International Journal of Digital Humanities (2021) 2:107–128
https://doi.org/10.1007/s42803-021-00036-1

ORIGINAL PAPER

Web-archiving and social media: an exploratory analysis

Call for papers digital humanities and web archives – A special issue of international journal of digital humanities

Eveline Vlassenroot[1] · Sally Chambers[2] · Sven Lieber[3] · Alejandra Michel[4] · Friedel Geeraert[5] · Jessica Pranger[5] · Julie Birkholz[6] · Peter Mechant[1]

Received: 5 October 2020 / Accepted: 18 April 2021 / Published online: 22 June 2021

Abstract
The archived web provides an important footprint of the past, documenting online social behaviour through social media, and news through media outlets websites and government sites. Consequently, web archiving is increasingly gaining attention of heritage institutions, academics and policy makers. The importance of web archives as data resources for (digital) scholars has been acknowledged for investigating the past. Still, heritage institutions and academics struggle to 'keep up to pace' with the fast evolving changes of the World Wide Web and with the changing habits and practices of internet users. While a number of national institutions have set up a national framework to archive 'regular' web pages, social media archiving (SMA) is still in its infancy with various countries starting up pilot archiving projects. SMA is not without challenges; the sheer volume of social media content, the lack of technical standards for capturing or storing social media data and social media's ephemeral character can be impeding factors. The goal of this article is three-fold. First, we aim to extend the most recent descriptive state-of-the-art of national web archiving, published in the first issue of International Journal of Digital Humanities (March 2019) with information on SMA. Secondly, we outline the current legal, technical and operational (such as the selection and preservation policy) aspects of archiving social media content. This is complemented with results from an online survey to which 15 institutions responded. Finally, we discuss and reflect on important challenges in SMA that should be considered in future archiving projects.

Keywords Web archiving · Social media archiving

✉ Eveline Vlassenroot
Eveline.Vlassenroot@UGent.be

Extended author information available on the last page of the article

1 Introduction

The aim of this article is to study the landscape of born-digital collections in heritage institutions and academia, in particular social media archiving (SMA) initiatives. This study is conducted in the context of the BESOCIAL research project that aims to develop a sustainable SMA strategy for Belgium. The research that led to these results is funded by the Belgian Federal Science Policy Office.

The aim of the article is threefold. First we aim to extend the most recent descriptive state-of-the-art article on national web archiving (Vlassenroot et al., 2019) by complementing it with a review of SMA practices. Secondly, we outline the legal, technical and operational (such as the selection and preservation policy) aspects of archiving social media content. To facilitate this review, a survey was conducted from July to September 2020, targeting cultural heritage institutions and their current practices in SMA. Thirdly, we want to discuss and reflect on important challenges in SMA that should be considered.

The article is structured as follows. The first section of the article 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 the second section, the methodology is described for the desk research, survey and synthesis. The results are presented in detail in the third section including the selection and preservation of, and access to, archived social media data, the tools used for SMA and considerations when using social media archives as a historical data source. In the final section the challenges for SMA initiatives are discussed, as well as the limitations of the empirical research and a future research agenda for SMA is posited.

2 Literature review

2.1 Social media

Coined in the nineties, it took until the mid-2000s for the phrase 'social media' to enter common parlance (Ortner et al., 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 et al., 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 function as representative of the "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 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

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 resources that exist in "no other format but the digital original", and are "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, n.d.-a) 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 social media archiving (SMA) initiatives. One of the pioneer projects in SMA is the Occasio project, launched in 1994 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 had been captured prior to this date, whereas the UK National Archives has archives of Twitter accounts dating back to 2008 in its collections (Espley et al., 2014; Hockx-Yu, 2014).

Also 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 allows us to document the past in ways we have never previously had the ability, as well as provide easy ways to archive. They are an invaluable resource for researchers to study human behavior and history as they provide clear records of communication (Ruths & Pfeffer, 2014). The logs, social media posts and related metadata allows us to document the past in ways we have never previously had the ability, as well as provide easy ways 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 so doing, 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 can be included under the umbrella of Article 10 of the Convention.[3] In particular, the Court added that providing citizens with Internet archives forms a substantial contribution to the preservation and the accessibility of news and information and constitutes also a valuable source for education and historical research.[4]

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 content, 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[5]; however, they go a step further by

[1] See for instance 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.

[3] ECHR (4th sect.), case of *Times Newspapers LTD (Nos. 1 and 2) v. The United Kingdom*, 10 March 2009, app. Nos 3002/03 and 23,676/03, §27; 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 20,148, app. Nos 60,798/10 and 65,599/10, §§90 and 102.

[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 23,676/03, §45.

[5] 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.

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.[6] 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 for 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 (Littman et al., 2018; Vlassenroot et al., 2019); this is often in contrast to the needs of practitioners (Venlet et al., 2018). In case of social media this metadata can be provided via Application Programming Interfaces (APIs).

Several metadata standards exist from which a common subset can be distilled, however, most tools which create metadata define descriptive metadata differently and mostly collect technical metadata. NISO lists 11 metadata standards in the cultural heritage field ranging from the storage efficient machine readable MARC format family developed in 1968 to several XML-Schemas and OWL ontologies like DDI and PREMIS developed in recent years. Whereas these standards cover different types of metadata, Dooley and Bowers (2018) reviewed existing metadata standards with respect to descriptive metadata and recommend the use of 14 data elements.[7] 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 & Dooley, 2018).

[6] 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".

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

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 e.g. the 'Social Media Research Toolkit',[8] or the wiki 'Social media data collection tools'.[9] A list of general web harvesting tools were collected by the Data Together initiative in 2018 in form of a collaborative spreadsheet (Hucka, 2017).

3 Methodology

The research methodology consisted of three phases. In the first phase, a secondary research approach (also known as desk research) was taken. This involved summarising, collating and/or synthesising documentation related to existing SMA projects. A number of archiving initiatives were selected and analysed in depth (see Table 1).

With regard to the selection of our sample of web archiving initiatives, a number of characteristics were taken into account:

- Web archiving initiatives that were included in PROMISE-project[10] - 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) is 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 main research question for this study was: how are national libraries and archives engaging in social media archiving (as an extension to Vlassenroot and authors (2019) on web archiving)? 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 regard 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 addition to the desk research related to SMA projects, a desk research study was carried out by computer engineers reviewing documentation and GitHub repositories.

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

[9] See: Freelon (n.d.). Social media data collection tools. http://socialmediadata.wikidot.com/.

[10] The criteria for the PROMISE Project included: established web archiving initiatives; web archiving initiatives in countries where both the national library and the national archives are involved in web archiving (as the PROMISE project was a collaboration between the Belgian Royal Library and State Archives, useful lessons could be drawn from countries where both institutions engage in web archiving); web archiving initiatives in countries with multiple official languages; web archiving initiatives in countries of different sizes; and a combination of web archiving initiatives relying on external service providers and initiatives that manage all aspects of the process in-house.

Table 1 Web-archiving initiatives that were considered

Country	Institution	Name	Abbreviation
Canada	National Library	Library and Archives Canada	LAC
Canada	Regional Library	Bibliothèque at Archives nationales du	BAnQ
Denmark	Royal Danish Library	Netarkivet	Netarkivet
Estonia	National Library	Eesti Veebiarhiiv	Eesti Veebiarhii
France	National Library	Bibliothèque nationale de France	BnF
France	National Audiovisual	Institut national de l'audiovisuel	INA
Hungary	National Library	National Szèchènyi Library	NSL
Ireland	National Library	National Library of Ireland	NLI
Luxembourg	National Library	Bibliothèque nationale du Luxembourg	BnL
New-Zealand	National Library	National Library of New Zealand	NLNZ
Switzerland	National Library	Webarchiv Schweiz	Webarchiv Schweiz
The Netherlands	National Library	KB Webarchief	KB
The Netherlands	National Archive	National Archief	NA
UK	British Library	UK Web Archive	UKWA
USA	University Library	George Washington University libraries	GWUL

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 store 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.

In the second research phase, a questionnaire which ran from July to September 2020 was sent to representatives from the aforementioned institutions. The aim of this survey was to address the gaps that remained in 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 had already been answered beforehand and the respondents were asked to verify them. 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. The answers to the questions that were obtained during the desk research and from the survey were integrated.

In order to respond to the research question and create an overarching view of the selected SMA initiatives, comparisons were drawn in an exploratory analysis based on these answers.

4 Results

Below we present the results from the survey. This section is structured as follows: 1) selection and collecting - outlining which social media platforms are being archived by

institutions, and keeping in mind legal aspects; 2) data management - this includes ingesting and storage of the data, as well as the data stewardship of managing the technical aspects of archiving; and tools used – this provides an overview of the different tools, strategies and practices of institutions; 3) access, use and reuse - how are the archives accessed and if and how they are used in research; and 4) preservation policies and practices.

4.1 Selection of content for social media archiving

Vlassenroot et al. (Vlassenroot et al., 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 regard to social media differed widely between institutions. Table 2 provides an update of this overview, including data from additional SMA initiatives (updated data is marked in bold). 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 2 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; celebrities, world leaders and politicians, among others, 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 2, 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).

As reported by Vlassenroot et al. (2019), social media accounts that are captured focus, in general 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) 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 is thus often lost (e.g. number of likes, retweets, comments; see also ‘exhaust data’ (McCracken, 2007), ‘read wear’ (Hill et al., 1992) or ‘attention metadata’ (Najjar et al., 2006)).

While most archiving institutions use a twofold approach for archiving regular web content – combining broad crawls (covering top-level domains) and selective crawls

Table 2 Overview of social media archived by web archives

Country	Abbreviations	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	Dailymotion, Vimeo SoundCloud
Hungary	NSL	No	No	No	Yes	No	Occasionally (currently in a pilot phase)
Ireland	NLI	No	Yes	Yes	No	No	Very limited amount of social media archiving (focusing their efforts on websites)
Luxembourg	BnL	Yes	Yes	Yes	No	No	
New-Zealand	NLNZ	Yes	Yes	No	Yes	No	
Switzerland	Webarchiv Schwiez	No	No	No	No	No	
The Netherlands	KB	No	No	No	No	No	WhatsApp thriller (story that consist of WhatsApp messages) + Started a social media archiving pilot in 2020
The Netherlands	NA	No	No	No	No	No	
UK	UKWA	Yes	Yes	No	No	No	
USA	GWUL	No	Yes	No	No	No	

(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. As such, the archiving institutions in our sample only use selective crawls to archive social media content.

Most often these 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 include numerous social network 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.

4.2 Framing by the law

The first important legal aspect to keep in mind when archiving social media content is their public or private character. This question, far from being simple, raises many controversies both among authors and jurisprudence. For some, all data and information exposed on social media can no longer belong to the private sphere and are intrinsically public (Manach, 2010; Pailler, 2012). To decide between private and public spheres, it is rather recommended to consider the factual circumstances and to make a case-by-case analysis using different criteria. In this respect, the "reasonable expectations test" – used by both the European Court of Human Rights and the Belgian Court of Cassation (Raepsaet, 2011) – is a good indicator. The idea is to ask whether the user can reasonably expect his right to privacy to be protected when publications that may contain "personal data" are disseminated on social media. Without going into detail, however, it should be noted that the right to privacy is not absolute and that interferences are allowed if the conditions of legality, legitimacy and necessity are met. Thus, if this "triple test" is respected, it is conceivable to select social media content that belongs to the private sphere.

The second important legal aspect is related to data protection law. In our previous contribution (Vlassenroot et al., 2019), we briefly pointed out the specificity of the GDPR in terms of archiving in the public interest.[11] This is particularly interesting for heritage institutions in their web and SMA activities. Although born-digital content often contains what is called "personal data"[12] and their preservation constitutes "processing"[13] this data, the European legislator has been attentive to the issue of heritage preservation. It should be noted that, in the context of social media archiving, many elements may constitute personal data: web and social media content, signed posts and comments, someone's cultural preferences (literary, cinematographic, musical or artistic tastes), opinions, comments and views expressed by natural persons on blogs and social medias, names and surnames of natural persons, personal and professional contact details of natural persons (postal address, email address, telephone number) both personal and professional of natural persons, bibliographical data, photos, religious or political beliefs, sexual orientation, economic income or standard of living, etc. Hopefully, the GDPR facilitates the work of heritage institutions by establishing a derogatory regime for personal data processing for archiving purposes in the public interest. We regret however that no legal definition of this particular

[11] See Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR), *O.J.*, 4 May 2016, L 119/1, art. 89, §§1 and 3.

[12] According to the GDPR, personal data means "any information relating to an identified or identifiable natural person". See GDPR, art. 4 (1).

[13] According to the GDPR, processing means "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means […]". See GDPR, art. 4 (2). Besides data processing taking the form of "long-term conservation", in the context of web and social media archiving, we can also cite other processing operations that fall under GDPR scope such as collection, consultation, use, communication, erasure and destruction.

processing exists. Recital 158 of the GDPR merely states that "*public authorities or public or private bodies that hold records of public interest should be services which, pursuant to Union or Member State law, have a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for general public interest*". To benefit from the derogatory regime, the activity of the preservation of archives in the public interest must be legally required by Union or national law. In our view, heritage institutions such as national libraries, national archives and some museums undoubtedly fulfil this condition. On the other hand, it is not the case for those who perform a cultural mission of preserving archives without relying on a legal obligation.[14]

In their web and social media archiving activities, heritage institutions that meet the conditions set out in Recital 158 will benefit from exemptions from certain key processing principles, from certain obligations imposed on them and on certain rights conferred to data subjects. It should however be stressed that these exemptions only apply for the processing operation of "long-term preservation" (i.e. archiving) of archives of "public interest", that is to say archives which have a certain cultural, heritage or historical value for society. Moreover, while some exemptions are directly applicable because they are provided by the GDPR, others are simply allowed by the GDPR but must be put in place by the European or national legislator in order to be applicable.

4.3 Data management and the tools used

In this section we discuss tools, file formats and metadata used by the surveyed institutions and we present preliminary results of ongoing work to compare social media capturing tools with respect to their features, availability and metadata capabilities. The current version of the comparison table is available on the BESOCIAL project page. It is based on existing work (Hucka, 2017) with the intention of contributing specific social media related insights back to the original author.

Although most institutions have similar archiving processes for both social media and websites, a few institutions mentioned issues with setups for capturing social media. Specific tools and scripts are used such as Twarc and the Social Feed Manager (SFM). In terms of data formats most institutions use the WARC format, but the native JSON formats from harvested social media providers, which may be stored separately or within the WARC files, are also stored by some institutions.

While most institutions report that they do not use a metadata standard to describe captured social media data, some institutions use customised schemas (i.e. the national libraries of Canada and New Zealand) or plan to explore the possibility of using a metadata standard. The National Library of Estonia is planning to adopt Dublin Core this year and the National Library of Hungary is already using a Dublin Core-based metadata scheme for selected websites but not yet for social media content.

However, even if no metadata standard is currently used, metadata stored in WARC files or created by harvesting tools is present, and, thus, could potentially be mapped to any metadata schema as Linked Data. According to the National Library of France they "*store response/request/metadata records in WARC files and also keep all*

[14] They can still benefit from another very similar derogatory regime: the one for historical research purposes.

configuration, log and report files from Heritrix to document the crawls precisely". Additionally the National Library of France reported to use many levels of documentation, such as a general collection policy, scoping and guidance notes and selection metadata like harvest frequency collected by a tool called *BCweb*.

Of the most common tools used: 4CAT, APIBlender, Brozzler, SFM, STACKS, DMI-TCAT, Twarc and Webrecorder (Conifer), all of them are open source, available via an open licence and cover common social media platforms, mostly however Twitter. Whereas most tools are still updated, which is crucial given that social media APIs are subject to change, the APIBlender tool appears to no longer be actively maintained. Several of these tools are modular and hence could be more or less easily extended for support of other social media platforms. Additionally these tools store both metadata about the collection and detailed descriptive metadata in form of API responses, which can satisfy an often stated provenance need in social media research. Whereas most of these tools store received JSON data in their own database structures or as files, SFM also uses WARC to store API requests and responses.

4.4 Access, use and reuse

4.4.1 Use of social media archives for research

Web and social media archives provide an invaluable resource for researchers to study human behaviour and history as they provide clear records of communication (Ruths & 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 (Alalwan et al., 2017; Cheston et al., 2013; Filo et al., 2015; Leung et al., 2013; 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; Pal et al., 2016; Rabina et al., 2013); 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 (Ashrafi-rizi & Kazempour, 2020; Cinelli et al., 2020). This includes digital ethnographies and qualitative research, as well as quantitative or statistical analysis of the social media material.

The results from the 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.

Vlassenroot et al. (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 4 shows that only a single institution (NLI) has their social media collection open and freely accessible online without any requirements for access. However, their

Table 3 The exemptions contained in the derogatory regime for personal data processing for archiving purposes in the public interest

Exemptions	Legal basis	Conditions
Purpose limitation principle	GDPR, art. 5,§1,b)	Implementation of appropriate safeguards for data subjects′ rights and freedoms (technical and organisational measures)
Storage limitation principle	GDPR, art. 5,§1,e)	Implementation of appropriate safeguards for data subject's rights and freedoms (technical and organistional measures) **AND** Personal data solely processed for archiving purposes in the public interest
Prohibition to process special categories of personal data	European Union or national law	Processing necessary for archiving purposes in the public interest **AND** European Union or national law shall proportionate to the aim pursued, respect essence of data protection right and provide suitable and specific measures for data subjects' rights and interests
Right to information in case of indirect data acquisition	GDPR, art. 14	Implementation of appropriate safeguards for data subjects' rights and freedoms (technical and organisational measures) **AND** Providing information likely to render impossible/seriously impair the fulfilment of the archiving purposes in the public interest
Right to erasure	GDPR, art. 17, §3,d)	Implementation of appropriate safeguards for data subjects' rights and freedoms (technical and organisational I measures) **AND** Processing necessary for archiving purposes in the public interest **AND** Implementation of the likely to render impossible/seriously impair the fulfilment of the archiving purses in the public interest
Right of access Right to rectification Right to restriction Right to data a portability Right to object Obligation of notification in case of erasure, rectification or restriction	European Union or national law	Implementation of appropriate safeguards for data subjects' rights and freedoms (technical and organisational measures) **AND** Exemption necessary to fulfil the archiving purposes in the public interest **AND** Implementation of the right likely to render impossible/seriously impair the fulfilment of the archiving purposes in the public interest

Table 4 Overview of access methods to the social media web archives

Country	Institution	Open & freely accessible online	Physical access on location	Requirement 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 from within the partner libraries)	Authorized users of the BnF (proof their identity and show their need to access the BnF collection for their research)
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)	Yes	No requirements
New Zealand	National Library	Yes (some content e.g. Twitter ID's)	Yes	Only for researchers who sign up with the reading room.
Switzerland	National Library	No social media archiving	No Social media Achiving	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 reader 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.

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 found a workaround and show the metadata publicly only, 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 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.

4.5 Framing by the law

From a data protection perspective, national legislators may determine appropriate safeguards with regard to access to archives containing personal data processed for archiving purposes in the public interest. Obligations vary depending on whether personal data are pseudonymised or not (Table 5).

Table 5 Appropriate safeguards introduced by the Belgian legislator for access to personal data processed for archiving purposes in the public interest

	Communication	**Dissemination**
Definition	Third parties who consult the archives are identified	Third parties who consult the archives are not identified
Pseudonymised data	Can be disseminated	Can be disseminated, except special categories of personal data listed in Article 9 of the GDPR
Non-pseudonymised data	Can be disseminated but heritage institutions must prevent data reproduction other than for personal data relations to criminal convictions and offences; if agreement with the initial data controller phobibits it; or if the data reproductiveother than in a handwritten form would be detrimental to data subjects' security **AND** Can be disseminated without this specific obligation if: consent; personal data have a close connection with public/historical nature of facts in which data subjects has been involved	Can't be disseminated, unless: consent; personal data made public by data subjects; personal data closely related to the subject's public/historical character; or personal data have a close connection with public/historical nature of the facts in which data subject has been involved

4.6 Preservation policies and practices

In order to gather an overview of the policies and procedures in place for the preservation of archived social media content, we dedicated a particular section of the survey to such questions.

When asked to briefly describe the preservation procedures in place, two respondents provided specific responses such as "WARC files are bit-preserved on several distinct physical locations in a digital preservation system" (BnL) or social media is harvested as WARCs which are integrated within the thematic web archiving collections, which are then organised, described and arranged being preserved on tape (LAC). Furthermore, the NLNZ noted that they take specific actions to make social media content more preservable, such as un-shortening shortened URLs included in Tweets and also capturing content that has been linked to from social media.

In response to this question, a number of organisations described their capture, harvesting or archiving procedures, rather than describing digital preservation procedures in particular. One respondent (NSL) noted that they currently do not have a procedure.

Respondents were also asked whether written preservation procedures were available for archived social media content. Seven responding institutions did not have a specific written preservation procedure for archived social media content. One organisation (BnF) mentioned that the same procedure is used as for the harvested web content. Another (LAC) mentioned that a specialised procedure is currently under development and would be included in their Strategy for a Digital Preservation Program. Furthermore, the NLNZ mentioned that they have internal documentation available.

Four respondents indicated that specific preservation systems or software is used: Preservica (BnL), Rosetta (NLNZ), a Hadoop cluster (UKWA) and SPAR (BnF). Two institutions are also planning to acquire such systems or software in the near future (LAC, NSL). When asked whether any other tools were used for long-term preservation of archived social media content, two respondents (KB & LAC) indicated that the content was both stored digitally on servers and also on tapes.

Regarding preservation standards or norms used for archived social media, four respondents (BnF, BnL, UKWA and LAC) indicated they use the WARC ISO 28500 standard and LAC also uses its predecessor, the ARC format. One respondent (NLNZ) indicated that no specific standards or norms are used for social media compared to other collections. Regarding format migration, the respondents who answered this question all mention that they do not routinely migrate formats of archived social media content to specific formats for preservation purposes.

Finally, respondents were asked what they consider to be the biggest challenges for preserving social media. Formats were cited the most often as the biggest challenge, including: the evolution of formats (BnF), format migration (LAC), the interrelationship between formats and accessibility (e.g. to be able to provide both user - and preservation - friendly formats that are also easily understandable) (NLNZ) and changes in the original format provided by the social media platform (INA). For two institutions, acquiring a preservation system and setting up the necessary procedures remain the most important challenge (KB, NSL). One institution (Netarkivet) sees the lack of standardisation as the biggest challenge: methods and APIs change so often that

it becomes necessary to maintain collection, access and preservation procedures for large amounts of social media data. Other elements that were mentioned related to preservation are technological evolutions and evolutions in hardware (BnF), accessibility (BAnQ) and Digital Preservation Planning (LAC).

5 Discussion and conclusion

In this paper 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 to understand the practicalities and legal aspects of this type of archiving for institutions. Some limitations in our research approach should be noted however. Firstly, during the analysis of the results it became clear that there is a lack of common understanding of certain concepts amongst the participants who filled out the survey (e.g. 'preservation formats' or 'preservation standards') and that definitions of these terms should have been provided in the survey itself in order to improve clarity. Secondly, due to the convenience sample we gathered, we cannot claim to have used a representative sample of SMA initiatives. Thirdly, our study is based on self-reported data gathered using an online spreadsheet; although this proved to be a quick and easy method to collect data on current SMA initiatives, it was also a 'rough' data collection method in the sense that there was no opportunity to pose follow-up questions, ask participants for clarification or request more details on certain answers. Other limitations such as reliability issues in online surveys and a possible self-selection bias need to be also taken into account (Gosling et al., 2004). Nonetheless, we hope that this study will be used as a point of departure for further research on SMA.

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 the inclusion of known limitations, there is also the limitation of the tools. Our findings show that much of the crawling is done through application programming interfaces (APIs). APIs limit the information that can be collected (i.e. maximum request per day) and limit its reuse (Morstatter et al., 2014; Thomson & Kilbride, 2015), which further limits access and knowledge to access questions of quality. In most cases these limitations are due to the unwillingness of some social media companies to allow research or archiving. There is a concern that this may result in implicit unrepresentative sampling which influences the external

validity of data samples (Birkholz et al., 2013; Gayo-Avello, 2016; Tufekci, 2014). External validity refers to when the conclusions drawn from the sample are applicable for an entire population (Lucas, 2003); thus, giving a user or researcher an indication of the generalisability of the research to a specific population. Without explicit knowledge of these selection criteria, researchers are limited in drawing conclusions and testing theories.

Moreover, another challenge is how to provide and ensure access to archives and under which copyright conditions (Zimmer, 2015; George Washington University Libraries, 2017; McCreadie et al., 2012). In particular, for researchers using these broad and large archives there remain questions around the quality of these data (Tufekci, 2014); largely this is about the (in) completeness of the material and related metadata (Milligan, 2019). Despite these known challenges that are inherent to current SMA processes, social media archives remain an under-utilised resource for humanities and social science researchers in particular.

Preservation practices proved to be diverse among those surveyed. According to the UNESCO definition, "digital preservation consists of the processes aimed at ensuring the continued accessibility of digital materials" (UNESCO, n.d.-b). 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. It would also seem that there were several interpretations of what is understood as 'preservation procedures'; with harvesting and archiving being closely connected processes and sometimes used interchangeably. The WARC standard was also named as a preservation standard even though the format was developed as a storage format, rather than specifically for long-term preservation. This may demonstrate a lack of common understanding of what is considered as digital preservation procedures or formats, and therefore a need for raising awareness about preservation of social media content.

Thus, future research should consider the need for metadata and different standards and tools to provide this metadata, Future directions towards interlinked knowledge graphs, such as envisioned by NISO (Riley, 2017), demand metadata of certain qualities. The FAIR principles (Findable, Accessible, Interoperable and Reusable) provide a framework for effective (research) data management and thus data stewardship. Although there are differences between FAIR and long-term preservation, e.g. long-term preservation is explicitly excluded in a European Commission's report on FAIR data (Collins et al., 2018), it still offers a useful framework for curation (Barwick & Thieberger, 2018) and can enhance locked up metadata in database records, JSON or WARC files (typical for captured social media data).

Also, future research should further encourage researchers to consider born-digital data as one of the many available resources alongside both digitised and analogue resources. A recent paper (Holownia & Chambers, 2020) considered whether the concept of 'library labs', as pioneered by organisations such as the British Library, and more recently, exemplified through the international Building Library Labs network (Chambers et al., 2019), could be considered ideal incubators for both increasing access to born-digital resources (such as web and social media archives) and encouraging their analysis alongside digitised and even analogue sources. By developing sustainable data curation workflows (Candela et al., 2020; Padilla et al., 2019), these 'Collections as Data' can be published as (FAIR) datasets within a 'labs' environment,

therefore increasing their visibility and potentially their take-up and usage. Initiatives like WARCnet have dedicated working groups on research data management across borders (Chambers, 2020; Rosenberg, 2020), which could further stimulate the publication of such datasets.

To conclude, when polled for the challenges they face, the respondent from the National Library of Canada stated "Social media perhaps even more than [the] web demonstrates that our old thinking and traditional approaches are becoming meaningless for digital". In light of this statement, we do hope that our article supports academics and professionals in the fields of archiving, library and information science in understanding better the fast-evolving field of SMA, and in general, that it contributes to the knowledge about the challenges involved in social media archiving.

Acknowledgements We would like to thank the survey respondents for sharing their experiences with web archiving and SMA. The respondents also gave their permission to their data being shared with the wider research community. The spreadsheet containing the source data can be consulted on the BESOCIAL project page. This is also a good illustration of how open the international web archiving community is and our commitment to Open Science.

References

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.

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

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

Barwick, L., & Thieberger, N. (2018). Unlocking the archives. In N. Ostler, V. Ferreira, & C. Mosely, *Communities in Control: Learning Tools and Strategies for Multilingual Endangered Language communities*. Proceedings of FEL XXI Alanena 2017. Hungerford, UK: Foundation for Endangered Languages.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Birkholz, J. M., Wang, S., Groth, P., & Magliacane, S. (2013). Who are we talking about? Identifying scientific populations online. In J. Li, G. Qi, D. Zhao, W. Nejdl, & H.-T. Zheng (Eds.), *Semantic web and web science* (pp. 237–250). Springer.

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

Chambers, S., Mahey, M., Gasser, K., Dobreva-McPherson, M., Kokegei, K., Potter, A, Ferriter, M. and Osman, R. (2019). Growing an international cultural heritage labs community. *Libraries as Research Partner in Digital Humanities, DH2019 Pre-conference Workshop, July 2019* 10.5281/zenodo.3271382.

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/

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*. European Commission.

Dooley, J., & Bowers, K. (2018). *Descriptive metadata for web archiving: Recommendations of the OCLC research library partnership web archiving metadata working group*. OCLC Research.

Espley, S., Carpentier, F., Pop, R., & Medjkoune, L. (2014). Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content. *Alexandria*, 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. http://socialmediadata.wikidot.com/. Accessed 8 September 2020.

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 University Libraries. (2017). Social feed manager. Building Social Media Archives: Collection Development Guidelines. https://gwu-libraries.github.io/sfm-ui/resources/guidelines. Accessed 8 September 2020.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 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. IFLA WLIC Libraries, Citizens, Societies: Confluence for Knowledge in Lyon. August 2014. http://library.ifla.org/999/1/107-hockxyu-en.pdf.

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.

International Institute of Social History (IISH). (2020). Occasio Digital Social History Archives. https://iisg.amsterdam/en/detail?id=https%3A%2F%2Fiisg.amsterdam%2Fid%2F collection%2FARCH04348. Accessed 24 July 2020.

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.

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 of Congress. (2017). Update on the Twitter archive at the Library of Congress. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf. Accessed 24 July 2020.

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

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.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory, 21*, 236–253. https://doi.org/10.1111/1467-9558.00187.

Macdonald, N. (2019). How the National Library preserves New Zealand's digital heritage. https://www.stuff.co.nz/national/115118698/how-the-national-library-preserves-new-zealands-digital-heritage. Accessed 8 September 2020.

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.

Manach, J. M. (2010). *La vie privée, un problème de vieux cons?* Fyp.

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. https://cultureby.com/2007/07/how-social-netw.html. Accessed 8 September 2020.

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.

Milligan, I. (2019). *History in the age of abundance. How the web is transforming historical research*. 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.

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

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). SAGE Publications Ltd..

Padilla, T., Allen, L., Frost, H., Potvin, Sarah, Russey Roke, E., & Varner, S. (2019). Final report — always already Computational: Collections as Data. 10.5281/zenodo.3152935.

Pailler, L. (2012). *Les réseaux sociaux sur internet et le droit au respect de la vie privée*. Larcier.

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

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010), Altmetrics: A manifesto. 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.

Raepsaet, F. (2011). Les attentes raisonnables en matière de vie privée. *Journal des tribunaux du travail, 10*, 145–185.

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

Riley, J. (2017). Understanding metadata: What is metadata, and what is it for?: A primer. *National Information Standards Organization*. https://www.niso.org/publications/understanding-metadata-2017

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). SAGE Publications Ltd..

Rosenberg, K. (2020). Working groups. https://cc.au.dk/en/warcnet/working-groups/. Accessed 8 September 2020.

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

Samouelian, M., & Dooley, J. (2018). *Descriptive metadata for web archiving: Review of harvesting tools*. OLCL Research.

Social Media Data Scholarship. (2020). Social Media Research Toolkit. https://socialmediadata.org/social-media-research-toolkit/. Accessed 8 September 2020.

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.

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.

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

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

Treem, J. W., & Leonardi, P. M. (2012). Social media use in organizations: Exploring the affordances of visibility. *Editability, Persistence, and Association. In Communication Yearbook, 36*, 143–189.

UNESCO. (n.d.-a). Concept of digital heritage. https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital-heritage Accessed 14 September 2020.

UNESCO. (n.d.-b). Concept of digital preservation. https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital- preservation. Accessed 7 September 2020.
UNESCO. (2003). Charter on the preservation of the digital heritage. https://unesdoc.unesco.org/ark:/48223/pf0000179529.locale=en Accessed 14 September 2020.
Venlet, J., Stoll Farrell, K., Kim, T., Jai O'Dell, A., & Dooley, J. (2018). *Descriptive metadata for web archiving: Literature review of user needs*. 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. *Int. Journal of Digital Humanities, 1*, 85–111.
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.ss
Zimmer, M. (2015). The twitter archive at the Library of Congress: Challenges for information practice and information policy. First Monday, 20(7).

Affiliations

Eveline Vlassenroot[1] · Sally Chambers[2] · Sven Lieber[3] · Alejandra Michel[4] · Friedel Geeraert[5] · Jessica Pranger[5] · Julie Birkholz[6] · Peter Mechant[1]

1 Imec-mict-UGent, Ghent University, Ghent, Belgium

2 Ghent Centre for Digital Humanities, Ghent University, Ghent, Belgium

3 Imec - IDLab - UGent, Ghent University, Ghent, Belgium

4 Université de Namur, NADI/CRIDS, Namur, Belgium

5 Royal Library of Belgium, Brussel, Belgium

6 Department of Literary Studies (WeChangEd), Ghent University, Ghent, Belgium