the establishment of a web archive of press articles that the Court for the first time[3] specified that "[...] Article 10 guarantees not only the right to impart information but also the right of the public to receive it. In the light of its accessibility and its capacity to store and communicate vast amounts of information, the Internet plays an important role in enhancing the public's access to news and facilitating the dissemination of information in general. The maintenance of Internet archives is a critical aspect of this role and the Court therefore considers that such

[1] See ECHR (2nd sect.), case of *Ahmet Yildirim v. Turkey*, 18 December 2012, app. no 3111/10, §54.

[2] European Convention for the Protection of Human Rights and Fundamental Freedoms, adopted at Rome the 4th November 1950, art. 10, §1: "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers […]".

[3] The Court subsequently repeated this principle in subsequent judgments. See in particular ECHR (4th sect.), case of *Wegrzynowski and Smolczewski v. Poland*, 16 July 2013, app. no 33846/07, §59; ECHR (5th sect.), case of *M.L. and W.W. v. Germany*, 28 June 20148, app. nos 60798/10 and 65599/10, §§90 and 102.

archives fall within the ambit of the protection afforded by Article 10".[4] In particular, the Court added that providing citizens with Internet archives afford a substantial contribution for the preservation and the making available of news and information and constitutes also a valuable source for education and historical research.[5]

However, even if SMA initiatives have a particular resonance in terms of fundamental rights' protection, they still involve competing interests that should be considered. Alongside the interest of scientists, researchers and society at large in accessing archived contents, there are the interests of other stakeholders such as copyright holders, people involved in producing, the owners of websites or social media pages or (national) cultural heritage institutions.

Implementing SMA initiatives obviously involves the same legal considerations as the web[6]; however, they go a step further by raising additional legal issues compared to those of web archiving. Here, we can think of the ambiguous relationship between social media and the right to privacy protected by Article 8 of the European Convention on Human Rights.[7] In that respect, when it comes to archiving social media, we must be attentive to the question of whether the content posted on social media belongs to the private or public sphere. This question, which is at the heart of many controversies in the jurisprudence, is crucial to assess a possible violation of the privacy of persons targeted by publications on social media. In addition, the right to privacy is a greater concern for social media, than web pages; specifically aspects related to image right or e-reputation are much more sensitive on social media than on web pages.

2.5 Metadata standards for effective data management

Archiving and mastering the volume, variety and velocity of data on social media platforms demands high-quality metadata to, among others, allow effective (research) data management. The National Information Standards Organization (NISO) defines several types of metadata: descriptive metadata to find and understand resources, administrative metadata which can be of technical, preservation or digital rights nature, structural metadata to describe relationships between resources and markup languages which integrates content with metadata to express other structural or semantic features (Riley, 2017).

There is a strong need for provenance metadata on different levels for archived web content (Venlet et al., 2018) for both basic users and scholars (Vlassenroot et al., 2019; Littman et al., 2018); this is often in contrast to the needs practitioners (Venlet et al., 2018). In case of social media this metadata can be provided via Application Programming Interfaces (APIs). Several metadata standards exist from which a common subset can be distilled, however, most tools which create metadata define descriptive metadata differently and mostly collect technical metadata. NISO lists 11 metadata standards in the cultural heritage field ranging from the storage efficient machine readable MARC format family developed in 1968 to several XML-Schemas and OWL ontologies like DDI and PREMIS developed in recent years. Whereas these standards cover different types of metadata, Dooley et al.

[4] See ECHR (4th sect.), case of *Times Newspapers LTD (Nos. 1 and 2) v. The United Kingdom*, 10 March 2009, app. nos 3002/03 and 23676/03, §27.

[5] See ECHR (4th sect.), case of *Times Newspapers LTD (Nos. 1 and 2) v. The United Kingdom*, 10 March 2009, app. nos 3002/03 and 23676/03, §45.

[6] For archiving the web we must pay attention to the distribution of missions, roles, competences and responsibilities between national cultural heritage institutions in charge of web preservation, the definition of the "national web", the copyright, the sui generis right on databases, the right to data protection, the authenticity and integrity of online content, and the issue of illegal or harmful online content.

[7] European Convention for the Protection of Human Rights and Fundamental Freedoms, adopted at Rome the 4th November 1950, art. 8, §1: "Everyone has the right to respect for his private and family life, his home and his correspondence".

(2018) reviewed existing metadata standards with respect to descriptive metadata and recommended the use of 14 data elements.[8] These elements are applicable both on collection and on item level. Although these 14 elements largely overlap with Dublin Core, they are meant to be standard-neutral. Although no minimum set is required, Title and URL are the absolute minimum and Collector, Creator, Date and Description are strongly recommended. In practice, descriptive metadata is defined differently by different platforms and tools, but that most tools provide technical metadata as WARC is an often used file format to store captured web content (Samouelian et al., 2018).

Several commercial tools for social media harvesting exist, but also various open source solutions have been developed to monitor, capture and store social media content. For lists of social media research tools, including data collection and archiving tools, curated by researchers see: the 'Social Media Research Toolkit'[9], and the wiki 'Social media data collection tools'.[10] A list of general web harvesting tools were collected by the Data Together initiative in 2018 in form of a collaborative spreadsheet (Hucka, 2017).

[8] Recommended elements: Collector, Contributor, Creator, Date, Description, Extent, Genre/Form, Language, Relation, Rights, Source of description, Subject, Title and URL.

[9] Social Media Data Scholarship. (2020). Social Media Research Toolkit. https://socialmediadata.org/social-media-research-toolkit/.

[10] Freelon, D. (n.d.). Social media data collection tools. http://socialmediadata.wikidot.com/.

3. Social Media Archiving Initiatives

The main research question for this task was to determine how national libraries and archives are engaging in social media archiving (as an extension to Vlassenroot et al., (2019) on web archiving). Thus, there was a need to take stock of the current SMA initiatives. This was accomplished through: 1) a secondary research or desk research, 2) a questionnaire, and 3) validation and synthesis by means of in-depth interviews. This resulted in the summarisation, collation and synthesizing documentation related to existing SMA projects. This desk research took place in fall 2020 and resulted in the following list of archiving initiatives (see Table 1. List of Web Archiving Initiatives). A number of characteristics were taken into account:

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

The resulting list of initiatives served as a starting point for investigating social media archiving activities, with the assumption that those institutions that were already active and had experience in web archiving would also possibly have SMA's.

Table 1. List of Web Initiatives

Country	Institution	Name	Abbreviation
Canada	National Library	Library and Archives Canada	LAC
Canada	Regional Library	Bibliothèque et Archives nationales du Québec	BAnQ
Denmark	Royal Danish Library	Netarkivet	Netarkivet
Estonia	National Library	Eesti Veebiarhiiv	Eesti Veebiarhiiv
France	National Library	Bibliothèque nationale de France	BnF
France	National Audiovisual Institute	Institut national de l'audiovisuel	INA
Hungary	National Library	National Széchényi Library	NSL
Ireland	National Library	National Library of Ireland	NLI
Luxembourg	National Library	Bibliothèque nationale du Luxembourg	BnL
New-Zealand	National Library	National Library of New Zealand	NLNZ
Switzerland	National Library	Webarchiv Schweiz	Webarchiv Schweiz
The Netherlands	National Library	KB Webarchief	KB
The	National Archive	Nationaal Archief	NA

Netherlands			
UK	British Library	UK Web Archive	UKWA
USA	University Library	George Washington University Libraries	GWUL

The web archives were studied from an operational, legal and technical point of view. The aim was to fill in the gaps and extend the information with regards to SMA in each of the institutions covering a) the selection, b) the social media archiving process itself, c) access to, and (re)use of the social media archive, d) preservation policy.

In the second research phase, a questionnaire which ran from July 2020 to September 2020 was sent to representatives from the aforementioned institutions. The aim of this survey was to address the gaps that remained on the specific initiatives following the literature review. Each of the participants were sent a personalised spreadsheet with questions and were asked to provide written replies. Based on the desk research some questions were already answered beforehand and the respondents were asked to verify this. Additionally, participants were also asked if the information in the spreadsheet could be shared with the broader international web archiving community in an open format.

The third and final research phase encompassed further validation and synthesis by means of in-depth interviews. Table 2 below shows with whom and when these interviews were conducted. The answers to the questions that were obtained during the desk research and from the survey were integrated. On the basis of which, comparisons were drawn in an exploratory analysis, in order to respond to the research question and to create an overarching view of the selected SMA initiatives. A summary of the main findings from these interviews are detailed below in Section 4.

Table 2: Overview of conducted interviews for WP1

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

	Institute (INA)		Jérôme Thiève	
Luxembourg	Bibliothèque nationale du Luxembourg (BnL)	18/11/2020	Ben Els & Yves Maurer	Sally Chambers, Sven Lieber, Jessica Pranger & Eveline Vlassenroot
New Zealand	National Library	24/11/2020	Gilian Lee, Ben O'Brien, Valerie Love & Ronda Grantham	Sally Chambers, Friedel Geeraert, Sven Lieber & Eveline Vlassenroot
Portugal	Arquivo.pt	10/12/2020	Daniel Gomes	Sally Chambers, Sven Lieber, Jessica Pranger & Eveline Vlassenroot
United Kingdom	British Library	30/11/2020	Nicola Bingham	Friedel Geeraert, Sven Lieber, Eveline Vlassenroot & Sally Chambers
United Kingdom	National Archives	4/02/2021	Tom Storrar, Claire Newing & Sarah Dietz	Sally Chambers, Sven Lieber, Fien Messens & Eveline Vlassenroot

4. Selection of content for social media archiving

4.1 Social media archived by web archives

Vlassenroot and authors (2019, Table 2) reported that a number of web archiving initiatives in their study included social media content in their collections; however, the policies with regards to social media differed widely between institutions. Table 3 provides an update of this overview from the interviews, including data from additional SMA initiatives (updated data is marked in bold). A more detailed analysis of these results can be found in the forthcoming publication of Vlassenroot and authors. The most notable change is the inclusion of Facebook, YouTube and Instagram by the National Library of France. Others are experimenting with adding social media content to their archives using small scale tests or in the context of collaborating with other institutions (e.g. the National Library of the Netherlands participates in WARCnet and projects such as TwiXL focusing on curating and making accessible Dutch language collections of social media and web data).

Table 3 shows that Twitter is the social media platform most often archived by the institutions in our sample, followed by Facebook and Instagram. This focus on Twitter is not surprising given that Twitter is being used as a communication tool in many industries and domains. The platform has increasingly integrated itself into daily life and functions as an effective communication system for breaking news; alongside which, celebrities, world leaders and politicians have been increasingly utilising Twitter to engage with the media and citizens. As a result, some archiving institutions (e.g. the National Library of Luxembourg) have chosen to focus archiving Twitter content as part of their ongoing, large scale archiving efforts.

In addition to the social media platforms listed in Table 3, a number of archiving institutions also collect and archive content from other social media channels such as Dailymotion, Vimeo or Soundcloud (e.g. the National Audiovisual Institute in France). Often content from these (slightly) less popular social media platforms is archived because it was embedded in a tweet or in a webpage that was archived by the institution earlier (e.g. the National Library of Hungary).

Table 3: Overview of social media archived by web archives (Vlassenroot et al., forthcoming)

Country	**Institution**	**Facebook**	**Twitter**	**YouTube**	**Instagram**	**Flickr**	**Other**
Canada	LAC	Yes	Yes	Yes	Yes	Yes	
Canada	BAnQ	Yes	Yes	No	No	No	
Denmark	Netarkivet	Yes	Yes	Yes	Yes	No	
Estonia	Eesti Veebiarhiiv	One page	No	No	No	No	Experimenting with archiving social media
France	BnF	**Yes**	Yes	**Yes**	**Yes**	No	
France	INA	No	Yes	Yes	No	No	Youtube, Dailymotion, Vimeo, Soundcloud

Hungary	National Library	No	No	No	Yes	No	Occasionally (currently in a pilot phase)
Ireland	National Library	No	Yes	Yes	No	**No**	Very limited amount of social media archiving (focusing their efforts on websites)
Luxembourg	National Library	Yes	Yes	Yes	**No**	No	
New-Zealand	National Library	Yes	Yes	No	Yes	No	
Switzerland	Webarchiv Schweiz	No	No	No	No	No	
The Netherlands	National Library	No	No	No	No	No	
The Netherlands	Koninklijke Bibliotheek	No	No	No	No	No	Whatsapp thriller (story that consists of Whatsapp messages)
UK	British Library	Yes	Yes	No	No	No	
UK	UKWA	No	Yes	Yes	No	No	
USA	GWUL	No	Yes	No	No	No	

4.2 Technical issues in selection: archivability of social media

As reported by Vlassenroot and authors (2019) social media accounts that are captured, in general focus on important people, organisations and events. This is done by archiving certain profiles or channels or by archiving content related to a certain hashtag. In some cases, for example when no hashtag is available, the results of specific search queries are stored. Only a few institutions (e.g. the National Library of France, NLNZ, UKWA) attempt to archive related social media data, such as the comments or the interaction data of a certain Tweet. This 'implicit' data that consumers of social media content produce (e.g. number of likes, retweets, comments, … see also 'exhaust data' (McCracken, 2007), 'read wear' (Hill, Hollan, Wroblewski & McCandless, 1992) or 'attention metadata' (Najjar, Wolpers and Duval 2006)) is thus often lost due to technical difficulties in capturing these on a large scale, e.g. Netarkivet: "One of the main issues in archiving social media is it's archivability; you can't get comments from Facebook for example, so it would be nice to get a WARC for Facebook with comments. That is probably one of the features the community would really like, that is high fidelity crawling of Facebook''.

4.3 Temporal issues in selection: speed and volume of social media

Most archiving institutions use a twofold approach for archiving regular web content – combining broad crawls (covering top-level domains) and selective crawls (for thematic or events based collections) (Vlassenroot et al. 2019). The nature of social media content (e.g. the volume, velocity and variety in which content is produced) necessitates another, more targeted approach, e.g. Netarkivet: "Broad crawls are not the solution, sometimes specific content is important, e.g. the

Danish prime minister when announcing the border will close because of COVID-19.”. The archiving institutions in our sample only use selective crawls to archive social media content.

A social media archiving approach needs to take into account the time-sensitive nature of social media archiving, as BnF remarks: “The difference with web archiving is mainly about temporality : we are faced with contents that appear often unexpectedly, according to the current news and events. These contents can rapidly reach a significant volume. (…) Contents on social networks can also disappear quickly : suppressions, blocked accounts, contents that become private...”.

Most often selective crawls focus on events, demonstrations or even emergencies and to a lesser extent on specific themes. For example, the National Library of France has set up a specific crawl dedicated to news events that includes numerous social networks accounts and content tagged with specific hashtags. Archiving these events requires a frequency that differs significantly from archiving regular web pages. The National Library of France therefore launches this crawl twice a day. Similarly, the National Library of Canada has been conducting event-based crawling since the inception of their programme in 2005. As it is difficult to anticipate or plan for archiving major events, their strategy shifted away from reacting and then documenting the event in motion, to an automated collection of news and social media content supplemented with curated archived content**:** “We collect all the topical hashtags from the media, we collect those daily (LAC)”. Next they can comb through the harvested hashtags (from whatever source) and “(…) we set all these hashtags to collect for a long period of time and then we analyse the traffic (so where is the majority of traffic being generated?). We use that information to limit the number of hashtags we collect, that is how we determine which are most prominent.” (LAC)

4.4 Spatial issues in selection

Next to the time-sensitive nature of social media archiving, there is also the issue of the non-existing spatial boundaries. Some institutions aim to identify if content can be considered as ‘national’ on social media before collecting the information, using various methods. The National Library of New Zealand therefore looks for hashtags that use NZ (e.g. Covid19nz), keywords that include NZ, tweets that are using geolocation codes. But as NLNZ remarks: “We realise that people overseas may also use these tags. Or the other way around, that some tags without NZ may trend for a time and the bulk of tweets comes from NZ. In these cases we might harvest those tags for a few days until it stops trending or we notice that it’s being used outside NZ. Also tweets from NZ may be missing if people use more generic hashtags like #Covid19 and don’t include NZ in their tweet.”.

4.5 Other issues in selection: semantics, nature and representativeness of social media

BnF takes a similar incremental approach with regards to hashtag selection, starting from a first hashtag, that the librarian identifies (e.g. using the platform's suggestions ("trends" for Twitter)). Next tweets harvested with this hashtag are examined to determine which other hashtags are used concurrently. Also variants (e.g. the plural form of the hashtag), and reinterpretations of official hashtags or the typos in the hashtags are considered. For BnF, the hashtag must be clearly linked to a precise (news) event, and they retain only the most relevant and least ambiguous hashtags: “For example, in the recent cases of Nice terrorist attack, hashtag #Nice was not necessarily the most relevant, since it was also used daily for all the tweets about this town (not to mention the English term "nice" that has nothing to do with the town).”

A large number of web archiving initiatives are developing procedures and methods to also select and archive the (embedded) content of social media posts, such as hyperlinks, images or embedded videos from various video sharing platforms. For example, INA is currently working on a solution to collect web pages that are linked in a Tweet.

Next to these technical issues related to capturing embedded content in social media posts and the sometimes ambiguous character of e.g. hashtags, another important issue that was raised during the interviews is that of representativeness. Social media is too often the 'battle field' with several different camps (in types of beliefs, political orientation, …) fighting each other. The challenge to reflect the plurality and diversity of viewpoints, and the different stances taken in the social media posts, is not an easy one and is strongly linked to the selection of hashtags and profiles which the social media archiving process will be based upon.

To tackle this challenge, the UK British Library is looking into the co-creation of collections (for example diaspora collections, LGBTQ collections etc), through the contacts and networks they would identify agencies, groups and individuals that would have a good view on what is happening. An example is the "boredom project" where they hosted a workshop with young people to see which content should be captured (Woolman, 2020). Another example is the "save a website" campaign where they accept nominations from the broad public if they are UK in scope. Nevertheless an important role stays allocated to the archivist for example an Armenian website was nominated but this website was taking a stance saying that the Covid-19 pandemic was fake, an anti-vaccination policy that was wrapped up in a website that looked like legitimate information. As an archivist "you have to think about public safety versus the need to archive" according to the British Library.

In order to ensure that the social media archive reflects the plurality and diversity of the real web, two approaches can be followed, labelled 'open' diversity and 'reflective' diversity (McQuail & Van Cuilenburg, 1983; Takens, Ruigrok, Van Hoof, & Scholten, 2010). Open diversity considers diversity as an equal (social media archive) representation of all possible categories. Reflective diversity argues that a social media archive should reflect the proportions in society (McQuail, 1992). Take for instance research on the diversity of political opinions in the news. From an open viewpoint, diversity would be evaluated as an equal representation of all voices in the political spectrum, while from a reflective viewpoint, evaluation of diversity would be based on the question to what extent these voices coincide with the current distribution of political opinions in society (Joris et al., 2020).

5. Analysis of Legal Frameworks

This section of the report provides an analysis of the national legal frameworks with respect to web and social media archiving. The first subsection provides an update on the legal aspects of the national initiatives studied during the PROMISE research project (Chambers et al., 2018), with a specific emphasis on provisions related to social media archiving. In a second subsection, analysis of new national initiatives which have been analysed, since the end of the PROMISE project.

5.1 Update on national initiatives studied during PROMISE research project

5.1.1 France

The French legal framework for the web legal deposit has not changed since the exhaustive analysis carried out during the PROMISE research project (Chambers et al., 2018, p. 28-35).

Legal deposit is still governed by Articles L131-1 to L133-1 and R131-1 to R133-1 of the French Heritage Code. In 2006[11], the French legislator enshrined the "web legal deposit" by introducing a deposit obligation for "signs, signals, writings, images, sounds or messages of any kind if they are made available to the public by electronic means".[12] There is no doubt that with such a wording, the scope of application of the web legal deposit in France includes social media contents. As the BnF clearly stated, "social networks are considered as websites".[13] However, two important precisions should be made. On the one hand, by "made available to the public", the French Heritage Code refers to "any communication, distribution or representation, whatever the process and the target audience, as long as the latter goes beyond the family circle".[14] On the other hand, personal correspondence and private spaces available on Intranet sites and on social media are excluded from the scope of the web legal deposit (Graff & Sepetjan, 2011, p.182). As a result, both BnF and INA only archive the "public" contents of social media, those whose access does not require any authentication. Social media contents that fall under the "private sphere" are therefore not included in the archived collections. For example, for the BnF, it is the criterion of "authentication" that makes it possible to draw the boundary between the public and private spheres for social media content: if the content is freely accessible it is considered as "public", whereas if authentication is necessary to access the content then it will be considered as "private".[15]

During the analysis carried out as part of the PROMISE research project, we detailed the various French legal criteria used to determine that a website will fall or not within the scope of web legal deposit (in particular, publication made in France or by a French publisher abroad, .fr top level domain, content produced in France, etc.). The question then arose as to how BnF transposed these legal criteria to the archiving of social media. On the one hand, with regard to the territoriality of a content or the nationality of the publisher or producer of the content, BnF indicated that social media platforms are generally international and do not allow content to be filtered using territoriality or

[11] This extension of legal deposit to the web was made possible thanks to the "DADVSI Law" which modified the French Heritage Code in 2006. See Loi n° 2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information, *JORF*, 3 août 2006.

[12] French Heritage Code, art. L131-2, al. 3. Personal translation.

[13] See results of the survey spreadsheet.

[14] See French Heritage Code, art. R131-1, al. 2. Personal translation.

[15] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

nationality criteria.[16] However, the problem is solved in a "pragmatic" way. Indeed, BnF indicates that since the algorithms of these major platforms use nationality and location criteria, the fact that BnF's collection robot/harvester is linked to a French IP address makes it possible to receive the vast majority of French content.[17] On the other hand, BnF also specifies that when it selects the social media accounts to be archived and a doubt remains as to the place of residence of the producer of the content, then it resorts to external sources to determine this.[18] This is done in particular through the press, academic sites and other social media that mention the place of residence, such as Linkedin.[19]

Within the meaning of the French Heritage Code, persons subject to web legal deposit are those who publish or produce signs, signals, writings, images, sounds or messages of any kind for the purpose of communicating to the public by electronic means.[20] The list of persons (whether natural or legal persons) likely to fall under this definition is therefore extremely broad. In order to implement the web legal deposit obligations, the French legislator offers BnF and INA a double option. They can either carry out the collection/capture themselves using automatic procedures or they can agree with the persons subject to the web legal deposit obligation on the collection methods.[21] It should also be remembered that the French legislator has been attentive to the information provided to publishers and producers of web contents. Indeed, Article L132-2-1 of the French Heritage Code obliges BnF and INA to keep them informed of the collection/capture procedures they have put in place to enable the web legal deposit. In order to comply with this information obligation, the BnF harvester robot identifies itself as belonging to BnF when collecting websites and social media content and refers via a hyperlink to the BnF website which contains additional explanations and a contact e-mail address to obtain further information if necessary.[22]

With regard to copyright aspects, a distinction should be made between the phase of collection of social media content and the phase of access to archived content. In order to facilitate the collection of "materials" subject to legal deposit, an exception to the reproduction right of copyright holders has been inserted by the French legislator. For the web legal deposit, the creation of such an exception has indeed proved to be essential since it is technically impossible to capture a web content without "reproducing" it in the sense of copyright legislation (Graff & Sepetjan, 2011, p. 179-180). The copyright exception thus covers acts of reproduction which are intrinsically linked to the web legal deposit (Graff & Sepetjan, 2011, p. 180).[23] It prohibits right holders from preventing BnF and INA from reproducing content (whatever the medium and process used) when such reproduction is necessary for collection, conservation and on-site consultation.[24] For access to archived collections, an exception to the "right of communication to the public" of right holders has also been provided for by

[16] See follow-up interviews.

[17] See follow-up interviews.

[18] See follow-up interviews.

[19] See follow-up interviews.

[20] French Heritage Code, art. L132-2, al. 1, (i). By “communication to the public by electronic means”, the French Leotard Law aims “any making available to the public or categories of the public, by a process of electronic communication, of signs, signals, writings, images, sounds or messages of any kind which do not have the character of private correspondence”. See Loi française n° 86-1067 du 30 septembre 1986 relative à la liberté de communication, art. 2, al. 2. Personal translation.

[21] French Heritage Code, art. L132-2-1.

[22] See follow-up interviews.

[23] *Ibid*., p. 180.

[24] French Heritage Code, art. L132-4 to L132-6. It should be noted that by "on-site consultation", French law refers only to consultation by accredited researchers on individual workstations reserved exclusively for their use.

the French legislator. Right holders cannot therefore prohibit BnF and INA from offering accredited researchers the possibility of consulting the web/social media archives on site on individual workstations whose use is strictly reserved for them.[25] No remote or online consultation/access of the collections is therefore possible (even for PhD researchers and even with an agreement); it can only be done in the premises of the BNF, the INA or one of the 20 regional partner libraries.[26] The results of the survey spreadsheet provide an interesting precision for the use of data for research purposes. The BnF offers possibilities for concluding agreements when a research team needs to launch data analyses/data processing treatments on the web archives collection. Such an agreement is necessarily concluded between the research institute and the BnF and the data processing treatments will necessarily take place in BnF's premises. The research team will not be able to share the data from the web archives but only the result of the data processing treatment, i.e. the "processed data".[27] Finally, the BnF does not provide to readers any possibility of reproducing the collections (whether by printing or downloading) without a written permission from the right holders.[28]

With regard to data protection, the BnF has created a specific email address (dpd@bnf.fr) to address these issues. If someone wishes to exercise his or her right to rectification or right to erasure ("right to be forgotten"), the BnF is technically able to "blacklist" an archived website from the access interface. If the request is duly justified after an analysis of the legal service, the content will no longer be accessible ("reachable") to readers. However, this content will always remain in the WARC file because the BnF does not delete or modify/rectify any data from the WARC files (Chambers et al., 2018, p. 34).[29] The BnF informed us that, at the end of 2020, it had still not received any request to exercise the right to rectification or the right to erasure on the basis of the GDPR.[30]

With regard to illegal and harmful content, the BnF had indicated in the interviews conducted for the PROMISE research project that they did not guarantee the legality of web archives but that filtering tools were used to prevent access to illegal contents and websites that would violate the law (Chambers et al., 2018, p. 34). On the other hand, in the results of the survey spreadsheet, the BnF indicates that illegal and harmful contents are not as such integrated in their selection policy but that they do not seek to avoid it. The BnF also specifies that no strategy for identifying such contents from

[25] French Heritage Code, art. L132-4 to L132-6. It should be noted that the individual workstations are equipped with access, search and processing interfaces provided by the BnF, INA or regional partner libraries. See French Heritage Code, art. R132-23-2 and R132-43.

[26] French Heritage Code, art. R132-23-2 and R132-43.

[27] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[28] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[29] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630. It should be noted, however, that in the context of the analyses carried out for the PROMISE research project, the BnF had indicated at the time that it applied access restrictions only on a case-by-case basis and on the basis of a court decision. See follow-up interviews.

[30] See follow-up interviews.

social media has been put in place. Finally, the BnF indicates that access to archived illegal and harmful contents will only be restricted in the event of a court decision.[31]

5.1.2 The Netherlands

As the KB and the National Archive of the Netherlands do not yet archive social media content and as there is still no legal deposit law in the Netherlands, no new legal data has been reported since the analysis carried out as part of the PROMISE research project (Chambers et al., 2018, p. 15-18, p. 21-23).

The only indication that the KB mentions in the survey spreadsheet is that, in the context of future archiving of social media content, they do not intend to archive content that is considered as illegal or harmful.[32]

5.1.3 Luxembourg

In Luxembourg, the legal deposit obligation was introduced in the law of 25 June 2004.[33] Article 10 very broadly defines the scope of legal deposit by referring to "publications of any kind, printed or produced by a process other than printing, whatever their technical production process, their medium, their publishing or distribution process [...] published on the national territory and offered for public sale, distribution or rental, or transferred for reproduction [...]".[34]

The Grand-Ducal Regulation of 6 November 2009 provides details on the modalities of legal deposit. Thus, among the categories of "publications" subject to legal deposit are those "without any material support made available to the public through an electronic network, in particular Internet sites and content, as well as all signs, signals, writing, images, sounds or messages of any kind, including [printed and graphic publications and digital publications on a physical support]".[35] Given the broad acceptance of this scope, social media content could be partially covered if the condition of "publication on the national territory" is met (Chambers et al., 2018, p. 46). To alleviate this problem, the BnL states that "however we decide in some cases to archive publications of Luxembourg citizens abroad, which broadens our scope beyond the legal deposit".[36] BnL takes care not to collect content that is part of the private sphere and determines for each "social media seed" whether the nature of the content published on the concerned profile meets a "public interest" criterion.[37] Furthermore,

[31] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[32] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[33] Loi luxembourgeoise du 25 juin 2004 portant réorganisation des instituts culturels de l'Etat, *Mémorial A120*, 15 juillet 2004.

[34] Loi luxembourgeoise du 25 juin 2004 portant réorganisation des instituts culturels de l'Etat, *Mémorial A120*, 15 juillet 2004, art. 10. Personal translation.

[35] Règlement grand-ducal du 6 novembre 2009 relatif au dépôt légal, *Mémorial A225*, 26 novembre 2009, art. 1 (3). Personal translation.

[36] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[37] ibidem

according to the BnL, the determination of whether social media content is public or private varies both according to the type of social media (for example, Twitter is by nature more "public" than Facebook) and according to the type of channel (for example, "pages" and organisations are more public than personal profiles and groups).[38]

With regard to copyright aspects, BnL indicates that it does not request any authorisation from right holders to collect content. Furthermore, Luxembourg law does not provide for any exception to copyright for acts of reproduction intrinsically linked to the "web legal deposit". Access to the web archives and social media content collections is done according to the restrictions imposed by copyright legislation, i.e. only in the premises of the BnL without the possibility of remote access.[39] The BnL nevertheless mentions that for research purposes it is possible to benefit from better access conditions provided that a contract is concluded which regulates the purposes, rights and risks linked to the use of collections.[40]

With regard to data protection, the approach taken by BnL is similar for all the digital content that forms part of its collections. This approach "is based on the presumption that capture and storage are exempt, only access can be restricted for legitimate reasons".[41]

With regard to illegal and harmful content that can be found on social media, BnL does not exclude them from its crawls and does not delete them from its archives. What is conceivable, for certain situations, is to restrict access to specific content from the "browsable web archive" and to allow accessibility only for specific purposes such as scientific research.[42] The BnL also believes that it is important to archive these particular forms of expression "since they might be unique to social media and are an undeniable part of the whole picture for any given topic or event".[43]

5.1.4 The United Kingdom

In the United Kingdom, legal deposit legislation has been extended to the web. The "Legal Deposit Libraries (Non-Print Works) Regulations 2013"[44] complements the "Legal Deposit Libraries Act 2003"[45]. This instrument determines which non-print works are subject to the legal deposit requirement and cites among others "work that is published online".[46] This is therefore likely to cover websites, social media and other types of online publications.

Three types of content are, however, expressly excluded from the scope of the web legal deposit legislation: works consisting only of sound recording and/or film and material merely incidental to it;

[38] ibidem

[39] ibidem

[40] ibidem

[41] ibidem

[42] ibidem

[43] ibidem

[44] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013. For more information on the modalities of the web legal deposit in the United Kingdom, we refer the reader to the complete analysis carried out in the framework of the PROMISE research project. See S. Chambers, E. Di Pretoro, F. Geeraert, G. Haesendonck, P. Mechant, A. Michel & E. Vlassenroot, PROMISE: final report Work Package 1 web archiving state of the art, 30 May 2018, pp. 53 to 58.

[45] Legal Deposit Libraries Act, 30th October 2003, Chapter 28.

[46] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 13, (1), (b).

works containing personal data AND[47] that are only made available to a restricted group of persons, and finally works that were published before the entry into force of the "Legal Deposit Libraries (Non-Print Works) Regulations 2013".[48] It therefore follows that social media content such as videos or posts containing personal data (which is mostly the case) made accessible only to certain people do not fall within the scope of the web legal deposit in the United Kingdom. Regarding the notion of "restricted group of persons", the guidance on the Legal Deposit Libraries (Non-Print Works) Regulations 2013 issued by the Department for Culture, Media & Sport gives interesting clarifications, especially for social media content.[49] We can see that "[…] *'restricted' means that the work is not generally available to all members of the public rather than the 'practical' barrier of registration that might make some works less immediately available. Therefore 'private' social networking content (e.g. 'protected tweets' to approved followers on Twitter, posts to 'friends' on Facebook, chat room discussions limited to a restricted group) would be out of scope of the regulations, but open access social networking pages, blogs and public comments added to articles are within scope*".[50] Furthermore, with regard to the difference between social media content belonging to the "public or private spheres", the British Library considers that a content is public when it can be accessed without a login.[51]

Regarding copyright aspects, the "Legal Deposit Libraries (Non-Print Works) Regulations 2013" determines a list of "permitted activities" in Sections 19 to 31.[52] Moreover, the "Copyright, Designs and Patents Act 1988" contains a specific provision for web legal deposit. Indeed, Section 44A specifies that there is no copyright infringement where a deposit library copies a work available on the Internet if the three following conditions are met: (1) the work is published on the Internet, is not mainly a sound recording and/or a film and is not content containing personal data and only accessible for a restricted group of persons; (2) its online publication or the person who has published it has a certain connection with the United Kingdom[53]; and (3) the copy of the work shall be made in accordance with the conditions laid down by the law.[54] With regard to access, the web legal deposit collection is only accessible within the premises of deposit libraries (and on their computers) and no

[47] Let us insist on the fact that two conditions must be met in order to exclude this kind of work from the scope of the web legal deposit legislation: on the one hand, it must be a content containing personal data and, on the other hand, this content must only be accessible to a limited group of people.

[48] See Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 13, (2), points (a), (b) and (c).

[49] The Department for Culture, Media & Sport (April 2013). Guidance on the Legal Deposit Libraries (Non-Print Works) Regulations 2013 (p. 17). Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/182339/NPLD_Guidance_April_2013.pdf. Last accessed on 22/09/2020.

[50] *Ibidem.*

[51] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[52] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Sections 19 to 31. On this point, see S. Chambers, E. Di Pretoro, F. Geeraert, G. Haesendonck, P. Mechant, A. Michel & E. Vlassenroot, PROMISE: final report Work Package 1 web archiving state of the art, 30 May 2018, p. 56.

[53] The certain connection is assessed as prescribed in Section 18 of the Legal Deposit Libraries (Non-Print Works) Regulations : if the domain name of the web page refers to the UK or any place therein or if the person who made the work available to the public has carried out the activities relating to the creation or publication of the work in the UK's territory; however, are excluded works which access is restricted to persons who are located outside of the national territory. See Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 18 (3).

[54] Copyright, Designs and Patents Act, 15th November 1988, Chapter 48, Section 44A ; Legal Deposit Libraries Act, 30th October 2003, Chapter 28, Section 10 (5) (a) ; Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 13 (3). It should be noted that the same provision exists for works protected by database law in the "Copyright and Rights in Databases Regulations 1997" in Section 20A (1).

wider access is allowed even for research purposes.[55] The conditions of access are particularly strict since deposit libraries are furthermore required to ensure that "only one computer terminal is available to readers to access the same relevant material at any one time".[56] The same social media content that has been archived can thus only be viewed on one screen at a time in each deposit library. Finally, before online content can be accessible to readers, an embargo of seven days must be respected.[57] The reason for the seven days embargo was that publishers were concerned that legal deposit copies (so the web archives) could compete with their own content. They wanted exclusivity in the accessibility of the content even if it was only for a few days.[58]

In addition, right holders may submit a request for deposit libraries to respect an embargo period before non-print works are made available. To this end, making works accessible must be likely to unreasonably prejudice the interests of right holders.[59] Such an embargo is valid for a period of 3 years, renewable as many times as the necessary conditions are met.

Regarding illegal and harmful contents that can be found on social media, the British Library applies a notice and takedown policy. Moreover, it does not especially seek to collect this kind of content and if some of them end up in their collections it is unintentionally.[60]

5.1.5 Denmark

In Denmark, the web legal deposit is covered by the Danish Act on Legal Deposit of Published Material.[61] Whereas previously the criterion for legal deposit was that of "printed materials", the criterion is now the one of "published materials".[62] As a result, digitally born content that is considered "published" (and thus "public"), such as web content, falls within the scope of legal deposit. Social media content therefore falls within this definition if they are freely accessible.[63]

Among the contents subject to legal deposit, the law lists the publications made available to the public through electronic communication networks.[64] According to the Danish law, the Royal Danish Library is thus allowed to download all Danish online content (Chambers et al., 2018, p. 70-71).[65] Unlike public online content, content that is considered as "private", i.e. content that is only made available via intranets or via closed extranets that are only accessible to a limited number of people,

[55] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). Web-archiving and social media: an exploratory analysis. INTERNATIONAL JOURNAL OF DIGITAL HUMANITIES, in press. The law defines the "computer terminal" as "a terminal on library premises controlled by the deposit library from which a reader is permitted to view relevant material". See Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 2 (1).

[56] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 23.

[57] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 24.

[58] See follow-up interviews.

[59] Legal Deposit Libraries (Non-Print Works) Regulations, 5th April 2013, Section 25. See also The Department for Culture, Media & Sport, *op. cit.*, p. 13.

[60] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[61] Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004.

[62] von Hielmcrone, H. (2011). Le dépôt légal au Danemark – Récents développements : le moissonnage des sites Internet. *Les cahiers de la propriété intellectuelle*, 2011/1, p. 73. Let us precise that by "published", we mean that the publication has been made legally available to the public.

[63] See results of the survey spreadsheet.

[64] Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §1.

[65] von Hielmcrone, *op. cit.*, p. 74. See also Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §8.

does not fall within the scope of the web legal deposit.[66] With regard to social media content, the Royal Danish Library states that it requests permission from social media to capture/collect contents.[67]

For reasons of both copyright and data protection[68], the Royal Danish Library takes care not to make its collections accessible unless very specific conditions related to scientific research purposes are met.[69] For research purposes, a remote access is possible for "material not readily available for acquisition".[70] Otherwise, a copyright exception to the right of communication to the public only allows access in the reading rooms. Nevertheless, the Royal Danish Library may enter into agreements with right holders' organisations in order to be able to implement wider access. Currently, following a specific agreement with right holders' organisations and in exchange for payment, this is for example the case for press articles which are accessible online in digital form for students of Danish universities.[71]

With regard to illegal and harmful contents that can be found in social media, the Royal Danish Library collects them and sometimes even specifically targets them in selection policies for documentation purposes.[72] When the Danish Royal Library decides to collect a particular type of content as part of a collection project, it ensures that the conservation decisions and the profiles and content selected are well documented.[73] Furthermore, the law foresees that legal deposit institutions must ensure that illegal contents that may be included in their collections are not accessible to unauthorised persons.[74] On the other hand, although the Royal Danish Library ensures that content deemed illegal is no longer accessible via the search index, this content will not be removed from its archives (it does not delete any content).[75]

With regard to data protection, when archived content contains errors or inaccuracies, it may be marked as "erroneous" and rectification is permitted.[76] Furthermore, when such inaccuracy/error concerns a natural person, correction or deletion must be made possible as soon as possible.[77]

5.1.6 Portugal

In Portugal, there is no web legal deposit legislation but only a "traditional" legal deposit for printed publications. However, Arquivo.pt has a mandate to acquire, preserve and safeguard the digital heritage. It therefore follows that in practice, even in the absence of web legal deposit legislation, web archiving still takes place thanks to this mandate (Chambers et al., 2018, p. 78-80). During a

[66] Ministry of Culture (n. d.). Bemaerkninger til lovforslaget (point 5.5.2.1). Retrieved from http://www.pligtaflevering.dk/loven/bemaerkninger.htm. Last accessed on 28/09/20.

[67] See results of the survey spreadsheet.

[68] Danish Act n° 1439 on Legal Deposit of Published Material of 22nd December 2004, §19 (3).

[69] See results of the survey spreadsheet.The Danish Royal Library works with a researcher agreement to provide data-level access.

[70] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[71] ibidem

[72] ibidem

[73] ibidem

[74] Danish Executive Decree n° 636 on the Disclosure of Published Material of 13th June 2005, §10, unofficial English Translation given by Jakob Moesgaard. See also Ministry of Culture, *op. cit.*, point n° 5.5.2.3.

[75] See results of the survey spreadsheet.

[76] Ministry of Culture, *op. cit.*, point n° 5.5.2.3.

[77] Ministry of Culture, *op. cit.*, point n° 5.5.5.

follow-up interview with the Arquivo.pt team in December 2020, it was confirmed that the situation has not changed since the PROMISE research project.

5.1.7 Ireland

Before 2019 in Ireland, legislation only existed for traditional legal deposit (publications on physical support and also extended to electronic publications) enshrined in the “Copyright and related rights Act 2000”.[78]

A review process of the legal deposit legislation has taken place and was completed at the end of 2019. During the revision phase, many voices were raised to include web legal deposit in the legislation and thus enable the preservation of online resources.[79]

Irish legal deposit legislation has been amended by the “Copyright and Other Intellectual Property Law Provisions Act 2019” and has been extended to some extent to digital publications.[80] Now, Section 198, (4A), (a) of the Copyright and related rights Act 2000 foresees that “where […] a digital publication is first published in the State by a publisher, [authorities having control of the *National Library of Ireland*, of the library of the Trinity College, of the University of Limerick or of the library of Dublin City University; as well as the Board of the British Library] may, by notice in writing given to the publisher, request that the publisher comply with this subsection […] as if the digital publication were a book referred to in that subsection and the publisher shall comply with that request […]”.[81] Let us precise that, in the scope of the Irish law, “digital publication” means “any publication published online or offline which is made available to the public in a medium other than print (including any publication in any digital or electronic or other technological form, but does not include any sound recording or film or any combination thereof)”.[82] These new provisions came into force on 2 December 2019. Although the wording of the definition of “digital publication” might have suggested that it covered websites and social media content, this does not appear to be the case. Indeed, H. Shenton indicates that “Last year’s Copyright and Other Intellectual Property Law Provisions Act extended the preservation of Irish digital publications to Irish copyright libraries. Websites, however, were excluded and so there is no legally mandated, comprehensive archive of the Irish web domain”.[83] It should be noted that the issue of including the archiving of the .ie domain name within the scope of the legal deposit legislation has been the subject of much back-and-forth in the legislative process.[84] Unfortunately, this was not the case in the final version of the Act and the new Section 198 of the Copyright and related rights Act 2000 is very narrow in scope. As E. O'Dell points out, “Section 198(4A) permits a copyright deposit institution to ask for a digital copy instead of a hard copy of a book due under copyright deposit. But its reach is very partial. It does not permit the institution to seek the digital copy as well as the hard copy, and often there are differences between the two that would make this desirable. It does not cover works that are exclusively digital; this is a

[78] Copyright and related rights Act 2000, n° 28 of 2000, Section 198.

[79] See in particular O’Dell, E. (2017, May 11). Legal deposit of digital publications. Retrieved from http://www.cearta.ie/2017/05/legal-deposit-of-digital-publications/. Last accessed on 5/10/2020. See Copyright Review Committee (2013, October 1). Modernising Copyright: The Report of the Copyright Review Committee (pp. 79 à 84). Retrieved from http://www.cearta.ie/wp-content/uploads/2013/10/CRC-Report.pdf. Last accessed on 5/10/2020.

[80] Copyright and Other Intellectual Property Law Provisions Act 2019, Section 29.

[81] Copyright and related rights Act 2000, n° 28 of 2000, Section 198, (4A), (a).

[82] Copyright and related rights Act 2000, n° 28 of 2000, Section 198, (4A), (b).

[83] O’Dell, E. & Shenton, H. (2020, May 1). Irish copyright law must enable digital deposit. Retrieved from http://www.cearta.ie/2020/05/irish-copyright-law-must-enable-digital-deposit/. Last accessed on 5/10/2020.

[84] *Ibid*.

yawning chasm now which will become steadily vaster as publishing moves increasingly (and more and more often, exclusively) online. It does not permit such an institution to harvest the .ie domain (even though the National Library is already doing so[85])".[86]

Finally, as a "weak compromise", Section 108 of the "Copyright and Other Intellectual Property Law Provisions Act 2019" indicates that "within twelve months of the enactment of this Act the Government shall bring forward a report on the feasibility of establishing a digital legal deposit scheme to serve as a web archive for .ie domain contents and advise on steps taken towards that goal".[87] According to some, however, it seems unlikely that a revision of the law will be concluded on this point by the end of 2020...[88] By the end of 2020, no further information about the envisaged revision in section 108 of the Copyright and Other Intellectual Property Law Provisions Act 2019 was available. The BE-Social team will continue to monitor this during 2021.

5.1.8 Switzerland

Although there is no legal deposit legislation at the federal level in Switzerland, the Swiss National Library has been given a legal mandate to "collect, list, preserve, make available and make known information that is printed or stored on media other than paper and that has a connection with Switzerland".[89] In view of the wording of the law, this mandate thus enables the Swiss National Library to preserve information regardless of their medium, and thus to include the preservation of online content within the mandate.

However, this task of information preservation is made more complex by the absence of federal legislation on legal deposit, which means that the Swiss National Library has to conclude agreements with Swiss publishers' associations.[90] Regarding online information such as websites, the Swiss National Library has chosen to use the opt-out approach by notifying the website owner of the future archiving of its content and presuming authorisation in the absence of any reaction.[91]

For Switzerland, we have not obtained any additional information compared to the results obtained for the PROMISE research project via the survey spreadsheet.

5.1.9. Canada

In Canada, Section 10 of the Library and Archives of Canada Act foresees the legal deposit obligation for publications made available in Canada.[92] Since the legal definition of "publication" does not contain any requirements as to medium or form, online publications fall within the scope of the legal deposit legislation. Indeed, Section 2 defines "publication" as "any library matter that is made available in multiple copies or at multiple locations, whether without charge or otherwise, to the public generally or to qualifying members of the public by subscription or otherwise. Publications

[85] In relation to the fact that the National Library of Ireland nevertheless archives the web, we refer to the analysis developed in the PROMISE project. See S. Chambers, E. Di Pretoro, F. Geeraert, G. Haesendonck, P. Mechant, A. Michel & E. Vlassenroot, PROMISE: final report Work Package 1 web archiving state of the art, 30 May 2018, pp. 86 to 87.

[86] O'Dell, E. & Shenton, H., *op. cit*.

[87] Copyright and Other Intellectual Property Law Provisions Act 2019, Section 108.

[88] O'Dell, E. & Shenton, H., *op. cit*.

[89] Loi fédérale suisse sur la Bibliothèque nationale suisse du 18 décembre 1992, art. 2. Personal translation.

[90] The Swiss law enables such agreements' conclusion. See Loi fédérale suisse sur la Bibliothèque nationale suisse du 18 décembre 1992, art. 3, §2.

[91] For more information, see the results of the PROMISE research project. See S. Chambers, E. Di Pretoro, F. Geeraert, G. Haesendonck, P. Mechant, A. Michel & E. Vlassenroot, PROMISE: final report Work Package 1 web archiving state of the art, 30 May 2018, pp. 110 to 111.

[92] Library and Archives of Canada Act, S.C. 2004, c. 11, Section 10.

may be made available through any medium and may be in any form, including printed material, on-line items or recordings".[93]

The Library and Archives of Canada Act allows for the adoption of regulations to complement the legal deposit framework on certain points, including the measures that must be implemented to ensure that publications other than "paper" (i. e. online publications) are accessible to the LAC.[94] The Legal Deposit of Publications Regulations thus foresees in this regard two types of measures for publishers. Before providing a copy of the publication, they must, on the one hand, decrypt the possible encrypted data and, on the other hand, remove/disable the possible security systems/devices that restrict or limit access to the publication.[95] When providing a copy of the publication, they must provide to the LAC the following elements: a copy of the software specifically created by the publisher necessary to access the publication, a copy of technical and other information necessary to access the publication and the descriptive data about the publication (title, creator, language, data of publication, format, subject, copyright information, …).[96]

In addition to this section specifically devoted to the legal deposit, the Library and Archives of Canada Act contains another interesting provision. Section 8 (2) of the Act allows the LAC, in order to preserve and acquire publications and archives, to conduct a representative sampling of the web.[97] This possibility covers "the documentary material of interest to Canada that is accessible to the public without restriction through the Internet or any similar medium".[98] This provision therefore allows for the preservation of online resources that are publicly accessible (without passwords or access restrictions), thus making it possible to include websites and public social media content in the scope. The Library and Archives of Canada considers that social media content clearly falls within the scope of Section 8 (2) of the Act.[99] It indicates moreover that social media community standards and terms and conditions are honoured where possible.[100]

With regard to copyright aspects, the Canadian Copyright Act specifies that the LAC does not infringe copyright when, on the basis of the Library and Archives of Canada Act, it implements the possibility conferred by Section 8 (2) to conduct a representative sampling of the web.[101] Thus, when the LAC makes a copy of a protected work in the context of web archiving, this act of reproduction (normally protected by copyright) does not violate the rights of the owners. The collection/capture of online content therefore does not require the authorisation of the right holders but the LAC mentions that it does request authorisation for the archiving of social media content from personal accounts.[102]

[93] Library and Archives of Canada Act, S.C. 2004, c. 11, Section 2.

[94] Library and Archives of Canada Act, S.C. 2004, c. 11, Section 10, (2), (b).

[95] Legal Deposit of Publications Regulations, 12th December 2006, SOR/2006-337, Section 2, (a).

[96] Legal Deposit of Publications Regulations, 12th December 2006, SOR/2006-337, Section 2, (b).

[97] Library and Archives of Canada Act, S.C. 2004, c. 11, Section 8, (2): " In exercising the powers referred to in paragraph (1)(a) and for the purpose of preservation, the Librarian and Archivist may take, at the times and in the manner that he or she considers appropriate, a representative sample of the documentary material of interest to Canada that is accessible to the public without restriction through the Internet or any similar medium".

[98] Library and Archives of Canada Act, S.C. 2004, c. 11, Section 8, (2).

[99] See results of the survey spreadsheet. In relation to the fact that social media content falls within the scope of legal deposit (Section 10 of the Act), the Library and Archive of Canada states: "We consider social media covered by the LAC Act Section 8 (2) and perhaps by legal deposit, but the authority for web archiving is clearer/stronger".

[100] See results of the survey spreadsheet.

[101] Canadian Copyright Act, L.R.C. (1985), ch. C-42, Section 30.5, (a).

[102] See results of the survey spreadsheet. It should be noted that this operational decision seems to us to have more to do with privacy considerations than copyright.

Nevertheless, it should be stressed that this exception covers only acts of reproduction and not acts of communication to the public (Laforce & Paré, 2011, p. 275). Although the LAC may benefit from such an exception, it nevertheless mentions that it "would honour takedown requests for goodwill".[103]

With regard to privacy and data protection, the LAC states that due to the scale of online content collection it is inevitable that "private information" will be collected. Nevertheless, the LAC states that it honours takedown requests/requests for removal that are justified.[104] However, when a right to erasure request is justified, the LAC renders the content inaccessible but it is still kept in their archives.[105]

With regard to illegal and harmful content online, the LAC specifies that this type of content is voluntarily collected in a purposefully targeted manner and that samples are collected for future research and/or context.[106] However, the LAC mentions that illegal and harmful content can also be collected on a larger scale.[107]

As for access, the LAC states that, once its portal is ready, its intention is to make the collected contents (about 75 tera) freely available via the web for generalised access. However, for social media content, it indicates that the conditions of access have yet to be determined.[108] On the other hand, with regard to the accessibility of collected illegal and harmful content, the LAC states that the debate is still open (in particular, also on the question whether access could only be limited to research purposes).[109]

5.1.10 Quebec

In Quebec, contrary to the situation that prevails at the national level for Canada, there is only a legislation for "traditional" legal deposit[110] but which therefore does not cover web legal deposit. Online publications such as websites and social media content therefore do not fall within the scope of Quebec legal deposit legislation. Publishers can therefore simply, on a voluntary basis (no obligation at the Quebec regional level[111]), deposit digital and online publications.

However, the National Library and Archives Quebec has also been given the general legal mission "to collect, permanently preserve and disseminate Quebec's published documentary heritage and any

[103] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[104] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[105] See follow-up interviews.

[106] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[107] Ibidem.

[108] Ibidem.

[109] Ibidem.

[110] For legal deposit provisions, see Loi sur bibliothèque et archives nationales du Québec, art. 20.0.1. to art. 20.12.1. For more information about traditional legal deposit in Quebec, we refer to the results of the PROMISE project. See S. Chambers, E. Di Pretoro, F. Geeraert, G. Haesendonck, P. Mechant, A. Michel & E. Vlassenroot, PROMISE: final report Work Package 1 web archiving state of the art, 30 May 2018, pp. 100 to 102.

[111] It should be noted, however, that publishers are also subject to Canadian national legislation (the Library and Archives of Canada Act), which partially resolves the gap in Quebec law. See M. Laforce & J.-P. Paré (2011), *op. cit.*, p. 265.

related document of cultural interest, as well as any document relating to Quebec and published outside Quebec".[112] It therefore uses this legal mandate to preserve Quebec's cultural heritage to nevertheless conduct web archiving activities in the absence of a web legal deposit legislation. To do so, the BAnQ ensures that it obtains the necessary authorisations from website owners (Laforce & Paré, 2011, p. 268; Chambers et al., 2018, p. 100-102). However, it does not request authorisation for social media content.[113]

With regard to social media archiving, the BAnQ mentioned an interesting information for us in the survey spreadsheet: "*We asked our legal department if we can archive Twitter content. We received confirmation about the conservation of Twitter. For the moment, we don't give access. From what we understand, it is a violation of Twitter's Terms of Service to share large tweet datasets, so institutions sometimes get around this by sharing "dehydrated" tweet IDs. Every tweet has a unique identifier tweet ID, e.g., "1281680702648590338" which *can* be shared freely without any problems. A "dehydrated" set of tweets is just a list of tweet IDs. If you provide access to one of these lists, a researcher could "hydrate" the list to create their own tweet dataset, provided the data is still available on twitter.com. Twarc has a hydrate tool that allows you to retrieve the data from Twitter if you have a tweet ID list. It also includes a tool to extract Tweet IDs ("dehydrate")*".[114] We can therefore see that at this stage the BAnQ collects/captures social media content for preservation purposes only. No access is currently made possible.

With regard to illegal and harmful online content, the BAnQ does not archive them: they are excluded from the selection policy (Chambers et al., 2018, p. 101-102).[115]

5.2 Analysis of new national initiatives compared to PROMISE

This subsection of the report covers an analysis of new national initiatives (with the exception of Belgian initiatives which can be found in the BESOCIAL Report: Belgian Initiatives), which have been analysed, since the end of the PROMISE project.

5.2.1 Estonia

In Estonia, the Legal Deposit Copy Act of the 15th June 2016 expressly foresees a web legal deposit in favour of the National Library of Estonia.[116] Indeed, Section 2 of the Law makes explicit reference to "web publication".[117] The aim of this Estonian law is, like other national legislation, to ensure both the creation, long-term preservation and public accessibility of publications forming part of Estonia's cultural heritage. In this respect, the law mentions "the most comprehensive collection of the publications which are essential to the Estonian culture […]".[118]

[112] Loi sur bibliothèque et archives nationales du Québec, art. 14. Personal translation.

[113] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[114] Ibidem.

[115] Ibidem.

[116] See Estonian Legal Deposit Copy Act of the 15th June 2016, §12 (4).

[117] For the purposes of the Estonian law, a "web publication" is defined as a "web publication made publicly accessible through a technical device or process". See Estonian Legal Deposit Copy Act of the 15th June 2016, §2 (2). See also Estonian Legal Deposit Copy Act of the 15th June 2016, §4.

[118] Estonian Legal Deposit Copy Act of the 15th June 2016, §1 (2). We can add that the legal deposit obligation applies to publications with the same content but in different formats or forms. See Estonian Legal Deposit Copy Act of the 15th June 2016, §2 (5).

Regarding the scope of the Estonian web legal deposit, the law only covers web publications that are made publicly accessible (within the meaning of copyright law). Furthermore, in order to fall within the scope of the web legal deposit, the web page must have been made accessible according to one of the following criteria: 1) on the .ee ccTLD or on another top level domain territorially linked to Estonia; 2) on another top level domain provided that the web publication concerned is essential for Estonian culture; 3) by a Estonian citizen, a legal person registered in Estonia or by a natural person staying in Estonia provided that the web publication concerned is essential for Estonian culture.[119] We can therefore see the very broad scope of the Estonian legislation on web legal deposit. However, two types of web publication are expressly excluded from the scope of application of the law: on the one hand, real-time streaming of web publication and, on the other hand, web publication that requires an unreasonable large amount of data for preservation regarding its content.[120] The law also provides that a regulation shall be taken to list the categories of web publications that contain negligible information for the Estonian culture and "ephemera publications"[121] in order to exclude them from the scope of the law.[122]

The National Library of Estonia archives social media contents on the basis of this web legal deposit. It only collects social media contents that are accessible without any login to the platforms (so only "public contents") and it does so without requesting the necessary authorisations from the right holders.[123]

The Estonian Legal Deposit Copy Act also determines in some detail the procedure to be followed in order to fulfil the legal deposit obligation. Firstly, for web publications, the person responsible for the deposit ("the depositor") is either its producer either the issuing body if it has reached an agreement with the producer.[124] Secondly, the law provides for a collection procedure that varies according to the type of web publications. On the one hand, for freely accessible web publications published on the .ee domain name, on domain names territorially linked to Estonia or on other domain names provided that the content is essential for Estonian culture, the National Library of Estonia will carry out the web archiving.[125] To do so, the National Library of Estonia uses a web harvester to download the website with all necessary elements for the display and recording in archives.[126] In the event that it does not succeed in web archiving on its own, the National Library of Estonia will send a request to the person responsible for the deposit ("the depositor") who will have to allow a copy of the concerned web publication.[127] This copy must be submitted to the National Library of Estonia within

[119] Estonian Legal Deposit Copy Act of the 15th June 2016, §2 (2).

[120] Estonian Legal Deposit Copy Act of the 15th June 2016, §3 (1) 4-5.

[121] "Ephemera publication" means according to the Estonian law "publications with low-volume informative contents and text or published in relation to one-time event and with short-term importance". See Estonian Legal Deposit Copy Act of the 15th June 2016, §3 (2).

[122] Estonian Legal Deposit Copy Act of the 15th June 2016, §3 (2). Estonian law states that the National Library of Estonia (with the involvement of sector experts and representatives from the various memory institutions) shall make a proposal for the determination of these categories of web publications and ephemera. See Estonian Legal Deposit Copy Act of the 15th June 2016, §3 (3).

[123] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[124] Estonian Legal Deposit Copy Act of the 15th June 2016, §5 (4).

[125] Estonian Legal Deposit Copy Act of the 15th June 2016, §7 (2).

[126] Estonian Legal Deposit Copy Act of the 15th June 2016, §7 (2).

[127] Estonian Legal Deposit Copy Act of the 15th June 2016, §7 (3).

20 days of the request.[128] On the other hand, web publications that are essentials for Estonian culture and that have been made accessible by an Estonian citizen, by a legal person registered in Estonia or by a natural personal stating in Estonia must be submitted to the National Library of Estonia by forwarding a copy thanks to the electronic depositing system.[129] In such a case, additional data (which will remain private[130]) are submitted to the National Library of Estonia together with the copy of the web publication: the name or the title of the issuing body, producer or co-producer or the name and surname of the natural person; the descriptive and the technical metadata, the right of use, the structure of the object and the relationships.[131]

With regard to access and use of legal deposit collections, the applicable principle is that of maintaining good preservation conditions. This means that the National Library of Estonia may allow access to and use of legal deposit copies if it considers that this does not risk their adequate preservation.[132] In addition, access via a digital copy should be the first option.[133] Obviously, this is not a problem for web legal deposit. As far as web publications are concerned, they can in principle only be consulted on site (except for governmental websites) in one of the "authorised workplaces" if the copyright holder authorises it.[134] The research purposes do not allow a broader access to web legal deposit collections.[135] Moreover, Estonian law also specifies that, in the event that a web publication has to be deleted on the basis of a court decision, the legal deposit copy is not deleted and access can be made via a request submitted to the National Library of Estonia.[136] The same rule also applies in the case of a court decision which prohibits to render a publication accessible to the public.[137]

As far as illegal or harmful content is concerned, the National Library of Estonia states that it collects such contents and does not identify them as such.[138] Regarding access to such content, the National Library of Estonia does not yet have a policy in place yet for the simple reason that most web archives are not public.[139]

[128] Estonian Legal Deposit Copy Act of the 15th June 2016, §9 (2).

[129] Estonian Legal Deposit Copy Act of the 15th June 2016, §7 (4).

[130] The law specifies that these data are not public. See Estonian Legal Deposit Copy Act of the 15th June 2016, §11 (4).

[131] Estonian Legal Deposit Copy Act of the 15th June 2016, §11 (3).

[132] Estonian Legal Deposit Copy Act of the 15th June 2016, §14 (1).

[133] Estonian Legal Deposit Copy Act of the 15th June 2016, §14 (2).

[134] Estonian Legal Deposit Copy Act of the 15th June 2016, §14 (5) and (6). Section 16 (1) of the law defines an "authorized workplace" as "a computer terminal designed for in-house use of a digital legal deposit copy by using of which the making accessible to the public of the legal deposit copy or recording to external data carriers shall be precluded by using technical and physical means". The law lists the five places where authorized workplaces are located: the Estonian Literary Museum Archival Library, the National Library of Estonia, the Library of Tallinn University of Technology, the Academic Library of Tallinn University, the University of Tartu Library. See Estonian Legal Deposit Copy Act of the 15th June 2016, §16 (2).

[135] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630. It should be noted, however, that there is a small enlargement regarding data protection in Estonian law in a well-defined assumption. Indeed, if a web publication containing personal data is no longer accessible to the public, it is still possible to submit a request to access it for scientific research purposes only. See Estonian Legal Deposit Copy Act of the 15th June 2016, §15 (3).

[136] Estonian Legal Deposit Copy Act of the 15th June 2016, §14 (7) and (8).

[137] Estonian Legal Deposit Copy Act of the 15th June 2016, §14 (9).

[138] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[139] Ibidem.

Finally, with regard to data protection, the Estonian law contains some clarifications. Firstly, it authorises the processing of legal deposit copies containing personal data for the fulfilment of the objectives followed by the Legal Deposit Act.[140] Secondly, on the basis of a duly justified request from the data subject, the making available to the public of personal data contained in a web publication copy must cease.[141] Thirdly, if the National Library of Estonia has not verified whether a web publication contains personal data when archiving it, it must then make the web publication in question accessible to the public by applying technical restrictions preventing the finding of the web publication by using a search with given name or surname of a natural person.[142]

5.2.2 Hungary

In Hungary, the web legal deposit legislation came into force on 1st January 2021. Thanks to a recent legal revision of the Hungarian cultural law, the Hungarian National Library is entitled to archive the Hungarian web.[143]

The detailed rules for web archiving are contained in a government decree.[144] Within this framework, the Hungarian National Library has to hold a list of web contents which includes, in addition to the web content, its unique identifier, its URL, the outstanding qualification, its name and the date of its listing.[145] For full web harvesting, the Hungarian National Library collects data from the homepage of the website up to 10 link depths. On the other hand, selected web contents are web contents that are archived because of their educational, scientific, cultural, social, historical or research importance in relation to specific events.[146] The government decree also provides for a comprehensive web harvesting twice a year and a harvesting of selected web contents on a quarterly basis.[147] In the latter case, the Hungarian National Library may archive the web closer together when this is justified for certain events due to their nature or topicality.[148] Among other things, the government decree missions the Hungarian National Library to check the quality of web content archived during harvesting.[149]

The government decree excludes from web archiving online radio, television, video, podcasts and other online audiovisual contents as well as hacked web contents infected by a virus or other malicious code.[150] It also provides for an obligation of cooperation between the producer of the web content and the Hungarian National Library. The content producer must, on the one hand, upon request make a written declaration regarding the availability of the notified web content in a publicly accessible archive and, on the other hand, grant access rights to the collection and copying of web content published by it for archival purposes.[151]

[140] Estonian Legal Deposit Copy Act of the 15th June 2016, §15 (1).
[141] Estonian Legal Deposit Copy Act of the 15th June 2016, §15 (1).
[142] Estonian Legal Deposit Copy Act of the 15th June 2016, §15 (2).
[143] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630. See also English summary of the Hungarian law provided by Márton Németh.
[144] See English summary of the Hungarian law provided by Márton Németh.
[145] See English summary of the Hungarian law provided by Márton Németh.
[146] See English summary of the Hungarian law provided by Márton Németh.
[147] See English summary of the Hungarian law provided by Márton Németh.
[148] See English summary of the Hungarian law provided by Márton Németh.
[149] See English summary of the Hungarian law provided by Márton Németh.
[150] See English summary of the Hungarian law provided by Márton Németh.
[151] See English summary of the Hungarian law provided by Márton Németh.

In terms of copyright aspects, the National Library can, thanks to the web legal deposit Act, carry out the acts of reproduction (the capture of web contents). The Hungarian National Library reports that it requests permission from website owners for archiving but that it does not yet ask permissions for social media contents.[152] On the other hand, as far as access to the web archives collection is concerned, these are only accessible on site on dedicated terminals in order to prevent unauthorised reproduction (copying) of the contents. Moreover, the research purposes do not allow wider access.[153] However, in the case of web contents of the government (both federal and local) and public institutions, as well as web contents created with the help of state subsidies, the Hungarian National Library may make such web archives publicly accessible.[154] For other web contents, the only way for the Hungarian National Library to allow public access (via the internet) to the web archives collection is to obtain individual permission from the copyright holders by means of a written agreement.[155]

With regard to illegal or harmful web contents, the Hungarian National Library does not have an established policy yet.[156]

5.2.3 New Zealand

In New Zealand, a web legal deposit legislation exists. Indeed, the National Library of New Zealand Act 2003 – which was adopted with a view to ensure the preservation of the documentary heritage and the accessibility of the collections to New Zealanders[157] – provides for the provision of copies of public documents[158] to the National Library. In this context, the Act expressly states that the provision of public documents "extends to Internet documents and authorises the National Librarian to copy such documents".[159]

The provisions relating to the legal deposit obligation ("provision of copies of public documents" under the terms of the New Zealand law) are contained in sections 29 to 43 of the Act.

[152] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). “Interview/Survey spreadsheet WP1 BESOCIAL”. Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[153] Ibidem.

[154] See English summary of the Hungarian law provided by Márton Németh.

[155] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). “Interview/Survey spreadsheet WP1 BESOCIAL”. Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630, and English summary of the Hungarian law provided by Márton Németh.

[156] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). “Interview/Survey spreadsheet WP1 BESOCIAL”. Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[157] National Library of New Zealand Act 2003, 5 May 2003, Section 3. See also Section 7 for the purpose of the National Library.

[158] It should be noted that New Zealand law defines "document" very broadly, covering all possible forms. See National Library of New Zealand Act 2003, 5 May 2003, Section 4: “document means a document in any form; and includes— (a)any writing on any material; and (b) information recorded or stored by means of any recording device, computer, or other electronic device, or any other device, and material subsequently derived from information so recorded or stored; and (c) a book, manuscript, newspaper, periodical, pamphlet, magazine, sheet of letterpress, sheet of music, map, plan, chart, painting, picture, etching, print, table, graph, or drawing; and (d) a photograph, film, negative, tape, or other device in which 1 or more visual images are embodied so as to be capable (with or without the aid of equipment) of being reproduced; and (e)a second or subsequent edition of any of the above”.

[159] National Library of New Zealand Act 2003, 5 May 2003, Section 3, (h).

First of all, under New Zealand law, three cumulative conditions must be met in order to be in the presence of a "public document".[160] First, this document must have been in minimum one copy issued to the public, made available to the public upon request or made available to the public on the Internet, whether or not there is any [physical, technical or mechanical] restriction on the persons acquiring or accessing the document. Second, it must be a document that was printed or otherwise produced in New Zealand by a resident of the country or by a person who has their principal place of business in the country. Third, this document must be protected by copyright or must be a legislative or judicial document not protected by copyright (bills, acts, regulations, bylaws, parliamentary debates, reports, judgements, …). We therefore note that documents published on the Internet fall within the notion of "public document". Furthermore, the law defines the "Internet document" as a "document that is published on the Internet, whether- or not there is any restriction on access to the document; and includes the whole or part of a website".[161] As far as the publisher of an Internet document is concerned, the law considers it to be "the person who has control over the content of the website, or part of the website, on which the document is located".[162]

Second, the Act allows the Minister (via a notice in *The Gazette*) to authorise the National Library of New Zealand to make copies[163] of Internet documents. Such copies may be made at any time and at the discretion of the National Library but in accordance with the terms and conditions as to format, public access or other measures specified in the Minister's notice.[164] The Minister has made use of this possibility and authorises the National Library to make copies of Internet documents.[165] The Act also allows the National Library to request, in writing, the assistance/help of the publisher to obtain a copy of the document and the publisher must offer assistance at its own expense within 20 working days of the written request.[166]

Third, the Act contains clarifications regarding the use of "public documents" within the National Library. Thus, Section 34, (2) of the Act allows the National Library in the broad sense (including its employees, contractors, agents of the chief executive) to possess, copy, store in electronic form (offline or online) and use any copy of deposited documents in order to comply with its duties.[167] For contents that are outside the scope of the legal deposit legislation, the National Library obtains the authorisation of the creator to collect and give access.[168] With regard to making documents available to the public, the Act foresees that the National Library may only make a maximum of three copies available to the public, whether at its premises or elsewhere.[169] In principle, in the absence of the

[160] See National Library of New Zealand Act 2003, 5 May 2003, Section 29, (1). The Act excludes from the concept of "public document", on the one hand, the "public records" within the meaning of the Public Records Act 2005 (except those made available to the public, i.e. those with an ISBN or ISSN number) and, on the other hand, reprints of document with the same content and form as a public document already deposited to the National Library. See point (d), i) and ii).

[161] National Library of New Zealand Act 2003, 5 May 2003, Section 29, (1).

[162] National Library of New Zealand Act 2003, 5 May 2003, Section 29, (1), definition of "publisher", point c).

[163] In this respect, the Act specifies: "make a copy, in relation to an Internet document, means to make a copy of the document for the purpose of storing and using it in accordance with this Part; and includes circumventing any technological protection measures which otherwise would prevent or hinder the copying, storage, or use of the document". National Library of New Zealand Act 2003, 5 May 2003, Section 29, (1).

[164] National Library of New Zealand Act 2003, 5 May 2003, Section 31, (3).

[165] National Library Requirement (Electronic Documents) Notice 2006, 2 May 2006, Section 8.

[166] National Library of New Zealand Act 2003, 5 May 2003, Section 33.

[167] National Library of New Zealand Act 2003, 5 May 2003, Section 34, (2).

[168] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[169] National Library of New Zealand Act 2003, 5 May 2003, Section 34, (3).

publisher's agreement, documents cannot be made available on the Internet. Nevertheless, there is an exception in the case where the deposited document has been made publicly available on the Internet by the publisher without access restriction or without restriction of use by the public. In this case, the National Library is authorised to make the document available for access and use by the public on the Internet.[170]

In relation to contacts with social media platforms, the National Library of New Zealand indicates that it has had relatively limited success in direct contacts with them. With regard to terms and conditions and community standards of social media, the National Library has chosen a risk-based approach in order to give priority to the preservation of the New Zealand published documentary heritage and to allow access to it.[171]

Concerning aspects relating to the right to privacy and data protection, the National Library of New Zealand gives priority to the capture of public social media content and ensures that access is restricted in the case of social media content to which access is restricted by the profile owner. In addition, it has developed takedown policies for social media content.[172]

With regard to illegal and harmful content that can be found on social media, the National Library of New Zealand points out that: "we recognise there is an ethical tension between the need for collections to reflect the true nature of social conversations New Zealanders were having, with the need to protect end users and the subjects of speech from possible harm, but we are still working through our policies in this area. Where there is a legal request to remove illegal content from the collections or restrict access we will work to do so".[173]

[170] National Library of New Zealand Act 2003, 5 May 2003, Section 34, (4).

[171] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

[172] Ibidem.

[173] Ibidem.

6. Tools to create social media archives

In addition to the desk research related to SMA projects, see Section 2 and 3 above, a second desk research study was carried out reviewing documentation and GitHub repositories to identify the tools used in social media archiving. A variety of tools for SMA exist, but we seek to answer to which extent each tool addresses challenges of a particular use case, e.g. which social media platform is supported and does the tool stores collection-level metadata? Therefore we reuse an existing comparison of regular web archiving tools and extend it with respect to social media capturing and analysis.

To gather the state of the art of social media archiving from the technical perspective we asked related questions to the surveyed institutions and performed a comparison of (social media) archiving tools ourselves. In the following we elaborate on which tools are used (section 6.1), if social media is harvested via Application Programming Interfaces (APIs) (section 6.2), how privacy and copyright are addressed (section 6.3), which metadata standards are used to describe the harvested data and the processes (section 6.4), and which features are currently missing according to the surveyed institutions (section 6.5). Additionally we present results of a tool comparison we performed (section 6.6), a summary is available online under a CC4.0 license[174]. We conclude by discussing our findings (section 6.7).

6.1 Which tools are used

Most institutions use a similar set of tools to harvest social media as they do to harvest websites in general. The archiving of web content usually starts by crawling websites resulting in files mimicking the original website as Web Archive (WARC) files (Vlassenroot et al., 2019). Several so called web harvesters exist which perform such archiving activities. Most institutions use general web harvesters such as Heritrix or existing infrastructure like Archive-IT to also create social media collections (listed in Table 4).

However, social media requires specialized tools due to its dynamic nature (Vlassenroot et al., 2019). Therefore a few institutions also experiment with specialized tools such as Twarc or Social Feed Manager (SFM) which crawl data from the social media providers APIs; customized scripts are used to further process collected data or are directly used to collect data.

Table 4: Social Media Archiving tools used by some of the surveyed institutions: look and feel harvesters such as Heritrix are quite often used, API crawling tools depending on the use case.

Institution	Tool	Comment
Canada	Archive-IT and Twarc, SFM	Twarc preferred over SFM (personal choice)
Denmark	Heritrix and Twarc	Twarc was only experimental
Estonia	Heritrix, Squidwarc	
France (BnF)	Heritrix	Additionally some specific social media configurations and external scripts to

[174] Social media archiving tools comparison,https://docs.google.com/spreadsheets/u/1/d/1nGuTC9Ww5yWZQ0wSUPPnIITMJBf1JyEaDOCO0Ve-O9U/edit#gid=0.

		complete the crawls
Hungary	Webrecorder	Harvests of Instagram
France (INA)	Own tool via API	Different modalities to collect Twitter data
Luxembourg	Heritrix and Brozzler	Brozzler is way slower than Heritrix but the quality is better
New Zealand	Heritrix, Twarc and Webrecorder	For twitter hashtag crawls, for Facebook Webrecorder was unsuccessfully tested
Portugal	Heritrix, Webrecorder, Browsertrix	Depending on the use case
British Library	w3act	w3act uses Heritrix
UK National Archives	MirrorWeb (commercial), Webrecorder, Browsertrix	MirrorWeb offers data from APIs. Webrecorder and Browsertrix for Instagram but only experimental

We can identify three types of tools: 1) regular web harvesting tools such as Heritrix, 2) API-based tools such as Twarc and 3) simulated Browsers such as Webrecorder. As seen in Table 4, only one institution uses a commercial archiving service, most other institutions use the web harvester Heritrix also for social media. Half of the institutions also use API-based harvesters (own tool, Twarc and SFM) and three experiment with simulated browsers (webrecorder, Browstertrix and Brozzler). Many institutions still consider social media archiving experimental and/or keep those collections separate from the regular web archives.

An interesting perspective is that there is no distinction between websites and social media, and thus a tool selection is use case specific. The Portuguese web archive: "it is not 'social media', it's about which kind of information you get". They suggest a bottom-up approach where stakeholders such as librarians or digital humanists first determine which information they want to archive and, very important, how many resources they have to collect these information. Only then in a second phase appropriate tooling can be selected based on the use case and available resources. However, such an approach might not be in-line with the objectives of each institution, the National Library of France: "our primary mission as a library is to harvest publications and not data, what you get from APIs is really raw data".

Despite the mode of crawling, there is also the possibility of content donation. Specific events exist in which people can donate their data for archives. Such a donation can also be asked for manually identified interesting content: the National Library of New Zealand refers to cases for "caution with art work on Instagram", in which they prefer to contact the account owners to donate their data via the Instagram account download (excluding other people's comments) instead of crawling it. This, however, is a manual process for which to the best of our knowledge no tooling exists so far, the

National Library of New Zealand for example provides instruction videos with screen captures to possible data donors.

6.2 Look & Feel vs API

Web content harvested via its HTML and CSS representation preserves the look and feel, in contrast information harvested from an API usually consists of raw data in a structured format. Both harvesting methods have their merits, i.e. preserved look and feel allows browsing of archived content by regular users in original context, whereas archived structured data facilitates linguistic analysis or general post processing of archived data. Even the look and feel is subjective, the National Library of France: "The question is which [look and feel], the one you have on your desktop, the one you have on the phone, the one you have on which application?"**.** Despite how social media is harvested, it was recognized important to harvest it somehow, the National Library of New Zealand: "As long as we get it in one form or another, it's fine".

As mentioned in the previous section, the selection of a tool might be use case specific which includes the look and feel, e.g. the National Library of France: "the page layout is a bit different when you come as a robot compared to a human". Most surveyed institutions are national libraries and it is fair to assume that users of their services would prefer the original look and feel of social media archives as they are not necessarily data scientists. Collecting data from an API might also not be their objective, the National Library of France: "our primary mission as a library is to harvest publications and not data, what you get from APIs is really raw data".

Although archiving the look and feel might be preferred it does not always give good results, an alternative are customized visualizations on harvested API data. Standard web crawling tools may give good results in certain cases, the Portuguese web archive: "we got information from Twitter directly from within Heritrix and the result was very good"**.** However, getting the look and feel is not always feasible. As mentioned, the dynamic content of social media and changes in their websites and APIs led to problems in the community and motivated the creation of dedicated social media archiving tools. The National Library of France: "we have very specific configurations to harvest social media properly. It does not come out of the box of Heritrix and sometimes we have external scripts to complete the harvest, e.g. for Instagram". Possible problems are mentioned by the British Library: "sometimes there are problems with the stylesheet of the page, the background can be lacking for example". One alternative to the original look and feel is a customized visualization of harvested API data, possible for example with functionality of the tool Twarc. The Royal Danish Library could imagine such a visualization which ideally would indicate within the visualization also its provenance and the UK National Archives already uses a customized visualization of Tweets they collect; the UK National Archives only archive tweets from official accounts where they are also copyright holders.

If data is harvested with the look and feel or via API also influences data storage for preservation. If social media is harvested in a regular web harvesting fashion resulting in WARC files, they technically can be treated the same way as a web archive and be preserved with the same mechanisms. Content harvested via the APIs is often treated as separate collections. For the National Library of New Zealand: "tweet IDs are an open dataset", thus in line with Twitter terms of use, but full versions of harvested data are kept as access copies which also become the archive master record.

6.3 Privacy and Copyright

In contrast to web archiving where content within a specific domain is considered for preservation, social media may specifically concern individuals privacy or copyright. It is common to have a take-down policy but institutions are also planning to look into pseudonymization.

Some institutions archive only governmental or other official accounts for which they have the right and obligation to archive it, others also collect country-specific hashtag-based collections. In both cases personal data of other individuals may end up in collections, the UK National Archives: "if you go through the tweets you can see the screen name of the people that the government is replying to, even though it is a one sided conversation". However, most institutions have a take-down policy where right owners can request to remove their content, in most cases this will then be removed from the search index but remains in the collections for data integrity.

Another approach to increase privacy is pseudonymization, the removal of identifiable information from records. Several institutions are interested in such approaches or are using it already to a certain extent, e.g. the National Library of New Zealand replaced user IDs with a hash value: "userid is kept but anonymized for users which had fewer than 5,000 follwers at the time of suspension".

6.4 Metadata standards

Libraries and archives work with metadata ever since. Several metadata standards to describe digital archival records exist. However, up to this point only a few institutions use metadata standards to describe social media content and even less to describe the archival process, although quality research and archives require provenance metadata (Littman et al., 2018).

Several institutions store the response/request metadata as provided by WARC files, other institutions have their own customized schemas or currently explore standards and want to create bib records for OCLC, use Dublin Core, MET, EAD or ISAD(G).

Social media data harvested from the API usually contains more item level metadata which can be used to automatically populate records described with common standards. Some institutions want to look into possibilities, e.g. the National Library of New Zealand which does not yet automatically populate their EAD records from crawled data but would like to investigate it.

6.5 Missing features

Several social media archiving tools exist which facilitate the task of archiving. Asked which features are currently missing we received answers concerning both the content and the harvesting workflow.

Libraries and Archives Canada and the Royal Danish Library miss functionality to capture Facebook posts including comments. Similarly, harvesting of comments from YouTube and in general tweets with all their contents would be appreciated by Libraries and Archives Canada. It has to be noted that especially comments are a tricky subject as they may represent personal data of people other than the seed, i.e. comments of citizens on a post from a public figure. Lastly, regarding missing features, the Portuguese web archive mentioned the harvesting of video, as it is very complex with different formats and video codecs.

From a workflow perspective, the National Library of France points out that "there is no single magic tool that would solve all your problems'". Indeed, it seems that social media archiving deals with a lot of trade-offs. For instance a trade-off between high-fidelity capture and scaling was reported by the

British Library when talking about “fantastic crawling results'' with the tool Webrecorder which unfortunately took very long and do not scale well: “tooling that is a bit of a halfway house between Webrecorder and something that is scaled up a little bit more”. Another trade-off between harvesting from API and from the web was mentioned by the French Audiovisual Institute (INA) when expressing their concerns regarding fragile API harvesting in which you “can be kicked off any time” if certain rate limits are reached. On a more detailed level, both the UK National Archives and the National Library of New Zealand would like to “capture short links as part of the workflow” respectively “unshorten tweet URLs on the fly”, a technique used to get an original URL by following the usually temporarily hosted shortened URL to retrieve the original longer URL.

6.6 Tool comparison

Several social media archiving tools exist, they vary in supported social media providers, usability and functionality. Following an existing comparison of web archiving tools, we compared social media archiving tools. In a first step we collected information about the following relevant social media archiving tools: 4CAT, APIBlender, Brozzler, Instaloader, TCAT, STACKS, Social Feed Manager, Twarc and Webrecorder. Based on a first assessment with respect to how active the tool is maintained and how well it fits to our use case, we selected the following five tools for preliminary testing on which we elaborate in this section: Brozzler, Instaloader, Webrecorder, Twarc and Social Feed Manager. Additionally we present a tool developed at CENTAL with the unique feature of exploring new content to crawl.

From the tested tools, Brozzler and Webrecorder are general purpose and harvest every website, and, thus, also allow the harvest of Twitter, Instagram and Facebook. Twarc and the tool from CENTAL are Twitter-specific and Instaloader is an Instagram-specific harvester. Out of the box SFM supports several social media providers as it reuses existing API harvesters, but for our use case it only covers Twitter out of the box, however, with development effort it could be extended for other social media providers.

We compare the tested tools based on their setup, configuration and how collections are created and monitored. Detailed descriptions of the tools with respect to these dimensions are available in Appendix A.

Setup

All tested tools can be set up with minimum programming experience and/or experience with docker containers. Brozzler, Instaloader and Twarc provide a command line interface out of the box with which harvests can be started easily and quickly with a single command. SFM and Webrecorder are more complex and consist of several components, however, both provide docker images and thus the setup is theoretically also possible using docker-compose and a single command. However, if something does not work as expected debugging requires a deeper understanding.

Configuration

All tools can be configured without changing code, either via configuration files or via web interfaces.

Creating a new collection (general)

Collections can be created with all tools, for the command line based tools each harvest is an implicit collection or collections can be defined as certain output folders. For SFM and Webrecorder explicit collections exist which are stored in a database, and, thus, also explicit provenance exists.

Creating an account-based collection

With all tested tools the harvesting of accounts can be performed by providing an account or account-list via configuration file or collection user interface.

Creating a keyword-based collection

All tested tools allow the creation of keyword-based collections, either by harvesting from specific API endpoints or search URL's of keywords.

Monitoring a collection

The command line based tools usually have to be monitored manually, for instance by manually or programmatically checking their output for error messages. Scheduling of harvester can be outsourced to e.g. UNIX Cron jobs which call the scripts in certain intervals. SFM provides a list of active and/or currently running harvests, if errors occurred they are visible partially in the UI or complete in the logs created by the respective docker containers. Webrecorder performs live harvesting and directly shows the progress of harvesting to the user recording.

6.7 Discussion on tools to create social media archives

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

Combination of tools to serve different use cases

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

FAIR archives through high quality provenance

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

7. Access to and use of social media archives

7.1 Scholarly use of social media archives

Web and social media archives provide an invaluable resource for researchers to study human behaviour and history as they provide clear records of communication (Ruth & Pfeffer, 2014). Despite the massive increase in studies using social media materials as a source from self-archived and curated data sets as evidenced by the increasing reviews of literature (Cheston, Flickinger & Chisolm, 2013; Alalwan, et al., 2017; Leung, Law, Van Hoof & Buhalis, 2013; Filo, Lock & Karg, 2015; Tess, 2013) and arguably its emerging subfields (Priem, et al., 2010; Sugimoto, et al., 2017); the use of publicly accessible national archives and large scale social media archives in scientific studies are only just emerging.

This research can be categorised into two types of work that would potentially fall under a mandate for social media archives at the national level (with the exception of personal and one off archives such as the Library of Congress' acquisition of a Twitter dump (Raymond, 2010)). This includes research investigating: 1) the social media use - behaviour and strategy, of government organisations, elected officials and so forth (Acker & Kriesberg, 2017; Rabina, Cocciolo & Peet, 2013; Pal, Chandra & Vydiswaran, 2016); and 2) notable/historical moments (Rogers, 2018; Le Follic & Chouleur, 2018) e.g. a social movement (Howard et al., 2011); accidents, natural disasters (Macdonald, 2019), COVID-19 crisis (Cinelli et al., 2020; Ashrafi-rizi & Kazempour, 2020). This includes digital ethnographies and qualitative research, as well as quantitative or statistical analysis of the social media material.

The results from our survey mimic this trend, where many institutions provide one-on-one support to researchers for accessing the social media archives. A few institutions are aware of researchers using the social media data they have collected. This is partly attributed to the fact that many of these initiatives have recently launched, or are doing relatively small or specific crawls. Netarkivet, BnF, INA, KB and GWUL are aware of research being done using the collections, some in cooperation with the institutions on research projects, others as scientific papers.

Other institutions such as the NLNZ established a so-called Digital Research Working Group to develop an organisational strategy to expand support for digital research using the Library's digital collections. This includes researchers accessing digital content in the reading room, researchers viewing/browsing/downloading content or metadata from online webpages and catalogues but also large-scale computational researchers in the humanities, sciences and other fields. While this working group is not responsible for developing or providing access to collections, the group works closely with and includes representatives from other parts of the Library with these responsibilities.

7.2 Access to social media archives

Vlassenroot and authors (2018, Table 3) underlined that access conditions to web archives differ widely between institutions; when looking at access conditions to social media archives we see there are some small differences in granting access.

Table 5: Overview of access methods to the social media web archives (Vlassenroot et al., 2021 - submitted)

Country	Institution	Open & freely accessible online	Physical access on location	Requirements to obtain access
Canada	National Library	No (portal is being relaunched)	No (portal is being relaunched)	Not applicable
Canada	Regional Library	No	No	Not applicable
Denmark	Royal Danish Library	Yes	Yes	Only for researchers (at universities) for specific research projects.
Estonia	National Library	No	No	Not applicable
France	National Library	No	Yes (but also fro, within the partner libraries)	Authorized users of the BnF (proof their identity and show their need to access the BnF collection)
France	National Audiovisual Institute	No	Yes	Candidates have to demonstrate a research purpose.
Hungary	National Library	No (except for the social media pages of the library itself)	No	Not applicable
Ireland	National Library	Yes (very limited amount of social media archiving)	No	No requirements
Luxembourg	National Library	No (very limited amount of social media archiving)	Yes	No requirements
New-Zealand	National Library	Yes (some content e.g. Twitter ID’s)	Yes	Only for researchers who sign up within the reading room
Switzerland	National Library	No social media archiving	No social media archiving	No social media archiving
The Netherlands	National Library	No	No	Not applicable
The Netherlands	National Archive	No	No	Not applicable
UK	British Library	Yes (small number)	Yes	A readers pass of a Uk Legal Deposit Library is necessary.
USA	University Library	Yes (some content e.g. Twitter ID’s)	Yes	The GWU community has full access to the data, non-GWU can access ID’s only.

Table 5 shows that only a single institution (NLI) has their social media collection open and freely accessible online without any requirements for access. However, their collection of social media is very limited as they are concentrating their efforts on websites. Other institutions like Netarkivet and

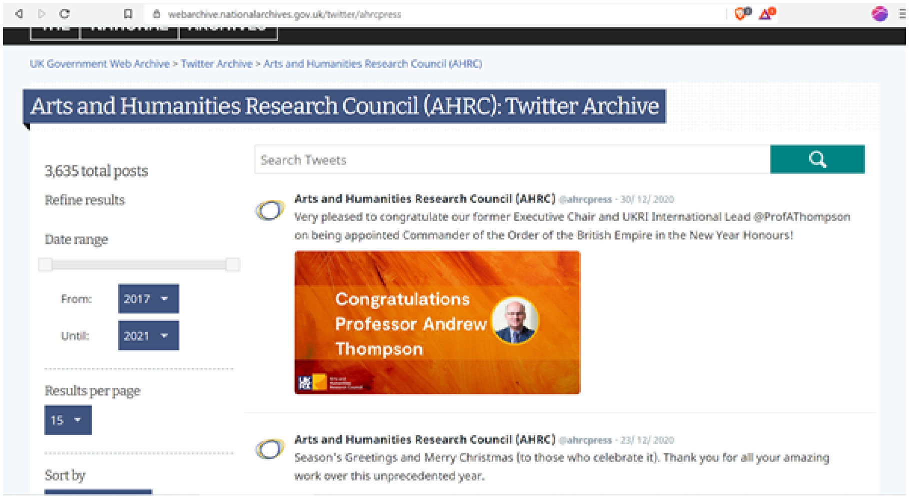
UKWA are also accessible online but some requirements are in place, such as for example being an accredited user or demonstrating a certain research purpose. Institutions such as BnF, INA and NLI only grant physical access on location to the social web archive. In most cases, the access restrictions are in place because of copyright reasons. Some institutions (NLNZ and GWUL) did find a workaround and show only publicly the metadata e.g. Twitter ID's.

A few institutions are still exploring how to grant access to these collections, e.g. LAC is planning to relaunch their discovery and access portal and is currently internally testing a new beta version of a new online interface (launch is planned in March/April 2021).

Also NSL is planning a service policy to grant access at the reading rooms. At the moment the access for the social media content at the NSL is restricted to the archive staff members, except for the social media pages of the library itself which are publicly available. To conclude we see three institutions that do not grant access to their archived social media content: BanQ, Eesti Veebiarhiiv, KB. NA and Webarchiv Schweiz are simply not collecting any social media content.

Institutions that do grant access to their archived social media content, provide a variety of access possibilities to the archived social media.

Most often, an online search interface is available that also allows thematic or topical browsing through the collection. A prototypical example of ensuring access to archived social media content in this way can be found on the website of the National Archives (UK) where tweets from Twitter profiles that are being preserved as part of the public record are published. Visitors can find archived tweets, or other social media content, from government organisations, the London 2012 Olympic games official channels and other accounts, and can consult them in such a way that the look and feel of the original Tweet is more or less preserved, including the embedded images but excluding interaction data (such as number of retweets or likes) (see Figure 1).



webarchive.nationalarchives.gov.uk/twitter/ahrcpress

UK Government Web Archive > Twitter Archive > Arts and Humanities Research Council (AHRC)

Arts and Humanities Research Council (AHRC): Twitter Archive

3,635 total posts

Refine results

Date range

From: 2017

Until: 2021

Results per page

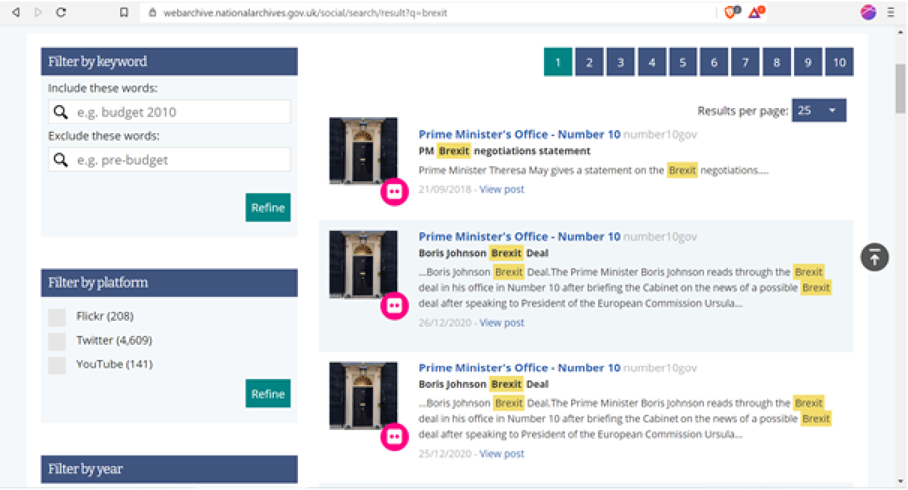
15

Sort by

Search Tweets

Arts and Humanities Research Council (AHRC) @ahrcpress · 30/ 12/ 2020
Very pleased to congratulate our former Executive Chair and UKRI International Lead @ProfAThompson on being appointed Commander of the Order of the British Empire in the New Year Honours!

Arts and Humanities Research Council (AHRC) @ahrcpress · 23/ 12/ 2020
Season's Greetings and Merry Christmas (to those who celebrate it). Thank you for all your amazing work over this unprecedented year.



Prime Minister's Office - Number 10 number10gov
PM Brexit negotiations statement
Prime Minister Theresa May gives a statement on the Brexit negotiations....
21/09/2018 - View post

Prime Minister's Office - Number 10 number10gov
Boris Johnson Brexit Deal
...Boris Johnson Brexit Deal.The Prime Minister Boris Johnson reads through the Brexit deal in his office in Number 10 after briefing the Cabinet on the news of a possible Brexit deal after speaking to President of the European Commission Ursula...
26/12/2020 - View post

Prime Minister's Office - Number 10 number10gov
Boris Johnson Brexit Deal
...Boris Johnson Brexit Deal.The Prime Minister Boris Johnson reads through the Brexit deal in his office in Number 10 after briefing the Cabinet on the news of a possible Brexit deal after speaking to President of the European Commission Ursula...
25/12/2020 - View post

Figure 1: chronological presentation of Tweets & overview of search results across different social media channels for 'brexit' https://webarchive.nationalarchives.gov.uk/.

Another approach to providing access to archived social media content is via so-called labs. For example, KBR's 'lab' the KBR Digital Research Lab works to facilitate data-level access to KBR's digitised and born-digital collections for digital humanities research. It supports the digital access of textual sources and stimulates the (re)use and research of these digital sources, often by going in direct one-to-one interaction with the researcher. Archived social media can be published as (FAIR) datasets within these 'lab' environments, therefore increasing the visibility of these datasets of archives social media content and potentially their take-up and usage. Offering data-level access means that these labs provide access to the underlying files (often in warc- or json-format), enabling a fine-grained level of access which facilitates data analysis. Typically these labs make a number of data and API services available for general use. Often also, a few example tools or scripts are made available, showing how the data might be used. For example, BnF is developing their data lab in their research library; a physical place within the reading room where a researcher will be able to come and work on data services and where he/she can request data-level access for research purposes.

Sometimes, archived social media content is made available via publicly available scripts. For example, in the case of the GLAM workbench (Sherratt & Jackson, 2020), resources shared as Jupyter notebooks are made available. These notebooks focus on data that is readily accessible and able to be used without the need for special equipment. They use existing APIs to get data in manageable chunks and are available for four particular web archives: the UK Web Archive, the Australian Web Archive (National Library of Australia), the New Zealand Web Archive (National Library of New Zealand), and the Internet Archive. Also, these tools and approaches can be easily extended to other web archives. Institutions such as Arquivo.pt have their own API that enables the refinement of queries (e.g. by special collection).

A final approach to providing access to archived social media content uses applications or dashboards to disclose (most-often) processed social media data. One example is the dashboard https://tweets.covid19misinfo.org provides information on the Covid19 discourse and on the Twitter bots active on this topic. Another example is the website http://www.politiekebarometer.be that tracks the number of tweets (and their sentiment) of Flemish political parties and politicians.

8. Analysis of preservation policies

Digital preservation is constantly developing on a practical and academic level. Within BESOCIAL, the aim of *Task 1.4: Analysis of preservation policies* is to provide an overview of the current status of preservation policies for social media content internationally. This analysis lays the groundwork for *Task 3.3* in which a preservation plan will be developed for archiving social media content at KBR. This report is structured in five sub-sections: 1) definition of concepts, 2) limitations, challenges and potential solutions, 3) institutions surveyed, 4) the analysis of the results, 5) preservation formats and 6) the conclusions accompanied by recommendations for social media archiving at KBR.

8.1 Definition of concepts

The survey and follow-up interviews carried out in the framework of *Task 1.1. Analysis of selection and access policies,* highlighted the importance of clearly defining all concepts related to web and social media archiving, especially those related to digital preservation. During our research it became clear that there is currently no real consensus on the meaning of concepts used, which could lead to misunderstandings that may be detrimental to the quality of the work and could also hinder inter-operational cooperation. This lack of harmonisation of definitions is also noted by the Digital Preservation Coalition, which in its glossary explains that the 'Digital Archiving' term is used in very different ways in different sectors (Digital Preservation Coalition, n.d. Glossary). The aim of this section of the report is to provide a clear definition of the concept of digital preservation which will be considered throughout this report.

The UNESCO website offers a practical definition of digital preservation: *"Digital preservation consists of the processes aimed at ensuring the continued accessibility of digital materials. […] Digital preservation can be seen as all those processes aimed at ensuring the continuity of digital heritage materials for as long as they are needed."* (UNESCO, n.d.)

This is the definition used in this report. The aim of digital preservation is to ensure the accessibility of content, which implies a reflection, among others, on the means of access to data and the way in which they are stored. UNESCO recommends, in particular, to reflect on the selection of the content to be preserved, on the inspection of archived material and on storage locations. *(ibid.*)

Furthermore, Bodleian Libraries add that digital libraries involve "*policies, planning, resource allocation (funds, time, people) and appropriate technologies and actions to ensure accessibility, accurate rendering and authenticity of digital objectives"* (Oxford LibGuides, 2020). They also distinguish between two types of digital preservation: bit-level preservation and logical preservation, as visualised in Figure 2 below. Both are intrinsically linked. Bit-level preservation involves a very basic preservation of the content, as captured. It may involve keeping copies, checking for viruses and updating storage media (Digital Preservation Coalition, n.d.). It is not a digital preservation mechanism per se, but it is essential for logical preservation, as it ensures the survival of the original digital material. Logical preservation (or format preservation, or active preservation) will ensure the durability of the digital objects over time. It consists of three stages: characterisation of the data (definition and understanding of the contents), planning of their use (identification of threats, planning of actions to be taken to counter these risks) and action on them (Oxford LibGuides, 2020).

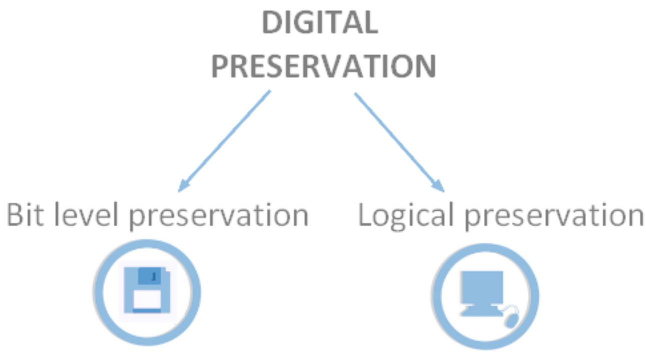


Figure 2: Two types of Digital Preservation

8.2 Limitations, challenges and potential solutions

Numerous studies and technical documents, published for more than a decade now, specify the limitations and challenges related to digital preservation. The list of challenges and potential solutions presented in Table 6 below is based on the guidelines of the Digital Preservation Coalition (Digital Preservation Coalition, n.d. Preservation issues). However, it is not exhaustive and focuses on the issues that are or could be encountered, in the short or long term, at KBR, at the level of archived data:

Table 6: Digital preservation challenges and potential solutions for KBR

1. **Preservation of data and their meaning:** archived data must be kept over the long term, without loss or damage	
Challenges	**Potential solutions**
- Media or format obsolescence. - Formats or media not supported by current computer systems or software - Files corrupted or damaged. - Communication problems, related or not to the network (Nilesh & Verma, 2012). - Accidental or unintentional deletion or modification of archived content. - Presence of viruses and malware (Nilesh & Verma, 2012).	- Inspection and update of storage media. - Multiple backups, in different geographical locations (**replication**). - Determining mechanisms for automatic control of file integrity. - Use **migration** from a format to another if necessary. - Use pre-defined **emulation** processes for obsolete software.
2. **Maintaining user trust in the data** (Merwood, 2020).	
- Risk of loss of authenticity or integrity of original data due to computer or human issues	- Availability of the complete history of data archiving and all manipulations carried out on the data (transparency).

(Merwood, 2000).	- Preservation of the original data in the form of a non-manipulable back-up copy (Nilesh & Verma, 2012). - Use of current best practices
3. Preservation of the data context	
- Loss of the context in which the original data were created or used. - Absence of established preservation standards, protocols and methods (Nilesh & Verma, 2012).	- Identification and capture of any necessary contextual information.
4. Digital expansion	
- Need to cope with the amount of data available (vs. technological and human limitations)	- Implementation of a well-defined selection process. - Consideration of technical constraints (e.g. storage) and their cost.

Digital preservation also implies other issues related to institutional organisation. The question of whether collection management is undertaken internally or is outsourced is frequently raised in the literature, as it involves quite different institutional management. This question also refers to the roles and responsibilities of the depositary institution, but also of any individual who will be confronted with these data. In this sense, some institutions, such as the National Library of New Zealand, have drawn up a charter for all users who may be confronted by this kind of archived data. This document *"codifies high-level expectations and responsibilities under which activities should be discharged by members of the community"* and is applicable to *"anyone who plays a role in preserving digital material"*, such as researchers, teachers of digital preservation, or developers (National Library of New Zealand, 2018).

Another aspect not to be neglected is resources, which refers both to the costs of preservation and the skills of the individuals responsible for it. Other challenges, directly related to laws and other legal aspects should also be mentioned. These aspects will be developed in *Task 1.2. on the analysis of the legal framework in Belgium and abroad*.

Furthermore, it is important to mention here that the terminology used to characterise the different documents related to digital preservation is often unclear. The terms 'digital preservation strategy', 'digital preservation policy' and 'digital preservation plan' refer to different realities but are used in different ways depending on the institution. The use of one name rather than another is often linked to the presence of certain types of documents within an institution. For example, if an institution has already developed documents relating to preservation strategies, regardless of what these cover, it is very likely that digital preservation documents will also be grouped under the heading 'strategies'. In order to make the best choice for KBR about terminology, we refer here to the UK National Archives' website (UK National Archives, n.d.) which specifies the use of digital preservation policies and strategies, and to the Government of Canada’s website (Government of Canada, 2020) which develops the issue of the digital preservation plan.

In general, the prevention and limitation of damage requires accurate and up-to-date documentation of collections and technical procedures. Among these documents, a clear digital preservation policy and a preservation plan are most important. In addition, an ethical charter and a preservation strategy may also be useful for KBR.

8.3 Institutions surveyed

As outlined in *Section 3* above, the social media archiving practices in various European and international institutions we examined through a survey (see Table 1) and follow-up interviews (see Table 2). Dedicated questions related to digital preservation practices for both web and social media archiving were included. For preservation policies specifically, an in-depth analysis of the websites of each institution was carried out, as well as an in-depth literature review. On the basis of this, we found additional institutions that are of particular interest in relation to preservation policies because of their inclusion in the literature or because other institutions referred to them (see Table 7). For each institution, the existence of specific preservation policies for digital content was identified. In order to prepare *Task 3.3 Development of a preservation plan for archived social media*, preservation strategies and plans were also identified. The literature related to the preservation of digital collections was also examined as exhaustively as possible. However, in general, it can be noted that these institutions do not systematically make this documentation available. In cases where these documents are not available online, it is important to consider whether they exist as documents for internal use within the institution.

The additional institutions of interest in relation to their preservation policies are:

Table 7: Additional institutions of interest related to preservation policies

Country	Institution	Abbreviation
Australia	National Library of Australia	NLA
Greece	Athens University of Economics and Business	AUEB
Japan	National Diet Library	NDL
United States of America	University Libraries, University of Washington	ULUW

8.4 Analysis of the results

8.3.1. Australia

a) National Library of Australia

The National Library of Australia launched its web archiving project, PANDORA, in 1996. It has been included in the Trove website since 2019 (Pandora. Australia's Web Archive, n.d.). In March 2014, the Australian Government Web Archive project also began. However, the various websites do not mention social media.

The PANDORA project website mentions that the content of the archive is preserved according to the rules defined by the National Library. The aim is to preserve, as far as possible and using best practices, the look-and-feel, as well as the content and titles (Pandora. Australia's Web Archive, 2020. Policy and practice statement.).

The National Library of Australia also has a Digital Preservation Policy, which is frequently mentioned in the literature as an example. Social media are not mentioned in the policy, but they may be included depending on the definition of collections in point 3. This document specifies the challenges of resource accessibility (point 4) and the institution's preservation policy (points 5 and 6). It is also specified that the National Library of Australia uses the OAIS system and is based on international standards and practices advised, for example, by PREMIS and the OPF. (National Library of Australia, 2013a) This document, although written in 2013, is a good working basis for the definition of the digital preservation policy at KBR.

The web page related to content selection and preservation also specifies that digital content is systematically associated with its metadata. Content, connections and context are of primary importance, and no changes are made to the original copy. It also mentions the formats and documents for which archiving is problematic: RealMedia, VRML, Shockwave and Quicktime VR documents. (National Library of Australia, 2013b)

8.3.2. Canada

a) Library and Archives Canada

The Library and Archives Canada (LAC) *Stewardship Policy Framework* (Library and Archives Canada, 2014) states the importance of long-term preservation, also for digital documents. Similarly, in their *Guidelines of file formats for transferring information resources of enduring values* (Library and Archives Canada, 2015), the LAC also specifies their use of international best practices for digital preservation. They also draw on their own experience in order to define the best practices to follow. This document also presents the 'preferred' and 'acceptable' formats for each type of content. In the survey results, the LAC states that they keep documents in the form of WARC files (limited holdings in ARC up to 2011, the rest in WARC and derivatives for a total of approx. 75 terabytes). These files are included in their thematic web collections, in which they are organised, described and preserved on tape. Social media is considered as part of the web archive. It is harvested in WARCs and integrated in the thematic web archive collections.

Currently LAC is working on developing a suit of next generation tools that is largely based on Preservica. There are already procedures in place for Digital Preservation (DP), however they are largely manual with limited automation. Concerning preservation LAC largely depends on technical infrastructure which comprises of servers linked by fibre, that lead to LTO6 taping using Commvault for the actual DP copies. They have two copies of the entire archive of all digital at LAC on tape which currently comprises around 14 petabytes.

In addition, LAC follows the international OAIS model and has planned a schedule for the development of a digital preservation strategy (Library and Archives Canada, 2017). Progress in digital preservation in the institution is described in a short article on the Digital Preservation Coalition blog (Lemay, 2019), but nothing is specifically written about social media. Although the focus is on digital preservation, social media are not currently considered in the available literature and documentation. However, in the survey results, LAC indicates that the development of a specific methodology for social media preservation is on its way.

b) Bibliothèque et Archives nationales du Québec

The Bibliothèque et Archives nationales du Québec (BAnQ)'s website provides a broad presentation of their collections and digital resources (Bibliothèque et Archives nationales du Québec, n.d.). Documents related to policies and procedures are available online. However, no documents related to digital preservation have been found there.

The preservation policy for heritage collections (Bibliothèque et Archives nationales du Québec, 2013) explicitly states that it does not concern digital content. Furthermore, the development policy of the universal collection explicitly states, in its point 3 concerning the scope of action: *« La Collection universelle de BAnQ comprend les ressources documentaires de toute nature et sur tous les supports »*. This definition is followed by a list of potential content, mentioned as non-exhaustive. Digital resources and documents are mentioned, without any information about websites or social media.

The 2016-2018 strategic plan (Bibliothèque et Archives nationales du Québec, 2016) mentions the importance of focusing on digital content, but does not present any practical strategy.

8.3.3. Denmark

a) Det Kongelige Bibliotek

The preservation policy related to Netarkivet (Netarkivet.dk, 2014a) does not mention social media content as such as material preserved by the institution. However, this kind of content can be included in the *"material collected from portals, where material is provided via other interfaces than web interfaces"* category, although the focus here is more on FTP interfaces. This document advocates the use of international standards and the preservation of metadata related to content. Netarchive also makes use of the bit preservation and uses a quality control system.

In the survey results, Netarkivet indicates using their own bit archive (including replicas, checksum replicas, etc). The social media data is preserved along with other data in WARC files. These files contain either API-retrieved JSON and resource files, or contain "standard" Web archiving files as harvested by the Heritrix web crawler.

Social media content is also absent in Netarkivet's preservation strategy (Netarkivet.dk, 2014b). This document specifies the information provided in the preservation policy, mentioning in particular the international standards used: ISO 14721, ISO 14873, ISO 28500, ISO 16363:2012 and DS/ISO/IEC 27 001. This document recommends to follow the international standards, and mentions that all data must be kept in their original format. The data are kept uncompressed to avoid loss of data. In addition, two replicas of the archived data are kept as well as a back-up copy. No virus or malware checks are done in this data, which is kept in its entirety by the National Repository Software.

During the interview that was conducted with Netarkivet, the lack of standardisation was put forward as one of the biggest challenges in the digital preservation field: "*The fact that methods, API's, etc. change so often that we end up having to maintain collection, preservation and access procedures for a very large amount of different social media data.*"

8.3.4. Estonia

a) National Library of Estonia

The National Library of Estonia's website, in its English version at least, includes very few details and only mentions that the institution's main objective is the preservation of digital content (National Library of Estonia, n.d.). The Eesti Veebiarhiiv website states that websites related to cultural heritage are archived via Heritrix and can be accessed via Wayback. It also mentions that audio-video streaming and the contents of scripts are difficult to archive (Eesti Veebiarhiiv, n.d.). There is no mention of social media. As the search engine of the National Library of Estonia can exclusively be used in Estonian it was not possible to find relevant information.

However, according to the Sirp website, an Estonian culture-oriented newspaper, the Estonian archive is currently collecting information related to the Coronavirus: social media groups, portals, blogs and YouTube videos related to Estonian life during the pandemic (Sirp, 2020). Social media therefore does not seem to have been forgotten in the Estonian collections, but no information about their preservation could be found. The Estonian laws on preservation (Riigi Teataja, 2017) and laws related to the Estonian National Library (Riigi Teataja, 2016) do not mention the preservation of digital content either.

8.3.5. France

a) Bibliothèque nationale de France

According to Bermes and Fauduet (2009), the Bibliothèque nationale de France has been working on the issue of digital preservation since 2003, five years after the creation of their digital department. The institution has a general preservation policy (Bibliothèque nationale de France, n.d., Politique de conservation), supported by a preservation charter and a digital preservation policy (Bibliothèque nationale de France, n.d., SPAR). All three are available online on the institution's website. The first document informs us of the application of the AFNOR/CN 46-10 preservation standards for documentary collections, linked to the TC 46/SC 10 and AFNOR/CNCBC standards for the preservation of cultural property. The digital preservation policy specifies the use of SPAR, the Distributed Archiving Preservation System, based on the OAIS principles (ISO-14721) and which aims to guarantee the long-term preservation of digital In their survey response, the institution specified that this preservation policy is similar for the archiving of social media and web archiving. The Bibliothèque nationale de France uses WARC files for preservation.

The SPAR (Système de Préservation et d'Archivage Réparti) system aims to perform digital archiving operations using, in parallel, multiple copies of documents to avoid loss and destruction. However, no document mentions the case of social media, since the procedure for harvesting social media falls under the web content harvesting procedure. It can be assumed that the place of these in the functional diagram of SPAR is linked to the "Web Archiving" part.

b) Institut national de l'Audiovisuel

At the Institut national de l'Audiovisuel (INA), format obsolescence is a very present issue, since the institution is mainly in charge of safeguarding and digitising data from television and radio. From 2012, the choice of a 'digital migration' was made, using in particular the JPEG format for professional archives (Institut national de l'audiovisuel, n.d.). JPEG 2000 is becoming the central format for archiving and digitisation. The INA does not have a digital preservation policy online. There is no document mentioning the preservation of social media.

8.3.6. Greece

a) Athens University of Economics and Business

The Athens University Library of Economics and Business and the DB-net team began web archiving in February 2010, focusing on the archiving of 78 websites (Web Archive of Athens University of Economics and Business, 2018). However, their site does not mention any information related to content preservation. The literature consulted did not provide more details on this point.

8.3.7. Hungary

a) National Széchényi Library

The Budapest Library launched a pilot project for web archiving in 2017, which within a year became the OSZK Web Archive project, which is still active (Németh, 2020). The project website mentions the preservation of born-digital data since the early 2000s, but web archiving was not considered until 2006 (OSZK Archívum, n.d., About this website). By browsing the archive, we can see that some Twitter, Instagram (mostly), Tumblr and Pinterest accounts are archived using Webrecorder (OSZK Archìvum, n.d., Archive of the National Széchényi Library's website). A page also mentions that some Instagram pages have been archived annually since February 1, 2020 (OSZK Archìvum, n.d., Thematic and genre-based harvests).

The library's website mentions a brief history of the project, but does not mention information about the preservation of digital data. The same can be said for the OSZK Web Archive website.

8.3.8. Ireland

a) National Library of Ireland

Everything related to the preservation of the contents of the National Library of Ireland refers to the Digital Repository of Ireland (DRI). The DRI provides access to documents relating to their collections, web archiving and digital preservation. Thus, the Q&A on web archiving (National Library of Ireland, n.d.) specifies that this archiving concerns websites and social media accounts, with the exception of Facebook, social media feeds and hashtags. The document also specifies that the institution archives Twitter, Instagram and YouTube accounts if they are publicly accessible, free and Irish or related to Ireland.

The DRI collection policy (Digital Repository of Ireland, 2015) specifies the type of data that may be included in the DRI collection. These are digitised or born-digital documents of social, historical, political, scientific or economic interest related to Irish culture. These documents must be accompanied by their metadata written in Irish or in English. The formats recommended by DRI include PDF/A, RTF, TXT, XML, WAV, BWAV and TIFF. The DRI also recommends that the metadata originally present with the original file be used and systematically saved. However, it encourages the creation of new metadata in order to facilitate the recovery and promotion of archived content. It also advocates the use of a standardised vocabulary (Digital Repository of Ireland, 2013).

The document entitled *Long-term digital preservation* (Digital Repository of Ireland, 2014), also made available by the DRI, briefly outlines the risks associated with the preservation of digital data. The DRI has set up short-term activities to ensure access to digital content. These activities consist of a series of validation processes that ensure the integrity of the archive, but also possible updates or migrations. Backup strategies are also considered, as is the use of DOIs (digital object identifiers). The long-term strategies developed by the DRI also imply the development of skills and the continuity of the institution's management, in terms of infrastructure, personnel, funds, etc. There is no mention of the preservation of social media.

8.3.9. Japan

a) National Diet Library

The National Diet Library launched a web archiving project in 2010, but it archived websites dating back to 2002 (National Diet Library, Web Archiving Project, 2014a). The project, called WARP, is explained in great detail on its website. It states that the institution archives websites of national institutions, prefectures, cities, municipalities, councils, administrative societies, universities, electronic magazines, etc. Social media are not mentioned (National Diet Library, Web Archiving Project, n.d.). The project statistics inform us of the total amount of data collected, as well as the number of data collected by type of formats. The image format represents 36.53% of the archived content, followed by HTML (21.25%) and PDF (26.25%) (National Diet Library, Web Archiving Project, 2020).

However, the institution reports long-term storage practices, as well as the distinction between data retention and logical recording. The first term refers to redundant storage using RAID technology, as well as the division of storage space. Logical storage refers to format migration and emulation for replication of the environment (National Diet Library, Web Archiving Project, 2014b).

8.3.10. Luxembourg

a) National Library of Luxembourg

The National Library of Luxembourg reported in the survey that they are using a digital preservation system in which WARC files are bit-preserved on physically distinct locations. After this step, Payloads inside WARC files are characterised and file types determined, but they are not migrated. Replay systems are kept up to date to enable higher fidelity. Nevertheless, the institution's website does not make available information on the preservation of collection. Currently, the website only focuses on a presentation of the Webarchive.lu project, which archives the Luxembourgish web. Nor does it mention any indications about preservation policies. Preservation information is implied by a reference to the websites of the IIPC, DPC, Internet Archive and WARCnet. The website does, however, mention the establishment of a shared platform for long-term digital preservation in cooperation with the State Information Technology Center and the National Archives (Bibliothèque nationale du Luxembourg, 2020).

Most of the collections of the National Library of Luxembourg are captured with Archive-It or by the Internet Archive, which means that the archive is stored either at the National Library or at Internet Archive. Their web archive, which also includes their social media mainly resulting from event crawls, is ingested into Preservica, as part of their long-term preservation programme.

8.3.11. New Zealand

a) National Library of New Zealand

The National Library of New Zealand's website briefly presents the New Zealand Web Archive project, launched in 2008 and whose results are available in the library's catalogue. It also mentions that the institution preserves websites, blogs and videos related to New Zealand culture. Social media is not discussed (National Library of New Zealand, n.d.). In 2018, a short brochure on digital preservation in the institution states that the Web harvest represents 1.57% of the current collection, without mentioning whether or not social media are present in the collection (Sajwan, Goethals & Wu, 2018). MacDonald (2019), mentions that the library began capturing social media (Twitter and Facebook) when it noticed that election campaigns were moving from websites to social media. Steve Knight's team worked on capturing tweets following specific events, such as the elections, the Kaikoura earthquake or the attack on the Christchurch mosque. This data should be in the Alexander Turnbull collection. The library's next project is to archive 100 Facebook profiles of citizens to see how New Zealanders are using social media.

The National Library of New Zealand makes available a wide range of documentation related to its collections, particularly the digital ones. Their *Digital preservation program* (National Library of New Zealand, n.d., Digital preservation programme) specifies that the digital or born-digital content is kept at the library's digital preservation repository, the National Digital Heritage Archive (NDHA) which was established in 2008. Digital materials collected by the library include websites, digital publications, e-books, photographs, cartoons, music, videos, oral history and email. The reference model used is the OAIS system.

The *Digital preservation strategy* (National Library of New Zealand, 2011) establishes the theoretical principles related to digital preservation, which will then allow to specify the preservation policy and concrete actions. A very interesting document for the BESOCIAL project at KBR is the *Digital preservation policy manual* (National Library of New Zealand, 2012) distributed by the National Library of New Zealand. Although this document is a work-in-progress, it already includes the necessary policies in terms of technical aspects, anti-virus, data backup and risk management. A set of operating rules is listed. The Preservation Management Policy specifies the rules and principles necessary for the preservation of digital content. It details four basic principles: preservation actions must not affect original objects, the importance of copying, the integrity of digital content must be preserved as well as its authenticity. Preservation actions must also be transparent. Twelve operating rules are then enacted. Although it is already eight years old, this document represents an excellent working basis for KBR. It is completed by an ethical charter (National Library of New Zealand, 2018), published for anyone involved in the preservation of digital content. In addition, in their survey response, the institution specified the use of specific actions to ensure the preservation of social media, such as the conversion to long version of reduced URLs in tweets and the capture of content that has been linked to publications.

Among its digital projects, the New Zealand institution has set up a Bulk Ingest Pipeline project to rework the import of digital data, which is arriving in ever-increasing quantities. It also plans to work on a new version of the NDHA and works closely with the National Library of the Netherlands on the Web Curator Tool (WCT), which is used for web archiving programs in both institutions.

In the survey, The National Library of New Zealand indicated that they use Rosetta as a system for preserving their online content. Additionally, the Library also has some content (inter alia WARC file format) that is temporarily stored in Archive-It. In the future they plan to copy this data to their own preservation repository.

On the basis of the consulted documents and the survey results, it seems that the National Library of New Zealand is the institution that currently presents the greatest progress in terms of strategies and policies for the preservation of digital data. One of their biggest challenges is getting complex data collections into formats and data structures that are understandable and usable for researchers and also preservation-friendly. Although their social media collections are treated the same as their other collections, it is an excellent source of inspiration for the BESOCIAL project.

8.3.12. Portugal

a) Arquivo.pt

Arquivo.pt does not publish a preservation policy or strategy on its website. However, digital preservation seems to be a concern to the institution. On their website a list of recommendations is made available regarding the preservation of published data (Arquivo.pt, 2017). These are tips for current websites, with a view to ensuring the preservation of their content over time. Some considerations also seem to be linked to a long-term preservation vision, notably the importance of attributing to each content a specific URL that is persistent over time. In addition, Arquivo.pt provides training modules in Portuguese on how the website works, on publishing content for preservation, on automatic processing via APIs, and on website preservation (Arquivo.pt, 2020). During the interview, it was noted that the archiving of social media is not Arquivo.pt's focus, however, they do make their best effort within the scope of their web-archiving activities. Furthermore, they emphasised that the web was intended as a social platform from the beginning, with many websites including interactive features. However, they also noted that social media is starting to migrate to apps outside of the web.

Arquivo.pt's digital preservation infrastructure operates on different levels. Firstly, capturing the websites before they vanish. Secondly, copying the data to two geographically distant nodes. Finally, ensuring that the information is accessible. Without access the information is dead. Use of the web archive enables both quality assurance and helps to sustainability of the services. Without users the service would be shut down. As preservation standards Arquivo.pt makes use of WARC files. Besides that they do not comply with specific standards -because of the limited resources -, but only if they are not in-line with the internal workflow used. The institution does support the Memento protocol for their patching feature.

8.3.13. The Netherlands

a) Koninklijke Bibliotheek

The website of the 'Koninklijke Bibliotheek' in The Hague currently mentions the work done on web archiving, without mentioning the work on social media archiving. However, the list of websites archived in June 2019 mentions the institution's Facebook page as an archived item. This is the only example mentioned.

In terms of preservation, the Koninklijke Bibliotheek mentions on its website that it works closely with the IIPC (National Library of The Netherlands, n.d.), but no other information could be found.

b) Nationaal Archief

The preservation of digital content at the 'Nationaal Archief' in the Netherlands seems to have been widely considered since 2015, with the drafting of a preservation policy based on the OAIS model. However, no information about social media could be found.

The model used by the National Archives is therefore based on the OAIS model. Their preservation policy (National Archives of the Netherlands, 2015) states : *"the NA also applies an open standard to opening up and providing access to digital archive documents (EAD). The relationship between the substantive metadata and the digital file will be guaranteed by means of a unique identifier"*. It also emphasises the importance of metadata, which must be linked to its original content. Articles 21 to 26 refer to generalities related to the preservation of digital content, without specifying the type of data to which this refers.

Furthermore, the National Archives have no limitations in terms of numbers or types of formats for the data deposited in the e-depository. However, the institution has chosen to limit the number of formats to open standards, in the interests of long-term preservation. The distinction between data types is as follows: audio, database, text document, image, presentation, spreadsheet, vectorized image and video (National Archives of the Netherlands, 2016).

8.3.14. United Kingdom

a) British Library

The British Library publishes many documents for the public on its website, including all of their policies and procedures (British Library, n.d., Our policies and procedures).

A document on legal deposit, entitled *Joint Collecting framework for UK legal deposit 2015-2020* (British Library et al., n.d.), mentions the aims of preservation of digital content (websites, e-books and e-newspapers), but does not specify any information on social media as such. A practical document provides some brief additional information on the preservation of digital content, stating: "*For born digital material it is not possible to distinguish between retention and preservation. Preservation of the intellectual content must begin at the time of acquisition, or before in order to ensure sustainable access. The lack of longevity of storage media, together with the inevitable obsolescence of retrieval hardware and software pose significant digital preservation challenges"* (British Library. Preservation Advisory Centre, 2013). This document refers in a note to the Digital Preservation Handbook of the Digital Preservation Coalition. The latter is also mentioned when the issue of storage of digital items is addressed.

Furthermore, no document mentions a digital preservation policy. Nevertheless, a preservation strategy document, entitled Sustaining the value. *The British Library Digital Preservation Strategy 2017-2020* (British Library, 2017) presents the action plan for this period, defining the objectives and action strategies for the preservation of digital material. According to this document, issues related to preservation policies will be reflected upon as early as 2017-2018, but no record of these are available on their website. The determination of the preservation policy for digital material therefore seems to have remained at the draft stage, or has simply not yet been published. To date, the British Library's website mentions a digital preservation strategy, but no digital policy or action plan.

In the survey conducted in autumn 2020 the British Library indicated that they collect their content with Heritrix and store everything according to preservation standards in WARC files. As the library does not distinguish between social media archiving, their archiving procedure is the same for both.

b) The National Archives

Like the British Library, the National Archives of the United Kingdom publishes on its website numerous documents for the public and professionals, including all its policies (The UK National Archives, n.d. Our policies) and a list of recommended formats (The UK National Archives, n.d., File formats for transfer). Their preservation policy (The UK National Archives, 2018) states that the institution preserves both physical and digital material. Digital collections (without specifying what exactly this term covers) are kept in the Digital Records Infrastructure (DRI), to which access is regulated. Public access to digital content is provided through Discovery, an online catalogue. The preservation policy states: *"they [= digital content] are available either in their original, or, where possible, in a more accessible format"*.

Another document presenting the institution's digital strategy mentions the specificity of digital content and underlines the preservation issues it creates : "there is no long-term solution to the challenge of digital preservation" (The UK National Archives, 2017a). At the National Archives, the focus is on images, documents (Word, Excel), e-mails, video and mixed data (including websites and Twitter). This is one of the few explicit references to the presence of social media in institutional collections. The document, entitled Digital Strategy, highlights the strategies that will be implemented in the institution, suggesting that this has not yet been developed. However, this strategy envisages the development of many tools for digital preservation, and also foresees the preservation of digital recordings in their original format, using emulation for content that has become obsolete. This document also specifies the importance of informing users of changes made to the original document, for example when an email is transferred to PDF format. The contextualization of digital recording is also of particular interest.

The presentation document of the UK Government Web Archive (UKGWA) project, published in 2017, outlines the project and presents a technical guide to web archiving (The UK National Archives, 2017b). This document provides thoughts on how social media can be added to the archive. It states again that the content of social media such as Twitter, Facebook, FlickR and YouTube are a real challenge for web archiving (The UK Archives, 2017b, p. 12). In the interview conducted in February 2021 the National Archives mentioned the use of the tool Preservica for their social media data with especially a focus on file transfer.

In general, the preservation policies proposed by the National Archives of the United Kingdom are developed with precision. The few mentions of social media suggest both their inclusion in their digital collections but also their consideration in the drafting of preservation policies. The published documents take into account the description of the issues, the long-term vision, using a precision that goes beyond the description of simple objectives for action plans. However, at present, no specific information on the preservation of archived content from social media could be found.

8.3.15. United States of America

a) George Washington University Libraries

These libraries' website mentions the existence of a department related to the preservation of documents (George Washington Libraries, n.d., Preservation). The Web Archiving Program collects websites and web content related to the libraries' collections and is available on the Wayback Machine and the Archive-It website since 2015 (George Washington Libraries & Academic innovation, n.d.).

Their *Digital Stewardship Program* (*ibid.)* is an initiative that ensures the long-term stewardship of digital materials created by institutions, including academic publications by students and faculty members and specialized cultural heritage collections. The digital content is included in a strategy for its long-term preservation. Several levels of preservation services (entitled Tier 0, 1 or 2) are considered by the library, depending on the source, format, historical value and access restrictions associated with the material (George Washington Libraries, n.d., Digital Stewardship). However, more specific information on preservation policies could not be found at this moment.

b) University of Washington, University Libraries

This library's *Digital Preservation Policy* (University of Washington, University Libraries, 2017) is relatively succinct but it effectively summarises the various points to keep in mind when writing such a document. It presents eight principles necessary for digital preservation: standards-based, high-quality metadata, technically robust, authenticity, collaboration (and community-minded processes), legal compliance, currency and sustainability.

The institution also has a list of preferred formats for archiving and preservation. However, social media are not presented as such (University of Washington, University Libraries, n.d.).

8.3.16. Switzerland

a) National Library of Switzerland

The digital collections to Swiss National Library concern born-digital collections and digital data. The institution's website specifies the content of the born digital collections: these are academic digital publications, commercial digital publications, official digital publication of the Confederation and websites. Social media are not mentioned (Swiss National Library, 2018)

No document related to digital preservation made available by the institution informs us of the practices currently implemented.[175] The web page entitled Basic principles (preservation) explains that the digital archives are preserved on the basis of the OAIS model, which became an ISO standard in 2002. It also specifies that copies are used for documents that are too fragile (Swiss National Library, 2020).

On the other hand, a study conducted by D. Burda in 2017 provides further details on the Swiss institution's long-term view of digital preservation. The aim here is a reflection on the long-term preservation of digital date. The emphasis is made on the importance of collaboration and global coordination. The article underlines the importance of preserving digital information. It also highlights the presence, at present, of unresolved issues, in particular related to legality, responsibility for preservation, etc. The article also emphasizes the importance of the preservation of digital information. It provides an overview of current and planned activities for long-term preservation in Swiss institutions and is, to date, the most detailed document found on the issue of digital preservation.

8.5 Consideration of preservation formats

A data format, in computing, refers to the way in which a type of data is encoded as a sequence of binary elements (or bits). It defines the structure and type of data stored in a specific file, and in particular allows interoperability between software programs (TechTerms, 2011).

At present, there is no commonly accepted standard data format for social media archiving (Naets, 2018). Furthermore, social media platforms have not published internal preservation plans (Thomson, 2016). To date, and similarly to all digital collections (images, texts, videos, etc.), institutions decide, on a relatively adhoc basis, on the format used for the preservation of digital content. At best, institutions publish lists of recommended formats for digital preservation in their institution, see for example, the University of Washington or the National Archives of the United Kingdom. However, it is important to note that none of the list of formats consulted for this report refer directly to social media. Overall, institutions do not yet mention specific formats for the preservation of archived social media content. The content is systematically separated per document type, e.g. images, videos, texts, audio content, etc. It is therefore also interesting to consider the technical aspects of social media archiving in order to determine whether there is an ideal format for the preservation of this specific type of content.

Based on the literature (e.g. Naets, 2018) distinguishes three formats that could be used for social media preservation: JSON, CSV and TEI formats. The CSV format allows data to be kept as a text file, without embedded data. The TEI format comes from the consortium of the same name and is already used for the preservation of social media data. Currently, Twitter delivers data in JSON formats through APIs, while other platforms use XML format instead (Thomson, 2016). Unlike Naets, Thomson (2016) considers that JSON and XML are not suitable formats for the long-term preservation of social media, because these formats do not preserve the contextual information linked to any social media post, such as websites or other media.

In the BE-Social survey, some questions specifically focused on the preservation formats used. Four institutions (BnF, BNL, UKWA and LAC) specified that they use the WARC ISO 28500 format. In its

[175] A German document, entitled *Konservierungsleitlinie Schweizerische Nationalbibliothek*, was consulted, however it focused more on the preservation of analogue collections including digitisation, rather than born-digital materials.

preservation policy, the UK National Archives states that the formats *"are available either in their original, or, where possible, in a more accessible format"* (The UK National Archives, 2018). The Library of New Zealand in turn said that it does not currently use any standards or norms. Format migration, as discussed in Section 8.1 of this report, is not currently being implemented by any of the survey respondents.

It is also important to question the scientific basis of the allegedly sustainable formats. As computer science is evolving at breakneck speed, the question of the choice of formats is absolutely essential in order to avoid, or rather to limit as much as possible, the resort to emulation or migration of formats in the coming years. The consensus on so-called 'preservation' formats does not yet seem to be mature, as evidenced by the debate on the PDF/A format, frequently considered by heritage institutions as the recommended format for the preservation of textual documents. This format has been questioned by Klindt in his 2017 article entitled 'PDF/A considered harmful for digital preservation' (Klindt, 2017).

8.6 Conclusions and recommendations for KBR on preservation

The long-term preservation of social media involves many challenges (Thomson, 2016). S. Jeffrey (2012) adds that the long-term preservation of social media is not just about backing up data, but about making content fully usable and accessible for decades to come. This long-term view on the usability of archived content implies a number of challenges, which KBR will have to consider, some of which were highlighted by participants in the BE-Social survey. The question of formats, their evolution, migration and their interrelation with data accessibility featured frequently, as well as the problems linked to changes in formats and capture methods used by social media platforms (e.g. lack of standardisation). Two other institutions also mentioned the difficulty and need for clear procedures and standardised mechanisms. Accessibility and technological developments are also a significant challenge.

It must be noted that, generally speaking, social media are not included in the digital preservation policies of the various institutions surveyed. However, social media have to be considered as an entity in its own right, distinct from the web. Thus, according to S. Thomson (2016), the techniques used for archiving the web will not bear fruit in the case of social media archiving, because in this case it involves information flows and interconnected data. Social media data involve intrinsic and external content. According to S. Thomson, the only solution to preserve the external content involved in the elements published on social media is to preserve this content simultaneously with the rest. This opinion is also shared by H. Hockx-Yu (2014). Furthermore, the preservation of social media content implies, on the one hand, the preservation of content and metadata and, on the other hand, the preservation of embedded media and URLs (Thomson, 2016). It is on the basis of these findings that KBR needs to consider this emerging collection of archived social media.

More surprisingly, it is important to note that many institutions do not yet have digital preservation policies. However, some respondents to the survey indicated that they have preservation-related software: Preservica for BNL, Rosetta for NLNZ, Hadoop for UKWA or SPAR for BnF, but this is currently far from wide-spread.

Moreover, the relationship between digital preservation policy, preservation strategy and preservation plan is rather difficult to establish. Section 8.2 of this report has attempted to clarify the objectives of each of these documents and how they are linked. Ideally, each institution should have all three, directly related to their digital collections, as well as an ethical charter. All of them are important and necessary and should ideally be drawn up for KBR once decisions on selection, access and technical and operational aspects have been taken.

M. Pennock's article (2020) on preservation policies summarises an important element to conclude this report on digital preservation. This article highlights the fact that some institutions, such as British Library, do not yet have preservation plans for their entire digital collections because they consider that they do not yet need them. M. Pennock also points out that the preservation plan can reflect different realities and can be difficult to implement in practice. From the research undertaken for this report, developing a preservation plan is important for KBR in order to define clear practices and, above all, to plan the operations for the future social media collections specifically, and digital collections more generally. KBR will also have to position itself on what they understand a preservation plan to be. In other words, whether KBR's understanding is in line with that of the British Library or that of other institutions that have set up a preservation plan, such as the National Library of New Zealand, it is important that this reflection be carried out.

The work carried out under *Task 1.4: Analysis of Preservation Policies* has thus allowed for identifying best practices, and other practices that should undoubtedly not be followed. On the basis of this analysis, some initial recommendations for the development of an effective preservation program at KBR are proposed here:

Table 8: Recommendations for the development of a preservation program at KBR

Domain	**Recommendations**
Administrative	- Have a clearly defined content selection process. - Draft and publish a digital preservation strategy which defines KBR's long-term vision for digital preservation. - Draft and publish a digital preservation policy, which defines KBR's medium-term objectives for digital preservation and which ensures the transparency of the institution in terms of digital preservation. This policy should be regularly reviewed and updated as necessary. - Draft and publish a digital preservation plan, which clearly specifies the actions to be undertaken to ensure the long-term preservation of KBR's digital collections. This is the operationalisation of both the medium-term objectives, as outlined in the digital preservation policy, and the contributes to achieving KBR's long-term vision for digital preservation, as outlined in the digital preservation strategy. - Draft and publish an ethical charter, gathering the guidelines to be followed by each individual brought into contact with these digital archives. - Having clear procedures for each process that must be undertaken on the data (**transparency**)

Technical	- Do not limit to a single backup of the data (**replication**): back up the data in multiple ways and in different geographical locations. Among these backups, keep one of the raw data, without any subsequent modifications or additions. - Having a mechanism to control the obsolescence of formats and files kept in the collections - Having a mechanism for automatically checking the integrity of the files. - Having processes for **migrating** files from one format to another. - Having predefined **emulation** processes for obsolete software. - Availability of the complete history of data archiving and all manipulations carried out on the data (**transparency**). - Preserve the necessary contextual data (**metadata**) according to a pre-established selection process.
Organisational	- Having competent staff, capable of keeping up to date with the latest developments in technical, operational and legal terms. - Taking into account technical constraints, such as storage and its cost.

Documents such as preservation strategies, policies and plans can be drawn up at KBR using the best examples, namely the National Library of Australia and the National Library of New Zealand. The article of D. Burda (2017), produced as part of the study of digital preservation in Switzerland, is also an example to be taken into consideration. Following the Handbook edited by the Digital Preservation Coalition also seems to be wise advice (Digital Preservation coalition, n.d.).

9. General conclusion and discussion

This report details the tasks completed during WP1 of the BESOCIAL project. In this report we sought to provide a review of the state of art of SMA in the context of web archiving institutions, both by providing a literature review of previous work, as well as a survey and in-depth interviews to understand the practicalities and legal aspects of this type of archiving for institutions in practice. This report details the international initiatives around SMA, a separate report details the Belgian specific initiatives.

As documented here in this report, social media archiving, similar to web archiving, is occurring in a number of institutions worldwide, in an effort to document and archive records of online communication. There are diverse definitions of what encompasses social media, and thus the scope of social media archiving initiatives vary in scope and size, as well as the ways to preserve them. This was confirmed in our research which combined desk research, a survey and interviewing a selection of global SMA initiatives. It should be noted, due to this convenience sampling we cannot claim to have interviewed a representative sample of SMA initiatives. In addition, all of the information was self-reported by the institutions, thus is privy to some unintended bias; although this proved to be a quick and easy method to collect information the state of current SMA initiatives.

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

These selections are a challenge for SMA initiatives as despite efforts to be transparent about their crawling activities and including known limitations, there is also the limitation of available tools. Our findings show that much of the crawling is done through application programming interfaces (APIs). APIs limit the information that can be collected (i.e. maximum request per day) and limit its reuse (Morstatter, Pfeffer and Liu 2014; Thomson and Kilbride 2015), which further limits access and knowledge to access questions of quality. There is a concern that this may result in implicit unrepresentative sampling which influences the validity of this information for future use and particularly for generalising the findings Thus, it is important for institutions to be careful to document and be transparent about selection processes.

In additional to these practical challenges of selection of crawls and tools, there is also a challenge of how to provide and ensure access to archives and under which copyright conditions (Zimmer, 2015; George Washington University Libraries, 2016; McCreadie, Soboroff, Lin, Macdonald, Ounis & McCullough, 2012). The ways in which institutions provide data level access to their social media collections varies from scope and size. The national legal frameworks for accessing social media information also varies.Preservation practices proved to be diverse among those surveyed. It is clear that preservation of social media content is currently on the back burner in many institutions. Format migration is still a nascent field, even though the native file formats will at some point in the future have to be migrated to preservation formats. There remains a lack of common understanding of what is considered as digital preservation procedures or formats, and therefore a need for raising awareness about preservation of social media content and specifically long-term preservation.

To conclude, a growing number of initiatives are actively archiving social media, in addition to web archiving initiatives. This report details the scope of 9 interviewed initiatives and serves to document the current state of the field in regards to the theory and practice of social media archiving. Despite the challenges detailed in this report, there are also many opportunities for learning how to accurately archive and preserve this currently under utilized information as records of our (recent) past.

10. Bibliography

Acker, A., & Kriesberg, A. (2017). Tweets may be archived: civic engagement, digital preservation and Obama White House social media data. *Proceedings of the Association for Information Science and Technology*, *54*(1), 1-9.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, *34*(7), 1177-1190.

Ariel, Y., & Avidar, R. (2015). Information, interactivity, and social media. Atlantic Journal of Communication, 23(1), 19-30.

Arquivo.pt. (2017). Recomendações para a publicação na Web de informação preservável. Retrieved from https://sobre.arquivo.pt/pt/recomendacoes/.

Arquivo.pt. (2020). Formações acerca de preservação da Web. Retrieved from https://sobre.arquivo.pt/pt/ajuda/formacao/

Ashrafi-rizi, H., & Kazempour, Z. (2020). Information typology in coronavirus (COVID-19) crisis: a commentary. *Archives of Academic Emergency Medicine* *8*(1), e19.

Benkler, Y. (2006). The Wealth of Networks: How Social Production Transforms Markets and Freedom. London: Yale University Press.

Bermès, E., & Faudet, L. (2009). The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. UC Office of the President: California Digital Library. Retrieved from https://escholarship.org/uc/item/6bt4v3zs.

Bibliothèque et Archives nationales du Québec. (2013). Politique de conservation des collections patrimoniales. Retrieved from https://www.banq.qc.ca/a_propos_banq/mission_lois_reglements/lois_reglements_politiques/politiques_procedures/politique_conservation_coll_patrimoniale/index.html.

Bibliothèque et Archives nationales du Québec. (2016). Plan stratégique 2016-2018. Retrieved from https://www.banq.qc.ca/a_propos_banq/acces_a_linfo/plan_strategique/.

Bibliothèque et Archives nationales du Québec. (n.d.) Politiques et procédures (pratiques opérationnelles). Retrieved from https://www.banq.qc.ca/a_propos_banq/mission_lois_reglements/lois_reglements_politiques/politiques_procedures/.

Bibliothèque nationale de France. (n.d.). Politique de conservation. Retrieved from https://www.bnf.fr/fr/politique-de-conservation.

Bibliothèque nationale de France. (n.d.). SPAR (Système de Préservation et d'Archivage Réparti). Retrieved from https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti.

Bibliothèque nationale du Luxembourg. (2020). Innover grâce au numérique. Retrieved from https://bnl.public.lu/fr/services-professionnels/innovation/Inno_num.html.

Birkholz, J. M., Wang, S., Groth, P. T., & Magliacane, S. (2013). Who are we talking about?: identifying scientific populations online. In *Chinese Semantic Web Symposium & Web Science Conference Proceedings* Springer Publications.

Bishoff, L. (2010). Digital Preservation Plan. Ensuring Long Term Access and Authenticity of Digital Collections. *Information Standards Quarterly, 22*(2), 21-25. Consulté à l'adresse https://www.niso.org/niso-io/digital-preservation-plan-ensuring-long-term-access-and-authenticity-digital-collections.

British Library et al. (n.d.). Joint Collecting Framework for UK Legal Deposit, 2015-2020. Retrieved from https://www.bl.uk/britishlibrary/~/media/bl/global/legal%20deposit/joint-collecting-framework-for-uk-legal-deposit.pdf?la=en.

British Library. (2017). Sustaining the value. The British Library Digital Preservation Strategy 2017-2020. Retrieved from https://www.bl.uk/britishlibrary/~/media/bl/global/digital%20preservation/bl_digitalpreservationstrategy_2017-2020.pdf.

British Library. (n.d.). Our policies and procedures. Retrieved from https://www.bl.uk/about-us/freedom-of-information/5-our-policies-and-procedures.

British Library. Preservation Advisory Centre (2013). Building a preservation policy. Retrieved from https://www.bl.uk/britishlibrary/~/media/bl/global/conservation/pdf-guides/building-a-preservation-policy.pdf.

Burda, D. (2017). Digitale Langzeitarchivierung in der Schweiz. Ergebnisse einer Studie im Auftrag der Schweizerischen Nationalbibliothek. Berner Fachhochschule, E-Government-Institut. Retrieved from https://www.nb.admin.ch/dam/snl/de/dokumente/kommission_nb/studien_und_berichte/Digitale%20Langzeitarchivierung%20in%20der%20Schweiz.pdf.download.pdf/20170922_LZA_Studie_Abschlussbericht6862.pdf

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. Journal of Information Science. https://doi.org/10.1177/0165551520950246

Chambers, S. (2020). Web-archives for Open Science: how FAIR can we go? WARCnet kickoff meeting, 4-6 May 2020. https://cc.au.dk/en/warcnet/presentations/kickoff-meeting-2020/

Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Mechant, P., Michel, A. & Vlassenroot, E. (2018). PROMISE: Final Report Work Package 1 web archiving state of the art, 30 May 2018.

Cheston, C. C., Flickinger, T. E., & Chisolm, M. S. (2013). Social media use in medical education: a systematic review. *Academic Medicine*, *88*(6), 893-901.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F. & Scala, A. (2020). The covid-19 social media infodemic.https://arxiv.org/abs/2003.05004

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksons, L., & Wittenburg, P. (2018). Turning FAIR into reality. Final report and action plan from the European Commission expert group on FAIR data. Brussels: European Commission.

Digital Preservation Coalition (n.d.). Preservation issues. Retrieved from https://www.dpconline.org/handbook/digital-preservation/preservation-issues

Digital Preservation Coalition. (n.d.). Glossary. Retrieved from https://www.dpconline.org/handbook/glossary#D

Digital Repository of Ireland. (2013). Fact sheet: Metadata and the DRI. Retrieved from https://repository.dri.ie/catalog/bz60sj10d.

Digital Repository of Ireland. (2014). Fact sheet: Long-term Digital Preservation. Retrieved from https://www.dri.ie/sites/default/files/files/Fact%20Sheet%20No%204%20Long%20term%20digital%20preservation%20ver%201.pdf.

Digital Repository of Ireland. (2015). Collection policy. Retrieved from http://dri.ie/sites/default/files/files/DRI-collection-policy-april2015.pdf.

Dooley, J. & Bowers, K. (2018). Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Dublin, OH: OCLC Research.

E. Graff & S. Sepetjan (2011). Le dépôt légal en France. Les cahiers de la propriété intellectuelle, 2011/1.

Eesti Veebiarhiiv. (n.d.). Abi. Retrieved from http://veebiarhiiv.digar.ee/abi.php.

Espley, Suzy & Carpentier, Florent & Pop, Radu & Medjkoune, Leïla. (2014). Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content. Alexandria. 25. 31-50. 10.7227/ALX.0019.

Filo, K., Lock, D., & Karg, A. (2015). Sport and social media research: A review. *Sport management review*, *18*(2), 166-181.

Freelon, D. (n.d.). Social media data collection tools. Retrieved from http://socialmediadata.wikidot.com/.

Gayo-Avello, D. (2016). How I Stopped Worrying about the Twitter Archive at the Library of Congress and Learned to Build a Little One for Myself. https://arxiv.org/abs/1611.08144.

George Washington Libraries & Academic innovation. (n.d.). Digital services. Retrieved from https://lai.gwu.edu/digital-services.

George Washington Libraries. (n.d.). Digital Stewardship Service Catalog. Retrieved from https://library.gwu.edu/digital-stewardship-services.

George Washington Libraries. (n.d.). Preservation. Retrieved from https://library.gwu.edu/content-management/preservation.

George Washington University Libraries. (2017). Social Feed Manager. Building Social Media Archives: Collection Development Guidelines. https://gwu-libraries.github.io/sfmui/resources/guidelines. Accessed 8 September 2020.

Gosling, S.D., et al., Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. American Psychologist, 2004. 59: p. 93-104.

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. (1992). Edit Wear and Read Wear. Paper presented at the ACM Conference on Human Factors in Computing Systems (CHI'92), New York City.

Hockx-Yu, H. (2014). Archiving Social Media in the Context of Non-print Legal Deposit. Paper presented at IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 107 - National Libraries. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. Retrieved from http://library.ifla.org/id/eprint/999

Holownia, O. and Chambers, S. (2020). Supporting research use of web archives: a 'labs' approach. Digital Humanities in the Nordic Countries (DHN2020), 20th-23rd October 2020, National Library of Latvia, Riga.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Mazaid, M. (2011). Opening closed regimes: What was the role of social media during the Arab spring? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2595096

Hucka, M. (2017). Comparison of web archiving software. https://github.com/datatogether/research/tree/master/web_archiving. Accessed 8 September 2020.

Institut national de l'audiovisuel. (n.d.). Plan de sauvegarde et de numérisation. Retrieved from https://institut.ina.fr/institut/statut-missions/plan-de-sauvegarde-et-de-numerisation.

Joris, G., De Grove, F., Van Damme, K., & De Marez, L. (2020). News diversity reconsidered : a systematic literature review unraveling the diversity in conceptualizations. JOURNALISM STUDIES, 21(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59-68.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, 54(3), 241–251.

Klindt, M. (2017). PDF/A considered harmful for digital preservation. iPRES2017.

Le Follic, A., & Chouleur, M. (2018). La collecte des médias sociaux, un enjeu pour la constitution des collections de dépôt légal du web à la Bibliothèque nationale de France. In A. François, A. Roekens, V. Fillieux & C. Deraux (Eds.), *Pérenniser l'éphémère. Archivage et médias sociaux* (pp. 109-124). Louvain-la-Neuve: UCL.

Lemay, F. (2019). Current State of Digital Preservation at Library and Archives Canada. Digital Preservation Coalition (DPC) Blog post. Retrieved from https://www.dpconline.org/blog/idpd/current-state-of-digital-preservation-at-lac.

Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A Literature review. *Journal of travel & tourism marketing*, *30*(1-2), 3-22.

Library and Archives Canada. (2014). Stewardship Policy Framework. Retrieved from https://www.bac-lac.gc.ca/eng/about-us/policy/Documents/stewardship-policy-framework.pdf.

Library and Archives Canada. (2015). Guidelines on File Formats for Transferring Information Resources of Enduring Value. Retrieved from https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf.

Library and Archives Canada. (2017). Strategy for a digital preservation program. Retrieved from https://www.bac-lac.gc.ca/eng/about-us/publications/Documents/LAC-Strategy-Digital-Preservation-Program.pdf.

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

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, https://doi.org/10.1111/1467-9558.00187.

MacDonald, N. (2019). How the National Library preserves New Zealand's digital heritage. Stuff Blog post. Retrieved from https://www.stuff.co.nz/national/115118698/how-the-national-library-preserves-new-zealands-digital-heritage.

Macnaught, B. (2018). Social media collecting at the National Library of New Zealand. IFLA WLIC Transform Libraries, Transform Societies in Kuala Lumpur. August 2018. http://library.ifla.org/2274/1/093-macnaught-en.pdf.

Marwick, A. E. (2010). Status update: Celebrity, publicity and self-branding in Web 2.0 (PhD. Dissertation). New York: New York University.

McCracken, G. (2007). How social networks work: the puzzle of exhaust data. Retrieved from https://cultureby.com/2007/07/how-social-netw.html.

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

Mckinney, P. (2018). National Library of New Zealand. Code of ethics for digital preservation. Retrieved from https://natlib.govt.nz/files/digital-preservation/Draft-Code-of-Ethics-for-Digital-Preservation.pdf.

McQuail, D. (1992). Media performance: Mass communication and the public interest. Thousand Oaks, CA, US: Sage Publications, Inc.

McQuail, D., & Van Cuilenburg, J. J. (1983). Diversity as a media policy goal: A strategy for evaluative research and a Netherlands case study. I*nternational Communication Gazette*, 31(3), 145-162. doi:10.1177/001654928303100301

Merwood, H. (2020). Do you trust our digital archive?. Open Preservation Foundation Blog post. Retrieved from https://openpreservation.org/blogs/do-you-trust-our-digital-archive/.

Milligan, I. (2019). History in the age of abundance. How the web is transforming historical research. Montreal & Kingston: McGill-Queen's University Press.

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased?: assessing the representativeness of Twitter's streaming API. In WWW '14 Companion: Proceedings of the 23rd int. Conference on World Wide Web (pp. 555–556). New York: Association for Computing Machinery.

Naets, H., Techniques de collecte et d'archivage de Twitter, dans François, A., Roekens, A., Fillieux, V. & Derauw, C., dir., *Pérenniser l'éphémère. Archivage et médias sociaux*, Louvain-la-Neuve, 2018, p. 215-237.

Najjar, J., Wolpers, M., & Duval, E. (2006). Attention Metadata: Collection and Management. World Wide Web conference at Edinburgh, Scotland, 23-26 June 2006.

National Archives of the Netherlands. (2015). Preservation policy. Retrieved from https://www.nationaalarchief.nl/sites/default/files/field-file/National%20Archives%20of%20the%20Netherlands%20preservation%20policy.pdf.

National Archives of the Netherlands. (2016). Preferred formats National Archives of the Netherlands. In view of sustainable accessibility. Retrieved from https://www.nationaalarchief.nl/sites/default/files/field-file/National%20Archives%20of%20the%20Netherlands%20preferred%20and%20acceptable%20formats.pdf.

National Diet Library Web, Archiving Project. (2020). Statistics [in japanese]. Retrieved from https://warp.da.ndl.go.jp/info/WARP_statistic.html.

National Diet Library, Web Archiving Project. (2014).What is a web archive? [in japanese]. Retrieved from https://warp.da.ndl.go.jp/contents/reccommend/mechanism/mechanism01.html.

National Diet Library, Web Archiving Project. (n.d.). Questions and answers [in japanese]. Retrieved from https://warp.da.ndl.go.jp/info/WARP_qanda.html#01_01.

National Library of Australia. (2013). Digital Preservation Policy 4th Edition. Retrieved from https://www.nla.gov.au/policy-and-planning/digital-preservation-policy.

National Library of Australia. (2013). Preservation Intent – Selective Web Harvesting: NLA Digital collections: Statement of Preservation Intent. Retrieved from https://www.nla.gov.au/content/preservation-intent-selective-web-harvesting.

National Library of Estonia. (n.d.). Development projects. Retrieved from https://www.nlib.ee/en/content/development-projects.

National Library of Ireland. (n.d.). What is Web Archiving?. Retrieved from https://www.nli.ie/getAttachment.aspx?Id=504f8c97-1aec-4370-b7d5-b184f6472e47

National Library of New Zealand. (2011). Digital Preservation Strategy. Retrieved from https://natlib.govt.nz/files/digital-preservation/Digital_Preservation_Strategy.pdf

National Library of New Zealand. (2012). Digital Preservation Policy Manual. Retrieved from https://natlib.govt.nz/files/digital-preservation/digital-preservation-policy-manual.pdf

National Library of New Zealand. (2018). Code of ethics for digital preservation. Retrieved from https://natlib.govt.nz/files/digital-preservation/Draft-Code-of-Ethics-for-Digital-Preservation.pdf.

National Library of New Zealand. (n.d.). Digital preservation program. Retrieved from https://natlib.govt.nz/collections/digital-preservation/digital-preservation-programme.

National Library of New Zealand. (n.d.). Whole of domain web harvest. Retrieved from https://natlib.govt.nz/publishers-and-authors/web-harvesting/domain-harvest.

National Library of The Netherlands. (n.d.). Usability in the long term. Retrieved from https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/usability-in-the-long-term.

Németh, M. (2020). From pilot to portal: a year of web archiving in Hungary. International Internet Preservation Consortium (IIPC) Blog Post. Retrieved from https://netpreserveblog.wordpress.com/2020/06/26/a-year-of-web-archiving-in-hungary/.

Netarkivet.dk. (2014). Policy for long term preservation of material collected for the Netarchive (Netarkivet) by the Royal Library and the State and University Library. Retrieved from http://netarkivet.dk/wp-content/uploads/2016/05/Netarchive_Preservation_Policy_eng_FOR_PUBLICATION.pdf.

Nilesh, S., & Verma, S. (2012). Digital Preservation, Issues and Their Strategies for Libraries: A Modern Era (online version). Retrieved from https://www.academia.edu/28653621/Digital_Preservation_Issues_and_Their_Strategies_for_Libraries_A_Modern_Era

Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge. An introduction to the special issue. *Telecommunications Policy, 39*(9), 745-750.

Ortner, C., Sinner, P., & Jadin, T. (2018). The History of Online Social Media. In N. Brügger & I. Milligan (Eds.), The SAGE Handbook of Web History (pp. 372-384). London: SAGE Publications Ltd.

OSZK Archívum. (n.d.). About this website. Retrieved from https://webarchivum.oszk.hu/en/for-users/short-description/.

OSZK Archìvum. (n.d.). Archive of the National Széchényi Library's website. Retrieved from https://webarchivum.oszk.hu/en/webarchive/sub-collections/archive-of-the-websites-of-the-national-szechenyi-library/.

OSZK Archìvum. (n.d.). Thematic and genre-based harvests. Retrieved from https://webarchivum.oszk.hu/en/webarchive/sub-collections/thematic-harvests/.

Oxford LibGuides. (2020). Introduction to Digital Preservation: What is Digital Preservation?. Retrieved from https://libguides.bodleian.ox.ac.uk/digitalpreservation/whatisdp

Padilla, T., Allen, L., Frost, H., Potvin, Sarah, Russey Roke, E., & Varner, S. (2019). Final Report - Always Already Computational: Collections as Data. http://doi.org/10.5281/zenodo.3152935

Pal, J., Chandra, P., & Vydiswaran, V. V. (2016). Twitter and the rebranding of Narendra Modi. *Economic and Political Weekly, 51*(8), 52-60.

Pandora. Australia's Web Archive. (2020). Policy and practice statement. Retrieved from http://pandora.nla.gov.au/policy_practice.html

Pandora. Australia's Web Archive. (n.d.) Front Page. Retrieved from https://pandora.nla.gov.au/.

Pennock, M. (2020). 'So, have you got a Preservation Plan for that?'. Open Preservation Foundation Blog post. Retrieved from https://openpreservation.org/blogs/so-have-you-got-a-preservation-plan-for-that/.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010), Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto. Accessed 8 September 2020.

Rabina, D., Cocciolo, A., & Peet, L. (2013). Social Media Use by the US Federal Government at the End of the 2012 Presidential Term. *Alexandria, 24*(3), 73-93.

Raymond, M. (2010). How tweet it is!: Library acquires entire Twitter archive. Library of Congress blog. Retrieved from: https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/. Accessed 8 September 2020.

Riigi Teataja. (2016). Eesti Rahvusraamatukogu seadus (lühend - ERRS). Retrieved from https://www.riigiteataja.ee/akt/106012016008.

Riigi Teataja. (2017). Säilituseksemplari seadus (lühend - SäES). Retrieved from https://www.riigiteataja.ee/akt/107072016001.

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer. National Information Standards Organization. https://www.niso.org/publications/understandingmetadata

Rogers, R. (2018) Periodizing web archiving: biographical, event-based, national and autobiographical traditions. In N. Brügger & I. Milligan (Eds.)*, The SAGE Handbook of Web History* (pp. 42-56). London: SAGE Publications Ltd.

Rosenberg, K. (2020a). Web ARChive studies network researching web domains and events. https://cc.au.dk/en/warcnet/about/ Accessed 8 September 2020.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063-1064.

Sajwan, S., Goethals, A. & Wu, C. (2018). Preserving the past and present for the future: Digital Preservation Programme at the National Library of New Zealand. Brochure. Retrieved from https://natlib.govt.nz/files/digital-preservation/NDHA-Booklet-Oct18-2.pdf.

Samouelian, M., & Dooley, J. (2018). Descriptive Metadata for Web Archiving: Review of Harvesting Tools. Dublin, OH: OLCL Research.

Sherratt, Tim, & Jackson, Andrew. (2020, June 15). GLAM-Workbench/web-archives (Version 0.1.1). Zenodo. http://doi.org/10.5281/zenodo.3894079

Sirp. (2020). Eesti veebiarhiiv kogub infot koroonaviirust kajastavate veebilehtede kohta. Retrieved from https://www.sirp.ee/s3-pressiteated/eesti-veebiarhiiv-kogub-infot-koroonaviirust-kajastavate-veebilehtede-kohta/.

Social Media Data Scholarship. (2020). Social Media Research Toolkit.Retrieved from https://socialmediadata.org/social-media-research-toolkit/.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037-2062.

Swiss National Library. (2018). Digitally born collections. Retrieved from https://www.nb.admin.ch/snl/en/home/collections/digital-collections/born-digital-collections.html

Swiss National Library. (2020). Basic principles (preservation). Retrieved from https://www.nb.admin.ch/snl/en/home/information-professionals/preservation/basic.html

Takens, J., Ruigrok, N., van Hoof, A. M. J., & Scholten, O. (2010). Old ties from a new(s) perspective: Diversity in the Dutch press coverage of the 2006 general election campaign.

Communications-European Journal of Communication Research, 35(4), 417-438. doi:10.1515/comm.2010.022

TechTerms (2011). File Format. Retrieved from: https://techterms.com/definition/file_format.

Tess, P. A. (2013). The role of social media in higher education classes (real and virtual). A literature review. *Computers in human behavior*, *29*(5), A60-A68.

The UK National Archives. (2017a). Digital strategy. Retrieved from https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf.

The UK National Archives. (2017b). The UK Government Web Archive. Guidance for digital and records management teams. Retrieved from http://www.nationalarchives.gov.uk/documents/web-archiving-technical-guidance.pdf.

The UK National Archives. (2018). Preservation policy. Retrieved from http://www.nationalarchives.gov.uk/documents/preservation-policy-june-2018.pdf.

The UK National Archives. (n.d.) File formats for transfer. Retrieved from https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer/.

The UK National Archives. (n.d.). Our policies. Retrieved from https://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-policies/.

Thomson, S. D. (2016). *Preserving Social Media*. DPC Technology Watch Report 16-01 February 2016.

Thomson, S. D., & Kilbride, W. (2015). Preserving social media: The problem of access. *New Review of Information Networking, 20*(1-2), 261–275.

Treem, J. W., & Leonardi, P. M. (2012). Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association. In Communication Yearbook, vol. 36, pp. 143-189.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. https://arxiv.org/abs/1403.7400.

UK National Archives. (n.d.). Developing a digital preservation strategy and policy. Retrieved from https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/developing-a-digital-preservation-strategy-and-policy/.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2003). Charter on the Preservation of the Digital Heritage. https://unesdoc.unesco.org/ark:/48223/pf0000179529.locale=en Accessed 14 September 2020.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (n.d.). Concept of Digital Preservation. Retrieved from https://fr.unesco.org/themes/information-preservation/digital-heritage/concept-digital-preservation.

University of Washington, University Libraries. (2017). Digital Preservation Policy. Retrieved from https://www.lib.washington.edu/preservation/preservation_services/digitization-and-digital-preservation/digital-preservation-policy.

University of Washington. University Libraries. (n.d.). Preferred File Formats. Retrieved from https://www.lib.washington.edu/preservation/preservation_services/digitization-and-digital-preservation/preferred-file-formats.

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

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1), 85–111.

Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). "Interview/Survey spreadsheet WP1 BESOCIAL". Ghent University, Ghent, Belgium, April 2021. https://docs.google.com/spreadsheets/d/1ZM4LTkwhmXjHDArJObzpoKMQLL7HOhYzAB3tCoZdGNs/edit#gid=1321624630.

Web Archive of Athens University of Economics and Business. (2018). About us. Retrieved from http://archive.aueb.gr/.

Web Archiving Project. (2014). Long-term storage of web archives [in japanese]. Retrieved from https://warp.da.ndl.go.jp/contents/reccommend/mechanism/mechanism08.html.

Woolman, Anna. (2020). Introducing: The Boredom Project, https://www.britishscienceassociation.org/blog/introducing-the-boredom-project.

Zellier, J.-D. (2018). La mise à disposition des archives de Twitter par la Library of Congress. In A. François, A. Roekens, V. Fillieux & C. Deraux (Eds.), Pérenniser l'éphémère. Archivage et médias sociaux (pp. 125-134). Louvain-la-Neuve: UCL.

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

Appendix A. Social Media Harvesting Tools

As described in section 9, we compared several existing social media harvesting tools. In the following we provide more detailed information for each tool: its features, setup, configuration, creation of a new collection, creation of a new account-based collection, creation of a new keyword-based collection and monitoring of a collection.

A.1 Brozzler

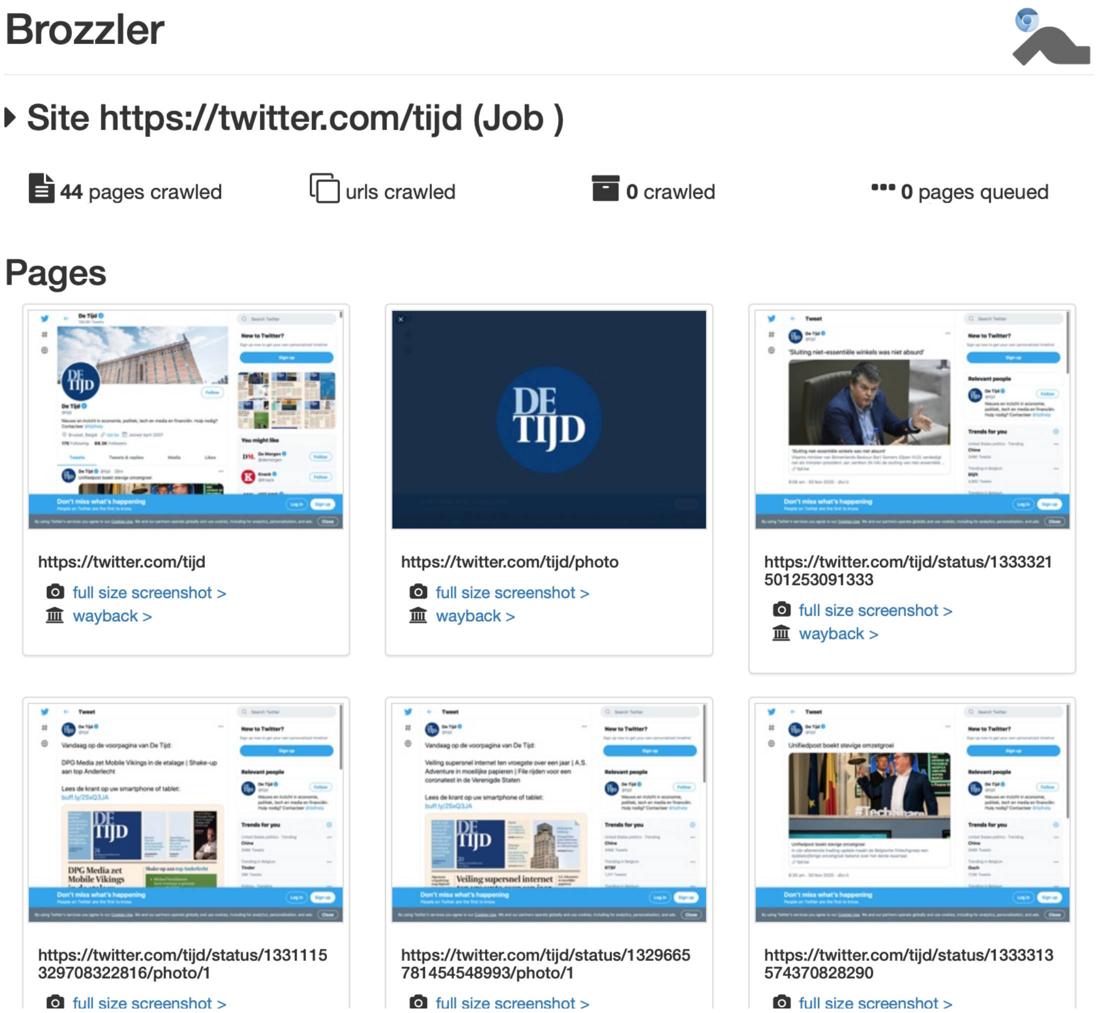
Brozzler, from the Internet Archive, was developed to harvest websites with dynamic content, it is a decentralized web crawler which uses a browser to harvest websites.

Features

The main feature of Brozzler is that instead of following hyperlinks and downloading the files, like regular web harvesters do, it records interactions between web servers and web browsers. Thus, it resembles more how a human would experience websites. Since it harvests via the web no API keys need to be provided, however, if content requires login, those login information have to be provided.

Setup

Brozzler is a Python tool. It has to be started from the command line and works with a RethinkDB database and warcprox (a proxy creating WARC files from captured HTTP requests) which need to be provided separately. Additionally, Brozzler has a dashboard as a user interface to monitor the status of crawls and to playback harvested websites, see figure A1.1. Docker can be used to quickly start all necessary Brozzler components.



Brozzler

Site https://twitter.com/tijd (Job)

44 pages crawled urls crawled 0 crawled 0 pages queued

Pages

https://twitter.com/tijd
full size screenshot >
wayback >

https://twitter.com/tijd/photo
full size screenshot >
wayback >

https://twitter.com/tijd/status/1333321501253091333
full size screenshot >
wayback >

https://twitter.com/tijd/status/1331115329708322816/photo/1
full size screenshot >

https://twitter.com/tijd/status/1329665781454548993/photo/1
full size screenshot >

https://twitter.com/tijd/status/1333313574370828290
full size screenshot >

Figure A1.1: Harvest results shown by the Brozzler dashboard, screenshots provide a preview of the harvested content. However, in our tests we noticed that the harvested websites often contain “something went wrong” error messages although the screenshot looked okay.

Configuration

A YAML configuration file is used to configure crawls for Brozzler.

Creating a new collection

A new collection is created using a YAML configuration file and the command line.

Creating an account-based collection

All collections are specified using YAML configuration files, many options exist, such as the metadata field which is actually not used by Brozzler but can be used to specify additional metadata. Besides that multiple seeds can be given which should be crawled, see listing 1.

```
id: twitter_standaard_tijd
time_limit: 900 # seconds
ignore_robots: false
warcprox_meta: null
metadata: {}
```

```
seeds:
  - url: https://twitter.com/destandaard
seeds:
  - url: https://twitter.com/tijd
```

Listing 1: A Brozzler configuration to fetch two public pages. Similar to the URL.

Creating a keyword-based collection

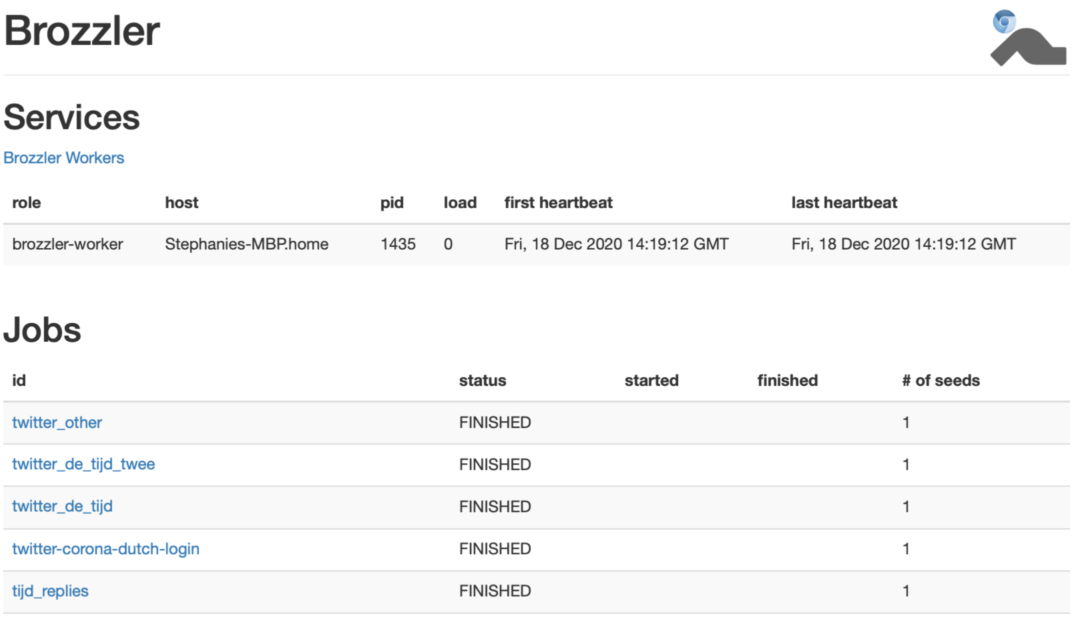
Similar to an account-based selection a YAML file is used. The difference is that the search term is a query and that a username and password are provided, see listing 2.

```
id: twitter_covidbe_login
time_limit: 900 # seconds
ignore_robots: false
warcprox_meta: null
metadata: {}
seeds:
  - url: https://twitter.com/search?q=%23CovidBe
  - username: myusername
  - password: mypassword
```

Listing 2: A Brozzler configuration to fetch tweets of the search term #CovidBe. A username and a password (in this case for Twitter) are given which is used for authentication during the harvest which allows non-interactive harvesting.

Monitoring a collection

Brozzler comes with a dashboard showing the status of harvesting jobs, see figure A1.2. However, one cannot pause, stop or configure harvests from this dashboard.



Brozzler

Services

Brozzler Workers

role	host	pid	load	first heartbeat	last heartbeat
brozzler-worker	Stephanies-MBP.home	1435	0	Fri, 18 Dec 2020 14:19:12 GMT	Fri, 18 Dec 2020 14:19:12 GMT

Jobs

id	status	started	finished	# of seeds
twitter_other	FINISHED			1
twitter_de_tijd_twee	FINISHED			1
twitter_de_tijd	FINISHED			1
twitter-corona-dutch-login	FINISHED			1
tijd_replies	FINISHED			1

Figure A1.2: The Brozzler dashboard showing the status of jobs.

A.2 CENTAL tool

This is a tool developed at UCLouvain with the purpose to collect Tweets for linguistic analyses.

Features

The special feature of this tool is that it aims to collect tweets of a representative sample of a certain society. Therefore it starts from an initial account list and based on the followers extends this list. Collected tweets are stored in line oriented JSON format.

Setup

The tool is implemented in Python and works with the following directories:

- 01_accounts_server: contains a web server managing the twitter account names
- 02_politicians_screennames_injector: contains a script to send the initial seed list of politicians to the webserver for harvest
- 03_followed_screennames_injector: contains a script to send an initial seed list of persons to the webserver for harvest
- 04_followers_screennames_injector: contains a script to send a list of followers to the webserver for harvest
- 05_politiians_tweets_retrieval: contains a script that downloads tweets from politicians and stores them in the 99_retrieved_tweets directory.
- 99_retrieved_tweets

No particular installation is required but some perl modules/dependencies have to be installed

- Mojo::Base
- Mojo::Pg
- Mojo ::JSON
- Data::Dump
- Data::Dumper

- Mojolicious

Configuration

A PostgreSQL database needs to be configured storing the names of the followers accounts and their max_id to avoid duplicate downloads. Twitter API access keys have to be configured in directories 4 and 5.

Creating a new collection

The webserver and database have to be enabled and then the initial seed accounts need to be provided in directories 2 and 3. Tweets and retweets with all their metadata are then collected during specific times controlled by a Unix CRON job.

Creating an account-based collection

The main purpose is the creation of tweets from accounts and the followers of those accounts.

Creating a keyword-based selection

It is also possible to harvest tweets from all tweets with a particular hashtag or expression. But the API has its limitations, mainly that only data from the past 7 days can be harvested.

Monitoring a collection

The harvesting of tweets is scheduled by Unix CRON jobs and thus breaks can be scheduled and direct access to errors is possible, no particular monitoring is currently implemented.

A.3 Instaloader

Instaloader is a tool to harvest data from Instagram, namely pictures or videos together with captions and other metadata. No API keys need to be provided, however, for some content a login is required for which login information needs to be provided.

Features

This tool is specialized on Instagram content, therefore it can also be configured to harvest public or private profiles, hashtags, user stories, feeds, saved media, comments, geotags and captions.

Setup

Instaloader is a Python tool. It can be directly used from within the command line or it can be called as a library from within your own code.

Configuration

There is no configuration file per se to configure crawls, but parameters for the command line tool can be placed in a separate text file and be passed to the instaloader command line tool. Thus these separate text files can be seen as configuration files for different crawls.

Creating a new collection

A new collection is created by passing accounts as parameters to the command line tool, see listing 3.

Creating an account-based collection

The configuration shown in listing 3 would crawl different types of content such as videos and pictures, additional parameters can be used to narrow down the selection as shown in listing 4 where videos and pictures are excluded but comments will be included.

```
foo@bar:~$ instaloader de.tijd destandaard
```

Listing 3: A Instaloader command line call to fetch the provided profiles (more can be added).

```
foo@bar:~$ instaloader de.tijd --no-videos --no-pictures --comments
```

Listing 4: A Instaloader command line call to fetch posts from the account of the newspaper De Tijd, videos and pictures are excluded but comments will be included (comments are turned off by default).

Creating a keyword-based collection

Keywords have to be provided as a parameter, similar to an account-based selection. As shown in listing 5, keyword searches can be narrowed down by e.g. specifying a specific location ID prefixed with "%". In the shown example, posts with hashtag "#covid_19" from the location "Brussels, Belgium" will be collected, additionally a username for login is provided which might be necessary for some content; a password prompt will open to ask for the password of the provided account.

```
foo@bar:~$ instaloader --login username %213633143 "#covid_19"
```

Listing 5: A Instaloader command line call to fetch #covid_19 posts from the Instagram-internal location ID %213633143 which is Brussels, a username for login is provided.

Monitoring a collection

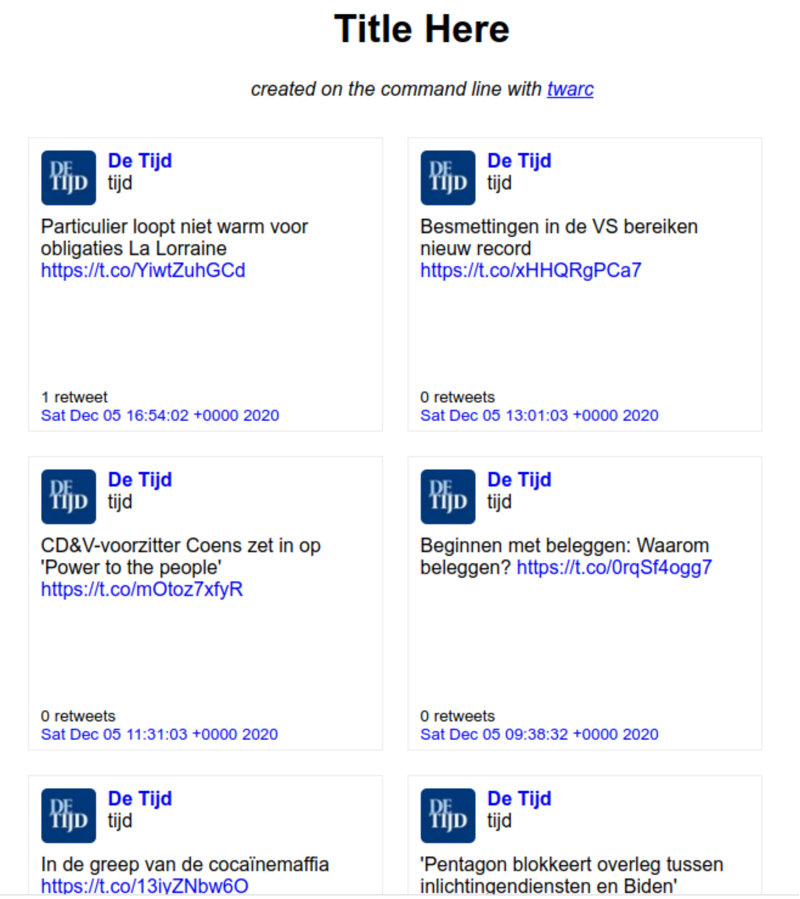
The command line tool returns a status indicating how many posts are available and how many of those have been downloaded already.

A.4 Twarc

Twarc, from the DocNow project, is written in Python and is able to harvest content from Twitter via their APIs, thus API keys are necessary which can be requested from Twitter.

Features

Twarc facilitates harvesting from Twitter as it manages rate limiting and offers several utility scripts to further process tweets, such as visualizations of tweets in different ways (e.g. a rudimentary list of tweets in HTML as shown in figure A4.1), a separate project called twarc-report[176] provides even more possibilities, mostly related to visualizations with D3js. Besides visualization, Twarc also offers scripts to hydrate/dehydrate tweets.



Title Here

created on the command line with twarc

De Tijd
tijd

Particulier loopt niet warm voor obligaties La Lorraine https://t.co/YiwtZuhGCd

1 retweet
Sat Dec 05 16:54:02 +0000 2020

De Tijd
tijd

Besmettingen in de VS bereiken nieuw record https://t.co/xHHQRgPCa7

0 retweets
Sat Dec 05 13:01:03 +0000 2020

De Tijd
tijd

CD&V-voorzitter Coens zet in op 'Power to the people' https://t.co/mOtoz7xfyR

0 retweets
Sat Dec 05 11:31:03 +0000 2020

De Tijd
tijd

Beginnen met beleggen: Waarom beleggen? https://t.co/0rqSf4ogg7

0 retweets
Sat Dec 05 09:38:32 +0000 2020

De Tijd
tijd

In de greep van de cocaïnemaffia https://t.co/13iyZNbw6O

De Tijd
tijd

'Pentagon blokkeert overleg tussen inlichtingendiensten en Biden'

Figure A4.1: A rudimentary visualization of line-based JSON tweets collected and visualized by Twarc.

Setup

Twarc is a Python tool which can be called from the command line or as a library within your own code, similar to the Instaloader tool.

Configuration

The Twarc command line tool has a dedicated "configure" command which stores Twitter credentials, but no configuration file for the creation of collections exists. Similar to the Instaloader tool, command line parameters for a specific collection could be saved as a text file and provided when calling the tool.

Creating a new collection

[176] https://github.com/pbinkley/twarc-report

A new collection is created by passing accounts as parameters to the command line tool, see listing 6.

Creating an account-based collection

```
foo@bar:~$ twarc timeline tijd > tijd-tweets.jsonl
```

Listing 6: A command to create a collection of tweets from the timeline of the newspaper De Tijd, the data is stored in line-based JSON.

Creating a keyword-based collection

Different Twitter API endpoints can be used, among others Twarc can access the “search” and “filter” endpoint using the commands with the same name, as seen in listing 7.

```
foo@bar:~$ twarc search ‘#CovidBe OR #CovidBelgium’
```

Listing 7: A command to create a collection of tweets following the provided search term. Additional parameters such as --geocode or --lang can be provided to narrow down the search.

Monitoring a collection

There is no monitoring out of the box with Twarc, however, it creates log files which could enable another script to provide this functionality.

A.5 Social Feed Manager

Social Feed Manager (SFM) is a tool which offers harvesting of several social media providers, it is modular and builds upon existing harvesters and thus mainly provides a management layer with a user interface and extensive provenance information. Web harvesting via Heritrix used to be available from within SFM as well, but it was deprecated in version 1.12.0 released in 2018 due to some problems in scaling and error handling[177].

Features

The main feature of SFM is its modularity, it provides an interface to create and export social media collections. Furthermore it relies on existing harvesters while maintaining harvester-independent provenance information such as *who* created *when* a collection and *when* was it changed the last time by *whom* and *why*.

Setup

Due to the modularity the setup at first seems quite complex as several components need to be setup: the main UI component, the different harvesters, a PostgreSQL database, a UI consumer, a WARC proxy and most importantly a RabbitMQ message queue via which all other components

[177] https://gwu-libraries.github.io/sfm-ui/posts/2018-06-13-releasing-1-12

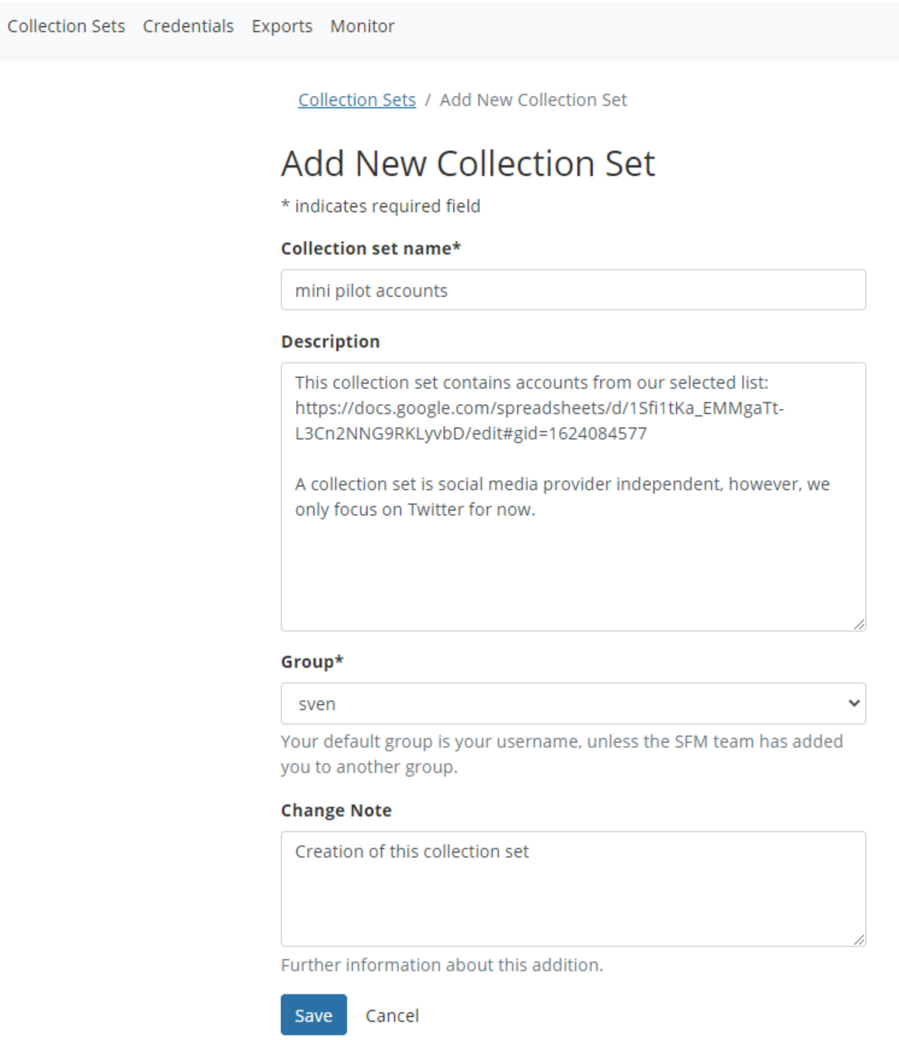
communicate. However, SFM is fully dockerized and thus can be set up in minutes with docker-compose configuration files for which examples are provided by the Team behind SFM as well.

Configuration

If docker-compose is used, the different components can be configured with provided environment variables,for which plenty exist in an .env file, including extensive comments. Actual harvesting can be fully configured using the user interface.

Creating a new collection

SFM knows the concept “collection set” which is a social media provider independent set for which a description and access rights can be configured, see figure A4. Part of such collection sets are collections which are associated with a specific harvester (configuration), e.g. a Tumblr harvester, a Twitter search or a Twitter filter. Each such collection can be configured to run in an repeating schedule, using specific API credentials, and seeds can be added which are harvester specific, e.g. a query string such as “#Covid19Be OR #CovidBelgium” for a Twitter search collection. Be aware that only one seed can be actively used by a collection.



Add New Collection Set

* indicates required field

Collection set name*

mini pilot accounts

Description

This collection set contains accounts from our selected list:
https://docs.google.com/spreadsheets/d/1Sfi1tKa_EMMgaTt-L3Cn2NNG9RKLyvbD/edit#gid=1624084577

A collection set is social media provider independent, however, we only focus on Twitter for now.

Group*

sven

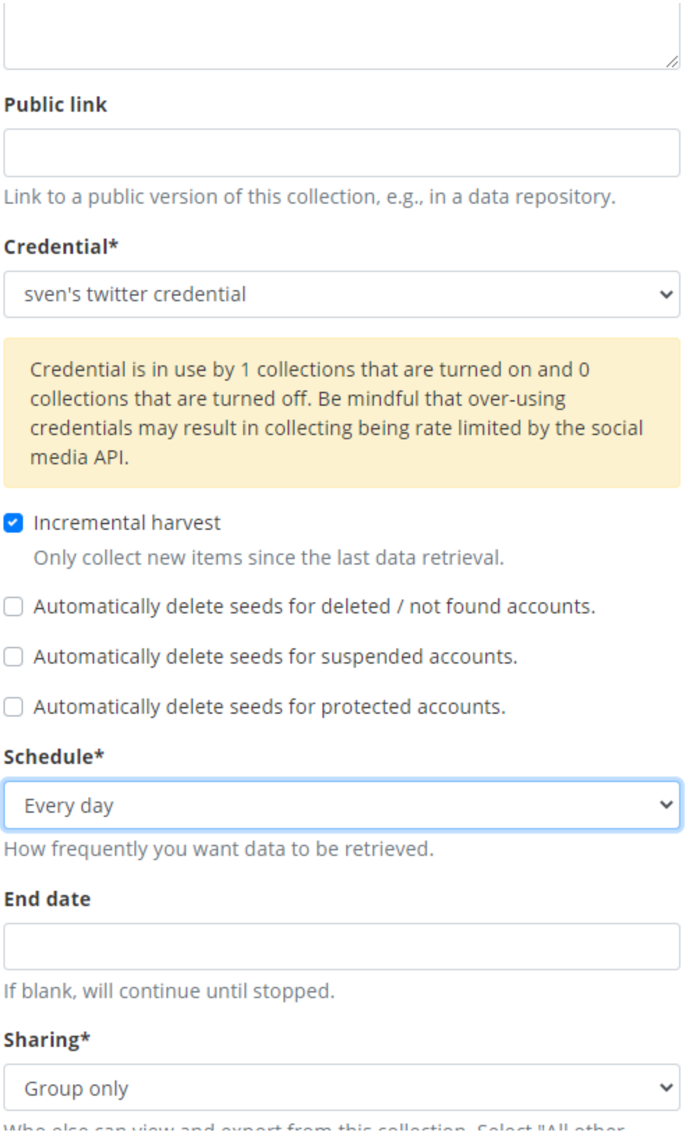
Your default group is your username, unless the SFM team has added you to another group.

Change Note

Creation of this collection set

Further information about this addition.

Figure A5.1: The creation of a new collection set in SFM which is harvester-independent and can contain several harvester-dependent collections.



Public link

Link to a public version of this collection, e.g., in a data repository.

Credential*

sven's twitter credential

Credential is in use by 1 collections that are turned on and 0 collections that are turned off. Be mindful that over-using credentials may result in collecting being rate limited by the social media API.

☑ Incremental harvest

Only collect new items since the last data retrieval.

□ Automatically delete seeds for deleted / not found accounts.

□ Automatically delete seeds for suspended accounts.

□ Automatically delete seeds for protected accounts.

Schedule*

Every day

How frequently you want data to be retrieved.

End date

If blank, will continue until stopped.

Sharing*

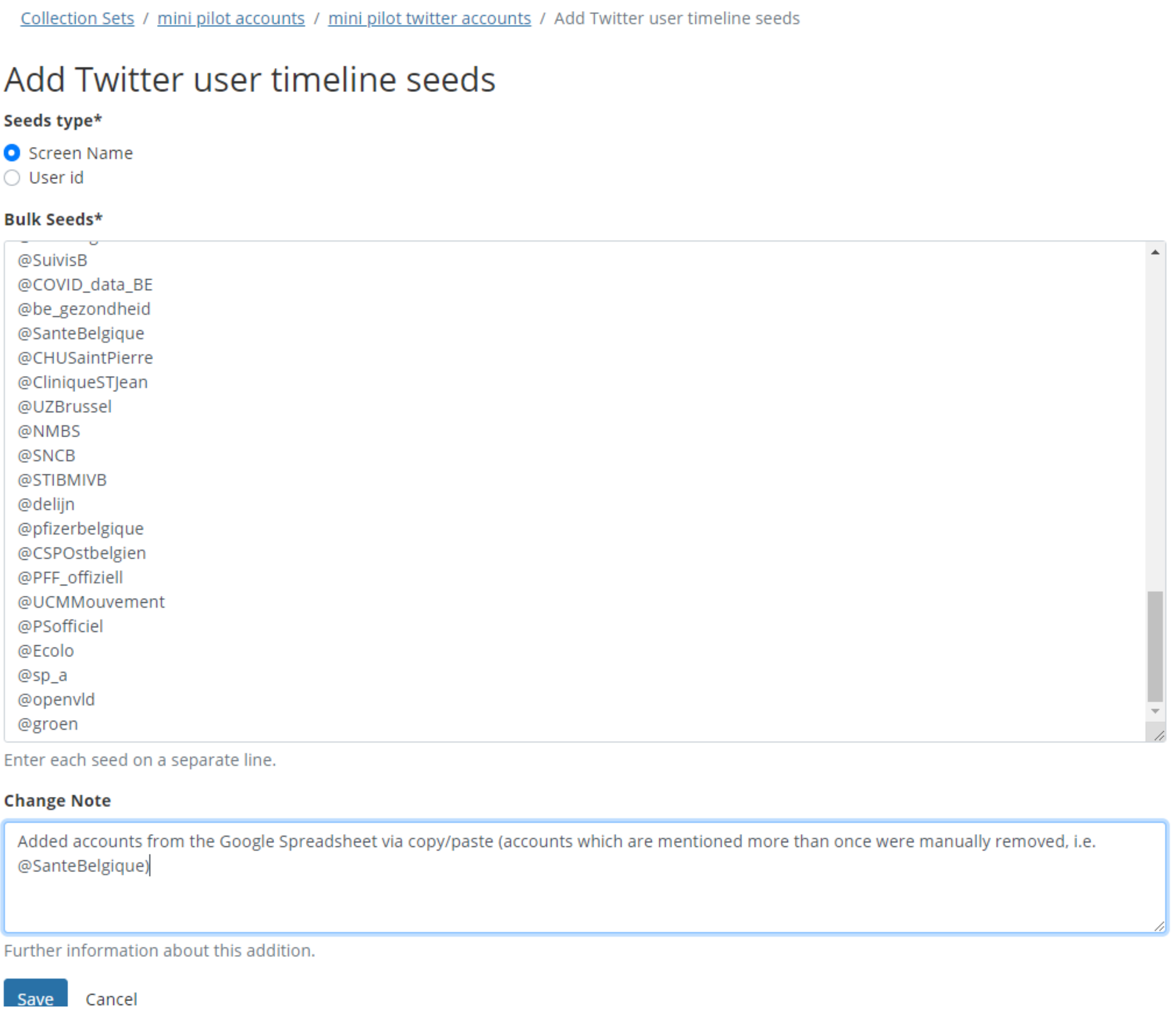
Group only

Who else can view and export from this collection. Select "All other

Figure A5.2: The creation of a harvester-specific collection in SFM: schedules and credentials can be configured as well as access rights. Additionally a warning is shown if the chosen API credentials are already in use by other collections.

Creating an account-based collection

For an account-based selection a “Twitter timeline” collection can be selected which will use Twarc to harvest tweets as JSON files from the Twitter API which will be wrapped in WARC files to preserve the HTTP request provenance. Accounts of such a collection can be specified in bulk as seeds, see figure A5.3.



Collection Sets / mini pilot accounts / mini pilot twitter accounts / Add Twitter user timeline seeds

Add Twitter user timeline seeds

Seeds type*

● Screen Name
○ User id

Bulk Seeds*

@SuivisB
@COVID_data_BE
@be_gezondheid
@SanteBelgique
@CHUSaintPierre
@CliniqueSTJean
@UZBrussel
@NMBS
@SNCB
@STIBMIVB
@delijn
@pfizerbelgique
@CSPOstbelgien
@PFF_offiziell
@UCMMouvement
@PSofficiel
@Ecolo
@sp_a
@openvld
@groen

Enter each seed on a separate line.

Change Note

Added accounts from the Google Spreadsheet via copy/paste (accounts which are mentioned more than once were manually removed, i.e. @SanteBelgique)

Further information about this addition.

Figure A5.3: Specify the seeds for an account-based collection in bulk using SFM: In this case Belgian Twitter accounts for a test of SFM harvesting.

Creating a keyword-based collection

For a keyword-based selection either a “Twitter search” or a “Twitter filter” collection can be made.

Monitoring a collection

SFM offers a dedicated monitoring tab in which currently running and previously executed harvests can be seen. For scheduled harvests per collection, error messages of the last run are also visible in a detailed view on a collections site.

A.6 Webrecorder (Conifier)

The Python tool Webrecorder comes with a user interface which allows to record browsing as a form to harvest the browsed content. A webservice exists, but a desktop version of the tool can also be downloaded from GitHub.

Features

This tool records what is shown on a website, thus all the dynamic content will be archived, it uses Browsertrix as its crawling system and makes use of remote browsers for the recording functionality.

Setup

This application consists of several components such as the UI, harvesting browsers, a database etc. For the desktop version a docker-compose configuration is available to quickly start all necessary components.

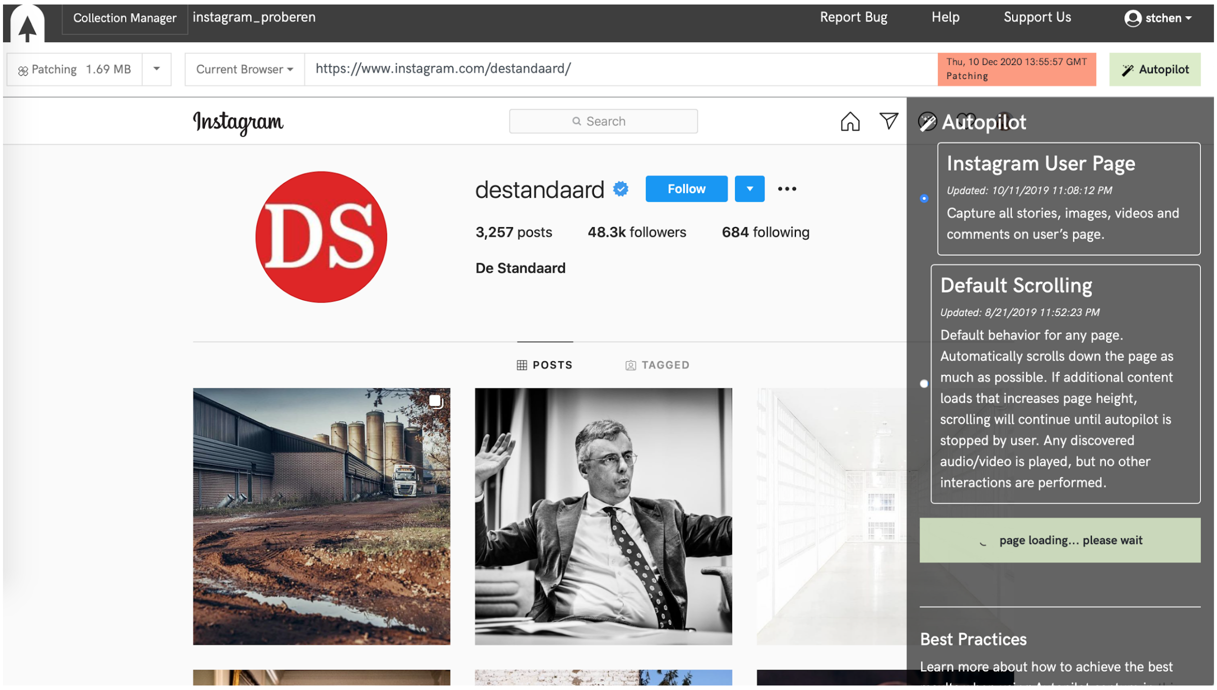
Configuration

All configurations of harvests are performed via a user interface in which collections are created and manual recording sessions, called captures, are started. Any login to potential social media providers happens manually, thus no API keys need to be provided.

Creating a new collection

Via the web interface a "new capture" can be created, a URL needs to be provided as well as a browser which will be used for harvesting. New private or public collections can be created/reused via a web interface when creating a capture, a name for the collection has to be provided. Later also descriptions can be added to a collection. When pressing a "start capture button", one browses to the selected URL and all user actions are recorded, e.g. scrolling down to capture the content. When pressing stop the capturing will stop.

An autopilot option exists which by default scrolls down until the end of the page, but can have specific behavior for certain social media providers, e.g. open tweets while scrolling down and harvest them too including replies.



https://www.instagram.com/destandaard/

Thu, 10 Dec 2020 13:55:57 GMT Patching

destandaard

3,257 posts 48.3k followers 684 following

De Standaard

Instagram User Page

Updated: 10/11/2019 11:08:12 PM

Capture all stories, images, videos and comments on user's page.

Default Scrolling

Updated: 8/21/2019 11:52:23 PM

Default behavior for any page. Automatically scrolls down the page as much as possible. If additional content loads that increases page height, scrolling will continue until autopilot is stopped by user. Any discovered audio/video is played, but no other interactions are performed.

page loading... please wait

Best Practices

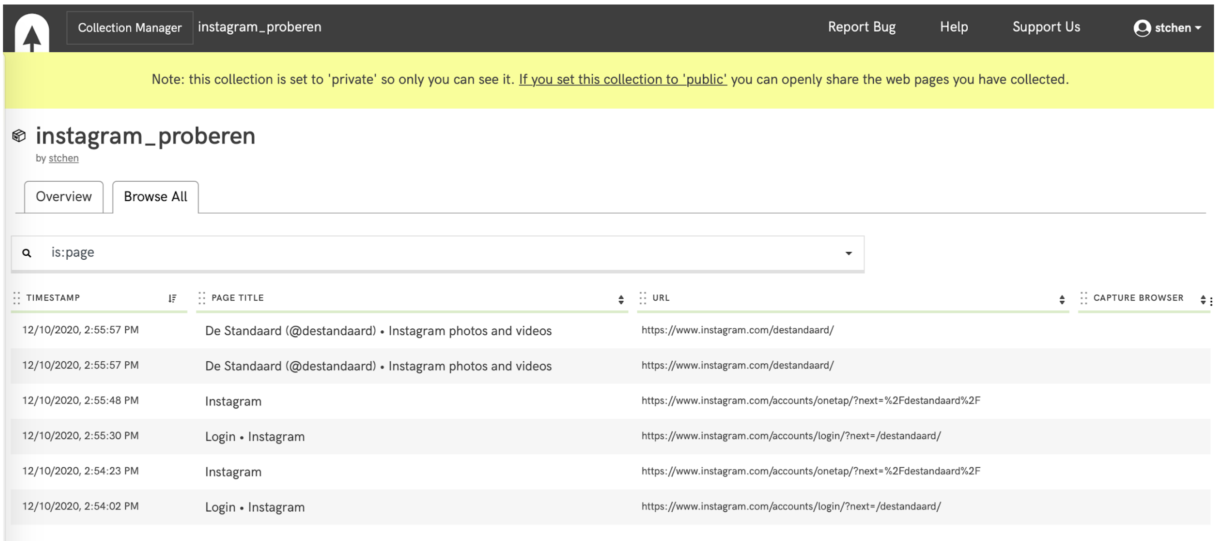
Learn more about how to achieve the best

Figure A6.1: Webrecorder during harvesting in "patching" mode where a specific website is captured and added to an existing capture.

Afterwards in a collection-manager view an overview of harvested pages can be viewed, missing content, e.g. when following a link on a recorded website which wasn't collected, can be "patched", see figure A6.1. So called "lists" can be created for each collection and thus the collection's content can be sorted in different lists, which can be set to be visible public or private.

A collection cover shows a preview of the collection, e.g. a list of texts from tweets together with a link to the live URL, see figure A6.2.

Collections can be downloaded as WARC files or be completed by uploading new WARC files.



Note: this collection is set to 'private' so only you can see it. If you set this collection to 'public' you can openly share the web pages you have collected.

instagram_proberen

by stchen

is:page

TIMESTAMP	PAGE TITLE	URL	CAPTURE BROWSER
12/10/2020, 2:55:57 PM	De Standaard (@destandaard) • Instagram photos and videos	https://www.instagram.com/destandaard/	
12/10/2020, 2:55:57 PM	De Standaard (@destandaard) • Instagram photos and videos	https://www.instagram.com/destandaard/	
12/10/2020, 2:55:48 PM	Instagram	https://www.instagram.com/accounts/onetap/?next=%2Fdestandaard%2F	
12/10/2020, 2:55:30 PM	Login • Instagram	https://www.instagram.com/accounts/login/?next=/destandaard/	
12/10/2020, 2:54:23 PM	Instagram	https://www.instagram.com/accounts/onetap/?next=%2Fdestandaard%2F	
12/10/2020, 2:54:02 PM	Login • Instagram	https://www.instagram.com/accounts/login/?next=/destandaard/	

Figure A6.2: The overview page of a collection created with Webrecorder.

Creating an account-based collection

For an account-based collection, a user only has to browse to account websites.

Creating a keyword-based collection

One has to browse to the search offered by the social media provider and search to see a list of search results. Then these results can be harvested with the recording feature.

Monitoring a collection

The harvesting happens live, i.e. people browsing, thus there is no job to be monitored programmatically, however, a log of what is harvested is created, see listing 8.

```
Viewed post CIZBf-yAXsp
Loaded 0 additional comment replies for post CIZBf-yAXsp
Loading post CIZBf-yAXsp comment replies
Viewed the contents of post CIZBf-yAXsp
Viewing post CIZBf-yAXsp
```

```
Viewed post CIdTzq9hjyx
Loaded 0 additional comment replies for post CIdTzq9hjyx
Loading post CIdTzq9hjyx comment replies
Viewed the contents of post CIdTzq9hjyx

Auto Captured Content:
Posts Captured: 12
Stories Captured: false
Highlights Captured: false
```

Listing 8: An example of log entries created by Webrecorder which are also shown during a harvest.