SCHOOL OF ADVANCED STUDY
University of London

Archiving Social Media: A Comparative Study of the Practices, Obstacles, and Opportunities Related to the Development of Social Media Archives

By

Beatrice Cannelli

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Digital Humanities Research Hub

Declaration

I declare that the work presented in this thesis is my own, and that information derived from published or unpublished work has been appropriately credited in the text and a list of references is given.

Beatrice Cannelli

Abstract

This thesis investigates challenges and solutions related to the development of social media collections. In the last decade, the cultural value of social media and the discussions generated on these platforms have been widely recognised, with archiving institutions increasingly looking into preserving this important resource. However, while archiving institutions have made significant progress in establishing practices related to the preservation of websites, social media data still presents significant challenges. Despite being an integral part of the World Wide Web, social media platforms have become a sort of separate ecosystem with unique characteristics, dynamics and technical aspects that increasingly separate them from traditional websites. While scholarship has explored web archiving practices at length, little research has been specifically dedicated to investigating challenges and approaches concerning the development of social media collections.

Using a combination of desk research and an exploratory online survey, this thesis begins by offering an overview of the state of the art, shedding light on the geographical location and stage of development of social media archiving initiatives, and uncovering imbalances and gaps in the international preservation of these sites. Moreover, the research identifies obstacles hindering the establishing of social media archives and collections in countries that are in the process of developing such initiatives and which are typically located in the Global North. Drawing from fieldwork and interviews with twelve web archivists from memory institutions archiving or planning to archive social media, this thesis provides a cross-national comparative analysis of the practices and solutions adopted for the collection of social media. The study is further enriched by two case studies focusing on archiving initiatives operating under electronic legal deposit legislation in the United Kingdom and France.

The comparative study demonstrates how the interaction of national legal frameworks, platforms' policies, technical problems, and lack of resources available heavily shape social media collecting activities and the granularity of collections, drawing attention to potential concerns of representativeness of collections at a national level. This thesis offers novel insights on the specific challenges that social media platforms pose to archiving institutions compared to other born-digital materials like websites. This research contributes to the advancement of social media archiving practices by offering guidelines and recommendations to support the practical development of future initiatives, extending beyond national libraries and archives.

Acknowledgments

This thesis marks the culmination of a rewarding yet challenging journey, made possible and enriched by the contributions of many individuals along the way. First of all, I would like to thank my wonderful supervisors, Professor Jane Winters and Dr Naomi Wells for their unwavering support and encouragement over these past four years. I deeply appreciate your invaluable feedback, insightful comments, and for generously sharing your knowledge and expertise, guiding me in refining this thesis and helping me grow as a researcher.

This research would have not been possible without the generous financial support of the London Arts and Humanities Doctoral Partnership (LAHP). I would like to extend a special thanks to the LAHP staff for their ongoing assistance and for offering additional learning opportunities throughout the course of my studies.

Thank you to the School of Advanced Study, the Digital Humanities Research Hub and the university professional services staff for creating such a welcoming and inspiring environment.

I would like to express my gratitude to the WARCnet for welcoming me and providing the unique opportunity to connect and engage with key researchers in the field of web archive studies. I am especially grateful for the financial support provided by WARCnet through the Short-term Network Stay programme, offering invaluable knowledge and insights that have significantly contributed to this research.

A special thank you goes to all the web archivists who kindly participated in this study, generously giving their time to share their valuable experiences with archiving social media and contributing despite the numerous challenges in preserving such important resource.

I would like to thank my family for their constant support and understanding throughout this journey. Special thanks to Mimi, my cat-assistant, for providing much-needed cuteness and cuddles. Last but not least, thank you to my partner for always being my number one supporter, for encouraging me whenever I doubted myself.

Table of Contents

List of Figures

List of Abbreviations

Institutions

ANZ	Archives New Zealand
BL	British Library
BNCF	National Central Library of Florence/*Biblioteca Nazionale Centrale di Firenze*
BNCR	National Central Library of Rome/*Biblioteca Nazionale Centrale di Roma*
BnF	*Bibliothèque nationale de France*
BnL	*Bibliothèque nationale du Luxembourg*
BNM	Marciana National Library of Venice/*Biblioteca Nazionale Marciana di Venezia*
INA	*Institut national de l'Audiovisuel*
KB	Royal Danish Library/ *Det Kgl. Bibliotek*
KBR	Royal Library of Belgium/*Koninklijke Bibliotheek van België*
LoC	Library of Congress
MoL	Museum of London
NLS	National Library of Scotland
NSZL	National Széchényi Library/Országos Széchényi Könyvtár
TNA	The National Archives (United Kingdom)
UKGWA	UK Government Web Archive
UKWA	UK Web Archive

Other

AoT	Archive of Tomorrow
APIs	Application Programming Interface
BCWEB	*BnF Collecte du Web*
BELSPO	Belgian Policy Office
DAFF	Digital Archiving File Format
IIPC	International Internet Preservation Consortium
LDUF	Legal Deposit User Forum
NPLD	Non-Print Legal Deposit
OTP –	One Time Password
ResPaDon	Network of Partners for the Analysis and Exploration of Digital Data/*Réseau de Partenaires pour l'analyse et l'exploration de données numériques*
SNSs	Social Network Sites
SUCHO	Saving Ukraine Cultural Heritage Online

TLD/[ccTLD]	(country code) Top Level Domain
URL	Uniform Resource Locator
UGC	User Generated Content
UNESCO	United Nations Educational, Scientific and Cultural Organization
WARC	Web ARChive file format
WARCnet	Web ARChive studies network
WACZ	Web Archive Collection Zipped

CHAPTER ONE
Introduction

1.1 Aims and Significance of the Thesis

In the course of this project, the social media landscape has undergone many changes,[1] demonstrating, in the wake of health and political crises, the central role these platforms have come to play in documenting events. I first began developing this project when events like the COVID-19 pandemic were mostly confined to dystopian novels. Instead, I found myself starting my PhD journey in the middle of a health crisis, when the solitude of consequential lockdowns left individuals reaching out to a variety of social media services to communicate with family and friends, trying to make sense of unprecedented events and, sometimes, finding solace through content generated on social media. As this thesis highlights, this specific event acted as a sort of turning point for the social media landscape, confirming the prominent position social platforms hold in daily communications and emphasising the cultural value of the testimonies and conversations taking place on these sites.

As a digital assemblage of data, objects and human interactions, social media represents "a massive data resource which [has] the potential to aid our understanding of patterns of social behaviour" (UK Data Forum, 2013, p.13). Social media holds unique cultural value, as it offers an unprecedented insight in to international, national and family history, social trends, politics and economic matters (Brügger, 2017b; Burkey, 2020; Fondren & Menard McCune, 2018). However, while there may be a diffuse misconception among users that the web and social platforms are there to stay, these fragile digital environments are not archives (Brügger, 2017a; Winters, 2017). Social media platforms in particular have demonstrated on too many occasions how anomalies, glitches and the highly competitive ecosystem of online service providers can put born-digital cultural heritage at risk of disappearing (Chokshi, 2019; Pankhurst, 2016; see Section

[1] The information included in this thesis related to social media platforms and their policies is accurate as of December 2023. Moreover, it is important to note that since Twitter was acquired by Elon Musk in 2022, the platform has been in a period of extreme flux which led to a rebranding as X. However, as the name Twitter is still widely used among the web archiving community and the interviews had been conducted before the rebranding, I decided to continue referring to this platform as Twitter throughout this whole thesis.

2.2.2). It is the significance and fragility of the content and interactions generated on these platforms as digital cultural heritage that has led archiving institutions to gradually start including social media material in their collections over the past decade (see Section 2.2). Nevertheless, Elisabeth Fondren observed how "the sheer amount of digital data, the speed with which people produce and share digital media, and the fast-paced technological environment […] all pose a challenge to institutions tasked with preserving knowledge and culture" (Fondren et al, 2018, p.33). In fact, some of the first social media archiving projects at an institutional level, such as the "Twitter Archive" at the Library of Congress or the UK Government Social Media Archive at The National Archives (see Section 2.3), have demonstrated the importance of curating and providing access to such diverse and valuable material. However, they have also brought to light numerous challenges related to the atypical, ephemeral and ever-changing nature of these new media. Similar archiving endeavours have drawn attention to the need to establish and apply specific archiving solutions and create ad hoc policies for social media data collection and curation. In the last decade, archivists and information professionals have come a long way in the practice of archiving websites (Costa et al., 2017; Pennock & Beagrie, 2013; Webber, 2020). However, capturing, preserving and providing access to social media data still presents significant challenges. As Thomson (2016) pointed out in one of the first reports attempting to address specific issues surrounding the preservation of social media, collecting data from social platforms is a relatively new practice that must contend with rapidly evolving technology and ecosystem. While web archiving practices have been explored in length (Bragg & Hanna, 2013; Costa et al., 2017; Masanès, 2006), little research has been dedicated specifically to approaches related to the development of social media collections. Since starting this thesis, the interest in investigating this matter – especially following the 2020 health crisis – has increased, as demonstrated by a number of studies illustrating practices at single institutions (Bingham & Byrne, 2021) or investigating approaches to developing national collection strategies (Chambers et al., 2021). In particular, Vlassenroot et al. (2021), drawing from results gathered in the context of the BESOCIAL project[2] (KBR), has provided an exploratory review of social media archiving initiatives, most of which were embedded within web archiving institutions.

Yet, the development of international standards and guidelines for sustainably curating social media material appears to be lagging compared to the broader web

[2] https://web.archive.org/web/20240503141538/https://www.kbr.be/en/projects/besocial/

archiving field, mostly due to the complexity and the ever-changing nature of social media sites. The primary aim of this thesis is to investigate the extent to which memory institutions within and outside the UK are archiving social media content a decade after the first collecting attempts. Previous studies have drawn attention to some of the issues related to social media archiving (see for example, Bergis et al., 2018; Fondren & Menard McCune, 2018; Helmond & Vlist, 2019; Thomson, 2016; Winters, 2020; Zimmer, 2015), however, most of the available scholarship appears to discuss these in combination with websites. While social media archiving inherits some of the challenges identified in the preservation of websites (Pennock & Beagrie, 2013), it comes with unique issues and concerns that need to be investigated independently. The incessant evolution, the rise and fall of social platforms, the introduction of new features and the ephemeral nature of social media content constantly bring new problems to the already rich plethora of challenges that legal deposit libraries and other memory institutions are called upon to unravel. Moreover, unlike archiving social media data for research purposes – usually conducted as a one-off effort within a specific timeframe – institutional archiving of social platforms has to deal with questions of long-term preservation, methods of acquisition, access and sustainability that set these sites apart from any other born-digital record (Bruns & Weller, 2016; Pehlivan et al., 2021).

Although it can be assumed that most national collecting institutions already capturing websites archive, to different extents, the scattered pieces of social media material, only a few of them appear to consistently collect this material, establishing solid, specific workflows to manage and preserve them. This thesis expands the catchment population examined in aforementioned studies and captures an updated snapshot of memory institutions, including organisations that have not yet started to archive social media but plan to. Given the key role social media has had in supporting communications during critical events, this study will consider the potential influence that events unfolding between 2020 and 2023 had on the emergence of new archiving initiatives. Moreover, this thesis seeks to identify the obstacles that memory institutions planning to initiate their own social media archives encounter at the start of their projects and how these may hinder their establishment.

Through a transnational comparative analysis of existing, well-established archiving initiatives, this thesis identifies obstacles faced and practices implemented, and explores the opportunities related to the development of social media archives. This analysis is complemented by two case studies focusing on social media archiving

institutions operating within the context of electronic legal deposit legislation in the United Kingdom and France. These case studies provide insights into the limitations imposed and archiving solutions implemented to adapt to different legal frameworks, drawing examples from the direct experiences of the respective web archiving teams. The results of this comparative study contribute to understanding the long-standing issues and latest concerns faced by memory institutions seeking to preserve social media content. Overall, this study contends that archiving institutions need to critically distinguish between websites and social media material in their archiving activities, given their inherently different nature, and emphasises the importance of developing specific archiving strategies for this type of material. The comparative analysis of approaches and solutions adopted by various institutions to tackle archiving problems unique to social media presented in this thesis, has allowed for the identification of guidelines and recommendations, which provide future initiatives with the knowledge necessary to critically develop social media collections. Recommendations include strategies to ensure representativeness of national collections and increase awareness as well as engagement with these important born-digital, web-based resources.

1.2 Scope of the Study

Web and social media archiving is to be considered as any form of "deliberate and purposive" preservation of web-based material, which can be carried out on both "macro" and "micro" levels (Brügger, 2017b, p.79). This thesis focuses on the "macro" level, exploring the collecting practices conducted by national archiving institutions with the intent of preserving the cultural heritage generated on social platforms (Brügger, 2017b). It is worth mentioning here the concerns raised over the years about the accuracy of the term "web archive" and the extent to which web archives can be considered, in fact, as archives (Brügger, 2017a, 2018b). In particular, Brügger (2017a) observed the misleading nature of a term that is often used to refer to collections mostly preserved within libraries. He proposed the term "webrary" – that is "a collection of Web publications" – as a potential and more suitable alternative name to indicate institutions that preserve content made publicly available, and therefore "published", on the web (Brügger, 2017a). Nevertheless, the terms "web archive" and "web archiving" have been widely accepted by the web preservation community since they were first coined in the 1990s and will be used in this thesis to indicate the institutions and the preservation

activities related to content made available on the web. While in this study I argue the need to operate a distinction between archiving strategies for websites and social media, the term "web archive/ing" will be used to indicate all the activities concerning the collection of web-based content, including social media, as this is broadly used to refer without distinction to websites and platforms' material by web archivists.

As will be illustrated in the course of this study, there is indeed a frequent overlap and lack of distinction between web and social media archiving practices within most archiving institutions. This, coupled with the fact that social media shares some of the challenges concerning the preservation of a wide variety of other born-digital material, necessitates emphasising that the focus of this thesis will be on the challenges highlighted by web archivists regarding the capture, curation and preservation of social platforms. While additional challenges that may affect the preservation of this data and the way social media collections will be studied in the future will be taken into consideration, it was deemed essential to maintain a focus on the obstacles which web archivists at national memory institutions perceive as crucial in determining the development of the resulting collections and the shape that they ultimately take. As a trained archivist with experience working in various archives in Italy and the UK, handling a wide array of traditional and digital records and the unique issues each context presents, I found it particularly important to capture web archivists' direct experiences. This does not mean that the archiving knowledge, for example, of other information professionals or researchers operating at private organisations should be considered of less value. Rather, to understand how the collective memory shared on social platforms is preserved at cultural heritage institutions, it is essential to explore web archivists' reasoning processes, and the limitations and concerns they face in their everyday activities. Moreover, a comprehensive understanding of professionals' first-hand experience with social media records enabled the compilation of guidelines that clearly reflect the needs of the archival community and can assist other professionals within the field.

Although a fundamental resource for everyone interested in web and social media history, the Internet Archive (IA) has been excluded from this research. The IA is a private digital library with a broad transnational scope, which sets this archiving initiative apart from other national cultural heritage institutions (Ogden, 2020). The latter are instead the primary focus of this thesis. In fact, the Internet Archive, as Masanès (2006) observed, has made its principal purpose that of archiving the whole web with no distinction between what may be considered of cultural value and what not (see Section 2.2).

Conversely, national memory institutions must meticulously select content from social platforms that has a certain cultural value, based on the scope and mission of their organisation.

Given the role that cultural heritage organisations have in ensuring the preservation – no matter the format – of records and artifacts that represent the collective memory of a certain area of the globe (Jacobsen et al., 2013), in this thesis I will use the term "memory institution" to refer to Libraries, Archives, Museums, Galleries and other institutions including Universities, whose mission is to manage and preserve documentary cultural heritage for the long term. While recognising the biases that may surround the preservation of mainstream histories in national archiving institutions (Charlton, 2017; Jimerson, 2006; Schwartz & Cook, 2002, see also Section 2.5), the term "memory institution" is used here to emphasise the role that these institutions play and the power they wield in preserving and shaping national memory. This thesis will, therefore, also discuss biases and factors that may influence the selection of material included in social media collections, illuminating opportunities that can support the development of more representative collections. In this context, it is essential to also clarify the use that will be made in this thesis of the concept of "heritage". As I will argue in this thesis, social media represents a unique type of born-digital heritage, which may not be perceived as such through the lenses of Authorised Heritage Discourse (AHD) — a term coined by Smith (2006) to refer to discourses that favour "monumentality and grand scale, innate artefact/site significance tied to time depth, scientific/aesthetic expert judgement, social consensus and nation building" (Smith, 2006, p. 11). Challenging this power/knowledge dynamics rooted in the Western Eurocentric tradition (Robinson, 2018), Smith (2006) defines "heritage" as a "multilayered performance" that is part of everyday life, and can be understood in different ways among a diverse variety of communities and countries over time. In Section 2.2.1, I will discuss how social media platforms, as "digital cultural heritage" (UNESCO, 2003), serve as virtual spaces for individuals to engage in discussions, reflect on past and present events, thus contributing to the production of a collective memory that goes beyond traditional performances of heritage and involves wider portions of society.

Finally, it should be noted that I have been a member of the Web ARChive Network (WARCnet)[3] for most of my PhD research, between 2020 and 2023. The

[3] https://web.archive.org/web/20240520221344/https://cc.au.dk/en/warcnet/about

WARCnet was a network of researchers and web archivists involved in the promotion of high-quality national and transnational research using content preserved in web archives. Being part of this network has enabled me to establish fundamental connections with web archivists and researchers dealing with the numerous challenges concerning web and social media archiving practices. In addition to facilitating communication with crucial web archiving initiatives, ensuring a smooth and responsive interviews invitation process, I was also awarded two consecutive travel grants (called Short-term Network Stays)[4] to conduct in-person interviews and fieldwork at the *Bibliothèque nationale de France* (BnF), the *Institut national de l'Audiovisuel* (INA), and the *Netarkivet* at the Royal Danish Library (KB). The STNS provided invaluable insights into the practices, challenges and administrative aspects related to social media archiving activities at the aforementioned institutions (see Section 1.4.5). In addition, in 2023 I joined the Legal Deposit User Forum (LDUF)[5] which has been established with the intent to amplify the voices and needs of legal deposit collections in the United Kingdom. While the work of the LDUF is still in its early stages, it has provided useful knowledge and a better understanding of various aspects related to access to e-legal deposit collections, including those preserved as part of the UK Web Archive at the British Library.

1.3 Research Questions

This thesis intends to explore the social media archiving state of the art and the diverse challenges faced by web archivists and information professionals in relation to the preservation of this unique and ephemeral type of born-digital record. The key research questions to be addressed are as follows:

- *How are cultural heritage institutions within and outside the UK dealing with challenges related to the preservation of social media after nearly a decade of activity?* What are the main legal and curatorial issues that institutions had to work on? What tools are being used

[4] The WARCnet Short-term Network Stays were intended to support research and knowledge exchange among the members of the network through short-term visits. A report of the two STNS can be found here:
https://web.archive.org/web/20230902214655/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Canelli_Reporting_from.pdf

[5] https://web.archive.org/web/20240301104936/https://libguides.tcd.ie/legal-deposit-user-forum

to archive social media? What are the issues faced and solutions adopted in relation to such tools?

- *What has been achieved so far in the field of social media archiving?* Where are the latest social media archiving initiatives located? How are they distributed geographically? When and why did cultural heritage institutions start to archive social media? What is the scope of their collections?
- *What are the barriers that prevent cultural heritage institutions from archiving social media?* How do national legal frameworks affect the development of social media archives?
- *What are the new concerns or obstacles related to the development of social media archives that need to be considered?* How have web curators been dealing with these new challenges? Do selection criteria take into consideration concerns about inclusivity and representativeness? Have critical events, such as the COVID-19 pandemic, impacted collection development?

1.4 Research Design

The first part of this section provides an overview of the methods used to collect data. A survey questionnaire was employed to map the distribution of current social media archiving initiatives and to identify participants for the interviews. Semi-structured interviews were subsequently used to gather in-depth information on practices and challenges related to social media archiving. Then, the rationale behind including two case studies is discussed, along with the reasons for selecting these specific institutions.

1.4.1 Comparative Method

A qualitative comparative method was adopted to answer the research questions identified for this study. This method was deemed to be the ideal approach to investigate the obstacles that archiving social media presents. Previous studies focussing on web archiving initiatives and the related challenges relied in equal measure upon qualitative research approaches or single case studies (e.g., Cadavid, 2014; Davis, 2014; Day, 2006; Lomborg, 2012; Masanes, 2006). Given the nature of social media content and the limited number of social media archives worldwide, a qualitative cross-national approach was perceived as a suitable method to gather meaningful data about current issues and policies. A qualitative comparative approach allows for the recognition of particular similarities

and the identification of differences between a selected group of interest and their experiences of a specific phenomenon (Ritchie et al., 2013).

Given that memory institutions around the world preserve records reflecting their country's or region's history and culture, it was important to consider the specific context of each institution when examining the challenges they face. Palmberger & Gingrich (2022) pointed out that comparison in qualitative analysis seeks to "achieve abstraction by doing justice to the context in which the different cases are embedded" (p.95). Context is indeed a critical aspect in this research, given the central role that, for instance, national legal frameworks have in shaping social media collection scope, selection practices and access. Moreover, a qualitative comparison facilitates going beyond the particularities of single cases and establishing a common ground among all the instances analysed, in order to extract significant aspects that may be applicable to a wide range of similar cases (Palmberger & Gingrich, 2022). The aim of this research is in fact twofold: first, to gather critical information by comparing the first-hand experience of web archivists and information professionals regarding the issues surrounding social media archiving; and second, to gain knowledge about solutions applied to specific issues to identify good practices and propose guidelines that could benefit current initiatives and the development of future social media archives. To clarify, this research does not intend to examine practices with the intent to criticise the quality of practices or solutions adopted. Rather it aims to understand the state of the art of social media archiving and attempts to offer an overview of well-suited solutions that have been adopted in regard to specific issues, in order to guide future archivists through the *mare magnum* of social media archiving.

Transnational, comparative research was carried out to capture the direct experiences of archivists in various countries, aiming to return a diverse picture of the different solutions adopted in different cultural and legal contexts to address the challenges encountered when collecting social media content. Furthermore, the uneven and sparse distribution of social media archiving initiatives underscores the need for cross-national comparative studies involving a range of diverse countries, especially when investigating the reasons behind the absence of international standards for preserving this type of born-digital record (see Section 1.2.3). It is worth noting how working with participants from different countries and languages, coupled with the additional filter often provided by the act of speaking of technical processes in a language that is not participants' native language, may lead to misunderstandings and inconsistent use of

terms. Nevertheless, the majority of the vocabulary and the tools used for web and social media archiving are predominantly created in the English language. Although the dominance of the English language in analogue settings still generates its own concerns and disadvantages (Drubin & Kellogg, 2012; Tsuda, 1998; Wolk, 2004; see also Section 3.4), the existence of a common language has certainly facilitated the exchange of concepts and information and the discussion of archiving issues between the interviewer and participants during the data collection process.

1.4.2 Case Studies

The use of case studies alongside a more general, cross-national comparison between participating institutions is intended to offer a detailed account of the problems and solutions related to two particularly significant instances. Creswell (2022) defined case studies as the research approach in which a researcher examines in depth one or more instances of a phenomenon. Mostly used in social science, case studies offer the opportunity to conduct a systematic and critical enquiry of mechanisms and to investigate the relation between causes and effects regarding a specific matter (Lougen, 2009; Simons, 2009). Moreover, another definition enunciated by MacDonald & Walker (1975) clarified that case studies allow for an examination of an instance in action, where the term "instance" is used specifically to imply the possibility for generalisation that comes from the study of a singular case. Also, this research approach allows the researcher to investigate a phenomenon within its context, in a real-life situation (Yin, 1994). The case study research design has proven to be suitable to investigate archival issues and has in fact been used in numerous previous studies and in doctoral theses examining aspects and practices related to archives, both traditional and born-digital.[6] Therefore, a case study approach was used to explore further the matter at hand, capturing the complexity and uniqueness of the social media archiving challenges faced, and the practices implemented at memory institutions within two specific legal contexts.

Due to time constraints, the research focussed on two case studies: the UK Web Archive at the British Library (BL) and the French e-legal deposit institutions, the *Bibliothèque nationale de France* (BnF) and the *Institut national de l'Audiovisuel* (INA). These

[6] See for example, Marshall, Jennifer Alycen (2007) *Accounting for Disposition: A Comparative Case Study of Appraisal Documentation at the National Archives and Records Administration in the United States, Library and Archives Canada, and the National Archives of Australia.* Doctoral Dissertation, University of Pittsburgh.

institutions were selected for their advanced archiving practices, established workflows, and innovative solutions to specific issues, making them valuable sources of information for this study. In particular, the decision to include the two French institutions was influenced by the distinctive distribution of the e-legal deposit mandate between them and the chance to visit their facilities in person, including the Bibliothèque François Mitterrand and INA's Training Centre at Bry-sur-Marne. The extensive information gathered through fieldwork, including informal conversations with various team members about their daily operations, justified a more in-depth analysis of these institutions. As for the UKWA at the British Library, it was selected as a key example for its comprehensive and high-quality information obtained from interviews and available documentation. The French and British instances were then contextualised and compared with the remainder of participants to gain a deeper understanding of the problems surrounding the development of social media archiving initiatives.

1.4.3 Survey and Site Selection

A comprehensive analysis of available scholarship and online resources revealed a significant gap in the knowledge concerning the discovery of social media archiving initiatives. This gap became evident due to the difficulties in finding an accurate and complete list of social media archiving initiatives that would go beyond the already existing web archiving institutions and the related legal deposit libraries and archives. In fact, while previous studies resulted in the creation of resources such as a Wikipedia page[7] listing numerous web archiving initiatives, including details about the year in which they had been established and the volume of content archived, no equivalent resource specifically exists for social media archives. Information related to social media platforms archived as part of broader web archiving efforts has simply been incorporated in the aforementioned list. Furthermore, aside from a few, well-established memory institutions, finding substantial information about archives or independent social media collections through a simple Internet search proved to be quite challenging. These discovery issues are particularly evident for smaller, newer initiatives that are not integrated into mainstream networks concerning internet or born-digital material preservation (e.g., the International Internet Preservation Consortium).

[7]https://web.archive.org/web/20231230001605/https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

For this reason, a survey questionnaire was designed to expand the scope of known projects and include social media archiving initiatives developed by memory institutions both within and beyond the GLAM sector. The aim was to identify any Library, Archive, Museum, Gallery, University, or governmental agency which was preserving social media as part of their collections. Moreover, the survey was extended to institutions that were in the planning stages, or merely considering the inclusion of social media in their digital collections. A short online survey was deemed to be the most suitable and effective way to reach out to a target population located in diverse geographical regions, and a specific type of participant (Evans & Mathur, 2005). The survey was created using Microsoft Forms. This tool was preferred amongst others available primarily for its compliance with the University's ethical standards and the latest data protection legislation. The data collected via Microsoft Forms was safely kept in the University's cloud storage and accessible only to the researcher.

Results from the survey were essential to gather information about the state of the art concerning social media archiving and to determine which memory institutions might be available for in-depth follow-up, and which could provide valuable insights into archiving practices and their direct experience with the issues related to the preservation of social media content. Preference was given to institutions with long-standing experience in archiving both websites and social media, with a particular focus on legal deposit institutions and those managing long-term social media archiving projects. To provide a diverse overview of social media archiving activities in different legal and cultural contexts, memory institutions from different areas of the globe archiving a broad range of social platforms were prioritised. This approach also considered that institutions operating within similar regions, such as Europe, might encounter comparable challenges, especially from a legal perspective. The selection of the sites was also influenced by personal connections established, for instance, through my participation in the WARCnet network (see Section 1.2), and the potential opportunity to visit those memory institutions in-person to observe their activity up close.

Moreover, based on responses to the online questionnaire I included a representative number of various initiatives, including those conducting pilot or fixed-term projects and institutions that are still in the planning phases of starting a social media archive. This allowed the investigatation of obstacles faced by memory institutions at any stage of preserving social media material. Therefore, in designing this study, I considered the social media archiving initiatives established or planned to be established at the

following memory institutions to be valuable and diverse sources of information. The memory institutions selected for interviews were: Archives New Zealand (ANZ), Arquivo.pt, British Library (BL), *Bibliothèque nationale de France* (BnF), *Bibliothèque nationale du Luxembourg* (BnL), *Institut national de l'Audiovisuel* (INA), Library of Congress (LoC), Museum of London (MoL),[8] National Széchényi Library (NSZL), Royal Library of Belgium (KBR), Royal Danish Library (KB), and The UK National Archives (TNA).

1.4.4 Triangulation and Data Sources

Qualitative data was collected using interviews to further expand the qualitative and quantitative information gathered via the survey and to explore the complexity of social media archiving challenges and practices. The triangulation of data refers broadly "to the use of multiple methodological resources" (Natow, 2020, p.2) and it offers the opportunity to know more about a phenomenon thanks to the use of more than one research method and the combination of several perspectives related to a specific experience or event (Kern, 2018; Moran-Ellis et al., 2006; Natow, 2020). Triangulation is widely considered as a valuable strategy that increases the validity of a study as it enables reflection of the complexity of a phenomenon by gathering and combining data generated by two or more methods (Moran-Ellis et al., 2006; Natow, 2020). Moreover, triangulation ensures accuracy, mitigating potential biases that may influence the outcome. Fielding (2012) pointed out that "mixing methods puts the findings from different methods into dialogue" (p.128), helping to reveal potential weaknesses and inducing the researcher to reflect on the limits of generalisation.

This study involves the comparison of multiple data sources, in the form of information gathered predominantly through survey and interviews across different countries, observations collected during in-person visits to some of the sites selected for this study, and additional data extrapolated from documents such as collection development policies or reports compiled by some of the participant institutions (Lieber et al., 2021; Messens et al., 2021; The National Archives, 2018). Documentation regarding the development of social media collections in the form of policies or reports was used to explore the research field and identify the main areas in which memory institutions were facing challenges. Most of this type of data was collected from institutional websites and scholarly papers published by members of social media archiving initiatives. This

[8] The Museum of London has been renamed as London Museum in 2024.

documentation provided the foundation for defining the type and extent of information to be collected through the survey questionnaire and was also crucial for formulating the questions for the interviews. Moreover, information gathered from official reports and other sources (e.g., blogposts, policies) was used to complement and compare with data collected during the interviews, in order to track any potential changes that have occurred over time concerning the development of social media archiving practices.

The exploratory survey was conducted with a trifold intent: to compile an up-to-date list of institutions collecting social media content; to provide a baseline to understand the state of the art and identify patterns and gaps in the distribution of such initiatives; and to gather essential information about each archiving project, to carry out an informed selection of potential interviewees based on initiatives' characteristics and location (See Section 3.2.1). This was combined with the main source of data, which consisted of interviews conducted with a selected group of twelve institutions (see Section 1.4.5), to capture the challenges they faced in performing social media archiving activities. Interviews provided an opportunity to shed light on the many efforts made by web archivists and the day-to-day obstacles they encountered when working with one of the most ephemeral born-digital items on the web, offering a unique perspective on archiving and preservation practices as well as on potential openings for development and improvement.

Fieldwork was conducted to provide additional context for some of the interviews. Due to time constraints, COVID-19 restrictions and the location of some of the interviewees, only three memory institutions were visited in-person. Visits to the libraries of France and Denmark were made possible thanks to two travel grants offered by the WARCnet Short-term Network Stays programme (see Section 1.2). The opportunity to travel to some of these sites and talk to members of the respective social media archiving teams allowed me to observe information professionals at work and enabled me to take into consideration social media archiving challenges within different institutional and cultural contexts. I spent two and a half days with each institution. The first part of each visit started with a tour of the building where each organisation was located (e.g., Bibliothèque François Mitterrand in Paris, the "Black Diamond" in Copenhagen) and the reading rooms where users can browse the web and social media collections. This provided an opportunity to consider the physical spaces occupied by teams and access terminals, offering insight into how the web and social media archive is integrated within the broader institution. Interviews with designated web archivists were

often complemented by presentations from other members of the respective teams. Beyond the opportunity to meet interviewees in person and observe their daily activities closely, the fieldwork provided valuable insights through informal conversations with other team members. Such discussions, coupled with observation, provided an opportunity to shed light on crucial aspects that might not have emerged during the limited time available during online interviews. This included examples about content selection, documentation practices, reflections on social media's value for society, the interactions with researchers and the complexity of communicating with social media platforms (see Chapters 5 and 6). Moreover, as with most e-legal deposit institutions, archived social media can only be accessed on-site at designated terminals in reading rooms by accredited researchers. Visiting the French and Danish institutions allowed me to gather essential information about modalities of access to their web collections, including the application processes to the web archives and user experience in navigating the archived web collections—insights that were made possible only through fieldwork (see Section 5.4.4 and 6.4).

1.4.5 Conducting Interviews

Semi-structured interviews were conducted with members of staff responsible for archiving content from social media. The selection of participants for the interviews was based on the results of the exploratory survey, following the criteria illustrated in Section 1.4.3. Based on the longevity and current stage of the initiatives surveyed, three types of respondents were identified: institutions that are running a long-term project, those that have been exploring the feasibility of archiving social media with a pilot or a short-term initiative, and collecting organisations that have started planning the development of a social media archiving project.

A total of 12 semi-structured interviews were conducted between April and September 2022 with professionals such as web archivists and curators dealing with archiving social media content. In addition, to provide a clearer picture of the challenges faced at each stage of the development of social media archives, including prior to the planning phase, follow-up emails were sent to the respondents who declared in their survey responses that they did not have any plans to embark on a similar archiving effort for the foreseeable future (Italy and Lithuania). Their responses were important to record as it helped to understand the obstacles and variables that may impact a project even before it begins. Moreover, follow-up emails were sent to long-standing institutions

previously interviewed in March 2023 to gather comments and information about the impact of the acquisition of Twitter by Elon Musk in 2022 and the consequent changes implemented on long-standing web and social media archiving initiatives (see Section 2.4.4).

When the project proposal for this research was initially submitted in 2019, the plan was to conduct as many in-person interviews as possible, with only a few online. However, due to the COVID-19 pandemic and the restrictions established between 2020 and 2021 to contain the spread of the virus, most of the in-person research was moved to virtual platforms, requiring a shift toward online interviews. This unprecedented situation has led most people globally to get used to the practice of meeting online over platforms such as Zoom, quickly becoming the norm and facilitating the transition from a predominance of in-person qualitative interviews to virtual platforms (Lobe et al., 2020; Oliffe et al., 2021). As Archibald et al. (2019) reports, participants in previous studies declared that they definitely prefer online interviews rather than in-person or telephone interviews. The versatility that interviews conducted online possess – especially when most of the participants in the interviews are scattered geographically, as in the case of this research – has led to the decision to seek, as much as possible, a balance between in-person and virtual interviews depending on the institutions' location. Thus, interviews were conducted in person where possible and correlated with short site visits. The in-person interview option was only offered to participating institutions based in the UK, and those situated in countries easy to travel to, depending on travel grants available (see Section 1.2).

The majority of the interviews (nine out of twelve) were conducted via Zoom. Zoom has been designated as the preferred service provider over others available because of its widespread usage. Its simplicity, user-friendliness, and high usability have been pointed out by multiple studies on the subject (Archibald et al., 2019; Lobe et al., 2020; Oliffe et al., 2021). In-person interviews are usually preferred as they allow the capture of non-verbal cues and the context in which the participant is located, especially when the interview is conducted at the site of interest, in this case memory institutions (Brinkmann, 2014). However, advancements in technology, such as the improved quality of webcams and reliable Internet connection, have enabled researchers to collect data of comparable quality and capture the same type of non-verbal communication via video conferencing as they would in person. Jenner & Myers (2019) compared interview research conducted using different methods and tools, finding that the use of synchronous online video

interviews does not reduce the quality of data collected nor adversely affect the rapport established between researcher and interviewer. In effect, tools such as Zoom appear to facilitate the forming of such connection, enabling the researcher to provide non-verbal cues and interact with facial expression during the interview, promoting natural and lively discussions (Archibald et al., 2019; Gray et al., 2020; Oliffe et al., 2021). Mediated interviews using video conferencing tools have the added benefit of overcoming the challenges of conducting research with geographically dispersed, remote participants - as in the case of this study - considerably reducing travel expenses and environmental impact while also being time efficient (Archibald et al., 2019; Jenner & Myers, 2019). In addition to the benefits highlighted so far, Zoom has also been identified as the preferred tool in comparison to other similar service providers available because of its straightforwardness, easiness to connect to – even in situations where Internet connection is not very stable – and for its privacy and security policies (Archibald et al., 2019; Zoom Inc., n.d.).

The semi-structured interview was deemed a suitable type of approach for this study since it combines the flexibility of an open-ended interview with the directionality and the focus of a survey (Schensul et al., 1999), helping the researcher to uncover personal experiences of participants, which is one of the aims of this study. The structure of the interviews followed the focus of the research questions, aiming to uncover all the areas of concern related to social media archiving. Topics for questioning included information about the institution, the history of the archiving initiative, the national legal framework, and the challenges faced from a legal, curatorial and technical perspective as well as methods of access to the archived material (see Appendix C). The phrasing of the semi-structured questions was adapted to suit the institution type and the project stage. This was made necessary by the potential diverse experiences and challenges memory institutions at different stages of their projects were facing. Thus, a similar set of questions was created for long-term and pilot/short-term projects, while a different one was specifically composed for projects that were still in the planning phase as they would have a diverse perspective on obstacles concerning social media archiving practices (see Appendix C). A specific set of questions was instead designed for the follow-up emails sent to the two institutions that declared in the survey that they did not have plans to archive social platforms. These questions focussed predominantly on understanding the reasons behind their answer (see Appendix B).

Interviews were mostly conducted on a one-to-one basis (except for interviews with the NSZL, KBR and LoC and those held in-person, as some of them included

additional contributions coming from multiple members of the respective archiving teams). All interviews were conducted in English as this is the language that is more consistently and widely used for communication worldwide. The only interview conducted in a different language was the short follow-up via email with the BNCF in Italy. Since Italian is my first language, in this instance, communicating in this language felt more natural and practical for this specific interaction. All the interviews were recorded, with permission, using an audio recording device for in-person interviews, and the recording tool made available by the service provider for those conducted on Zoom. Recording interviews was essential for me to fully focus on the interviewing process instead of having to take detailed notes, facilitating the natural flowing of conversation without having to break eye contact with respondents (Halcomb & Davidson, 2006; Nascimento & Steinbruch, 2019). Furthermore, recording the interviews helped to minimise bias and allowed for multiple reviews, particularly when language inflections or complex parts of the conversation required further clarification at a later stage. Following each interview, initial impressions and reflections about the interaction were noted down, alongside any memos regarding the context of the interview—especially when fieldwork was conducted. Memos included information about the building, reading rooms layout and access interfaces, as well as other data gathered through observation.

Following this phase of the research, interviews were transcribed verbatim in Microsoft Word documents, to facilitate the process of data analysis. Although the act of transcribing recorded audio in written text, especially when not automated, requires time to be completed, it certainly allowed for a deeper understanding of the interviews' content, constituting a first important step towards the interpretation and analysis of the qualitative data collected. Halcomb & Davidson (2006) explored the benefits of word-for-word transcriptions, highlighting that a verbatim record is essential to facilitate data analysis since it enhances the degree of closeness between the researcher and their data. The process of transcribing spoken words into written text proved to be particularly helpful in instances when it was necessary to further verify data accuracy, without having to listen to audio recordings again, or contact participants (Halcomb & Davidson, 2006). Lastly, the transposition of audio files into text documents was essential for the following content analysis phase, which was performed using the software MAXQDA.

1.4.6 Qualitative Data Analysis with MAXQDA

Data collected through interviews were analysed with the support of MAXQDA[9], a software for qualitative and mixed methods data analysis. As MAXQDA is structured around projects, firstly I created a new project named "Social media archiving" and then I proceeded to import into it each interview transcript as separate files. Memos were then added to each document to note down the context related to interviews plus any other aspect that could be relevant to the study and was deemed important to be recorded. Information included in the memos was later incorporated into Chapters 5 and 6 to provide further reflections and insights about the institutional context and practices implemented at the various institutions. Following recommendations from existing studies employing the MAXQDA software (Elaldi & Yerliyurt, 2017; Gizzi & Rädiker, 2021; Kuckartz & Kuckartz, 2002), two types of structural codes were created. The first type indicated the main themes discussed during the interviews: "General information", "Selection practices and challenges", "Technical challenges", "Legal frameworks and challenges", "Access", "Long-term preservation", "Future Developments", "Ethical Concerns". The second layer included coloured "highlight" codes to underline important quotes, and comments concerning specific platforms such as Twitter, YouTube, Facebook and Instagram (these two grouped under the name "Meta" that is the company that owns them). Each transcription was coded according to these thematic groups and highlights, as illustrated in Figure 1.

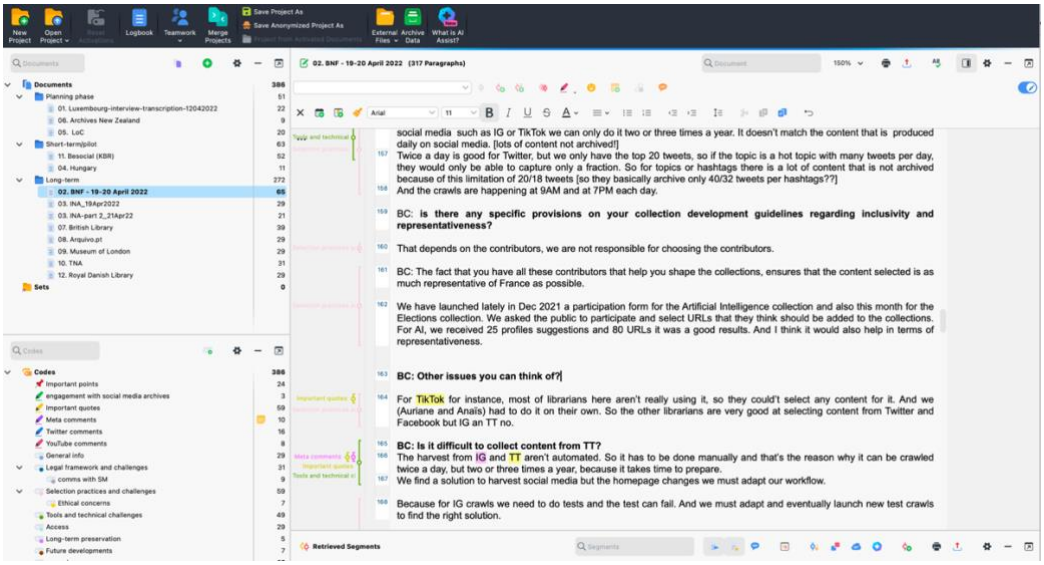

FIGURE 1: Example of interview coded in MAXQDA

[9] https://web.archive.org/web/20240906184332/https://www.maxqda.com/

After coding the documents, a content analysis was conducted on the qualitative data. The software MAXQDA offers multiple tools and features to visualise and support the analysis of the coded documents. However, for the purposes of this study, I used the Document Portrait tool, which allowed me to visualise the frequency of each code and to get a sense of the most recurring challenges mentioned by each participant. Particularly useful for the comparative analysis, which will be discussed in Chapter 6, was the ability to visualise all the portions of text from all the interviews related to a specific code. By clicking on one of the codes, for example "Selection Practices and Challenges", the system would open a pop-up window listing all the text segments marked with the selected code (Figure 2).

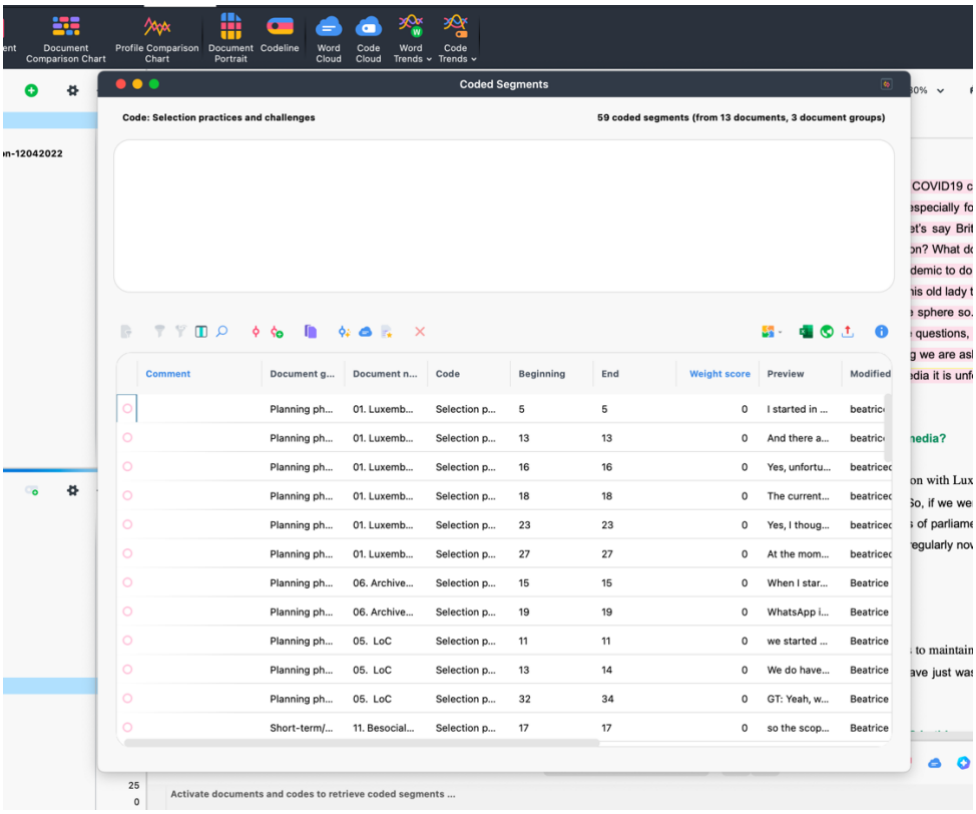

FIGURE 2: Coded Segments function

This feature provided a useful overview of the issues and examples mentioned by each interviewed institution, facilitating the comparison of similar challenges and identifying peculiar cases that could support the discussion of the cross-national comparative analysis (see Chapter 6).

1.5 Limitations

This study is not immune to some biases that are linked to the way the data collection was conducted. In regard to the survey, it is important to point out that not all the institutions worldwide that are currently or are thinking about archiving social media might have had access to the survey due to not being, for instance, subscribed to one of the mailing lists used for the recruitment of participants (see Section 3.1). In addition, a general survey fatigue and the possibility that invitation emails might have been moved automatically by spam filters into junk folders – and therefore missed - need to be considered. Nevertheless, while the sample population participating in the survey might be far from complete, the aim of the survey was never to compile a detailed census of such initiatives. Rather the aim was to provide, with all the duly made considerations, a suitable and as comprehensive as possible overview of the geographical distribution of social media archiving initiatives across the globe.

Moreover, the decision to use the English language for both the survey and the communication strategy (e.g., emails and Twitter) – a decision influenced by my affiliation with a UK-based university and a predominantly EU-based network – may have influenced the sample population demographics and limited the reach of the study to certain areas of the globe. In fact, language barriers, discoverability issues related to smaller archiving initiatives, and the network chosen to recruit participants, may have led to the involuntary exclusion of archiving initiatives or organisations located, for example, in Global South countries. To mitigate this issue, I identified through desk research additional institutions that could potentially be archiving or planning to archive social media. Moreover, by leveraging personal connections established over the years, I made efforts to reach out directly to national libraries in Asia, among other institutions, including Japan and Indonesia.

The limited amount of time allocated for interviews (60-90 minutes) to cover the wide range of challenges and solutions related to social media archiving, combined with the fact that most participants hold curatorial roles, may have influenced the depth and scope of the information collected, especially regarding technical aspects. This is not to say that participants did not provide satisfying examples or information about technical challenges. Rather, given their expertise in legal, curatorial, and access-related aspects, there was a predisposition to discuss these areas at the expense of strictly technical issues.

Finally, it is worth pointing out that the use of words such as "archive" and "archiving" may have discouraged other types of institutions that do not recognise

themselves as archiving institutions in the strict meaning of the term (see Section 1.2), including museums and galleries. In fact, as observed in one of the interviews (Aravani, Interview), the preferred term to describe preservation activities of artifacts that should be included in a museum is "collecting". Nevertheless, the use of the term "archiving" was deemed most appropriate in this context by dint of the fact that it has been broadly accepted and used in the field since the 1990s (Brügger, 2017a).

1.6 Thesis Outline

After introducing the scope and aims of the thesis and illustrating the research design strategy implemented to answer the research questions in Chapter One, Chapter Two will survey existing scholarship framing the context in which this thesis is embedded. It will explore the origin of social media, its evolution and fragility, and its cultural value as a means to facilitate the creation of collective memory. It will also provide an overview of key initiatives and events that have shaped the history of web and social media archiving, addressing concepts and research gaps related to the challenges and concerns of preserving social platforms.

Drawing from the results of a survey designed to gather essential information about archiving initiatives, Chapter Three will offer a global overview of the social media archiving landscape, identifying general gaps and trends in the preservation of social platforms worldwide. It will explore imbalances, power dynamics, representativeness concerns and obstacles that characterise the current landscape. It will also focus on the emblematic case of Italy to discuss potential challenges hindering the advancement of social media archiving activities in Global North institutions.

Having provided a general analysis of the state of the art, two case studies related to the collection of social media at legal deposit institutions in the United Kingdom (Chapter Four) and France (Chapter Five) will provide insights into specific challenges and practices implemented in two diverse cultural, institutional and archiving contexts. This will be followed, in Chapter Six, by a transnational comparative analysis of the obstacles and concerns raised by web archivists from twelve archiving institutions. This chapter will consider how the complex interaction of national legal frameworks, social media platforms' policies, institutions' decision-making processes and technical constraints influence the extent and granularity of the material preserved and the overall development of social media archiving initiatives.

Lastly, in Chapter Seven I will propose a set of guidelines and recommendations for developing future social media collections, concluding in Chapter Eight with further reflections on the findings and significance of the thesis, highlighting paths for future research.

CHAPTER TWO

Social Media, Collective Memory and Archives: A Literature Review

Over the last couple of decades, social media has radically changed the way people communicate online, rapidly becoming an integral part of our daily lives. Billions of people trust social media every day with their memories, sharing bits of their lives with friends and family and contributing to virtual debates on various topics. These new and ubiquitous means of communication have become widely employed to issue press releases and to ensure a greater demographic reach during election rallies, as well as often becoming one of the preferred media for artists to showcase their work, and an effective and direct way to communicate with their fan bases. Recent events such as the COVID-19 pandemic outbreak and the 2022 war in Ukraine have also seen social media not only establishing itself as a tool used by national governments for interacting with the public, promptly disseminating critical information regarding the emergency via official accounts; but also, having a central role as a digital environment where people could still connect, document events, share their experiences with each other and socialise when "social distancing" was put in place, for example, to limit the spread of the COVID-19 virus (Goldsmith et al., 2022a; Wong et al., 2020). While many studies have recognised the value of preserving content on the web both for the history of the web itself and as a primary source for numerous research fields (Brügger, 2017b; Brügger & Milligan, 2018; Brügger & Schroeder, 2017), the value and importance of social media content from a cultural and historical perspective needs to be addressed further. Because of the relative novelty of these platforms, researchers are still in the process of exploring and providing evidence of the potential value that these datasets may represent for future studies.

In this chapter, I will begin by defining social media and its role in daily communication and society, advocating for the recognition of the cultural value of the multitude of content shared on such new media. Significant recent examples will demonstrate how social media is profoundly intertwined with many aspects of our lives, society and contemporary history, enabling users to actively participate in the creation of a shared cultural heritage. As digital vestiges necessary to support the recollection of our collective memory, social media data must be archived, ensuring their preservation and access for future generations. Moreover, I will illustrate the central role that memory institutions play in the long-term preservation of these media and how these institutions

have evolved to allow space for web and social media materials. I will then offer a review of literature discussing legal, ethical, curatorial and technical challenges faced in the development of web archives and early attempts at archiving social media, underscoring areas that require further research. This will lay the necessary foundation and paint the complex research background in which my study is embedded, presenting a case for memory institutions, researchers working in digital humanities and, eventually, social media companies to support future steps toward new technologies, effective legal frameworks and policies that can actively encourage the development of further social media archives.

2.1 Defining Social Media

Social media has been defined as "the collective name given to Internet-based or mobile applications which allow users to form online networks or communities" (Hockx-Yu, 2014), including a wide range of web tools that permit users to create, publish and share their own content and interact with other members (Treem et al., 2016). Although one of the earliest forms of online social networking can actually be identified in Newsgroups and Bulletin Board Systems which appeared in the 1970s (Brügger, 2017b), the origin of the term "social media" can be traced back to the 1990s (Bercovici, 2010), when it was used for the first time to indicate a new medium that enabled social interactions between users on the web.

While the idea behind social media can be found long before the introduction of the so-called web 2.0 (Brügger, 2015; Ortner et al., 2019), the development of innovative tools, the shift in focus towards a more participatory culture and the added ability for users to easily generate content (O'Reilly, 2005) certainly laid the foundations for what we have learned to know and use as social media. In this regard, it has been observed how most of the single elements and functionalities used in the first social media services are far from new (Brügger, 2015). The novelty mostly lay in the way these functionalities had been mixed together to form structures in which users could interact, enabling them to fill their personal pages with content representing them as individuals and as part of a community (Brügger, 2015). As Cormode and Krishnamurthy (2008) observed in their article about the essential differences between web 1.0 and web 2.0, what separates the previous concept of the web from sites such as Facebook, Twitter or YouTube is that they potentially allow any user to participate in the creation of content in many forms (e.g.

photos, videos, text) and to add new, specific meanings when interacting with content created by other users by means of a set range of actions that include likes, comments, and re-shares. Moreover, web 2.0 appears to distinguish itself from its former version for the ability that users have to form connections (e.g., forwarding "friends" requests, subscriptions or updates) as well as facilitating communication between members of the same network, for example using the instant messaging systems internal to the platforms (Cormode & Krishnamurthy, 2008). In addition to this set of distinctive features, widespread access to the Internet and the development of mobile devices – in particular, smartphones – and application systems have further amplified the ubiquitous usage and dynamic nature of social media, leading to its growing popularity and wide circulation.

Since its appearance in the early 2000s, the expression "social media" has now commonly become part of the vocabulary and it designates many different web-based platforms and applications serving diverse – often overlapping – purposes such as social networking sites (SNSs), like Facebook and LinkedIn; microblogging sites such as Twitter; sharing user-generated content (UGC) in the form of videos (e.g. YouTube and TikTok) and photos (e.g. Instagram and Flickr); private messaging services, including WhatsApp and Telegram; as well as trade and marketing sites (TMS) such as Amazon and eBay. The distinction between social media categories, however, does not always appear to be clearly defined: the boundaries between these sites are often blurred and, for example, an SNS such as Facebook, originally created to facilitate networking interactions between users, with time has also become an important trade and marketing tool, while also encouraging users to share creative content (Dijck, 2013, p.8).

This study will mainly focus on SNSs and microblogging and UGC sites, since these are the types of platforms that have started to be archived as part of numerous web archives' collections around the world. This is mainly due to their consistent use among government agencies, public figures and the public, their unique cultural value, and the fact that the majority of the content shared on such platforms is for the most part public, setting them apart from private messaging services. Moreover, constraints related, for instance, to ethical and legal concerns, and the recent inclusion of social media content in the scope of some national digital legal deposit regulations, have further influenced the increasing archival effort towards this specific kind of social platform (for more about this regulation and the implications for social media archiving, see section 2.4).

2.2 The Cultural Value of Social Media

From the first social networking sites like Friendster (2002) or Myspace (2003) to the latest additions such as TikTok (first launched in China in 2016 and becoming global in 2019) and BeReal (2020) to merely "a flash in the pan" like Clubhouse[10] (2020), social media platforms have evolved into companies that have users' data at the core of their business models. The sheer amount of interactions, personal information and content shared everyday by users provides a source of great value for marketeers and other organisations interested in gaining, for instance, a better understanding of customers' behaviour and trends (Krombholz et al., 2012). To protect their profits, largely based on but not limited to the trade of data to third parties, social media corporations have become increasingly protective towards such information, so much so that they have been described as a sort of "walled gardens" since data is seemingly locked into their systems (Helmond et al., 2015; Thomson, 2016).

Yet that same data that is so valuable to social media companies, as it constitutes their economic fuel, also represents a form of digital trace left behind by billions of users in the fulfilment of their online lives and activities, consequently offering a privileged insight into contemporary history, societal and political trends. In this regard, scholarly research from different disciplines has recognised the importance of social media as a primary source to understand the present as well as find new paths to reinterpret past events (Burkey, 2020; Garde-Hansen, 2009; Henninger & Scifleet, 2016). Through social media, researchers have the unprecedented opportunity to study spontaneous conversations among individuals and gather insights into their daily lives on a large scale. Traditional historical research has always relied on fragments of information gathered from many different types of written records and media to reconstruct a certain epoch, and social media – together with the web – would offer a unique, varied source of information to future historians, depending, of course, on how promptly it is collected and properly preserved through time.

If, on the one hand, it may seem far-fetched – as Garde-Hansen (2009) pointed out in her chapter about digital memories and personal archives – to actually aspire to gain a full understanding of individuals' complex lives by simply analysing the multifaceted and disjointed elements posted, for instance, on their Facebook Walls; on the other hand,

[10] Clubhouse quickly gained popularity in 2020 during the first months of the COVID-19 pandemic. In April 2023, however, the company announced the layout of almost 50% of its employees due to, as per their statement, a shift in the market following the end of the pandemic.

it is evident that users' pages appear to be largely considered as personal databases, making social networks in this sense "a collection of collections and collectives"(Garde-Hansen, 2009, p.141). Moreover, researchers have been examining the usage of social media and the personal meaning that digital content on these platforms gains over time, unveiling the essential function it performs as part of users' digital belongings and as a fundamental support to memory practices (Allegrezza, 2019; Odom et al., 2010, 2011; Seyfi, 2017; Zhao & Lindley, 2014). In their study about social media as performance and exhibition, Zhao et al. (2014) observed how content shared on such platforms seems to gain meaningfulness and a distinctive personal value once it is not functional anymore to support the users' performance of identity. Thus, when the content no longer reflects a present status but conveys feelings regarding past events, it becomes memory. Seyfi (2017) added to this topic by investigating autobiographical memory and how this is influenced by sharing childhood photographs on social media. For Seyfi, some of these platforms appeared to play an important role in supporting the recollection of memories: for example, tools such as Facebook Memories allow users to see old posts shared years ago on a specific day, enabling users to reminisce about events that could potentially fade away with time (Seyfi, 2017). Furthermore, the digital objects and interactions that individuals collectively share on social media hold a significance that will help sustain the recollection of important events, capturing people's reactions and sentiment in a unique, detailed snapshot as has never been seen before.

2.2.1 Shared Memory, Shared Culture

Because of the role that social media plays in individuals' daily interactions and communication with government agencies and other institutions, the data produced on these platforms is an essential cornerstone to the future understanding of contemporary life and, foremost, forms part of the digital cultural heritage. Government officials adopted these tools to improve services and information offered to the public and groups of people that are harder to reach through traditional channels of communication, especially during health crises and natural catastrophes, or for security reasons. Social movements and activists around the globe have counted on platforms such as Twitter, Instagram or TikTok to denounce oppression and coordinate political or social protests. The #BlackLivesMatter (BLM) protests in 2020 that followed the killing of George Floyd (Hu, 2020), the 2019-2020 Hong Kong pro-democracy demonstrations (Stewart, 2019), or the 2022 war in Ukraine (Allyn, 2022; Makhortykh & Sydorova, 2017), for instance,

have shown the centrality of social media in helping to circulate useful information for protesters as well as trying to provide on-the-ground reports of events as they unfold in order to document facts and fight misinformation—especially where censorship or manipulation of news is in place (Runnacles, 2014). It is quite difficult then to envision future researchers approaching recent historical events without an essential piece such as the data produced on social media by a multitude of users: each comment and interaction made while reflecting on present or past events adds a specific value to the narrative, creating fragments that, linked together, constitute a shared collective memory.

Acting as "an arena of participation" (Simon, 2012, p.89), social media functions as a virtual space where diverse groups of people come together to engage, and share information and opinions in connection to specific historical events or pre-existing memories, emphasising the social character of the act of remembering together through digital platforms (Simon, 2012). Likewise, Tagg et al. (2016) described social media as a critical digital space where shared memory and cultural heritage is produced through the act of conversation, underlining how "digital interactions enable geographically-dispersed groups to come together to build up and disseminate a shared culture" (Tagg et al., 2016, p.2). In fact, one of the qualities of social media seems to reside in its ability to facilitate the exchange of and give public resonance to all those ideas and discussions that were once often only expressed verbally in private venues, consequently remaining hidden from written records (Farrell-Banks, 2020).

However, it has been observed how the design of such platforms – mainly based on dialogue, including those cases in which the reply or retweet is asynchronous – might open up the risk of having some voices exert a greater influence than others (Farrell-Banks, 2020). For example, Lutz (2022) noted how social media use seems to reproduce existing power dynamics and inequalities that are entrenched in offline societal, political and economic structures. Acknowledging the potential imbalance of influence within this environment is fundamental to fully understanding both content and engagement produced around certain events, as well as identifying any underrepresented or overrepresented groups (Lutz, 2022). Conversely, numerous studies have highlighted how social media has instead facilitated several minority groups to have their voice heard, also sustaining their active participation in the creation of a shared cultural heritage (Bergis et al., 2018; Farrell-Banks, 2020; Treem et al., 2016). Some marginalised communities, for example, have found in certain types of social media, such as YouTube, an ideal virtual space in which to share and communally document through user-generated recordings

examples of immaterial heritage, promoting recognition of forms of performance as expressions of their own culture.

The cultural importance of this specific type of heritage that is not associated with tangible goods has officially been recognised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) during the Convention for the Safeguarding of the Intangible Cultural Heritage held in Paris in 2003 (UNESCO, 2003b). The document produced under this convention defined intangible cultural heritage as all those "practices, representations, expressions, knowledge and skills […] and cultural spaces associated therewith" recognised by communities or individuals as an invaluable part of their cultural heritage (UNESCO, 2003b, p.2). Conversations, visual content, and interactions generated online by individuals and minority communities around certain critical events or aspects of their own culture have therefore found on social media platforms a virtual environment in which that immateriality can actually be captured in some sort of record. Pietrobruno (2013) observed how YouTube has been widely used as a sort of participatory and interactive repository, allowing users to generate and document cultural memories of practices (e.g., oral languages, ceremonies, dances and embodied knowledge), sometimes adding such evidences to pre-existing collections of intangible heritage, as in the case of the collections of videos available on UNESCOTV, the Organisation's official YouTube channel, which has become a means to preserve and disseminate officially recognised immaterial heritage (Pietrobruno, 2013).

As born-digital content enclosing a specific personal and cultural value, social media is, in effect, part of our digital heritage. This has been defined by UNESCO as all those "computer-based materials of enduring value that should be kept for future generations" (p.28), naming among the great range of resources that might fall under this definition all those "dynamic, informal interactions enabled by digital technology" (UNESCO, 2003c, p.30). Moreover, in the same document introducing "Guidelines for the Preservation of Our Digital Heritage: prepared in collaboration with the National Library of Australia", UNESCO (2003c) specified that the Internet as a whole – hence, including social media sites – deserves to be preserved as it can be considered a "priceless mirror of society" (p.5), an important piece of our (digital) legacy that needs to be passed on to future generations because of its value.

Still, the ephemeral component of social media continues to pose a risk to these digital items, owing to platforms not including clear plans for long-term preservation of content in their policies and the ease with which dialogues or digital objects on such

platforms can quickly be edited or deleted, often leaving no trace behind (Pietrobruno, 2013). Moreover, although social media platforms have been developed for a completely different purpose, through time, users have allowed these sites to function as a sort of repository for the memories and content individuals generate daily. They have become, using Brügger (2017b)'s definition, a spontaneous "living web archive", over which social media companies maintain, however, a strict control, determining who has access and how, to the billions of data and information stored in their databases. Nevertheless, the part that social media has come to play in facilitating the recording of the intangible and in the stratification of instances of shared collective memory, further underlines the value of these platforms. Its ephemeral trait poses, however, a grave danger that can result in future generations lacking essential tesserae to understand the complex mosaic that the 21st century happens to be.

2.2.2 Ephemerality and Fragility of Social Platforms

Along with its cultural and historical value, in the last couple of decades, social media and the web in general have also demonstrated on many occasions how fragile these digital ecosystems truly can be. Significant in this sense is a study conducted by SalahEldeen and Nelson (2012) to investigate the persistence of resources shared on social media. By analysing data collected between 2009 and 2012 regarding six different events that had a large response and engagement on social media (such as the H1N1 virus outbreak, Michael Jackson's death, Iranian elections and subsequent protests, Barack Obama's Nobel Peace Prize, the Egyptian revolution and the Syrian Uprising), they observed that about 11% of these resources had been lost after the first year of publishing, a number that rapidly increased to 27% after only two and a half years (SalahEldeen & Nelson, 2012). The study detected the existence of a linear connection between the time in which content had been shared on social media and the amount lost, also uncovering a similar pattern in relation to the time of sharing and the percentage of content collected by archives. With regards to the latter, SalahEldeen and Nelson (2012) calculated that only 20% of the content shared on SNSs is archived after the first year, and less than 47% after two years from its publishing.

The concerns raised by the limited persistence of shared resources on social platforms and websites seem to materialise in the difficulties encountered by researchers in gathering information and material on social media about past events. This phenomenon appears to be closely connected to the ephemerality of the communications

happening on the web, and in particular on social media where users can easily delete, and sometimes edit, resources they shared at any point in time, posing a threat to present and future generations of researchers. While in some cases content ephemerality on social sites may be linked to specific platforms' features, like the "Stories" on Instagram that are only available for 24 hours, there are instead instances where this is associated with precise and calculated practices such as the "proactive ephemerality" described by Ringel and Davidson (2020,p.3). In their article, the authors examined the ways and motives behind journalists' decision to periodically mass delete their tweets. What stood out from the interviews was that journalists perceived their tweets as belonging to a certain moment in time, as a sort of "live performance" occurring in an ephemeral environment over which they often felt the need to exert their control (Ringel & Davidson, 2020). Both manual and automated periodic deletion of content like the one described further stress how what is available on social media at a certain moment in time might be eventually deleted later —although some of this content can always be captured with a screenshot by other individuals before it disappears and thus continue to persist somewhere else on the web. If, on the one hand, deletion practices are legitimate and abiding by the "right to be forgotten" ruling (Bright, 2012; see also Section 2.4.2), on the other hand, these practices also represent a significant loss from an archival perspective, as they permanently exclude this material from collective memory and the historical discourse (Ringel & Davidson, 2020).

Moreover, social media does not seem to be immune to the fragility that characterises the digital world and the web. Platforms can be subject to decline and disappear, often overwhelmed by an ever-growing range of competitors or failing to adapt to new trends. A well-known example in this sense is the case of Friends Reunited, one of the oldest social network websites founded in the UK in the early 2000s. As per its name, Friends Reunited's original intent was to facilitate reunion and connection with old friends based on certain common information such as school, university, or workplace. After a dramatic growth, the website started to lose users, probably due to the rise of new services like Facebook. In 2016, the founder of Friends Reunited announced that the website would close down after 16 years of activity (M. Burgess, 2016; Pankhurst, 2016). Although users were given the opportunity to download personal photos stored on their servers, most of the data populating the platform disappeared with it. Like any other digital services, social media can also be vulnerable to more or less accidental glitches that can cause partial or complete permanent loss of users' data. This is, for example, what

happened to MySpace in 2019 when, after a faulty server migration, more than 50 million items of content including songs, photos, and videos uploaded before 2016 were lost, leaving no trace behind except for a banner at the top of the website's homepage inviting users to retain backup copies of the materials they intend to upload on the platform (Binder, 2019; Chokshi, 2019; Hern, 2019).

These are just a few examples of the huge amount of data silently disappearing every day for different reasons. In this context, Richardson's (2020) quote "we can only study what we save" resonates even more, calling for specific plans and policies to ensure the preservation of this important resource. Many scholars have mentioned in their studies the risks surrounding the fragile reality of born-digital resources – with a particular focus on the web and social media – painting critical scenarios that see future generations having to face several issues in reconstructing the present, very much similar to those encountered by historians approaching the Middle Ages; so much so that it earned the fairly controversial appellation of "Dark Age". Borrowing the term used by Terry Kuny (1998) to indicate the problems that might be encountered in ensuring the long-term preservation of digital artefacts at the beginning of the 21st century, Jeffrey (2012) envisaged a potential second "Digital Dark Age" that would, this time around, primarily affect the preservation of social media and collaborative websites (Jeffrey, 2012). As he points out, the first "Digital Dark Age" was only temporary thanks to good data management and new digital archival practices, thus the second one should also be temporary (Jeffrey, 2012). Certainly, national and local archives and libraries play a central and fundamental role in preventing this scenario, and some of them are working on developing effective strategies, policies, and tools to ensure long term preservation of these important, unprecedented resources.

2.3 Web and Social Media Archiving: A History

In the same way as monks in monasteries alongside the increasingly centralised royal administrations in the Middle Ages ensured the preservation of most of our knowledge by creating copies of numerous volumes and keeping them safe, in the last two decades following the rise and diffusion of social media platforms, some cultural heritage institutions committed themselves to guaranteeing the adequate collection and preservation of this important historical resource so that it can be passed on to future generations—despite the many challenges that such an endeavour entails.

Memory institutions, as per their definition (see Section 1.2), have been invested through time with the important task of keeping our collective memory and knowledge safe from any form of damage and loss, making sure it remains accessible in the long-term. The "Universal Declaration on Archives" endorsed by UNESCO in 2011 stated that these institutions have an active role to play in "the development of societies by safeguarding and contributing to individual and community memory" (ICA, 2011, p.2), enabling a better understanding of the past while documenting the present so that it can constitute a guide to future actions. The document also recognised the multiplicity of formats in which the unique and irreplaceable heritage that archives manage, and preserve are created, including electronic formats. Within this framework, cultural heritage organisations like the National Archives of the United Kingdom (TNA) have made it their mission to preserve "collections for posterity, to retain authenticity and value, to facilitate access, and to protect our current and future collections from risks such as deterioration, damage, loss, corruption or obsolescence" (The National Archives, 2018), recently specifying how this endeavour now concerns both physical and digital documents.

As an ephemeral and fragile born-digital material, social media content has seen a slow but steady increase over the last decade in the number of heritage institutions involved in its collection (see section 3.3), following the growing interest of researchers coming from different disciplines, and the value that these platforms may represent for history and the public. The interest that heritage institutions have demonstrated so far towards this particular type of media appears to gain a further and significant meaning when considering the essential quality attributed to archiving institutions, again, by the Universal Declaration on Archives: they provide authentic evidence of cultural and intellectual activities, mirroring at the same time the evolution of societies (ICA, 2011). Since the appearance of the firsts archons – Roman magistrates who represented the law and for this reason were entrusted with the safe keeping of documents that would be considered authentic and maintain their legal validity as long as they were kept in the archons' homes – archives, and by extension other memory institutions such as libraries and museums, have inherited this specific authority (Derrida, 2002). Thus, it is noteworthy that repositories and other cultural heritage organisations have been seeking more and more to capture social media. The archiving practices and the institution's authority help to award to the collected material an explicit validity while also ensuring its authenticity to a certain degree. Specifically, the act of preserving material from the live web and social media, according to Koerbin (2017), means creating a static "preservation artefact of what

was once (and may continue to be) the dynamic and live entity" (p.193). Therefore, web archiving practices ensure that from the moment social media content is collected, this becomes crystallised in time, as an instance of a flux that outside that archival context would otherwise potentially continue to change or disappear.

When examining the development and the, albeit brief, history of social media archiving, it is necessary to consider the fact that the practices that this activity entails are strictly intertwined with and a result of those adopted for the collection and preservation of the web. The term web archiving indicates all those activities and techniques required to make copies and preserve the online web and make it available to the public (Brügger, 2018b). Although the term has been the object of various discussions regarding its accuracy (Brügger, 2017a; see Section 1.2 for further details on the terminology), it has been widely accepted for decades now, supplanting any other proposed definitions, and it can be found today incorporated into the name of many national and transnational web collections like the UK Web Archive or the Internet Archive (Brügger, 2018b). The latter is the first web archiving initiative of its kind and one of the most ambitious ones in the field. The Internet Archive (IA) is a US-based non-profit organisation considered one of the largest digital libraries in the world (Ogden et al., 2017) with over 863 billion[11] web pages captured both within and outside the national domain. The IA began archiving the Internet in 1996, understanding the need to preserve a resource that at that time no one was saving, and making it accessible through the Wayback Machine[12], a service allowing users to visit archived versions of a website at an exact moment in time. Similarly, The National Library of Australia and The National Library of Sweden launched their own web archiving projects in the late 1990s: respectively, PANDORA[13] and Kulturarw3[14] (Arvidson et al., 2000; W. Smith, 1997). Following these first initiatives, many other memory institutions around the world began archiving the Internet, soon recognising common issues and the need for the development of standards.

Capturing single, static web pages eventually became an easier and more reliable process over the years. However, the ever-increasing complexity and dynamicity of the

[11] These figures relate to the amount of web pages collected by the IA up to December 2023.

[12] https://web.archive.org/web/20231218204358/https://archive.org/about/

[13] https://web.archive.org/web/20230306232533/http://pandora.nla.gov.au/
In March 2019 PANDORA content became part of the Australian Web Archive searchable in Trove.

[14] https://web.archive.org/web/20231123175308/https://www.kb.se/hitta-och-bestall/hitta-i-samlingarna/kulturarw3.html

early 2000s websites made capturing the web at scale, while also trying to preserve the "look and feel" and the interconnections established across sites, a challenging task which called for a joint effort of the growing web archiving community. In 2003, twelve institutions came together to form the International Internet Preservation Coalition[15] (IIPC) with the aim to "acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations" (IIPC, 2017). Thanks to the cooperation of international member organisations whose number increased to 35 in 2010, the Consortium has been fulfilling since then a central role in the development of new web archiving tools, specific workflows, and archiving file formats—such as the Web ARChive (WARC) file format (Costa et al., 2017; Mohr et al., 2008; Pennock and Beagrie, 2013; see also Section 2.6).

A survey circulated by Gomes et al. (2011) identified 42 web archiving initiatives created worldwide between 1996 and 2011, with a total amount of archived data reaching 6.6PB. This study highlighted how organisations involved with the preservation of the Internet often opted for a careful selection of the content archived, mostly due to the authorisation issues and thematic constraints imposed by institutions' collection development policies (Gomes et al., 2011). Certainly, the many challenges encountered by memory organisations, especially from a legal and technical perspective, have shaped most projects' structure and the type of data they were able to collect (Costa et al., 2017; Gomes et al., 2011; Pennock & Beagrie, 2013).

The last decade has seen important improvements in terms of technology and tools adopted, as well as a better understanding of the issues that had emerged up to that point (Pennock & Beagrie, 2013). Among the many successful web archiving initiatives driven by national archives and libraries worldwide, it is worth mentioning, for instance, The UK Government Web Archive[16] (UKGWA), established in 2003 as part of The National Archives of the United Kingdom, which is entrusted with the collection of all Public Records under the British Public Records Act including Government departments and bodies' online presence; or the Netarkivet,[17] that is the Danish web archive operated by the Danish Royal Library, created in 2005 – which covers all websites under the top level domain .DK and other publicly available online content published in Denmark,

[15] https://web.archive.org/web/20231216010323/https://netpreserve.org/

[16] https://web.archive.org/web/20231203204903/http://www.nationalarchives.gov.uk/webarchive/

[17] https://web.archive.org/web/20240213180301/https://www.kb.dk/find-materiale/samlinger/netarkivet

written by, addressed to and about Danish people. Despite the important achievements accomplished in the field, web archiving organisations like those mentioned above are still facing significant challenges, especially concerning access, searchability and long-term preservation, to which have been recently added a whole new set of challenges related to the collection of social and collaborative platforms (Pennock & Beagrie, 2013; Thomson, 2016).

As one entity is part of the other and both share similar characteristics, it appeared rather logical that the first projects meaning to systematically collect social media were often launched by national libraries and archiving institutions with consolidated experience in capturing the presence of government departments, special events and individuals of public interest on the web (Littman et al., 2018). As Littman et al. (2018) observed, social media content has been captured since platforms such as GeoCities or Blogger first appeared in the late 1990s. For instance, the Internet Archive has been capturing sites like Friendster or MySpace starting from the year in which they were founded, respectively 2002 and 2003 (Littman et al., 2018). However, social media material rapidly evolved into becoming something increasingly complex, so much so that collecting, preserving, and making it available to the public proved to be an even more intricate and opaque process than that required for other digital resources and online content. In fact, archiving social media inherited all those technical, curatorial and legal challenges already experienced by memory institutions when collecting the web, plus new issues that emerged because of the fleeting, highly interactive nature of this medium, the legal constraints imposed by both international and national legislation as well as platforms' terms and conditions (Hockx-Yu, 2014; Thomson, 2016).

In order to archive these particular types of data in a way that it is meaningful and useful to diverse research fields, memory institutions and researchers have been working to develop specific tools and ad hoc strategies that would ensure adequate capture, access and long-term preservation of this rich resource. Systematic endeavours to archive social media at scale, creating collections that would be as comprehensive and representative as possible, only emerged towards the end of the first decade of the 21st century. In 2008 the National Archives of the United Kingdom (TNA) in collaboration with the Internet Memory Foundation (IMF) devised a pilot project to develop a more effective method to capture complex social media content, in an attempt to align their current approach to web collection with the way in which the government was using the web (Espley et al., 2014; Newing, 2020). The project was limited to the collection of Twitter and YouTube

content published by UK central government departments. The decision to focus on these two platforms was mostly due to their public nature and for archivability reasons, despite the numerous technical challenges that collecting them still presented (Hockx-Yu, 2014; Storrar, 2014). TNA and the IMF together developed a solution for archiving social media which made use of Application Programming Interfaces (APIs). This method ensured the capture and replay of a higher volume of social media material than previous techniques, mitigating the scalability issues as well as ensuring better quality and allowing for the possibility to tailor the criteria of capture (Storrar, 2014). The social media archive was then officially released in 2014.

In 2010, the State Archives of North Carolina and the State Library of North Carolina expanded the scope of the North Carolina State Government Web Site Archives (WSA)[18] to include social sites' content (Littman et al., 2018). The WSA started to actively capture in 2012 platforms such as Facebook, Twitter, and Instagram, limiting the scope of collection to content produced and maintained by North Carolina state agencies (WSA, n.d.). Unlike TNA, the WSA chose to archive social media material using the Internet Archive's "Archive-it"[19], a subscription-based web archiving service that provides tools (e.g., Heritrix) and technical support for building and preserving web collections.

One of the early initiatives that has marked a milestone in the history of social media archiving for its uniqueness, is the Twitter Archive at the Library of Congress. In 2010, the US Library of Congress (LoC) in conjunction with the micro-blogging site Twitter announced that the memory institution would archive and manage every public tweet posted on the platform from March 2006 to April 2010 (Raymond, 2010). Based on the agreement, Twitter provided its historical archive that consisted of approximately 170 billions tweets, and also committed to delivering to the Library all future public tweets as they became available (Fondren & Menard McCune, 2018; Raymond, 2010; Zimmer, 2015). If, on the one hand, the project was praised for the promise that it represented for future research, on the other hand, it stirred many discussions on ethical issues and privacy, as well as questions about the value that this material really enclosed (Weller et al., 2014; Zimmer, 2015). In one of their very rare updates, in 2013 the LoC addressed the significant technical challenges they were facing in order to make this enormous collection searchable and accessible in a comprehensive way (Library of Congress, 2013). After a

[18] https://web.archive.org/web/20100527090351/http://webarchives.ncdcr.gov/aboutWSA.html

[19] More information about "Archive-it" can be found here: https://web.archive.org/web/20240323115608/https://archive-it.org/learn-more/

few years of silence where the archive continued to remain unavailable, in 2017 the LoC announced a drastic change in the collection practices for Twitter: starting from 2018 it would capture and preserve tweets on a selective basis (Library of Congress, 2017). In a white paper made available in concomitance with the update, the Library of Congress enumerated some of the reasons behind this sudden change of direction, including the dramatic increase in the volume of tweets since the agreement was signed. Moreover, the LoC underlined how the agreement covered the preservation of only-text tweets. Because Twitter content had become over the years more and more visual, excluding this essential linked material would limit the value of the collection itself (Library of Congress, 2017). Currently, the Twitter Archive remains embargoed as the LoC continues to work on solving accessibility and searchability issues.

Since these first few pioneering attempts, a growing number of memory institutions have started to archive or are at least considering the possibility of capturing social media content, especially those that are already actively archiving the web. Experiences such as the ones described, – especially the issues encountered for instance by the Library of Congress – prove how complex and challenging the task of planning, implementing, and making available to the public a social media archive is. Social media is difficult to archive on many levels: as mentioned, it inherits existing challenges concerning web archiving practices, while also adding to them new and complex ones. In the remainder of this chapter, I will review the main issues and concerns that have emerged from the initial efforts at archiving social media at an institutional level, highlighting the need for further research that explores advancements and new challenges faced by web archivists and curators after almost a couple of decades of testing and trials.

2.4 Legal and Ethical Concerns

Many are the legal concerns that information professionals must consider when archiving social media content, including restrictions set by digital legal deposit legislation, privacy and data protection laws, and copyright. Moreover, platforms' terms and conditions add a further layer of limitations influencing the scope and frequency with which institutions can access and harvest data on social sites, with consequences for selection practices and granularity of collections, as well as long-term preservation and access to archived social media material.

2.4.1 Ownership of Data, Intellectual Property Rights and Copyright.

The acquisition of (printed) archival material that possesses an enduring cultural value usually involves, at some point in its lifecycle, a transfer of ownership from its legal proprietor to the cultural heritage institution entrusted with its long-term preservation. Before it is ingested into a collection, the institution must seek an agreement - in the form of purchase, deposit or deed of gift - with the body, family, person or heirs that legally own the material (Arp, 2019). In a paper world, determining the legal ownership of a physical or digital object is generally a rather straightforward process regulated by Intellectual Property Rights (IPR) and private and inheritance law. Ascertaining ownership of content published on a website on the public web might prove to be more complicated, but, despite some exceptions, it does not usually represent a major challenge, especially if archiving institutions can access lists of websites registered within a domain name associated with a specific country (Stirling et al., 2012). Different is the case of social media where determining even with relative certainty to whom the data belongs can prove to be extremely challenging due to the involvement and stratification of a multitude of stakeholders (e.g., individuals, organisations, public or private institutions) who generate and re-share content on social platforms (Koščík & Myška, 2019; Marshall & Shipman, 2015). Most platforms' terms and conditions state that the content individuals generate and share on these sites – including any attached media such as videos, photos, or audio recordings – belongs to single users. However, by posting or submitting any type of content, users grant to social media companies a "non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods now known or later developed" (Twitter, 2021b) necessary for platforms to make the content available worldwide. While the influence that social media policies have on archiving activities will be discussed further in Sections 2.4.4 and 2.6, here it suffices to point out how such a statement establishes a multi-layered system where users retain their rights over content they create but at the same time allow social media companies to exert close control over the access to data generated on their platform. This also create a series of important ethical implications concerning users' unawareness of the way in which their information and content is used, potentially exposing the expression of their ideas to copyright and IPR infringement.

From an archival perspective, concerns about copyright and IPR may arise from collecting social media content at scale as this may likely include protected materials.

However, in countries like the United Kingdom and in some member states of the European Union, legislative frameworks tend to envisage special exceptions to copyright allowing national cultural heritage institutions to reproduce material for the "public good" or for academic, non-commercial purposes (see Section 6.1.3). Thus, memory institutions are allowed to collect copyright-protected content for the purpose of preserving national heritage or public records (Hockx-Yu, 2014). To mitigate copyright infringement concerns, some web archiving institutions request permission or notify the website owner upon collection. For example, when the UK Web Archive first started archiving the national web domain between 2008 and 2013 it adopted a permission-based approach, requiring archivists to contact rights-holders to formally request authorisation to make archival copies of their websites. Extremely impractical and time-consuming, this practice was mostly abandoned when the (Non-Print) Legal Deposit legislation was introduced in 2013, exonerating UK deposit libraries from the requirement to obtain permission to preserve (Schafer & Winters, 2021). The difficulties encountered with permission-based approaches were largely incompatible with the aim of scaling-up the number of websites archived and, in hindsight, it would have been even more unfeasible in the context of social media preservation. Given the multiple levels of ownership that characterise social media platforms and the sheer volume of ephemeral material generated daily on these sites, seeking out authorisation to archive for every single item of born-digital material shared on social sites would have proved to be practically impossible for many institutions.

In addition, it is worth noting that intellectual property rights may apply to social media on a two-pronged level, that is to the creator who generates content and the company that owns, for example, the design of the interface. Without specific legislative exemptions, institutions would have to request permission to make copies of content on social media to both parties, making the archival endeavour considerably more complicated and opaque than for websites (Hockx-Yu, 2014; see Section 6.1.3). The UKWA still opts for requesting permission in some rare instances as having the rights-holder's authorisation allows archivists to provide public off-site access[20] to some of the

[20] The UKWA offers two levels of access: public off-site access provides access to a limited selection of content for which the UKWA has received explicit permission from the rights-holder to make the archived content available online through the portal https://www.webarchive.org.uk/; on-site access to the full UKWA allows users to browse the complete collection of web archived materials only within the

archived material through the UK Web Archive portal (see Section 4.4.4). Nevertheless, capturing data on a large scale for long-term preservation purposes does raise a whole new set of ethical concerns. In particular, studies such as the one proposed by Marshall and Shipman (2017) about content ownership and media users' practices, shed light on concerns that surveyed individuals had about practices such as archiving social media for the "public good". Similar studies highlighted a strong need among participants to affirm their right to own and exert a certain control over their digital footprints for the long term (Marshall & Shipman, 2017; Shiozaki, 2021).

2.4.2 Privacy, Data Protection and the "Right to Be Forgotten"

Amongst the legal and ethical concerns that cultural heritage organisations must consider when archiving social media, there is the fine line that separates the right of individuals' privacy from memory institutions' mission to preserve information that is essential evidence about the recent past. Many studies have discussed the role that social media has in people's lives, investigating users' perception of these platforms as virtual spaces for the performance of the Self and the sharing of personal memories (Allegrezza, 2019; Seyfi, 2017; Zhao et al., 2013). However, because social media content is largely generated by private individuals, it can pose significant legal and ethical challenges concerning long-term preservation and particularly access to archived material as it may include and (accidentally) disclose sensitive or personal information.

Scholars across various disciplines have been using content shared on social media for research purposes for a long time now (e.g., Brügger & Schroeder, 2017; Lomborg & Bechmann, n.d.; Shane-Simpson et al., 2018; Steel et al., 2023). Yet, individuals may perceive inclusion in a collection within a cultural heritage institution as a more permanent act that could even lead to reshaping the way individuals interact online. In an article illustrating the challenges encountered while building the Twitter Archive at The Library of Congress (LoC), Fondren and Menard McCune (2018) duly observed how knowing that individuals' tweets might become part of history and consequently preserved forever could alter the way people use and communicate on social media. They also reported that some Twitter users received with concern the announcement of the agreement signed in 2010 between the Library of Congress and Twitter, raising questions about potential threats to their privacy (Fondren & Menard McCune, 2018). Additional examples of

premises of the six legal deposit libraries and at specific terminals available in the reading rooms (see Section 4.1).

similar concerns can also be found in some comments posted in reply to the Library of Congress's blog post celebrating the acquisition of Twitter's complete historical archive (Raymond, 2010). Moreover, in a study surveying users' conflicting perspectives on institutional social media archiving, Marshall and Shipman (2012) further emphasised how privacy appears indeed to be one of the main concerns for users. Specifically, the authors pointed out how personal physical archives have always been donated to institutions by explicit desire of their owners or heirs, whereas Twitter's gift to the LoC was made on the basis of the legal ownership granted to service providers by users when signing the platform's terms and conditions (Marshall & Shipman, 2012). In this sense, Marshall and Shipman talk about an "erosion of privacy" caused by the assumption that everything published publicly on the web is truly public and intended for posterity, complicating social media archiving practices even further (Marshall & Shipman, 2012, p.1).

The distinction between private and public on social media is indeed a rather opaque one. Burkell et al. (2014) observed that social sites are digital environments that are neither "public" nor "private", "open" nor "closed" per se, inhabiting a liminal space where these concepts blur. If, on the one hand, social media platforms are often considered as public spaces where individuals can express their ideas and represent their Self, on the other hand it is important to not ignore the fact that social sites are private companies to which users grant permission to dispose of the information generated on their platforms (see 2.3.1). The blurring of boundaries between what is truly private and the controversial meaning of "public" in a context such as that of social media sites owned by different types of legal entities has ethical implications that should not be ignored in the context of social media archiving. Nonetheless, most platforms allow users to manage their digital footprint, thus having a certain degree of control over who sees the information shared. This, of course, varies across platforms and is based on the main intent with which these were created. For example, on Twitter all newly created accounts and their posts are public by default. Users can modify accounts' privacy settings by choosing between the option to leave their tweets public or protect them. While the first-mentioned setting allows any user signed up to the platform to view and interact with the content shared, protected tweets are only visible to followers, unless blocked.[21] But, as "The Twitter Rules" explicitly remind users, Twitter's "purpose is to serve the public

[21] Twitter however warns users of the risks represented by capture and resharing of protected tweets. https://web.archive.org/web/20220128081538/https://help.twitter.com/en/safety-and-security/public-and-protected-tweets

conversation" (Twitter, n.d.) and therefore users appear to join the microblogging site with the main intention of interacting and participating in an ongoing public conversation, and perhaps for this reason only a handful of accounts are actually set to private (Fondren & Menard McCune, 2018). On other platforms designed with the aim of connecting friends, family and other acquaintances, such as Facebook, users, on the contrary, seem to favour semi-private account settings which enable only a community of "friends" – or even a selected few – to interact with their content (Burkell et al., 2014).

In connection with privacy concerns, another principle that archiving institutions in the European area must consider in relation to social media data is the so-called "right to be forgotten" (Rosen, 2012). Following the European Court of Justice ruling against Google Spain in 2014 and the debate it originated – mainly about, but not limited to, privacy, data protection, freedom of expression and the individual's right to be remembered on the web as they prefer or not at all – the European Union introduced in 2018 the General Data Protection regulation[22] (GDPR) (Bright, 2012; Judgment in Case C-131/12 Google Spain SL, Google Inc. V Agencia Española de Protección de Datos, Mario Costeja González, 2014; Peterson, 2015). The EU GDPR established new rules for how public organisations and businesses must handle clients' data, granting individuals more control over their personal information shared with third parties (MirrorWeb, 2018). Among other rules, the 2018 regulation allowed EU citizens to request the erasure of their personal data from search engines in the event that the information is or has become "inaccurate, inadequate, irrelevant, or excessive" (Bright, 2012; L. Cook, 2015).

In general, the implementation of GDPR in Europe – and its application in the United Kingdom as Data Protection Act 2018[23] – has not signified drastic changes for the archiving practices at cultural memory institutions, as it contains specific provisions for the act of "archiving in the public interest"[24] which have effects on the application of individuals' rights. In fact, Article 89 of the regulation makes specific provisions for exemptions from the rights expressed in Articles 15-21 (e.g. access, rectification, notification and data portability) as well as adaptations relating to processing personal data for archiving purposes in the public interest. Nonetheless, appropriate safeguards need to

[22] The complete text of the General Data Protection regulation can be found here: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[23] The Data Protection Act 2018 is the UK's implementation of the General Data Protection regulation (GDPR). See: https://web.archive.org/web/20231203012133/https://www.gov.uk/data-protection

[24] Data Protection Act 2018, Sch.2, part 6, para.28

be put in place to protect data subjects' rights. These safeguards include but are not limited to data minimisation and anonymisation.[25] Organisations archiving in the public interest are therefore required to have specific processes in place that ensure the adherence to GDPR's safeguards. Within this setting, an additional legal challenge is represented by the variety of national legislative frameworks, whose differences make archiving and provision of access to social media content for the long-term an arduous task, especially when attempting to develop common archival standards and practices.

2.4.3 Digital Legal Deposit Legislation

The adoption of the UNESCO "Charter on the Preservation of Digital Heritage" (UNESCO, 2003a) led a growing number of national libraries and archives to strongly advocate for the extension of existing national legal deposit laws – which require publishers to deposit a copy of any printed work published on a national territory at designated institutions – to content published online, including social media. Various studies have explored the promises and pitfalls of archiving born-digital materials from the web and collaborative platforms under national electronic legal deposit legislation (Arnold-Stratford & Ovenden, 2020; Dupont, 1999; Hockx-Yu, 2014; Larsen, 2005; Milligan, 2015; Stirling et al., 2012; Winters, 2020). As will be discussed later in this thesis (see Section 3.5 and 6.1), being appointed with a clear mandate to collect and preserve born-digital material that has an enduring cultural value allows national memory institutions to obtain vital resources and dedicated staff to efficiently archive, if not comprehensively, at least a selection of the ephemeral, historical traces publicly shared on web and social media platforms.

The gradual implementation in various countries across the world of electronic legal deposit legislation over the last decade certainly has represented an important step forward for the preservation of material published on the web. However, it has also brought to light several limitations that these regulations impose on web and social media preservation activities, concerning particularly restrictions in terms of selection criteria and access to the archived materials. The coming into force of the e-legal deposit legislation has sometimes revealed itself as a two-sided coin for many memory institutions and researchers seeking to engage with the archived web (Arnold-Stratford and Ovenden, 2020). As Milligan (2015) illustrated in his account following a visit at the British Library, the benefits associated with archiving under the e-legal deposit come with specific

[25] See Article 89(1), (EU) GDPR

restrictions that often affect access, use and reuse of the material archived. For example at the legal deposit libraries in the UK, web pages are statically rendered to prevent copyright infringement, all dynamic content is disabled and it is not possible to take photos of the computer screen as one would do for traditional archival material, nor copy and paste images or text; also, two users cannot simultaneously view the same archived web page (Milligan, 2015). Some of the identified barriers appear to stem from the fact that these regulations were often developed against the backdrop of a print-centric world, consequently struggling to adapt and keep up with the constant evolution required by a digital world that is rapidly changing (Arnold-Stratford & Ovenden, 2020; Gooding et al., 2019).

Moreover, at many legal deposit institutions, the source code of a web page is not available to researchers, although it would provide critical insights into the history of social platforms, technical infrastructures and business organisations (Helmond & Vlist, 2019; Winters, 2020). Paradoxically, as Winters (2020) noted, these restrictions have in some cases made it more difficult to use and significantly slowed down access to web archives, compared to collections of medieval manuscripts. Exceptional events such as the COVID-19 pandemic and the consequent closure of all memory institutions to the public for a long period of time have further highlighted the necessity of finding new solutions to extend the access to such rich collections outside libraries' reading rooms (see Section 6.4). For example, the Danish web archive allows users to remotely access the web archive, cut and paste content from an archived page, and provides dynamic rendering of web pages with accessible source code. However, these benefits are accompanied by stricter rules for obtaining access to the web archive, including restrictions to researchers associated with a Danish institution, and the requirement to submit a well-formed research proposal which, upon acceptance, grants access only to the portions of the web archive collections that are related to the accepted study (Schostag & Fønss-Jørgensen, 2012; see Section 6.1.3).

E-legal deposit legislation provisions significantly shape web and social media archives, determining the parameters within which institutions can select born-digital material. Often the scope of collections includes content made publicly available on websites belonging to national Top-Level Domains, or pertinent to the country where the archive is located. Hockx-Yu (2014) pointed out the many difficulties that emerged from trying to apply territoriality criteria to social media platforms and its content, making archiving social media at scale a particularly onerous task. Yet, little is known about the

impact that the differences in various national legal frameworks, coupled with the aforementioned limitations, have on the extent and type of social media platforms preserved at an institutional level and how this shapes the broader social media archiving landscape.

2.4.4 Social Media Policies

Social media platforms' terms and conditions impose an additional layer of obstacles to archiving institutions, further influencing the scope and granularity of the content archived. These obstacles are often linked to a set of limitations imposed by social media companies to protect their commercial interests (Thomson, 2016). Social sites such as Twitter or Facebook make a large portion of their profits by selling data to third parties interested in gathering information, for example, about consumer and voter behaviours (Lomborg & Bechmann, n.d.). For this reason, social media companies tend to limit the amount of information that can be accessed in a certain timeframe. For instance, Twitter's Developer Agreement and Policy (2020) set the conditions under which licensed material can be collected from the platform via APIs, including request frequency and the way data may be (re)used and accessed once acquired by a third party. Moreover, as has been pointed out in the "Preserving Social Media" report compiled by Sarah Day Thompson (2016) for the Digital Preservation Coalition (DPC), the Developer Agreement restricts the ability of researchers to share and transfer the data harvested through APIs to third parties, including any digital repositories and cloud storage.

The fact that memory institutions and other non-profit organisations are often subject to the same limitations imposed on commercial companies makes the social media preservation process even harder (Acker & Kreisberg, 2020; Thomson, 2016), although in 2017 Ahmed et al. (2017) noted that Twitter appeared to be more favourable to providing access to data compared to other platforms such as Facebook. In fact, following the Cambridge Analytica scandal (Bruns, 2019), where the personal data of Facebook users was collected and used for political advertising without their consent, the platform implemented several access restriction procedures. This event particularly influenced the ability of memory institutions to collect material from Facebook, partially explaining why this platform appears to be less archived than other sites of the same genre (see Section 6.1.4).

The severe limitations placed by social media organisations on their data through platform and developer policies are particularly challenging for collecting institutions that

are required by law to archive social media at scale. In an article offering an interpretation of Twitter's Developer Policy and the barriers it created not only for researchers but especially for archivists and librarians, Littman (2019) maintained that one of the main issues with these documents is their ambiguity. Due to such ambiguity and the lack of awareness on the part of social media companies regarding the needs of researchers and archivists, cultural memory institutions risk approaching social media archiving with extreme caution. This caution may slow down the archiving process and potentially lead to gaps in the historical record (Littman, 2019).

Over the years Twitter demonstrated that it had taken into consideration some of the frequent concerns raised by researchers, especially regarding the redistribution of content. In the 2017 Twitter Developer Policy[26] redistribution exceptions for academic institutions were only loosely mentioned, concerning in particular sharing limits of Tweet IDs for the sole purpose of non-commercial research (Twitter, 2017). However, in a general update published in the spring of 2020,[27] Twitter dedicated a separate section to the redistribution of content by academic researchers, specifying that scholars were granted the ability to share an unlimited number of Twitter IDs and/or User IDs for the purpose of non-commercial research (Twitter, 2017, 2020). The said section also clearly stated that an unlimited number of Tweet IDs can be shared "for the purpose of enabling peer review or validating your research" (Twitter, 2020), further highlighting the slightly increasing attention to academic research needs. Yet, there was no mention in the 2020 Twitter Developer policy of any exceptions specifically concerning cultural heritage organisations archiving social media, often leaving such policies to the interpretation of the individual institutions (Littman, 2019).

In this context, an additional challenge that needs to be addressed is posed by the constant and unpredictable amendments of social media terms and conditions. Perhaps the most emblematic case in this sense is the changes implemented after the acquisition of Twitter by Tesla CEO, Elon Musk, in October 2022 that brought a complete upheaval to the company. The change of leadership marked the "end of an era" not only because

[26] The web archived (05/01/2020) page is available here: https://web.archive.org/web/20200105002041/https://developer.twitter.com/en/developer-terms/policy

[27] The web archived page (18/04/2020) is available here: https://web.archive.org/web/20200418215448/https://developer.twitter.com/en/developer-terms/agreement-and-policy

of the rebranding as "X" but also for placing data behind a tiered access paywall. This particularly affected access to data for academic researchers – and a few institutions that had managed to be granted access – who were using the Twitter Academic API which was only launched a couple of years earlier (see Sections 5.4.6 and 6.1.4). Similar changes significantly impact the development of robust and long-lasting social media collection policies and workflows, compelling memory institutions to slow down and undertake scaled-down social media archiving efforts.

The complexity of archiving social media, combined with legal restrictions imposed by both national legal frameworks and social media platforms, alongside the lack of accountability on the part of the commercial digital platforms and services, pave the way for the loss of valuable information generated by users' activities on the web (Schafer & Winters, 2021). In the absence of long-term preservation plans or deposit agreements like the one signed by Twitter in 2010, memory organisations are left with the only option to save as much data as possible, especially when platforms are about to shut down or drastic changes are implemented, provided a notice period is given.

2.4.5 Additional Ethical Concerns

Archiving social media poses several ethical concerns, many of which often overlap and are sometimes even in contrast with legal stipulations. In the previous section I discussed some of the ethical implications originating from legal matters, including individuals' right to privacy and the right to erasure, and the blurred boundaries between private and public spheres on social media (see Sections 2.4.2 and 2.3.5). In addition to these concerns, users may not be fully aware of or adequately informed about how their data is being used by third parties, often due to blindly agreeing to platforms' terms and conditions when signing up (Bergis et al., 2018). This raises several ethical questions for collecting institutions, particularly concerning the extent to which memory institutions can be held accountable for any risks that may accidentally arise from providing access to such data. Although researchers are likely to anonymise data harvested from social platforms,[28] and service providers such as Twitter restrict researchers to only sharing Tweet IDs in a particular dataset or for a limited set of reasons related to non-commercial purposes, as

[28] It should be noted that attempts at further anonymising tweets, for example, by slightly altering the content of the post to avoid the user's identity being disclosed by searching the exact text on Twitter, may go against the platform's terms and conditions. Specifically, Twitter's "Display Requirements" states under the "Don'ts" to not "Modify post text" (Twitter, 2023b).

mentioned in Section 2.4, the risk of inadvertently disclosing personal information about individuals remains considerably high (Thomson, 2016; Thomson & Kilbride, 2015). Moreover, social media's interconnectivity poses a problem when combined, for instance, with geo-spatial data or other administrative data sets (Thomson & Kilbride, 2015). Even when data is anonymised, the conjunct analysis of datasets could be mined and used to intentionally or unintentionally reveal sensitive or personal information about users (Raad et al., 2016). For this reason, Thomson and Kilbride (2015) advocated for clear plans and the proactive development of strategies that could help memory institutions mitigate such concerns in the future.

Since requesting permission to archive content from social media users is neither feasible, sustainable nor required by institutions operating under a specific national legal mandate, the dilemma about whether or not to archive, and how to do this ethically even when collection is limited to content publicly available on the web, persists. Bergis et al., (2018) observed that "what is legal is not always ethical, and what is ethical is not always legal" (p.11), while Milligan (2018) similarly pointed out the need to separate what is legal from what is ethical. In an ethnographic study about the Archive Team and their attempt to preserve the "Not Safe for Work" posts on Tumblr, Ogden (2021) described the Archive Team's tendency to "archive first, ask questions later" (p.7), emphasising that they rarely ask for permission whenever they archive what is perceived as endangered content on the web. In the name of saving content from oblivion, the "rogue archiving" tenets of practice of the Archive team have underlined the need for more flexible mandates for national web and social media archives, in order to enable them to preserve content for posterity (Ogden, 2021). While this approach has the benefit of challenging current policies and restrictions imposed by service providers and legislation on web and social media archives, it certainly raises further ethical concerns regarding what content users genuinely desire to see preserved in the long-term and how they wish to access it, especially when this concerns activists and minority groups. The latter prompts additional reservations about the potential misuse of material captured from the web and social platforms, including surveillance by law enforcement (Bergis et al., 2018). A number of studies have begun to explore the complex ethical challenges associated with collecting activists and marginalised communities on social media (Bergis et al., 2018; Velte, 2018). Nevertheless, further research is required to support memory institutions' professional understanding of how to perform preservation activities that could ethically ensure representativeness of social media collections while also seeking the involvement of

marginalised or displaced communities, providing them with the opportunity to choose how and what stories to tell.

2.5 Selection

In archival science, the term "selection" describes the decision-making processes to identify and appraise the cultural and historical value of records, in any format and media, that should be retained and added to a repository (Ham, 1993). Appraisal and selection are crucial steps in the development of archival collections. Research about appraisal has a long history, with many studies discussing the implications and the very meaning of a process that Cox and Samuels (1988) defined as the "archivist's first responsibility" (p.29). Cook (2011) offered a comprehensive overview of the evolution and diverse perspectives surrounding the practice of archival appraisal and selection, and how archivists have come to a general agreement – even in the digital age, where initially the physical space limitations no longer seemed to apply and keeping everything suddenly appeared possible (Werf & Werf, 2014, p.9) – that, apart from few exceptions, it is not possible nor sustainable to keep all records ever produced, making selection indispensable.

This is even more true in the context of social media archives, where selecting from and preserving the sheer amount of content created every day on these platforms has been recognised since the beginning as one of the greatest challenges (Milligan et al., 2016; Thomson, 2016). This is partly due to legal constrains, storage limits and the costs associated with maintaining it, as well as technological constraints in terms of indexing and providing access to the archived material (Fondren & Menard McCune, 2018; Hockx-Yu, 2014). Fondren and Menard McCune (2018) pointed out that among the setbacks encountered by the Library of Congress in the management of the Twitter Archive, there were the difficulties in processing and providing access to a dataset of unprecedented proportions. Building on this point, it is important to note that preserving every single born-digital object generated on social media could potentially risk diminishing its overall significance.

Moreover, in light of privacy and data protection laws discussed in section 2.4.2, selection becomes essential to sift through relevant content, ensuring that what is archived fulfils the purpose of "public interest" and supports the future recollection of contemporary events and societal trends. At an institutional level, the criteria for selection are shaped by the interaction of the institution's own collection policies, national legal and

technical frameworks, and resources available at individual organisations. This research intends to investigate the interaction of these factors and how this influences the extent of social media material that can be sustainably preserved in the long-term (see Chapter 6).

In a DPC report exploring the challenges of web archiving, Pennock and Beagrie (2013) described the many issues stemming from defining the scope of social media collections compared to websites. In particular, the authors noted that, for example, platforms like Twitter are not merely made up of single posts disconnected from one another, but rather elements within interlinked conversations: capturing a single tweet without its retweets, comments and quotes would mean preserving only partial conversations (Pennock & Beagrie, 2013). The conversational aspect that characterises social media platforms not only raises questions about the potential incompleteness of the archived material, but also generates concerns in terms of clearly defining the parameters of collection. Thomson (2016) mentioned the complicated process of identifying materials relevant, for example, to a specific country, as interactions and conversations can extend and overlap across different accounts and combine exchanges in different languages. Provenance is a significant concern for social media appraisal, especially for institutions that operate under legal deposit regulations, which establish strict national boundaries. Recent studies have underlined the challenges associated with ascertaining provenance at scale on social sites, especially when geographical data provided by platforms are not entirely reliable (Bingham & Byrne, 2021; Graham et al., 2014).

Although social media archiving inherits some of the curatorial challenges already identified for websites (Hockx-Yu, 2014), preserving these platforms comes with its own peculiar issues stemming from their dynamic nature, and the intricate myriad of interactions that carry specific meanings in the conversations unfolding in these digital environments (Brügger, 2017b; Bucher & Helmond, 2017). For this reason, workflows and archiving strategies developed for website collection, as Lieber et al. (2021) observed, "cannot be adjusted to sustainable social media harvesting workflows out of the box" (p.3). Only a small number of studies have explored sustainable strategies to archive social media (Vlassenroot et al., 2021), with some of them focussing on collection from specific platforms like Twitter (Pehlivan et al., 2021; Schafer et al., 2019), while others have investigated how socio-technical infrastructural aspects may shape web collections (Ben-David & Amram, 2018; Maemura, 2023; Maemura et al., 2018; Ogden, 2020).

Nevertheless, practices related to the institutional development of social media collection policies and the obstacles that may prevent the establishment of a social media archiving initiative are still largely understudied. Moreover, in the web archive community, there have been calls for greater transparency about selection parameters for web archives (Leetaru, 2015; Thomson & Kilbride, 2015) as understanding the inherent mechanisms and factors influencing curatorial choices is essential to critically engage with archived materials, enabling researchers to identify the reasons behind potential biases and gaps in the resulting collections.

Like any other archive, web and social media collections are not neutral and reflect structures of power and biases that are intrinsic to the archiving organisations and collection processes (Velte, 2018). Bonacchi and Krzyzanska (2019) pointed out the epistemological weight that archivists and curator have in the decisional processes that guide the selection of materials. The "archival turn" which, using Ketelaar (2017) words, "entailed a move from archives as sources to archives as epistemological sites and the outcome of cultural practices" (p.228), has encouraged memory institutions to reflect on how power dynamics and legal obligations shape collections, trying to mitigate biases and uncover stories that might have been concealed over time. As discussed above, social media provides an unprecedented platform for marginalised or displaced communities to represent themselves and voice their concerns (see Section 2.4.5). Yet, there is a need to examine how the many challenges that web archivists face in archiving social media at scale affect the overall representativeness of collections and whether adequate solutions have been adopted to provide a faithful representation of contemporary society and events on social platforms (see Section 6.2.1). In particular, participatory practices have been increasingly incorporated into web archiving workflows to complement seed lists compiled by web archivists for broad and selective crawls, although these practices may still be influenced by digital inequities (Schafer & Winters, 2021). In this regard, Cui et al. (2023) emphasised the importance of gaining a better understanding about the effectiveness of participatory approaches in reshaping and redistributing power dynamics in digital environments, including social media.

2.6 Technical Background

The ephemeral and dynamic nature of social media platforms poses several technical problems to archiving institutions (Thomson, 2016; Webster, 2015). Web and social

media archiving can take many forms. For example, Brügger (2018) identified seven methods of acquisition employed to capture the web.[29] Of these, web crawlers and Application Programming Interfaces (APIs) are the most common approaches used in the web archiving community to collect social media.

Maemura et al. (2018) described web crawling as the process that "captures the web without direct access to the files stored on a server" (p.1224). As it can be automated to a certain extent, this form of web archiving enables the preservation of the web in a more systematic and scalable way (Brügger, 2018, p.81). Essentially, web archivists compile a list of URLs (called a "seed list") corresponding to the addresses of websites that the institution wants to capture, which is then inserted into the software. Next, the software contacts the web server corresponding to each URL, requesting and downloading content and resources linked or embedded in that specific page (Brügger, 2018b; Maemura et al., 2018; Masanès, 2006). Based on the depth of crawl established at the beginning of the archiving process – that is the number of levels from the starting URL the software is allowed to go to follow the hyperlinks (Brügger, 2018b) – the crawler can identify any links on the page, and follow and capture them as well (Masanès, 2006). Once all the elements and content have been retrieved for a specific URL, the crawler moves to the next URL in the seed list.

The crawling software that has been widely used across the web archiving community is Heritrix, a web crawler created by the Internet Archive,[30] the non-profit US-based digital Library that started archiving content from the web in 1996 (see Section 2.3). Almost a decade ago, Hockx-Yu (2014) outlined the tests and trials that national institutions underwent during the early attempts to archive social media using Heritrix, identifying issues in capturing the numerous dynamic elements that social media pages are composed of, inevitably affecting the overall quality of the snapshots (Pennock & Beagrie, 2013). In order to mitigate some of the aforementioned limitations the Internet

[29] The seven forms of web archiving identified by Brügger (2018) include: "(1) making an image, (2) making a screen movie, (3) downloading individual files, (4) web crawling, (5) collecting web material […] through an application programming interface (API), (6) collecting the web that has been taken off-line and preserved unchanged, and (7) collecting the web as presented in other media types, such as books, film, and television" (p.80).

[30] Further information about Heritrix is available here: https://web.archive.org/web/20240303080916/https://github.com/internetarchive/heritrix3

Archive developed new tools like Brozzler,[31] which promised to "record interactions between servers and web browsers as they occur, more closely resembling how a human user would experience the web"(Archive-It, 2022). Another software employed in the capture of social sites is Rhizome's Webrecorder[32] (rebranded in 2020 as Conifer), a high-fidelity capture tool that combines aspects of web crawling with screen recording techniques. Webrecorder allows the creation of collections of recorded copies of websites by capturing all of the elements loaded on a web page upon the interaction of a user (Brügger, 2018b). Unlike Heritrix, Webrecorder can interact with the dynamic content on the page, including embedded media and buttons, providing information about context and the user experience of browsing content on these platforms (Thomson, 2016; Thomson & Kilbride, 2015). Nevertheless, archiving attempts such as the one described by Bingham et al. (2020) highlighted the extensive efforts, time and staff required to operate and ensure the smooth running of the Webrecorder application, making its use not scalable (see Section 4.4.3). In 2022, the International Internet Preservation Coalition in collaboration with Webrecorder announced the development of the "Browser-based Crawling For All" project (IIPC, n.d.). Building on the existing Browsertrix-cloud,[33] the project aims at contributing to the further development of the browser-based high fidelity crawling system by extending its ability to capture complex, dynamic websites and social media platforms which cannot be adequately captured with existing crawling tools like Heritrix or Webrecorder. However, to date this tool is still in the testing and development phase.

The other form of social media archiving involves the use of APIs for the collection of social media data. Many studies have investigated the technical challenges of using APIs for accessing information generated on social sites, but only a few have explored the implications surrounding API use for cultural heritage preservation at scale (Acker & Kreisberg, 2020; Littman et al., 2018; Pehlivan et al., 2021). Using APIs allows institutions to access and get raw data in formats (e.g. JSON and XML) that allow computational processing (Thomson & Kilbride, 2015), although, as Pehlivan et al. (2021)

[31] https://web.archive.org/web/20240301073100/https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler

[32] In this thesis I will use the name "Webrecorder" as is still widely the name with which it is still widely referred to. More information about Webrecorder is available here: https://web.archive.org/web/20231204201050/https://conifer.rhizome.org/

[33] https://web.archive.org/web/20230811173251/https://github.com/webrecorder/browsertrix-cloud/

noted, it may not be enough for archiving at an institutional level as tweets, for example, can contain other items such as images, videos or URLS that need to be preserved as well. Moreover, as has been highlighted in 2.4.4, social media companies impose restrictions on access to data through their official APIs which can affect completeness and raise questions about the representativeness of the information collected due to the opaqueness of the sampling parameters (J. Burgess & Bruns, 2012; Pehlivan et al., 2021). Moreover, APIs have to be considered as a "black box" due to the opaqueness of its data sampling (Driscoll & Walker, 2014; Pehlivan et al., 2021). Yet, only a few studies addressed how such constraints may influence the overall representativeness of and the formation of gaps in national social media collections (e.g., Bergis et al., 2018; Caswell et al., 2017; Chambers et al., 2021). Understanding and documenting potential biases in social media collections and how these are linked to archiving processes is crucial for the research community to critically approach the archived material.

Archiving methods and technical constraints also impact access to the archived material. For archiving institutions, this means finding solutions to replaying and making searchable the data collected either via APIs or through web crawlers. Moreover, it is worth underscoring that the final result when replaying webpages and social media does not always match the live web (Garg et al., 2023). As "reborn digital material" (Brügger, 2012), archived social media content may have disappeared or been edited since it was last collected from the live web, and thus may not exactly mirror its ephemeral original (Brügger, 2018b). In addition, memory institutions face several challenges when replaying social media, including missing features and information which may be relevant to researchers (e.g., embedded videos, comments or number of likes) as well as difficulties in reconstructing the complex layout of social platforms due to curatorial decisions or technical constraints (Chambers et al., 2021). While a few case studies have addressed challenges associated with replaying and providing access to archived social media material from specific platforms (Garg et al., 2023), there is a need to document more comprehensively the challenges, decision-making processes and aspects related to providing (and maintaining) access to these resources. This is also essential to inform and improve awareness of and foster scholarly engagement with archived social media.

Conclusion

This chapter began by framing the role that social media has come to play in today's communication practices, (re)presentation of the Self and as a means to facilitate the creation of a shared cultural memory. The intertwined interactions of billions of individuals who have been using social sites to discuss and document events that have left an indelible mark, especially in recent years, possess a unique cultural and historical value. As this material represents irreplaceable evidence of the recent past and presents a "mirror to society" (UNESCO, 2003c, p.5), social media platforms must be preserved to ensure that this invaluable resource remains accessible and available to future researchers despite its fragile and ephemeral nature. Cultural heritage institutions such as national libraries and archives hold a key position in ensuring the safeguarding of this material.

The second part of this chapter retraced the history of the first social media archiving attempts within web archiving initiatives. Although both sources share similar archiving challenges, some of which are connected to their born-digital nature, social media has proven to be inherently different and particularly problematic to capture since the very first archiving efforts. Drawing from scholarly literature, I illuminated milestones and setbacks that have characterised the social media archiving scene, identifying gaps and areas that can benefit from further research. In the final part of this chapter, I laid out the legal, ethical, curatorial and technical background against which this research is positioned. The overview provided in this chapter emphasises the need for additional research concerning the challenges and archiving practices specifically related to social media as a distinct resource from more traditional websites, particularly because of the ever-changing and ephemeral nature of these platforms.

The following chapter will discuss the results of a survey I conducted with the intent of gaining a deeper understanding about the current state of the social media archiving landscape, tracing trends and dynamics stemming from the geographical distribution of social media archiving initiatives.

CHAPTER THREE
Mapping Social Media Archiving Initiatives

This chapter offers an overview of social media archiving initiatives across the globe using data gathered through an exploratory web-based survey that was conducted between October 2021 and February 2022. The survey was developed with the intention of gaining a better understanding of the state of the art concerning social media archiving activities, capturing essential information that would help to identify trends and gaps in the current landscape. Previous studies, such as the one conducted by Gomes et al. (2011), provided a global overview of web archiving activities. More recently, Vlassenroot et al. (2021), in the context of the BESOCIAL project, identified in 2020 a few national memory institutions that extended their web archiving efforts to social platforms. Building on these, the present exploratory study aims to capture an updated snapshot of the current social media archiving landscape, identifying initiatives at different stages of development, including memory institutions that are still planning or only considering archiving social media content. After illustrating the reasoning behind the sampling strategy and the structure of the survey, I will discuss results from the survey, exploring aspects and imbalances that characterise the advancement of the social media archiving phenomenon. Furthermore, I will reflect on power dynamics, barriers and concerns related to the representativeness of the digital cultural heritage generated on social media and preserved to date. In the final section of this chapter, I will investigate potential barriers and reasons behind the absence of social media archiving initiatives in some of the surveyed countries. In particular, I will concentrate on the Italian case as an illustrative example, exploring its peculiar web archiving history, obstacles and legal context.

3.1 Sampling Strategy

The dissemination of the questionnaire was conducted in multiple stages, involving diverse means of online communication and personal connections established with organisations and web curators working in different memory institutions (see section 1.4.3). A link to the online questionnaire, introduced by an invitation to participate in the study, was firstly distributed via email using mailing lists that were selected either because of their relevance to the subject at hand, or because many of their members were web

curators/archivists or working in archiving institutions. Moreover, I sought responses from a single representative of each institution in order to avoid duplicates.

Most of the participants in the survey indeed came from the emails sent using internal mailing lists directed to members and associates of, for instance, the International Internet Preservation Consortium[34] (IIPC), the Web Archive Studies Network[35] (WARCnet) and Archives Portal Europe[36] (APE). The survey was also shared on Twitter as it appeared to be, among others, the preferred social media platform by professionals working in the cultural heritage sector and by scholars. As Sibona & Walczak (2012) explained, using Twitter for recruiting participants allows the researcher to reach a wide range of geographical areas in a cost-effective and timely manner. Although Twitter is not consistently used in all countries by the target population, many web archiving institutions seem to be active on the microblogging platform. In order to reach the target population, appropriate hashtags – such as #Academictwitter, #Socialmediaarchiving and #webarchiving – were used. While sharing the link to the survey on Twitter allowed the study to be acknowledged by a good portion of professionals involved with social media archiving, the actual number of participants recruited via Twitter was lower than those completing the survey after receiving a direct invitation via email. It needs to be noted, however, that participants coming from Twitter appeared to mostly be archiving initiatives originating in institutions with no previous web archiving experience.

As the number of social media collecting institutions was anticipated to be rather small, and mainly part of archives and libraries, a combination of mailing lists, posts on Twitter and targeted, follow-up emails to selected institutions was considered to be the most effective strategy to recruit participants for this survey. Follow-up emails inviting respondents to complete the survey were sent based on secondary research, including Google searches, analysis of conference papers, and national library and archive websites as well as the study of collection development policies publicly available online in which the intention to preserve social media was mentioned. To this latter group of targeted emails must be added those sent to a list of institutions following a form of snowball sampling initiated by participants in the survey itself. The questionnaire included an optional section where participants were invited to nominate other social media archiving initiatives of which they were aware or would like to see included in the study. This

[34] https://web.archive.org/web/20231203011111/https://netpreserve.org/

[35] https://web.archive.org/web/20231201074047/https://cc.au.dk/en/warcnet/about

[36] https://web.archive.org/web/20231201223514/https://www.archivesportaleurope.net/

approach proved to be essential to discover small, independent initiatives established outside national archiving institutions which would otherwise have been difficult to locate because of the limitations discussed in section 3.2.

3.2 Survey Structure

The online questionnaire was designed to gather essential information about existing social media archiving initiatives and institutions planning to develop a collection of content harvested from social platforms. The questions in the survey were formulated to primarily answer the research questions (see section 1.3):

> "*What has been achieved so far?* Where are the latest social media archiving initiatives located? How are they distributed geographically?"

The goal of the first part of this research question was to identify social media archiving activities worldwide, as well as provide an overview of the gaps and imbalances that may arise from such distribution and their repercussions on the representativeness of the shared collective memory on social platforms preserved at an institutional level.

The online form consisted of a first introductory section where all the relevant information relating to the study was set out for potential participants, followed by a consent form including questions related to privacy and data protection. The main section of the survey was a mix of nineteen closed and open questions aiming to explore the type and activities of participants' social media archiving projects (see Appendix A). As the main scope of this initial phase of the study was to assess the status and location of social media archiving initiatives worldwide, participants were asked firstly to declare the country in which they were based and the name of the memory institution they worked for. After clarifying the type of institution, selecting from a closed set of possible answers ("Archive", "Library", "Gallery", "Museum", "Private institution", "Other"), and their role within their institution, respondents had to indicate whether they were already archiving or were planning to archive social media. In case the response was negative, participants were invited to provide a brief comment on the reason why their institution was not planning to archive social media content, and this would be the end point of the survey. If the answer was instead affirmative, they were asked to specify the status of their project, whether it was a long-term project, pilot/fixed-term project or if they were still in the planning phase. Next, they had to indicate in the designated box the year in which

their institution started archiving social media. Since social media archiving is often widely considered an inherent part of web archiving, a question was dedicated to the assessment of whether the social media archiving project was indeed part of a broader, pre-existing web archiving initiative.

Finally, participants in the survey were invited to select all the social media platforms their memory institution was already archiving or was planning to archive. They were provided with a list of service providers including "Facebook", "Instagram", "Twitter", "TikTok", "WhatsApp", "LinkedIn", "Snapchat", and "Other". The latter was followed by a box in which they could specify any additional platforms that were part of the scope of collection at their institution. Two optional questions closed the survey: as mentioned above, one allowed participants to provide any details of other social media archiving projects (no matter the size and stage of the collection) that they were aware of and would like to suggest for this research; alternatively, they could use the box to give any other information they thought would be relevant for this study. In the last box, they could provide their institutional email address in case they agreed to be contacted for further clarifications or a follow-up interview.

3.3 Exploratory Survey Results

The data collection via survey was declared closed at the beginning of February 2022 and, by that date, a total of 33 valid responses were collected. For transparency, of the 35 responses received, one was excluded as it was a duplicate from an institution that had already completed the form but submitted by a different curator; and another one was not included in the final count since it was from an independent researcher and not a memory institution, and therefore considered out-of-scope for this study.

In order to guarantee as complete an overview as possible of the location of current social media archiving initiatives, including those that are still in the planning phase, I added to the map (Figure 3) the 33 initiatives collected via survey alongside thirteen projects identified through desk research. The latter group included national libraries, archives and museums whose official websites or collection development policies mentioned social media archiving activities to any degree, and other organisations (e.g., universities, research institutes, and Non-Governmental Organisations (NGOs)) that have been systematically capturing social platforms. Attempts made to contact these institutions to encourage their participation in the study were, unfortunately, unsuccessful.

Additionally, it should be noted that the list of countries illustrated in Figure 3 excludes institutions that focus solely on archiving websites and not social media, unless they provided direct responses to the survey.

Although the list of social media archiving projects presented in this chapter is certainly far from complete, with probably many other similar archiving efforts in existence that the survey did not capture for various reasons (a more extensive analysis of the limitations of this study can be found in section 1.5), the list of countries identified in this study still provides an overall and significant indication of the geographical distribution of such initiatives across the globe.

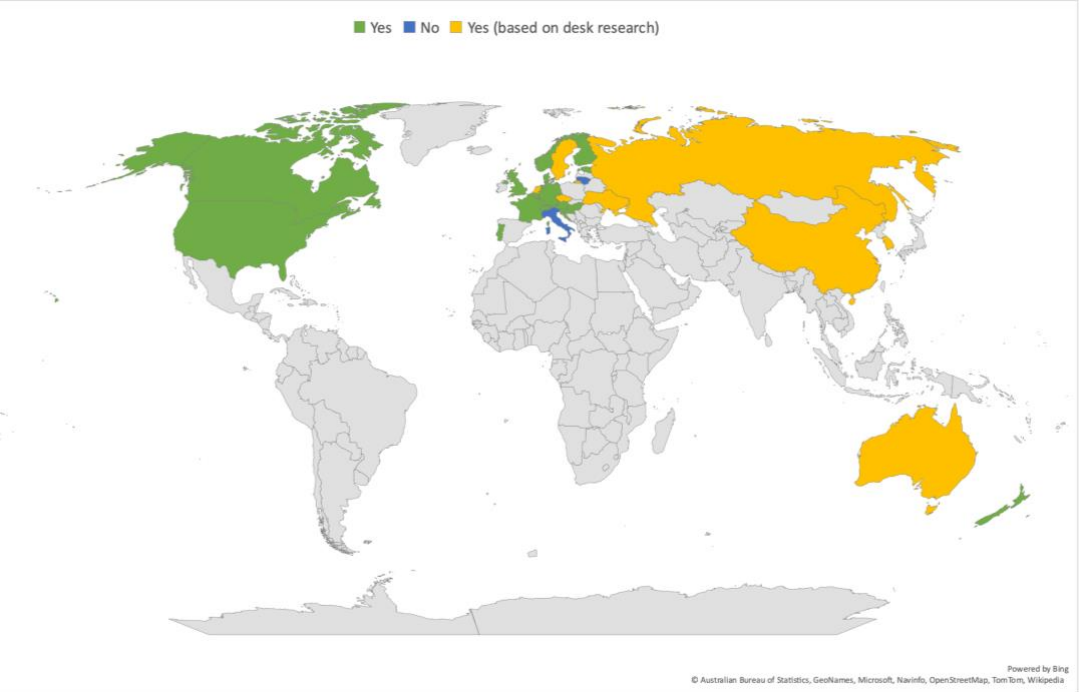

FIGURE 3: Location of social media archiving initiatives

As illustrated in Figure 3, social media archiving initiatives appear to be mainly located in the Northern Hemisphere,[37] with the majority of them concentrated in the European area, the US and Canada. In Russia, the non-governmental organisation "National Digital Archive of Russia"[38] has been preserving since 2021 websites and other digital materials

[37] Results gathered via survey included institutions situated in the following Northern Hemisphere countries: United Kingdom, Germany, Hungary, Italy, Lithuania, Luxembourg, Norway, Portugal, Slovenia, Switzerland, USA. Conversely, Czech Republic, Ukraine, Russia, Netherlands and Sweden were added based on secondary research.

[38] More information about the National Digital Archive of Russia can be found here: https://web.archive.org/web/20240220152256/https://ruarxive.org/about/intro

on the Russian-speaking web domain, including social media content, at risk of disappearing due to the lack of long-term preservation initiatives in Russia. It is also worth mentioning here the German-language Twitter archiving initiative carried out at the National Library of Germany[39] in collaboration with the Science Data Centre for Literature, which was prompted by the uncertainty surrounding the fate of the Twitter platform after it was purchased by Elon Musk (see also section 2.4.4). The Southern Hemisphere[40] sees New Zealand and Australia as two of the few countries in this region developing consistent collections of social media content. In Asia, the city-state of Singapore stands out for its efforts to work towards the development of its own social media archive. Also, the National Library of China announced in 2019[41] that it would start collecting billions of posts shared on the Sina Weibo platform, which is one of the most popular microblogging sites in the country, although no further updates on the development of the project have been released up until the time this PhD thesis was being written.

Only two memory institutions, located in Italy and Lithuania, completed the survey stating that at the moment they are not planning to archive social media content. Both institutions provided brief explanations about the reasons why their country does not plan to preserve this type of born-digital item for now, and these will be examined below (see section 3.5).

[39] The initial call for a crowdsourced effort to download Twitter content in German can be found here: https://web.archive.org/web/20230301195028/https://www.dnb.de/EN/Professionell/Sammeln/Sammlung_Websites/twitterArchiv.html

[40] Results from the survey included institutions based in the following Southern Hemisphere countries: New Zealand, Singapore. Conversely, China, Australia and South Korea were added based on secondary research.

[41] The article announcing the National Library of China's archiving project can be found at this link: https://web.archive.org/web/20220725143541/https://news.cgtn.com/news/3d3d674d79677a4e34457a6333566d54/index.html

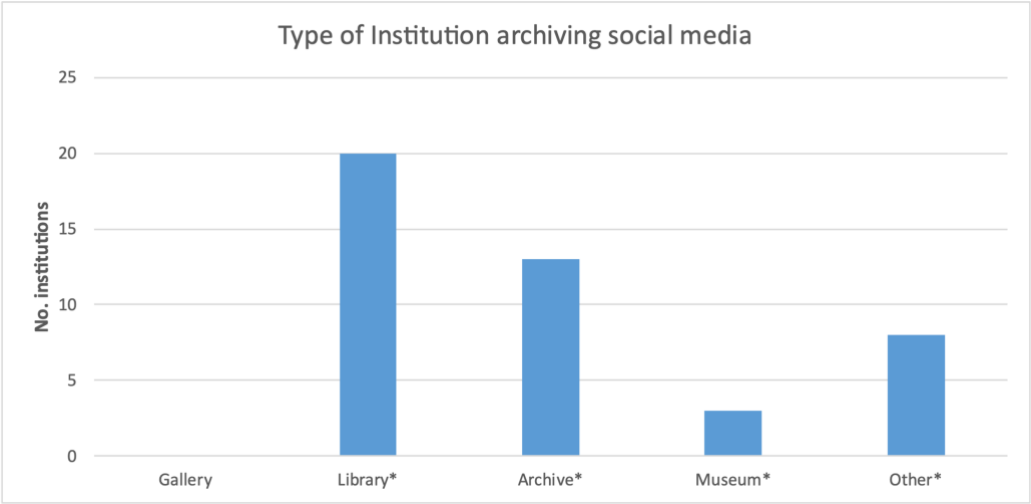

FIGURE 4: Type of institutions archiving social media [* includes data based on desk research]

One of the aims of the survey was to examine the type of memory institutions that are currently involved with social media archiving (Figure 4). As expected, the majority of social media archiving initiatives originate in libraries and archives, as it is inherent to their role and mission to preserve national history and cultural heritage, especially when records, including certain types of social media content, are received and preserved under national legal deposit legislation. Interestingly a fair number of respondents came from different areas of the GLAM sector and other institutions such as Universities and other governmental (and not) agencies. This is significant as it shows how the wide usage of social media, and the cultural and historical value that these platforms have come to gain in diverse sectors, have led many different institutions to include such items in their preservation strategies. Moreover, governmental agencies tend to collect social media not only for historical reasons but even more so to respond to the need for transparency (Bertot et al., 2010). Conversely, no responses from memory institutions defining themselves as galleries have been collected.

Unfortunately, this survey has not captured – although it was one of the intentions – any responses coming from community or activist archives, which might have provided additional insights on the challenges faced or solutions adopted by this specific type of archive. Moreover, it was rather difficult to identify any of these archives through desk research. Independent archiving initiatives, especially those flourishing in the context of activist causes or marginalised communities, are difficult to locate through a simple web search. Google's websites ranking system and geographical location further filters the type and number of potential community or activist social media archiving initiatives that can

be identified through a search engine query (Kliman-Silver et al., 2015). Websites that aggregate and promote community archives, such as the Community Archives and Heritage Group (CAHG)[42] for the UK and Ireland offer an optimal gateway. However, efforts to reach out to some of the community archives listed in the CAHG directory were unsuccessful. The already large scope of this research did not allow to dedicate additional specific time necessary to build effective relationships with groups linked to community archives. Moreover, factors such as the potential lack of volunteers or dedicated personnel managing archives' communication channels, or the risks associated with visibility of marginalised or activists' groups may have discouraged these initiatives to respond to the survey. Nevertheless, given the important role these type of archiving initiatives may have in contributing to the preservation of social media, future research could focus on exploring ways to bridge this gap (see Section 8.2).

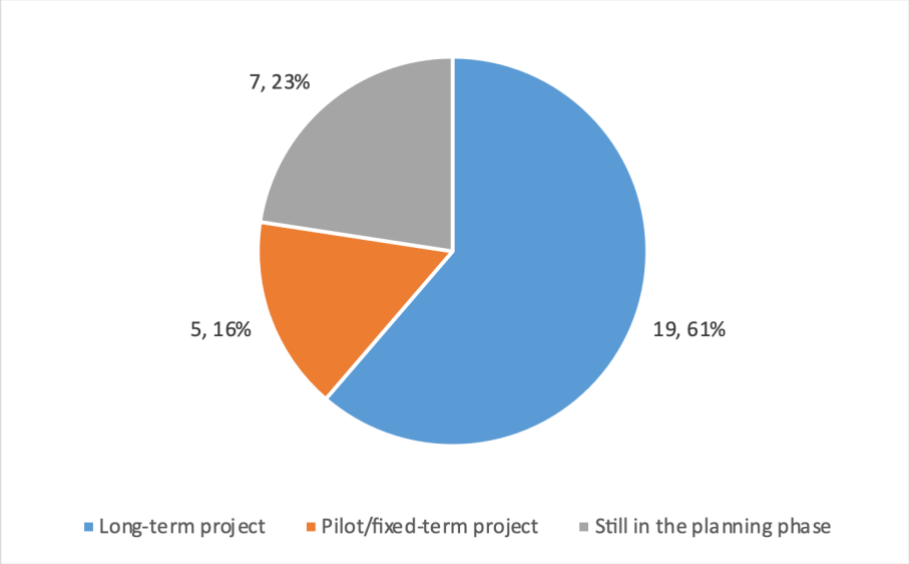

FIGURE 5: Status of current international social media archiving initiatives

The online survey sought to determine the type and status of participants' social media archiving projects. This information was essential to define the state of the art of social media archiving activities worldwide and to identify potential participants for the second phase of this research, which involved data collection via interviews. More than half of the institutions that completed the survey are currently running *long-term* projects (61%,

[42] The Community Archives and Heritage Group (CAHG) has also devised an interactive map showing the location of all community archives in the UK and Ireland: https://web.archive.org/web/20240114005505/https://www.communityarchives.org.uk/interactive-map

19), while the remainder are distributed between those working on *pilot/short-term* initiatives (16%, 5) and those that are still developing their own social media archiving project (23%, 7), as seen in Figure 5.

FIGURE 6: Social Media Archiving Initiatives: Year of Establishment

The analysis of the responses to the question about the year in which each memory institution started consistently archiving social platforms painted an uneven and fluctuating image of the history of social media archiving (Figure 6). One of the longest running initiatives appears to be the project established in 2007 by the Danish Royal Library as an extension of the Danish web archive project, followed by the British Library (2010), the Bibliothèque nationale de France (2012) and the collection developed at the Museum of London on the occasion of the Olympic games (2012). The year 2014 saw the creation of multiple social media collections at the same time, such as The UK National Archives, the Institut National de L'Audiovisuel in France, and The National Library of Finland. Similarly, in 2017 different types of projects began archiving social media in Canada, Luxembourg, Norway and Portugal. From 2018 up until now, a wide array of memory institutions has started developing their own social media collections, and many of them appear to be concentrated in 2020, with the Covid-19 pandemic probably being the main trigger event for attempting to start archiving these sites.[43]

[43] An example in this sense is the attempt to archiving social media content made by the National Library of Hungary in 2020 in the context of a COVID-19 pandemic special collection mentioned in WARCnet paper titled "Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary". The attempt was unsuccessful at the time but in a recent interview I

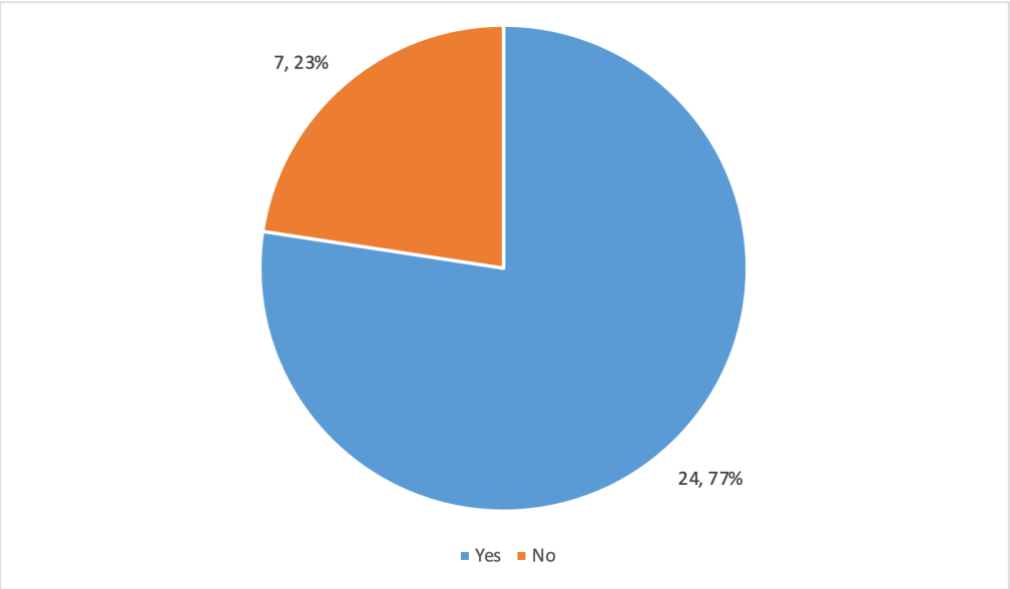

FIGURE 7: Number of social media archiving initiatives that are part (or not) of pre-existent web archiving projects

As expected based on previous exploratory studies (Vlassenroot et al., 2021), more than 77% (24 out of 31) of responses came from initiatives that were collecting social media as a natural extension of pre-existing web archiving projects (Figure 7). Conversely, 23% (7) of participants declared that their social media archiving project has been developed independently from any previous web-related initiative of the sort. The latter result is particularly meaningful as it seems to reveal a new trend in the development of social media corpora, especially in recent years.

conducted with the Hungarian web archive in April 2022 (discussed in Chapter 6) they stated that they are now working on finding solutions to further develop their social media collection. The WARCnet paper related to the 2020 interview is available here: https://web.archive.org/web/20240222101627/https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_Hungary.pdf

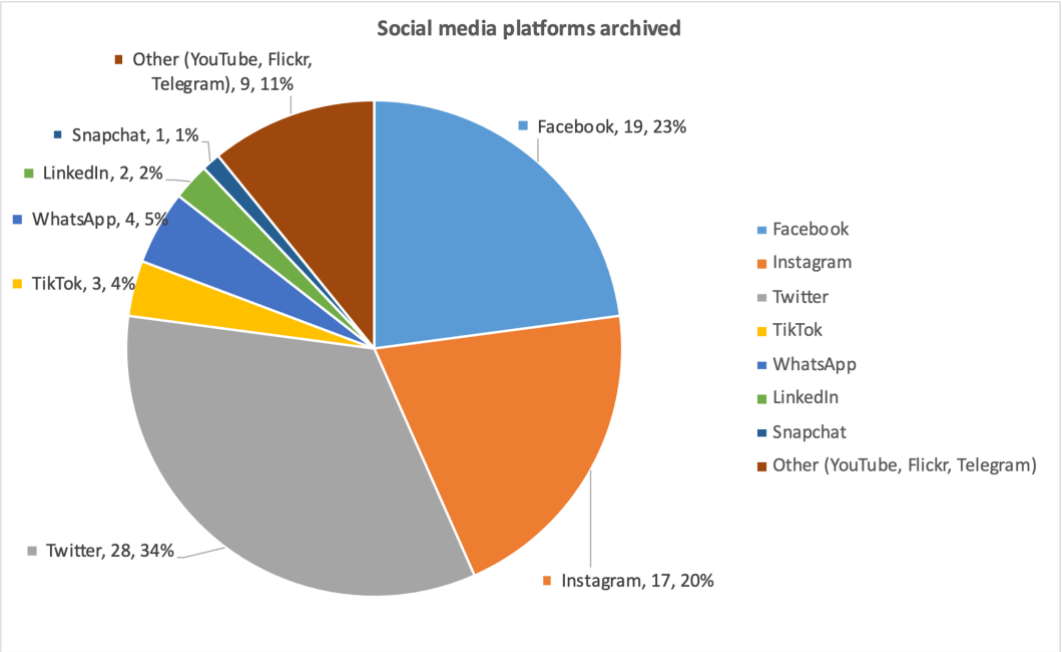

FIGURE 8: Most frequent social media platforms archived

Results from the survey illustrated in Figure 8 also confirmed the tendency for social media archiving initiatives to collect predominantly content from Twitter (34%, 28 out of 31), followed by Facebook (23%, 19) and Instagram (20%, 17). Interestingly, a fair amount of institutions seek to preserve material and conversations from WhatsApp (5%, 4); similarly, TikTok (4%, 3) seems to have been recently drawing the attention of various participant memory institutions, especially after becoming very popular around 2020-2021. Conversely, only a minority declared that they harvested content from LinkedIn (2%, 2) and Snapchat (1%, 1), the latter probably made even more difficult by the exceptional ephemerality of its content. Lastly, some respondents also indicated other social media platforms preserved by their institutions, such as YouTube, Vimeo, Telegram, Flickr, Tumblr.

3.4 Archiving Collective Memory on Social Media: Imbalances, Representativeness Concerns and Trends

Data from the survey and desk research revealed how the geographical distribution of social media archiving initiatives, including those that are still in the planning phase, appears to be heavily concentrated in the portion of the planet known, from a geopolitical perspective, as the Global North. As expected, the majority of these initiatives are located in North America, the European area, and Oceania where some of the pioneering projects

concerning the preservation of the web first appeared toward the end of the 1990s (see section 2.3). This is not to say, however, that similar archiving efforts do not exist in the Global South but rather that these may be less visible than well-established initiatives in the Global North. Limitations to this study described in section 1.5 identified some of the potential reasons behind the lack of representation of social media archiving initiatives from the Global South. Aspects such as the language used, and the strategy adopted to disseminate the survey may have influenced the capacity to reach such initiatives. While a comprehensive analysis of the geopolitical and infrastructural barriers that may be associated with the absence or scarcity of social media preservation activities in the Global South is beyond the scope of this thesis, it is crucial to acknowledge the complexity of existing power dynamics as well as cultural and technological divides, and how these can affect the development of social media archiving initiatives in certain areas of the globe (Chowdhury, 2002; Lutz, 2022). In this regard, a study conducted by Colin-Arce et al. (2023) exploring the impact of linguistic, social, economic and technical barriers on the development of web archiving initiatives in South America, revealed how the combination of all these obstacles often results in the creation of sparse, very small-scale collections of websites which are difficult to locate (Colin-Arce et al., 2023). Moreover, the scarcity of web archiving initiatives in this region can lead to a significant reliance on the Internet Archive and other international institutions situated in the Global North for the preservation of the web history of the Global South (Colin-Arce et al., 2023). This raises ethical questions about archiving practices that Lor & Britz (2004) defined as "benevolent", although they may risk being perceived as patronising toward countries of the Global South. Instead, efforts should focus on empowering these countries to preserve their own stories and online cultural heritage. In the case of social media archiving, where even the longest-running and well-established web archiving projects have to tackle a wide array of issues and are learning to navigate the complex process of preserving this new and ephemeral archival material, the aforementioned South-North archiving dynamics could potentially be further exacerbated.

Still, the geographical distribution and limited number of archiving projects actively involved in the collection of culturally and historically relevant content from social media, indicates that the collective memory generated on these platforms by the interaction of billions of users every day and currently being preserved is only a fraction of the real volume. Moreover, the national remit of most of the existing social media archiving initiatives further restricts the range of archived material (see section 6.1), and

their Global North location inevitably favours, for example, the preservation of specific platforms over others that are more popular or developed only for specific regions in the Global South. Differences in terms of platforms' popularity, their availability only in certain regions, and whether there are tools and expertise to adequately capture them, can generate gaps in the shared collective memory preserved on a global scale—gaps that will become increasingly difficult to fill in the future because of the ephemeral nature of social media materials. Recent studies have emphasised the need to raise awareness of web and social media archives across institutions in the Global South (Colin-Arce et al., 2023; Rockembach, 2017), the necessity to enhance mentorship and collaboration between institutions from the South and North, alongside a joint effort to make available translated versions of web archiving tools and training to facilitate the establishment of local web and social media archiving activities (Colin-Arce et al., 2023; Schafer et al., 2016).

Among the respondents, national archives and libraries are by far the most represented type of institution currently involved in the development of social media archives. As mentioned, the survey has revealed that the widespread usage of social media across different sectors has prompted various categories of organizations, including private ones, to pursue the preservation of social media content as an integral part of the documentary by-product of their daily activities. Among the responses collected through the survey, interesting examples in this sense are the historical archive of a political party in Germany, or the Institute of Social History in Belgium which decided to incorporate social media content into its collection in order to preserve its history across all media and platforms. These cases demonstrate how awareness of the historical value of material on social media is increasing, cutting across sectors, and determining the necessity to preserve it for the long-term in organisations other than national libraries or archives. Moreover, archiving such material becomes an obligation when it needs to be preserved for transparency and accountability reasons, at an organisational or governmental level.

Conversely, the survey did not capture any responses specifically from Galleries. Although this might be due in part to the survey not being able to reach the network of galleries that may actually deal with material from social platforms, an analysis of the websites of some of the most prominent galleries in the world has led to the conclusion that social media appears to be mostly used in this context to engage with potential visitors and as a means of communication rather than as items to preserve (Habelsberger & Bhansing, 2021). While it can be hypothesised that galleries preserve their social media accounts to some degree, the systematic creation of a collection of social media is

probably not among the current priorities of this type of institution. As collections and artworks are increasingly experimenting with including items such as those gathered from the web,[44] it cannot be excluded that at some point in the future even galleries might need to develop their own collection of works of art created in non-traditional or new media.[45] An interesting case in this regard is a collaborative project between the University of Manchester and a number of partners, including the Manchester Art Gallery, called the "Manchester Together Archive"[46]. Despite being at a standstill waiting for more funding as of 2022, the project focuses on materials related to the 2017 Manchester Arena terrorist attack.[47] The archive includes physical and digitised objects from the spontaneous memorials grown around the city after the event and a selection of content from Twitter (e.g., hashtags related to the terrorist attack) and Facebook. While the Manchester Art Gallery led on the collection of materials from the memorials, it is unclear whether it played a direct role in capturing social media content. Besides, in the FAQs section of the archive's website it is clarified that the "The Manchester Together Archive is distinct from the Gallery's art collection".[48] Nevertheless, this project offers a valuable example of how galleries might begin to engage in broader social media collection efforts.

In the attempt to paint a clearer picture of the current social media archiving landscape and to anticipate potential developments in the near future, the survey captured the stage which these archiving projects were at. The vast majority of respondents described their archiving efforts as "long-term" — which means projects that have been and will be archiving social media material for the foreseeable future, with no time

---

[44] See, for example, one of the artworks which are part of the ongoing exhibition */INB4/* at *Rhizome's Net Art Anthology*, that is based around social media such as Facebook. Available here: https://web.archive.org/web/20231207100454/https://anthology.rhizome.org/inb4

[45] This was confirmed in a conversation with the Conservator for time-based media at Tate Gallery, in an email sent on 5 January 2022: "At the moment Tate has no artwork using social media in the collection, we have only started to engage with web-based art recently".

[46] The "Manchester Together Archive" is being carried out by Dr. Kostas Arvanitis (University of Manchester) in partnership with with Manchester Art Gallery, Archives+, Manchester City Council, and Museu d'Historia de Barcelona (MUHBA). More information is available here: https://web.archive.org/web/20240903162933/https://mcrtogetherarchive.org/

[47] From the Manchester Together Archive FAQs page: "On 22 May 2017 a homemade bomb was detonated as people were leaving the Ariana Grande concert at the Manchester Arena. Twenty-two people were killed and hundreds were injured". Available here: https://web.archive.org/web/20240905105923/https://mcrtogetherarchive.org/faqs/

[48] ibidem

constraints, demonstrating the widespread interest in preserving social media content for the long term. As for respondents who run "pilot/fixed-term" projects, it needs to be recognised that, of these, only a few will probably translate into long-term archiving activities, with no assurance as to what will happen to the preservation of and access to content acquired during the trial period (see also section 6.1.3). The same applies to the valuable knowledge acquired during exploratory tests, especially when the curators involved in such projects move on to other roles or organisations. As for those institutions that are still in the planning phase, on the one hand, some appear to be waiting to receive additional fundings in order to start curating or integrating an initial collection of social media material harvested, for example, in response to a crisis, into what can be envisioned to be a long-term archiving effort; on the other hand, other respondents have been planning to start over, after previous problematic attempts at archiving social media, and are now searching for the most efficient ways to begin sustainably preserving this content again (see section 6.3).

Based on the data collected, the history of social media archiving has seen the initial few attempts at capturing these sites appearing first within pre-existing web archiving projects. As social media gradually emerged on the web scene, gaining great popularity, various archiving institutions – mostly national libraries and archives – started following the evolution of these sites. Such institutions then began to carry out harvesting tests to experiment on the actual archiving feasibility of social media, and especially its sustainability in the long-term, obtaining often mixed results. However, the survey also registered an increasing number of social media archiving initiatives which have been independently developed or plan to be developed separately from any pre-existing web archiving project. The reason behind this tendency can perhaps be found in the evolution of social media itself. It appears in fact that even though social media is clearly part of the Internet, these social platforms have started to gradually carve out for themselves a clear and well-defined space within the World Wide Web, becoming a sort of micro-cosmos with their own (community) rules and inherent dynamics. Thus, with the rising awareness of the central role these platforms play as means for interacting online as well as the cultural value they hold, interest has also been increasing among diverse memory institutions with no previous experience in web archiving in preserving collections of born-digital content created on social media. Moreover, if matched with the year in which these autonomous social media archiving initiatives were launched or intend to be launched, the above-mentioned tendency becomes even more significant: the majority of

them appear indeed to have started developing an archive of social media content (or are planning to) only recently, mostly between 2017 and 2022, and in connection with special events, such as terrorist attacks or global health crises. Furthermore, the fact that most of these independent archiving initiatives have been described by respondents as 'long-term projects' - therefore intended to remain in operation for the foreseeable future - further validates the cultural and historical significance associated with these born-digital records, as well as the commitment to their safeguarding for future generations.

Finally, the survey findings confirmed Twitter as the most archived social media platform in general, with Facebook and Instagram following close by. However, when looking at the popularity of platforms in 2023 worldwide (Statista.com, 2023b), Facebook appears to be the most popular social site, with almost 3.03 billion monthly active users; while its Meta companion, Instagram, sits in fourth place with 2 billion monthly active users in the year mentioned above. Conversely, Twitter can be found instead only in twelfth position, counting around 666 million monthly active users (Statista.com, 2023b). Numerous studies have analysed the profound impact that Twitter has had in terms of people participating in global discourses, activism, and information dissemination, especially through official authorities' channels (Brock, 2012; Florini, 2014; Holmberg & Thelwall, 2014), and what an important virtual environment it is for cultural and political expression. However, the fact that the microblogging site does not appear to be the most used platform in many countries, even though it is the most archived one, certainly raises questions about the actual representativeness of national social media collections.

On the one hand, a wide array of technical challenges and legal restrictions imposed by platforms such as those owned by Meta (see section 2.4.4 and 2.6), and on the other hand the relatively easier access to data allowed by Twitter Inc. until recently, especially through its official APIs, have been the main reasons leading many memory institutions to focus on preserving content from the latter platform. Undoubtedly, the current state of imbalance, in favour of Twitter, at least allows memory organisations to preserve material related to historically and culturally significant events from across the first two decades of social media usage. Nevertheless, the concrete difficulties concerning the inability to sufficiently diversify the existing social media collections successfully by including content from a wider variety of social platforms has generated a set of risks that might have profound repercussions on the representativeness of social media collections (see section 6.2.1). Also, as mentioned, recent developments in the administrative history of Twitter Inc. have made even more patent both how fragile these digital ecosystems are

and how heavily profit-oriented these platforms actually are (Paul & Milmo, 2022). In October 2022, Elon Musk acquired Twitter, making a series of changes to the way the microblogging platform had operated until then (see also Section 2.4.4). After a few months of internal restructuring and trying to find ways for the platform to become profitable (e.g. making available monthly subscriptions to obtain a blue tick next to the username and other benefits), in February 2023 Twitter announced the end of free access to the Twitter API and the establishment of a paid tier-based system (Twitter Dev, 2023a). While the full extent of the effects of such a move on memory institutions is still in the process of being assessed at the time this thesis is being written, it is easy to presume that it will have a heavy impact especially on smaller memory institutions with low budgets and on academics who used the free API for research purposes (see also section 5.5 and 6.1.4).

Among the institutions participating in the survey, a few also indicated that they are in the process of or planning to collect material from TikTok. This data reflects the growing interest generated by the Chinese social network after rising to popularity during the COVID-19 pandemic. During the first lockdown, TikTok was indeed inundated by millions of short videos documenting activities and users from any age group creating content in the wake of following viral trends to avoid boredom (Kendall, 2021). Although the interest in this platform continues to increase among memory institutions, many are the challenges related to the collection of TikTok and thus the number of archiving efforts in this sense is still quite low.

Among the social media platforms archived, it is also worth mentioning Telegram which is not among the most frequently archived platforms yet. However, the essential role played by Telegram as one of the main channels to disseminate critical, official information during the first few months of the 2022 war in Ukraine, has led a group of researchers from the Centre for Urban History in Lviv (Ukraine) to respond quickly and to initiate in a time of crisis an effective archiving effort focussed on selected Telegram channels. On their website, researchers involved in the development of the "Telegram Archive of the War"[49] declared that the purpose that prompted this project was "to collect and organize information flows that may quickly disappear due to their short-lived digital nature", creating then a collection of born-digital material with the aim of documenting

[49] Further information about the "Telegram Archive of the War" can be found here: https://web.archive.org/web/20230918165138/https://storymaps.arcgis.com/stories/0af72de4b008461bb441fc62fffb9f8d

Ukrainians' lives and testimonies about the war shared on Telegram. The Telegram Archive of the War is just one example of the preservation activities carried out by the above-mentioned autonomous social media archiving initiatives, highlighting the important role these may play in saving content that is probably not in scope for many of the longest-running archiving institutions but essential to document local historical events.

When reflecting on survey results about the type and frequency with which social platforms appear to be currently archived, there are some aspects that are worth considering further or underlining. The maturity of born-digital archiving initiatives, particularly those based in the US and European area, alongside the national focus of many collections governed by national legal deposit regulations, and the numerous restrictions imposed by major social media platforms on data harvesting at scale, collectively shape the type of social platforms currently collected. When looking at the data, it is also important to remember that, as mentioned, not all social platforms are popular or used to the same extent in all countries, resulting in inevitable gaps and imbalances in current national collections. On a global scale, such gaps risk being accentuated even further as there are some social media platforms that are only used in certain areas of the globe that currently appear to either have no plans to start preserving these sites, or simply do not have the resources yet to initiate a social media archiving initiative, exposing them to the risk of leaving no evidence of their online foot-print behind.

These are important elements to consider when reflecting on the real extent of the number of platforms excluded, on the virtual completeness and actual representativeness of social media collections created thus far on a national level, and also when considering the future and overall persistence of the collective memory created on social sites. Ephemerality of social media content is a well-known problem (SalahEldeen & Nelson, 2012), as explained in section 2.2.2, which emphasises even further the necessity for more memory institutions worldwide to start acting now in order to ensure the preservation of important material at risk of disappearing from social platforms, especially on those that are less mainstream. Fragments of content created on less archived social sites, or those not archived at all, will perhaps persist on the Internet under a different form (e.g. screenshots) or through other sources. Nevertheless, the current state raises questions regarding the preservation and representation of silences, marginalised voices, content produced worldwide and especially in Global South

countries, as this content is and will be essential for interpreting and studying events beyond the Global North and national borders. The emerging phenomenon of social media archiving initiatives being developed independently from previous web archiving efforts and, often, in organisations other than national libraries and archives, represents an opportunity to shape collections targeting specifically non-mainstream platforms, while also seeking to keep a record of silences and marginalised histories. This trend also indicates that an activity such as social media archiving, which has often started as a natural extension of web archiving, has been steadily defining itself as a separate archiving experience with its own dynamics and challenges, which requires the development of *ad-hoc* solutions and archiving practices in order to further advance.

The following section will explore the barriers that often delay or even prevent memory institutions from even thinking about starting to develop a social media archiving initiative. The mentioned obstacles emerged from a brief exchange of follow up emails with survey respondents who declared no plans had been made at their institution to start collecting material from social sites.

3.5 Some Considerations on the Barriers and Concerns Preventing Memory Institutions from Planning the Development of Social Media Archives

In order to assess the type of issues and concerns that may preclude or hinder memory institutions from even beginning to consider creating a social media archive, a short, written follow-up interview was conducted via email with the only two institutions that responded negatively to the survey question enquiring about plans to archive social media (see Appendix A, Q11).

The follow-up questions included an invitation to discuss the main challenges or concerns that were preventing them from initiating social media archiving activities at their institution; whether their national legal framework included provisions regarding the preservation of content published on the web, and specifically on social media; and if they had ever received requests or enquiries from members of the public or researchers regarding any social media collection activities at their institution (see Appendix B). Responses to the questions were received in January 2023.

The interviewees described a number of elements that hindered the creation and the beginning of a potential planning phase at their institution and in their country, in particular, the uncertainty surrounding the selection of content from social media, the lack

of specific standards or clear guidelines for the collection and preservation of material, the complexity of the legal barriers coupled with the absence of explicit legal deposit regulations about social media preservation at a national level, and the unlikelihood of establishing contact with social media platforms. These elements of uncertainty have been influencing the development – or rather the non-existence in this case – of social media archiving initiatives in the institutions surveyed.

Selecting items to be preserved among the sheer amount of content created and shared daily on social media can be quite a daunting task, especially for institutions that do not have familiarity with preserving similar born-digital material published on the web. In particular, even when content of interest is identified, questions still remain, for instance, in relation to how to deal with accounts of public figures who perform governmental duties: "Should we preserve it as private or institutional documents […especially when] the post on social media is not associated with institutional policies" (Bujokas, e-mail). One institution expressed reservations about whether to adopt for social media content the same selection criteria used for traditional archival records. Moreover, the interconnectedness of the activities happening on social sites may be a cause of further uncertainty due the great amount of data generated daily, from which to select valuable content and sift through the less meaningful, out-of-scope, or seemingly more private material (e.g., family pictures in politicians' social media profiles) (see Section 6.1.2).

These are all valid concerns that originate in the relative novelty of social media content as an archival record, which makes evident how the problem perhaps resides in the lack of clear and widely accepted guidelines that could enable institutions new to web archiving practices to tackle the preservation of this important resource: "For now the main obstacles are that there are no established practices on how to preserve social media. What formats should be used?" (Bujokas, e-mail). Navigating the different formats available, some of which were specifically developed to preserve websites and not social media data, is indeed a further source of uncertainty for many institutions that has profound repercussions for whether an organisation will actually develop a social media archive or not. One respondent noted that when preparing a proposal for a project of the likes of a social media archive, it is essential to be able to provide an accurate estimate of the resources, collection and preservation strategies necessary to initiate and sustain the project through time. Assessing the resources required for a similar archiving endeavour constitutes an additional issue, as often it is a matter of finding a balance between the

basic requirements to keep the project going and funding available in the cultural heritage sector.

A factor that would probably drive more funding towards institutions, allowing them to develop social media archives, could be the ability to demonstrate the growing interest of the wider public and academic community in studying such content. Unfortunately, based on survey findings and follow-up questions, none of the institutions that are currently not archiving social media have so far registered queries about social media archiving activities and access to a potential existing collection. The one institution located in Italy had received, however, a few enquiries from private organisations asking whether the memory institution provided an on-demand service for archiving social media channels on their behalf.

Both respondents to the survey follow-up questions identified national legal frameworks and the lack of specific regulations for non-print legal deposit as one of the main barriers. Legal deposit regulations often pertain to printed material and certain types of digital records, and, in many countries, these have yet to be updated in order to reflect the latest technological advancements. Modifying such regulations is essential, as it would enable national memory institutions to obtain a legal mandate to archive public content available on the web and social media. While copyright laws and other legal matters (such as the GDPR legislation in the European area) are still reasons for concerns in many memory institutions, the absence of updated national legal deposit legislation including content published on the web, social media and any future development of the sort is probably one of the key motives behind the scarcity of social media archiving initiatives in countries that might appear to have sufficient resources or infrastructures to develop them.

Both institutions recognised as an obstacle to the preservation of social media the struggle to establish any kind of communication or agreements with service providers. Agreements following the model of the one signed in 2010 between Twitter Inc. and the Library of Congress (see section 2.3) would greatly facilitate the harvesting process, especially where platforms impose strict limitations on the amount of data crawled. Moreover, in countries where legal frameworks have provisions for the preservation of content published on the web, the way clauses are phrased, the insertion of too specific descriptions of items to be preserved under existing laws and/or delays in approving related regulations create ambivalent circumstances that prevent the capture of specific material on the web and social media. That will remain the case at least until up-to-date

legislation is ratified. A good example in this sense is the case of Italy, where the Biblioteca Nazionale Centrale di Firenze [National Library of Florence] (BNCF), despite being able to count on very limited funding and resources, has been harvesting and preserving a small selection of Italian websites since 2006. The following section will briefly explore the Italian case, where a mixture of legal technicalities, political dynamics and the scarcity of funding often destined for the cultural heritage sector has created a challenging environment for the development of any social media archiving initiative.

3.5.1 The Italian Case

The Italian case appears to be particularly interesting for the purposes of this study, as it constitutes an optimal example of how national legal frameworks and the absence of clear legal deposit regulations concerning social media archiving, and consequently of a legal mandate that would enable institutions to collect public material online at scale, represent some of the main issues that impede the development of similar archiving initiatives. The Italian case appears even more significant as Italy - represented by the *Biblioteca Nazionale Centrale di Firenze* (BNCF) - appears among the twelve founding members of the International Internet Preservation Consortium (IIPC),[50] which is evidence of the interest in participating in the preservation of the history of the Internet and, specifically, the Italian web domain. Unfortunately, probably due to lack of resources and budget funding, the BNCF had to leave the Consortium after only a few years of involvement.

In Italy, the BNCF in collaboration with two other national libraries – the *Biblioteca nazionale centrale di Roma* (BNCR) and the *Biblioteca nazionale Marciana di Venezia* (BNM) - has taken up the task of preserving and providing access to selected types of digital documents published on the web that hold cultural value, including a limited number of websites and digital copies of doctoral theses.

A first prototype of the project called *Magazzini Digitali* [digital repositories] was implemented in 2006 in compliance with art. 37, para 2, of the D.P.R. n. 252/2006,[51] which provided for some voluntary tests for the legal deposit of born-digital documents

[50] The full list of the IIPC founding members can be found here: https://web.archive.org/web/20040603043437/http://netpreserve.org/about/members.php

[51] *Regolamento recante norme in materia di deposito legale dei documenti di interesse culturale destinati all'uso pubblico*, D.P.R. 3 maggio 2006 n. 252. Available at: https://web.archive.org/web/20231004103222/https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.del.presidente.della.repubblica:2006-05-03;252!vig=

transmitted through the Internet. This was followed in 2010 by a letter signed by the BNCF and BNCR (both serving as legal deposit libraries at a national level) and BNM (which functions as a regional deposit library) that sanctioned the creation of an infrastructure for the digital legal deposit. The test run was executed between 2012 and 2017 with a positive outcome, thus laying the foundations for the establishment in 2018 of a service for the harvesting and preservation of websites at the BNCF, following the example set by other archiving efforts carried out in national organisations in Western countries. As mentioned in one of the answers to the survey and clarified in the response to the follow-up questions, the BNCF does not, however, yet have an official legal mandate to systematically archive the Italian top-level domain. The complex set of legal matters and political dynamics surrounding the absence of an enforced digital legal deposit regulation is indeed one of the main reasons behind the non-existence in Italy of web and social media archiving initiatives similar to those already established in other parts of Europe. When the new legal deposit legislation was introduced in 2004,[52] article 4, letter r) included among the type of records that were to be collected under legal deposit: "documenti diffusi tramite rete informatica […]"[53] [documents transmitted through the web…].

The following regulation, issued in 2006,[54] specified in article 37 the methods for acquisition of electronic documents of cultural value, giving priority, among others, to digital material related to scientific publications (e.g. doctoral theses), and those related to public authorities. Unfortunately, the implementing regulation envisaged in 2006 by art. 37 para. 1 has not been issued yet, thus making legal deposit of digital records not yet mandatory. As mentioned, art. 37 para. 2 provided, however, for the establishment of the project *Magazzini Digitali*, a digital repository for digital copies of doctoral theses, eBooks, eJournals and some websites.

As for the legislation specifically related to the preservation of digital public records and acts, the *Codice dell'Amministrazione Digitale* [Digital Administration Code][55]

---

[52] *Norme relative al deposito legale dei documenti di interesse culturale destinati all'uso pubblico*, L. 15 aprile 2004 n. 106. Available at: https://web.archive.org/web/20240424120230/https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2004-04-15;106!vig=

[53] All translations from Italian language are my own.

[54] *Regolamento recante norme in materia di deposito legale dei documenti di interesse culturale destinati all'uso pubblico*, D.P.R. 3 maggio 2006 n. 252.

[55] *Codice dell'Amministrazione Digitale*, D. Lgs. n. 82/2005.

included a far too specific definition of what is to be intended as a *digital record*, hindering any web preservation activities even for the public sector. In article 1, letter p) of the D. Lgs. n. 82/2005, modified by D. Lgs. n. 179/2016, letter d),[56] the following definition is provided: "Documento informatico: il documento elettronico che contiene la rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti." [Digital record: an electronic document that includes the digital representation of acts, facts and data judicially relevant.]. This is too specific a definition which seems to only marginally consider the potential evolution of the digital world and completely excludes any material produced by government agencies in the fulfilment of their duties that might be created and shared online, such as websites or social media channels. These legal technicalities and the delays in approving the above-mentioned regulation, which would enable the BNCF and its partners to start archiving consistently the web and social media sites, have led to an impasse that explains in part the lack of social media archiving activities at an institutional level (or any plans to start one in the near future) in this country, although it cannot be excluded that small-scale, independent social media archiving initiatives may exist in Italy outside of national archiving institutions. Still, the ratification of the legal mandate to archive this born-digital material would certainly establish the basis for the BNCF to also receive adequate funding to gather the necessary tools, professional expertise and resources for finally developing and being able to sustain such archiving efforts through time.

Conclusion

The findings of the survey highlighted how the geographical distribution of social media archiving initiatives is the result of the combination of the economic and technical divide and power dynamics that characterises the dichotomy between the Global North and Global South. In the Global North, libraries and archives are the main type of institutions taking up the challenging task of preserving social media, although in recent years an

Available at: https://web.archive.org/web/20240227034336/https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82

[56] D. Lgs. n. 179/2016. Available here: https://web.archive.org/web/20240602210656/https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2016-08-26;179

increasing number of institutions across different sectors, including museums, universities and other small archives have started to or plan to archive social media. Yet, this should not be interpreted as a complete lack of initiatives of this sort in the rest of the globe; rather, it may be an indication of discoverability issues that are related for example to the small scale of the projects, the disconnection from known networks or the language used to disseminate the survey. The cultural value that social media holds as a result of the activities carried out by organisations throughout their existence and, in some cases, as essential records for ensuring transparency, is at the basis of the inclusion of social media in many digital preservation strategies at an institutional level, no matter the sector to which these institutions belong. This interest is confirmed by the fact that most of the participants in the survey were involved in "long-term" archiving efforts to preserve social media material for future generations. Questions about the fate of collections created within "pilot/short-term" initiatives after the end of the test period remain open, as many of them do not translate into long-term archiving activities. Nevertheless, data from the survey showed a fair number of participants actively considering or planning to start collecting social media at their institution, suggesting an increased interest in the preservation of social media. While most social media archives established so far have originated within pre-existing web archiving projects, the last five years has seen a trend inversion. Among the respondents, more and more social media archives seemed to be initiated (or plan to be) separately from any previous web archiving activities. Due to the rapid evolution of social sites in the two decades since their first appearance, these platforms have been able to carve out a distinctive space for themselves within the world wide web, and archiving initiatives appear to be following the same tendency.

Survey findings shed light on the type of social platform collected, indicating Twitter as the most archived one in the majority of the institutions participating in the study. Social media collections appear to be heavily shaped by national legal frameworks, limitations imposed by service providers as well as technical issues that affect the type of social site institutions are able to capture with an acceptable degree of success. Although archiving initiatives in institutions located in the Global North have made advances in tools and strategies for harvesting and preserving social media, persisting challenges and limitations concerning certain platforms have led to their exclusion from existing collections. The fact that the most archived platforms often do not enjoy the same popularity in all countries, and others are only used in countries that are currently not archiving any type of social media, raises concerns about the representativeness of

national collections and the fate of the shared global collective memory generated on social media. Nevertheless, the emerging trend that sees a growing number of autonomous social media archiving initiatives from previous web archiving experiences as well as outside legal deposit institutions, constitutes an opportunity to shape new and diverse social media collections.

However, smaller and independent archiving initiatives can prove to be more difficult to locate, especially because these may be less likely to become members of international networks (e.g., IIPC). For this reason, it can be assumed that there are many other independent initiatives actively preserving social media to some extent, that may not have been identified through the survey conducted as part of this research. Nevertheless, these autonomous social media archiving initiatives could represent an opportunity to archive material from less archived platforms, to capture the stories of marginalised communities and activists or to pass down snapshots of events related to Global South countries.

To gain a deeper understanding of the social media archiving landscape, follow-up interviews via email were conducted with respondents whose institutions declared that they had no current plans to archive these platforms at the time of completing the survey. Responses to the follow-up questions highlighted some of the challenges and concerns that seem to prevent Global North institutions initiating similar archiving efforts. Among the main problems mentioned, it is worth noting concerns regarding selection criteria, deletion of content from live sites and the lack of standards and best practices that may guide institutions throughout the planning phase. However, the prime obstacles appear to be identified in a series of legal issues such as the absence of legal deposit legislation clearly including social media and future developments of the web, copyright concerns and the inability to establish any type of agreement with social platforms. The case of Italy was used as evidence of how the combination of legal impediments and lack of resources is profoundly affecting the decision of whether to proceed with the development of a social media archive.

The following two chapters will focus on case studies of three institutions, two located in France and one in the United Kingdom. The case studies were selected on the basis of responses to the survey and following interviews and fieldwork carried out at the below-mentioned institutions (see Sections 1.4.2 and 1.4.4). These case studies will provide a more in-depth analysis of practices, challenges and solutions implemented in

relation to social media archiving at the British Library (Chapter 4), the Bibliothèque nationale de France and the Institut national de l'Audiovisuel (Chapter 5).

CHAPTER FOUR

Archiving Social Media Under Electronic Legal Deposit in the United Kingdom: A Case Study of the British Library (BL)

This chapter includes the first of two case studies that aim to illuminate in greater depth practices developed and challenges faced by archiving institutions in the development of a social media archive under electronic legal deposit legislation. As mentioned in Section 1.4.2, the first case study will focus on the obstacles encountered and practices implemented for the collection of social media material related to the United Kingdom by the web archiving team at the British Library (BL), one of the six legal deposit libraries involved in the long-term preservation of the UK web domain and social media. This chapter begins with an overview of social media archiving activities at the BL – the lead contributor to the UK Web Archive (UKWA) – and how these came to be.

This is followed by a section dedicated to the legal framework in which the project is embedded, as it is essential to understand how legal restrictions and limitations have been shaping the development of the social media collection and the modalities of access to the archived material. The chapter continues with a section illustrating the type of tools utilised to collect content from social platforms and the frequency with which content is crawled. The last portion of the chapter discusses challenges encountered when archiving social media, providing examples (where available), and illustrating potential solutions adopted by web curators and archivists at the British Library in the attempt to mitigate and solve such issues. The chapter closes with a brief paragraph concerning future developments and points of interest that will need to be addressed in the near future in order to ensure social media archiving progresses at the institution. Comments provided in this case study are based on the interview conducted on 23 June 2022 with Nicola Bingham, Lead Curator of the UK web archiving team. In addition, this chapter is enriched by information extrapolated from documentation made available on the BL website (e.g. collection development policies) and blog, the UKWA portal, articles appeared on other websites and platforms (e.g. netpreserve.org or YouTube), conference presentations and papers published by BL curators regarding their web and social media archiving activities.

4.1 Overview of the Project

The UK Web Archive (UKWA) is a consortium of the six UK legal deposit libraries – that is the Bodleian Libraries (Oxford University), British Library, Cambridge University Libraries, National Library of Scotland, National Library of Wales, Trinity College (Dublin) – who are working together to select, collect, and preserve, among other published items, copies of any UK related material digitally published, including websites and social media.

The origins of the UKWA can be traced back to 2002, when a study was undertaken following the request made early that year by the Joint Information Systems Committee (JISC) and the Wellcome Trust to produce a report on the feasibility of establishing a web archive in the UK (Bingham & Byrne, 2021; Webber, 2020). In an article discussing the origins of the UKWA, Bailey & Thompson (2006) explained how at the beginning a consortium of six UK institutions (The National Archives, The British Library, JISC, the national libraries of Scotland and Wales and the Wellcome Library) came together to collaborate on a trial to selectively archive UK websites. The collaboration of these six institutions led to the creation of the UK Web Archiving Consortium (UKWAC) whose main aim was to implement the archiving initiative. The UKWAC team looked into the many technical issues and curatorial concerns, which proved to be particularly challenging back then when only a handful of organisations in the world, such as the Internet Archive and the National Library of Australia, were active in this field (Webber, 2020). At the beginning, the UKWAC decided to use a version of the Pandora Digital Archiving Systems (PANDAS)[57] software adapted to respond to UK requirements. The PANDAS system had been developed by the National Library of Australia (NLA) after unsuccessfully searching for existing commercially available systems. The implementation of PANDAS provided a web-based, end-to-end archiving workflow and a managed environment in which material could be collected and managed, ultimately allowing the UKWAC to easily and quickly start archiving websites (Bailey & Thompson, 2006).

The UK Web Archive was officially launched in 2005, when the six UKWAC partners, including the British Library, started collecting copies of websites published in the UK for long-term preservation and access on a selective basis (Webber, 2020). The implementation of the Non-Print Legal Deposit regulation in 2013 (see Section 4.2)

[57] Further details on the Pandora Digital Archiving Systems (PANDAS) can be found here: https://web.archive.org/web/20230228035812/https://pandora.nla.gov.au/pandas.html

appointed the six UK legal deposit libraries to take responsibility for the preservation of non-print material published on- and offline, including websites. Since then, the aim of the UKWA project has been that of capturing the entire UK web domain at least once a year (Bingham et al., 2020). Fulfilling its legal deposit duties, the British Library has been observing the evolution of the web alongside its partners, seeking to preserve the UK's digital cultural heritage on the web. Thus, when the first social platforms such as Facebook were launched, rapidly becoming an integral part of UK society's daily communications, the British Library sought to incorporate them into the UKWA. While continuing to collect web pages, the BL began gradually experimenting with the capture of these dynamic sites around 2010. The first attempts at acquiring social media focused mostly on official accounts of public figures (e.g., politicians) and organisations. Excluded from the capture were profiles and private content (e.g., private chats) shared by individuals or behind login gates. One of the first experiments of the sort was conducted on material published on Facebook and Twitter during the UK General Election in 2010.[58] As for many other web archiving institutions, general election campaigns represent unique occasions to be documented and captured, since they may be of interest and an object of scientific analysis from a political, historical, cultural and sociological perspective (Bingham et al., 2020). The test, however, produced mixed results, as discussed in Section 4.4.2. After Facebook, further archiving tests were carried out by the BL on other social media platforms, such as Flickr and Instagram, drawing attention to the many technical difficulties, and legal and ethical concerns associated with the collection and preservation of such material. The British Library implemented in 2016 the first collection policy specifically dedicated to archiving social media, better defining the type of content shared on such platforms that was to be considered in scope for the UKWA.

To date, the British Library collects any and all UK based websites, including all sites comprising the UK top level domain such as .UK, .SCOT, .WALES, .CYMRU and .LONDON. It also includes any other website hosted on servers located in or whose owner resides in the UK territory. As reported on the UKWA portal,[59] websites are mostly

[58] A few examples of archived material from Facebook and Twitter collected on the occasion of the UK General Election in 2010 can be found here: https://web.archive.org/web/20220528014046/https://www.webarchive.org.uk/en/ukwa/collection/2422

[59] UKWA website, FAQ section: https://web.archive.org/web/20220123234843/https://www.webarchive.org.uk/en/ukwa/info/faq#what-is-non-print-legal-deposit

collected during the broad annual UK domain crawl, while only a selection of websites and platforms are archived more frequently depending on the rate at which these sites are updated (e.g., news sites or selected accounts on Twitter are archived on a daily basis). When it comes to social media platforms, the BL predominantly collects content on a selective basis. Moreover, social media is often preserved as part of special or thematic collections created by BL curators or in collaboration with external curators about important events that influence and shape UK society, for example general elections or health crises (Byrne, 2017; see also Section 4.4.2). The British Library mainly focuses its collection efforts on Twitter as it has proved to be relatively easy to archive and is often understood to be a much more "public" platform compared to others such as Facebook, which is instead perceived by users as a site where content is mostly shared among a more or less close circle of "friends" and only a small number of accounts are actually fully publicly accessible (Burkell et al., 2014; see also Section 2.4.2). The UK Web Archive preserves, up to 2023, a little over 4000 Twitter accounts related to public figures, some of which have been archived over a fixed period of time – as in the case of accounts related to a general election – whereas other profiles are archived on an ongoing basis, with a crawling frequency set at "daily".

4.2 Legal Framework

From a legal perspective, a turning point in the web collection activities at the British Library and its partners was represented by the enactment of the Non-Print Legal Deposit regulation[60] (NPLD) in 2013, which gave the British Library and the other five nominated legal deposit libraries the legal mandate to capture, preserve and make accessible content from the UK web domain to researchers and generations to come.

The history of legal deposit legislation for printed publications in the UK has been constellated with various Acts dating back to the "Act for the Encouragement of Learning" in 1710.[61] It was however only in the twentieth century that a series of Acts, starting with the "Copyright Act" in 1911, established the current arrangements for the

[60] Available at:
https://web.archive.org/web/20240828184923/https://www.legislation.gov.uk/uksi/2013/777/contents/made

[61] Available at:
https://web.archive.org/web/20230226215036/https://avalon.law.yale.edu/18th_century/anne_1710.asp

deposit of printed publications at the British Library (Gibby & Brazier, 2012). When publications began to be made available in a medium other than print (for example, CD-ROMs or e-journals) between the 1980s and 1990s, discussions regarding the need for an update of the legal deposit legislation that would include, for instance, electronic publications started to arise. Gibby & Brazier (2012) accounted for the complex and difficult road that led to the first step toward the inclusion of material published on new media, both on- and off-line: the Legal Deposit Libraries Act 2003,[62] which created the fundamental framework for subsequent secondary legislation on the matter to be introduced, extending "the system of legal deposit progressively and selectively to cover various non-print media as they develop".[63] Nevertheless, it took almost ten years of public consultations after the passing of the 2003 Act to finally see the NPLD regulation come into force in 2013 (Arnold-Stratford & Ovenden, 2020). Following the implementation of the NPLD regulation 2013, the designated six legal deposit libraries were legally required to collect and preserve a wide variety of material published digitally, including webpages.

As observed in similar legal deposit legislation that has come into force over the last couple of decades in other European countries, the NPLD regulation includes a generic description of what should be considered as a publication on the web. Using a broad definition of what type of born-digital material published, specifically on the web, needs to be considered in scope for collection under legal deposit has revealed itself as a determining factor in the development of social media archiving initiatives as it leaves room for the inclusion of any future and unpredictable evolution of existing technologies and formats (see also Section 6.1.2).

As prescribed by the NPLD and further explained in the "Guidance on the Legal Deposit Libraries (Non-Print Works) regulation 2013" published by the Department for Culture Media and Sport,[64] the British Library and the other five legal deposit libraries can

[62] Available at: https://web.archive.org/web/20240222072851/https://www.legislation.gov.uk/ukpga/2003/28/contents

[63] Legal Deposit Libraries Act 2003 – Explanatory Notes. Available here: https://web.archive.org/web/20240829162226/https://www.legislation.gov.uk/ukpga/2003/28/resources

[64] The full-text of the "Guidance on the Legal Deposit Libraries (Non-Print Works) regulations 2013"

harvest material made publicly available online, excluding, however, work consisting only of "a sound recording or film or both" and content which "contains personal data and which is only made available to a restricted group of persons"[65]. The NPLD regulation 2013 further clarifies that a work published on the internet needs to be treated as published in the UK when it is "made available to the public from a website with a domain name which related to the United Kingdom or to a place within the United Kingdom", and when it is published on the web by "a person and any of that person's activities relating to the creation or the publication of the work take place within the United Kingdom". It is thanks to such a broad definition and, in particular, to the inclusion of content published by people and as a result of their activities when these take place on UK soil, that it has been possible for the British Library and its Legal Deposit Library partners to actually include in their harvesting activities material published by public accounts on social platforms or in any other future development of similar media. In this way, the NPLD regulation 2013 enables Deposit Libraries to collect content from social media platforms which otherwise would be excluded, as the domains of these sites are mostly .COM and located in countries other than the UK. Moreover, the NPLD regulation 2013 (and reiterated in the "Guidance") appears to be taking into consideration the substantial size of a potential collection resulting from content archived from the web, clarifying methods to be adopted for this archiving endeavour:

> The regulations make it mandatory that the delivery of the work in response to the request must be by way of automated response from the website to the web harvester. (*Guidance on the Legal Deposit Libraries (Non-Print Works) Regulation 2013, section 3.3*)

The national legal deposit libraries are, therefore, allowed the use of automated web harvesting software to collect online material that is in scope, easing in this way the workload required for such archiving effort (Arnold-Stratford & Ovenden, 2020).

---

is available here:
https://web.archive.org/web/20230803193640/https://www.gov.uk/government/publications/guidance-on-the-legal-deposit-libraries-non-print-works-regulations-2013

[65] The Legal Deposit Libraries (Non-Print Works) regulations 2013 no. 777, part 3, regulation 13. Available at:
https://web.archive.org/web/20240828144924/https://www.legislation.gov.uk/uksi/2013/777/regulation/13/made

4.3 Technical Framework and Frequency of Capture

At the British Library web and social media archiving is done with the Internet Archive's Heritrix (version 3) software,[66] which is made available under a free software licence. Heritrix was first released in 2002 and was specifically developed to capture content from the World Wide Web. This web crawler can capture archival copies of a high number of websites and subsequently ingest them into the UKWA (Rossi et al., 2022). Content captured through Heritrix is stored as a Web ARChive (WARC) file format which has been recognised as an ISO[67] standard for the preservation of web archived content since 2009. The WARC file format can store and aggregate multiple archived digital resources and related information into a single archive file, instead of having to manage numerous smaller files. Developed from a revision of the ARC file format that was used by the Internet Archive since it first began its harvesting activities, the WARC file allows for the inclusion of additional information related to the digital resource archived, such as assigned metadata.

Curators at the British Library have also been testing alternative tools to Heritrix that could archive content from social media more effectively and with a higher level of fidelity. During the 2019 General Election, for instance, the BL experimented with the Rhizome service called Webrecorder[68] for the capture of content from selected social media accounts. Webrecorder is an open-source web archiving tool that can create an interactive copy of web pages that are being browsed, including content that is traditionally more difficult to capture, for example, dynamic content and interactions such as clicking on "read more" buttons, play videos or scrolling a page.

Ultimately, the frequency with which the BL captures social media is mostly decided based on the technical setup of social media sites and web crawlers. For social media content archived using Heritrix, the British Library usually sets up the frequency on a daily basis so that a snapshot of in-scope accounts can be harvested at least once a

---

[66] The name *Heritrix* comes from the Latin word "*heres*, *heredis*" which means "heiress" because, as explained by its creators, the intent of this crawler is "*to collect and preserve the digital artifacts of our culture for the benefit of future researchers and generations*". See *Introduction* section available here: https://web.archive.org/web/20231111232745/https://github.com/internetarchive/heritrix3. See also section 2.6.

[67] ISO 28500:2017 - Information and documentation — WARC file format, available here: https://web.archive.org/web/20230314232822/https://www.iso.org/standard/68004.html

[68] https://web.archive.org/web/20231211202322/https://conifer.rhizome.org/

day. Challenges related to the implementation, usage, and fidelity of Heritrix and Webrecorder experienced by the British Library web archiving team will be discussed in the following section.

4.4 Challenges

This section includes examples of the challenges encountered by the British Library web archive team from a legal, curatorial, ethical and technical perspective. It also discusses solutions implemented to tackle the identified issues or concerns and practices established by the British Library regarding social media archiving. As emerged from the interview, the main concerns and challenges gravitate towards certain limitations imposed by the legal framework - which also affects the type of content that is currently being archived - ethical concerns, technical issues in terms of scalability and fidelity of captures, and the need to find a balance between resources available and those actually required for such an archival effort.

4.4.1 Legal Issues

As discussed in sections 2.4.2 and 2.4.5, the blurred boundaries between the concepts of "public" and "private" on social media, along with the ethical concerns arising from users potentially not being fully aware of how social platforms utilise the data generated through their interactions, complicate archiving practices. I previously mentioned how, for example, the (UK)GDPR regulation includes specific exemptions for "archiving in the public interest". Still, memory institutions archiving material from social platforms like the British Library have to take into consideration these intertwined legal and ethical concerns when selecting content for harvesting. As the BL Lead Curator pointed out during the interview:

> Because of the nature of social media, we don't know what the person who's posting their content to their social media account intends by doing that. Do they think that the tweet is going to go out on the Internet to be widely seen… is it to be considered published? Or do they consider social media more like a communication tool, a messaging service? […] They don't intend that the National Library is going to come and archive this content and preserve it, so the law is not very helpful in this respect.

Similar concerns leave librarians to act to the best of their knowledge to find a way to balance both legal and ethical issues. When it comes to social media, it is extremely

challenging to determine the intention of individual social media users, particularly in discerning whether posting content without any privacy restrictions on their account aligns with a clear understanding of the platform's terms of use. As Bingham observed: "It's difficult with social media, isn't it? Because it's a grey area between what's off private and what's considered published and for that reason, we always review social media accounts before we archive them". In addition, concerns related to the "right to be forgotten" (see Section 2.4.2) clearly amplify the legal and ethical considerations regarding whether to proceed with the collection of private individuals' content. However, bits of conversations or thoughts shared on social platforms by private individuals hold an enduring cultural value, which is likely to increase in the future as it may help shed light on contemporary times. These considerations add an extra stratum of elements for librarians to reflect upon when selecting material from social media. Nevertheless, the absence of up-to-date and standardised guidelines specifically targeting social media collection practices, to which librarians and archivists can refer when in doubt, is currently limiting the extent and the granularity of collections.

A complex layer of restrictions stemming from the Non-Print Legal Deposit regulation 2013 further complicates the archiving of social media at the BL. As Bingham observed, the existence of regulations that enable web and social media archiving at a national level creates the necessary condition for memory institutions to lawfully collect such material without incurring sanctions. However, the existing legal framework paradoxically imposes a series of constraints that have profound implications on the composition of social media collections archived under NPLD legislation and on the accessibility of the archived material to the wider public. In fact, in the UK, the legal deposit framework strictly regulates the type of content that the British Library, as the lead legal deposit institution, is mandated to collect. As illustrated in section 4.2, the parameters set by the UK legal deposit regulation 2013 limit the collection of material published on the web to content that is UK-related or sites that are part of the UK top-level domain. However, given the "transglobal nature of social media and the web", as the BL Lead Curator noted during the interview, it often comes down to curators to decide what constitutes UK content on social media. In fact, if establishing with certainty whether a website pertains to the UK top Level domain – meaning URLs comprising extensions such as .UK, .SCOT, .WALES, .LONDON[69] – and therefore whether sites

[69] UKWA, FAQs: "How are websites selected?". Available here: https://web.archive.org/web/20220301013122/https://www.webarchive.org.uk/en/ukwa/info/faq/

are in scope is rather straightforward, for social media it is quite the opposite (Hockx-Yu, 2014). The majority of social platforms have a .COM domain name, with most of them storing data in centres located, for instance, in the US, the Republic of Ireland or Germany. As observed by Hockx-Yu (2014), the most popular social media sites such as Facebook, Instagram, Twitter and YouTube (Statista.com, 2023b), all use a .COM domain name and therefore do not match the NPLD territoriality criteria. Fortunately, the NPLD allows for the inclusion of content made publicly available on the web by UK-based individuals or organisations (Hockx-Yu, 2014, 2015).

However, ascertaining provenance of content or accounts in the context of social media at scale can still prove to be rather challenging, particularly since geographical information provided by platforms may not be completely reliable (Graham et al., 2014; see also section 2.5). For this reason, as Bingham explained, the BL has opted to manually assess and select social media material:

> We can't automatically say that [social media platforms with a .COM domain name] are in the UK, and therefore are in scope. So, what we have to do is that a curator, a human, has to actually look at social media account and say, "yes I can see that this is the social media account of this UK person, of this UK individual" and curators have to add an extra note to declare that that's the case, so this means that we can't archive social media at scale.

The manual selection and the extra steps required to determine and declare the pertinence of material to the territoriality requirement mean that, thus far, the British Library is not able to archive social media at scale. Hence, they can only archive a limited number of Twitter accounts.

Moreover, the NPLD regulation explicitly excludes from the legal deposit collection criteria content that consists exclusively or predominantly of a film, sound recording, or both, allowing collection of audio-visual material only in the case where this is embedded in predominantly text-based material. The reason for this can be perhaps found partly in the historical origins of the legal deposit legislation itself, which traditionally focused on preserving printed text-based works and was later extended to include born-digital, mainly text-based published materials, as explained in section 4.2. Gibby & Brazier (2012) further suggested that the exclusion of film and sound recordings was likely due to the unforeseen changes and explosion of born-digital mixed media content online after the implementation of the 2003 Act, which the Non-Print Legal Deposit legislation did not anticipate at the time of its drafting. Consequently, social media

such as YouTube or Spotify whose content is indeed predominantly in the form of videos and sound have been excluded from the UKWA (Hockx-Yu, 2014). Building on this, it can also be inferred that, if the BL were to consider adding other platforms to their collections, social media like TikTok, which primarily focus on the sharing of short-form videos, could be potentially excluded from the UK Web Archive as well. It is also worth noting that the legislation in force does not specifically address the language of the content archived. As will be discussed in section 6.2.1, this parameter would have presented several issues for the British Library, given the ubiquity and wide usage of English among individuals located outside the UK territory.

The UK legal framework has important repercussions on the way content is made available to the wider public. Milligan (2015) described at length the extent to which the restrictions imposed by the regulations currently in place affect access to the UKWA. The NPLD appears indeed to be perpetuating mechanisms and restrictions that had been established with the consultation of physical printed works in mind. For physical copies, the BL receives, for example, a copy of all the books published in the UK and, when a book is requested for consultation, only one library user at a time can consult the requested book where only one copy is available in the British Library repositories (Milligan, 2015). Likewise, based on the NPLD provisions, the British Library and its partners cannot provide access to the same archived webpage to two users at the same time for items in the UKWA (Milligan, 2015). Indeed, the "NPLD Guidance" (2013) made available by the Department of Culture, Media and Sport provides:

> Each deposit library may only display the same non-print work at one computer terminal at any one time (regulation 23).

One of the solutions the British Library and its UKWA partners have adopted to mitigate this issue is to make available multiple captures for the same webpage archived at different moments in time (but close to the version of the website the user seeks to consult) to circumvent the restriction. Again, this is a limitation resulting from the current legislation. In order to extend access to the wider public and make all archived web pages or social media accounts freely available on their website, the deposit libraries would need to seek the explicit permission of the single content owner. While this is still a difficult and long process that requires time and resources even for a single webpage, it clearly becomes a virtually impossible task to achieve in the context of social media, due to the sheer amount and stratified structure of its content and ownership (see section 2.4.1).

The few tweets that can be viewed online through the UKWA portal are mostly from public organisations. However, a quick look at the URLs reveals that most of those tweets come from social media feeds embedded in websites other than the original social platform. This is exemplified by the "Go London competition", an initiative organised in 2010 by the UK National Health Service (NHS) in partnership with the Greater London Authority and Transport for London in order to get Londoners to be more physically active. The limited number of tweets accessible on the UKWA website appear not to have been captured directly from twitter.com, but from the "Go London" Social Media Chatter website section.[70] Interestingly, in this way the UKWA has managed to save tweets published in 2013 that are not available anymore on the @GoLondon Twitter account, whose latest tweets appear instead to date back only to 2010. Conversely, legal deposit resources archived directly from twitter.com and other social media can only be accessed on-site through the terminals available in the reading rooms of one of the six legal deposit Libraries, unless the UKWA has obtained permission from the content owner (see also section 4.4.4). The following section will analyse some of the effects that the legal framework has on the selection of material to be archived, the practices implemented to ensure representativeness of the collection and additional ethical concerns that emerged from the interview.

4.4.2 Selection Practices and Ethical Concerns

As illustrated in the previous section, national legal frameworks have a profound impact on the shape social media collections will ultimately take. Once the scope has been delimited to capture social media material related to the UK or created and publicly shared by UK-based individuals or organisations, there is still an enormous number of born-digital items from which librarians can select. Often technical issues and difficulties prevent archiving activities from certain platforms to the point of completely excluding them from collections, in order to avoid unnecessary waste of resources and curators' time. For instance, as mentioned in the introductory section of this chapter (see Section 4.1), one of the first attempts at archiving social platforms at the British Library involved content from Facebook for the BL special collection dedicated to the General Election of 2010. Facebook is one of the most complex social sites to archive at the time this thesis

[70] https://www.webarchive.org.uk/wayback/archive/20130414062719/http://go.london.nhs.uk/social-media-chatter/author/admin/

is being written, mostly due the limitations imposed on automated harvesting by its owner, Meta Platforms Inc, as reported in previous studies (Hockx-Yu, 2014; Vlassenroot et al., 2021). While the first attempts at archiving this platform were considered more or less satisfactory in 2010, Facebook has since become increasingly difficult to collect content from over the years, compromising archiving activities on this platform. In this regard, Bingham observed that "there was a period of time where Facebook wasn't as locked down as it is now, and we were able to archive at least a little bit of that". In an update published on the UKWA blog in January 2018,[71] curators explained that the last complete capture of content from Facebook dated back to mid 2015. After that date, and that is after the Cambridge Analytica scandal (Bruns, 2019), Facebook began requiring users to log in even to view content that had previously been publicly accessible. Since then, the UKWA team is no longer able to archive Facebook accounts. This is partly due to the fact that all they would have been able to capture was the Facebook login page, and partly because the BL only considers in scope content that is public, meaning material that is not behind a login wall (Byrne, 2017). Similarly, attempts at collecting official accounts of public figures or organisations from Flickr or Instagram resulted in poor quality results, so much so that the British Library decided to exclude the platforms from future archiving endeavours. This was due to a combination of technical issues and restrictions imposed by social platforms to prevent crawler access (Byrne, 2017).

The social media platform where the BL has achieved better results, both from a technical and curatorial perspective, is Twitter. As emerged from the survey results discussed in Chapter 3, Twitter is indeed one of the most archived platforms, at least among the surveyed Global North institutions, due to its service provider's more permissive approach, allowing crawlers to collect data from their site "in accordance with the provisions of the robots.txt file" (Twitter, 2022). Furthermore, in Section 2.4.2 it was emphasised how Twitter's purpose "is to serve the *public* conversation" (Twitter, n.d.), which frames the activities and discourses happening on this platform as "public", unless individuals have stated otherwise by setting their account as "private". Thus, Twitter's inherent "public" nature clearly aligns with the BL's collection policies and the NPLD legislation mandate.

The development in 2016 of the first BL policy dedicated specifically to social media archiving laid the foundations for more precise criteria for the selection of content

[71] https://web.archive.org/web/20230306173107/https://blogs.bl.uk/webarchive/2017/04/the-challenges-of-web-archiving-social-media.html

to be included in the UKWA. Nevertheless, given the unfeasibility of determining the UK pertinence of social media platforms based solely on the domain name, as mentioned above, and what is to be (ethically) considered as content "made publicly available", curators have to carefully evaluate social media accounts one by one:

> To address this [challenge] what we do with social media is that we look at individual's social media accounts/profiles on a one-to-one basis. We don't automatically point the crawler at Twitter or any other social media account because we have to decide if it's UK in scope or if it is or not a […] private account.

Having to assess whether each of the prospective accounts of interest matches the BL selection criteria has a profound impact on the number of profiles that can be assessed by single human curators. In fact, curators' daily workload certainly does not involve only appraisal tasks, and the resources currently available to capture the ever-increasing amount of content on social media are not able to cover the sheer amount of work that this activity would require. For this reason, the quantity of the social media material currently collected by the British Library is limited, especially if compared to the number of websites archived and made available on the UKWA. Moreover, the risk of capturing individuals' accounts and consequently infringing their right to privacy, coupled with the impracticality of assessing the provenance of every single Tweet and obtaining individuals' permission to archive posts at scale, likely influenced the British Library's decision to exclude comments, retweets and content collected through hashtags from the scope of the UKWA. In this regard, Bingham commented:

> We archive Twitter accounts, but we don't archive hashtags because for a hashtag we can't say with certainty the provenance of the tweets, so they are probably international in scope, except in a couple of exceptional cases. For example, we archived the hashtag #Brexit [… as] we could say that #Brexit was pretty much related to the UK.

On social media, and particularly on Twitter, hashtags have been used to mark topics in a post and to help discover events or discussions that users might want to follow or participate in (A. Cui et al., 2012). Scholarship has highlighted the additional function associated with the distinctive and meaningful use of hashtags in the formation of, and to signal the affiliation to, online communities (Alfano et al., 2022) as well as in the construction of ad hoc publics (Bruns & Burgess, 2011). When using hashtags, users

participate in an on-going global discussion where it would be extremely difficult to filter tweets coming from a single country without accidentally incurring the risk of archiving content that originates elsewhere. Thus, the British Library only archives hashtags for which it can be established with certainty that they are related to the UK. However, when building collections around similar topics with ample, often politicised resonance, it is important to consider factors such as international interference, bots and the presence of fake British accounts. Any potential biases should be transparently disclosed in the collection description or documentation. Nevertheless, the case of the hashtag #Brexit, mentioned by Bingham, constitutes an appropriate example of selection carried out within the parameters set by the NPLD legislation.

As explained in section 3.4, there are numerous concerns regarding the degree of representativeness of national collections due to the abundance of material shared on social media, along with the legal constraints and technical difficulties surrounding social media archiving practices. Still, appraisal and selection practices implemented by memory institutions can help mitigate these potential imbalances within national collections. As far as concerns social media archiving at the British Library, curated special collections included in the UKWA often contribute to the harvesting and inclusion of social media content that would not otherwise be archived. For instance, it has been mentioned how the BL does not include Facebook in its collection development policy because of the difficulties of capturing content from this site. Nevertheless, certain relevant Facebook accounts are still being documented – although not archived – as part of some of the thematic collections in an attempt to provide adequate context and at least traces of information about material that could not be technically captured. As the BL Lead Curator for the web archive explained:

> We don't include Facebook in our collection development policies because it's a bit of a waste of resources, but […] the curators that are building special collections will often record Facebook accounts that they had wanted to include but we couldn't actually include. Because we think that that information could be useful for researchers further down the line who might want to understand what we had intended to crawl but didn't actually crawl.

The UKWA has around 137 collections that can be browsed on the web archive portal under the title "Topic and Themes":[72] most of the content from social platforms such as Facebook (up until 2015) or Twitter can, however, only be viewed on site. Among the thematic collections available on the UKWA website, there are different topics that frame events, aspects of life in the UK and national culture, such as "Politics & Government", "Science, Technology and Medicine" and "Society and Communities". As per the UKWA description, "Topics and Themes are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field".[73] The British Library can indeed count on a large number of library staff from various departments. This, combined with the contribution of external researchers and collaborators who curate special collections for the UKWA, enables active suggestions for the inclusion of public social media accounts belonging to organisations or public figures like journalists or politicians in the archive.

In addressing the representativeness of collections within the UKWA, both internal and external curators have strived to develop collections that portray the diversity and cultural history of the various regions of the United Kingdom. Over the years, the British Library has taken into serious consideration concerns surrounding collections' representativeness. In this regard, the BL strategy was refreshed in 2020 in response to *Black Lives Matter* and other anti-racism movements that were making their voice heard all over the globe during that year, mainly through social media. Such recent events have certainly accelerated the development of specific policies to ensure that the British Library's collections and, in this specific case, web archive collections, are as representative as possible of all the layers that constitute UK society and culture:

> There's also a lot of work that the library is doing about decolonising collections as well as staff are targeted by the library to be more inclusive, to include more ethnic diversity in our collections, to be more representative across regions in the UK so that we represent all areas of the UK [so that] we're not just focused on metropolitan areas or the South-east.

During the interview the BL Lead Curator for the web archive offered a few examples of approaches implemented for the development of more inclusive UKWA special

[72] UKWA, Topics and Themes webpage: https://web.archive.org/web/20220416193839/https://www.webarchive.org.uk/en/ukwa/category

[73] ibidem

collections. One of these is to seek the collaboration of curators who identify as being within specific groups or communities that the British Library wishes to include, or upon suggestion from external researchers. This participatory approach is exemplified by the special collection dedicated to the LGBTQ+ community which has been in development for a few years now. The BL Lead Curator noted how:

> For our LGBTQ+ collections we had curators who were also representatives of those communities but who were qualified information professionals at the same time and that worked really well because those curators had a bit more of a relationship with people in those communities, they had networks, they had contacts but they were also kind of in the best place to say what sort of subsections, what we should talk about, what voices we should represent in the collections.

Collaborating with curators who are both qualified information professionals and representatives of specific communities can considerably help to mitigate potential biases and create collections that reflects the plurality of voices and aspects of certain groups of society. Social media and its mode of functioning can indeed further emphasise issues of representation (Lutz, 2022). Algorithm dynamics, the tendency to follow a closed circle of accounts based on personal interests, and the sheer number of various small communities on different social sites, pose significant challenges when selecting content related to major events and broader themes on social media (Milan, 2015). Co-curation projects, such as the one described, can help tackling challenges surrounding representativeness of collections through collaboration with groups or individuals and their networks to identify content that should be included in the web archive (see section 6.2.2).

However, it is worth noting how even co-curated collections are still not immune to criticism by members of the very same communities. An example in this sense is the controversy that arose in regard to the "LGTBQ+ Lives Online" special collection. As Bingham explained:

> Curators wanted to reflect [the] full experience of gay and trans people living in the UK and one of the things that they deliberately wanted to include was incidents of transphobia or homophobia, so they included some organisations that were espousing those kinds of views.

Because of the presence of content and accounts that had transphobic and homophobic views, which had been included to comprehensively record the LGBTQ+ community

experience, the British Library had to deal with a few complaints. Critics questioned the inclusion of potentially harmful content, suggesting that its presence in the web archive could be interpreted as an endorsement of such negative behaviours by the UKWA, prompting calls for its removal. Nonetheless, as the BL Lead Curator highlighted during the interview, researchers and information professionals responsible for curating the special collection argued that excluding such content would fail to reflect the real, full experience of gay and trans people living in the UK. Also, this material would be, unfortunately, essential in terms of preserving and conveying a truthful representation of today's social media experience. In order to reiterate the fact that adding similar, potentially controversial content to the web archive is not to be considered as an endorsement of similar behaviours or views, the British Library and the other five NPLD libraries added a statement to the UKWA Frequent Asked Questions page. The statement declares that inclusion in the UK Web Archive does not imply support from Deposit Libraries for those views and that:

> Websites are reflected for the benefit of future researchers; they are intended to be reflective of contemporary UK life or for their relevance to a particular collection.[74]

Moreover, curators of the LGBTQ+ special collection and the individual who critiqued the inclusion of controversial content all agreed that the collection should be accompanied by a content advisory signalling the presence of material that may cause emotional distress to users. In this context, it is worth mentioning the important work done by the Archive of Tomorrow (AoT)[75] project in exploring and proposing content advisory language. Carried out in partnership with UK libraries and with the support of the British Library, the Archive of Tomorrow project sought to improve access to web archived material for health researchers, by building a pilot collection on public health information within the UKWA. Having to deal with material related to health, the AoT suggested applying content advisories at a sub-collection level to inform users about the potential inclusion of sensitive or harmful content, thus encouraging users to approach and use the collection

[74] UKWA, FAQs: "Is the inclusion in the web Archive an endorsement?". https://web.archive.org/web/20230911181903/https://www.webarchive.org.uk/en/ukwa/info/faq/#is-inclusion-in-the-web-archive-an-endorsement

[75] Further information about the Archive of Tomorrow project can be found here: https://web.archive.org/web/20240306121818/https://www.nls.uk/about-us/working-with-others/archive-of-tomorrow/

with care (Libraries, 2023). Similarly, the BL has included a statement in the description field of the "LGBTQ+" collection sub-section called "Health and Community",[76] which reads:

> This collection contains websites that reflect the full experience of LGBTQ+ people in society today and as such may contain websites that may be considered transphobic or homophobic. It is a principle of our collection development policy, underpinned by legal deposit regulations, to include, uncensored, everything that is published online in the UK for the benefit of future researchers.[77]

Since the description box, both for the general collection and subsections, has a limited number of characters (about 500 in total), the Web Archiving Team has also made available a longer collection scoping document including more details about the development of the LGBTQ+ special collection, in order to be more transparent about the selection decisions taken.

Likewise, the British Library had to face a few other challenges in terms of representativeness for one of its long-standing collections dedicated to Black and Asian communities, culture and the history of the presence of such communities in the UK. In this case, the criticism directed to the UKWA was about the title assigned to the collection "Black and Asian Britain". It was pointed out that this label appeared to not fully reflect the existing multifaceted diversity and the way people belonging to these groups might prefer to identify. While curators noted that the title of this or other similar collections had been assigned with the intent of facilitating the work of researchers and to make finding responses to potential queries easier, the BL web archiving team stated its commitment to work in concert with the BL anti-racism project[78] to revise these labels.

---

[76] The "LGTBQ+ Lives Online", subsection "Health and Community", can be found at this link: https://web.archive.org/web/20221218193114/https://www.webarchive.org.uk/en/ukwa/collection/3087

[77] Ibidem

[78] The Anti-Racism project (ARP) was established at the British Library in 2021 with the aim of "enacting a generational shift" so that the BL could become "a more representative and diverse organisation that is welcoming and empowering for everyone". Further information is available here: https://web.archive.org/web/20230205132255/https://blogs.bl.uk/living-knowledge/2021/03/towards-an-action-plan-on-anti-racism.html
See also: https://blogs.bl.uk/files/bl-race-equality-action-plan-jan-2022.pdf

An additional point that emerged during the interview concerning the inclusivity of the web archive collections is related to the online presence of minority groups. Bingham pointed out how many alternative voices are not adequately represented on websites. The Lead Curator continued by explaining that buying a domain, building a website, and then maintaining it through time requires skills and resources that not many communities possess, especially smaller ones. Hence, many marginalised communities find it easier to create and manage a page, a group, or an account on social media such as Facebook or Twitter.

However, since the UKWA collection development policy excludes from collection material that is shared within closed forums or groups (e.g., Facebook groups that require you to login or receive an invite to join), this means that most of this content cannot be archived by the BL, resulting in a lack of representation of those voices in the web archive collections. Nevertheless, as emphasised in section 2.4.5, a complex array of ethical questions, potential risks or heightened possibilities of harm arise when archiving content generated by members of marginalised groups, requiring careful consideration. In terms of moderating content added to the web archive, Bingham also explained how sometimes the material harvested can include content that, for example, is only suitable for an adult public or content that might be considered offensive. This is due to the difficulty of monitoring all that is crawled from the Internet and the sheer amount of material included in the web archive. Also, defining what is to be considered as "offensive" can be rather challenging, as it might vary through time and from one individual to another. Still, in order to mitigate the issues that may arise from users encountering content perceived as offensive while browsing collections, the UKWA terms and conditions state that by accessing the web archive "[users] acknowledge that the information on the Archive does not reflect our views and opinions".

Finally, in addition to the strategies described above to tackle the challenges related to representativeness of collections, the BL and its Legal Deposit Library partners have initiated the "Save a UK Website" campaign. Recognising the enormous endeavour that archiving the national web domain and, although limited for now, content from social media involves, the BL accepts suggestions about UK websites or published online material that falls under the scope of the NPLD. Although archivists and web curators will always play a central role in the selection and preservation of items to include in the UKWA, similar crowdsourcing initiatives have been revealing themselves as positive

examples of how web archives can engage with the wider public and seek users' help to build more diverse and representative collections (see Section 6.2.2).

4.4.3 Technical Obstacles

The British Library has been archiving the UK web domain for almost two decades now, and the first attempts at collecting this specific type of born-digital material involved the use of tools that were primarily developed for webpages—webpages that were, however, not as dynamic or interactive as social media platforms. The British Library and its Legal Deposit Library partners have been using and experimenting with a wide array of tools in order to best capture social media content.

As mentioned in section 4.3, the main crawling tool used at the BL for both web and social media archiving is Heritrix version 3. Heritrix works well with large-scale, bulk crawling, especially of static pages, but is not designed for high-fidelity captures of content on social media. While Heritrix generates fairly good results with Twitter,[79] maintaining the way tweets look on the live web and providing a rich set of metadata for the archived items, it demonstrated limitations on this and other social platforms. Specifically, it is unable to capture and interact dynamically with the content on a page. For example, it is not able to play videos, scroll pages or click on "read more" buttons. As a result, it only partially collects content, generating gaps in the archived material. In this sense, Bingham noted:

> Heritrix isn't a high-fidelity crawler, it doesn't capture interactions in a browser at all and a lot of social media is set up so that it relies on dynamic loading of the content, for example when the user scrolls down — Heritrix can't capture that kind of interaction.

The ability to capture dynamic content and interactions is essential to harvest content comprehensively and render a realistic representation of how social media looks and functions. Maintaining the "look and feel" is not only important for present researchers but will be essential for future generations who will need to understand how these platforms worked and the experience that people had while using them. Moreover, when archiving content from an in-scope account, the inability to interact with specific objects on the page, in particular buttons that allow the expansion of hidden sections and thus

[79] This statement was true up until 2022 when the interview was conducted. In section 6.1.4 I will discuss recent Twitter changes and how these affected the archivability of the platform.

the visualisation of posts that were published before the crawling process, could lead to extensive gaps in the material preserved. This issue would be especially evident for those accounts that publish large number of tweets or posts on a daily basis (see also Section 5.4.3). The technical barriers arising from the use of Heritrix v3 make archiving social media consistently difficult, which is one of the reasons why these platforms are currently only collected in a very limited amount at the British Library.

To address these gaps and find tools that could complement the work done by Heritrix and archive social media in all of its different layers of interactivity with a higher degree of fidelity, the web archiving team at the British Library has been testing new tools, such as Webrecorder (see Section 4.3). Regarding Webrecorder, the BL Lead Curator explained:

> One of the most promising [tools] is the Webrecorder and it works by launching a browser session in which the user can navigate a website and all of the interactions that they make by clicking around the website are recorded in Webrecorder.

A first, promising experiment using Webrecorder was conducted on the occasion of the UK 2019 General Election. In a blog posted on the UKWA Blog, (Bingham et al., 2020) offered a detailed description of the process implemented and issues faced when using Webrecorder for the capture of social media content during and immediately after the General Election. The day after polling day, the BL web archiving team ran the tool on twelve of their office computers, employing the Webrecorder autopilot function to capture selected accounts across different social media platforms. Although the experiment obtained very good results in terms of fidelity, using Webrecorder was regarded as "very resource intensive" (Bingham et al., 2020). In fact, the Webrecorder's automated function, which automatically clicks and plays links available on the selected page, did not work well on all accounts and a manual capture was required to complete the set of data. Bingham explained that: "The curators were having to sit in real time and navigate through a website so we could only archive a very small number of social media accounts with Webrecorder."

To this must be added issues related to the autopilot working on some computers and not on others, problems with IP addresses and difficulties with getting the app restarted once it crashed (Bingham et al., 2020). Additional complications emerged during the auditing of the WARC files and the quality assurance process. In order to speed up the collection process, the BL decided to run several parallel Webrecorder apps on

multiple computers so that each machine could focus on one social media account. Thanks to the way the British Library IT system works, one BL user logged into several computers running multiple Webrecorder apps at the same time, so that the resulting WARC files could be uploaded to the same user's cloud storage (Bingham et al., 2020). Associating more than one Webrecorder job per user, however, complicated the auditing process. In particular, the problem lay in locating completed WARC files among the many incomplete files produced (Bingham et al., 2020). The lack of resources and the time-consuming nature of having dedicated curators manually operating and ensuring that tools like Webrecorder capture only relevant content pose additional challenges. This becomes especially problematic when dealing with a large volume of content to be recorded, or when the archiving of an event is time sensitive or occurring within a limited timeframe. Moreover, Bingham drew attention to some risks associated with using tools like Webrecorder, including the potential accidental archiving of material that should not be included in the UKWA, such as login details or private information:

> For example, you could archive login information, or subscription information might be recorded. Or you could go too far into a social media account and archive potentially private information like messenger chats. So, we need to have a lot of user training for people that use Webrecorder.

Moreover, Bingham explained that the web and social media archiving workflow is usually automated: once the crawl is complete, the WARC files are automatically written to storage, then indexing processes are run on the content collected to finally make it available from the storage cluster. However, when using Webrecorder, the risks of capturing content that is private or out of scope necessitate the BL web archive team's manually verifying the WARC files resulting from the crawl before ingesting them into the UKWA collection. This obviously adds an extra step which is outside the British Library's standard workflow. The BL Lead Curator explained:

> We have to download those WARC files somewhere, ingest them separately onto our storage, [and] index them. All of those processes have to take place semi-automatically because we haven't engineered those tools to be a part of our workflow as yet and we hope to develop our work force.

Because of the additional steps, time and manual work required to use Webrecorder, combined with the manifold technical issues related to Heritrix's inability to archive social media to a good degree of fidelity, the British Library can only employ these tools on a

limited amount of social media content at present. In terms of quality assurance for content archived with Webrecorder, the process appears to be fairly straightforward as curators can replay the recorded copy using the replay tool included in the Webrecorder app. Bingham also highlighted that Webrecorder generates reports for all the assets of the URLs crawled so that curators can obtain server status codes and check for any missing elements. In case of missing elements, Webrecorder has a function to conduct a patch crawl to capture and integrate those components.

As for Heritrix, due to the sheer amount of items collected using this tool, the limited resources and time available do not allow for quality assurance on each and every social media account archived. Nevertheless, Bingham recounted that, for example, when archiving Twitter accounts with Heritrix, they encountered a technical error that was mirrored across all the accounts collected. The crawler was returning random languages – sometimes French, Turkish or Arabic. This was happening because Heritrix was navigating through Twitter's language drop-down menu and crawling different language versions of the site. In similar cases, they usually adopt a sampling approach to quality assurance, as the technical issue would be consistent across all the Twitter accounts crawled.

4.4.4 Accessing the UKWA

The easiest way to access the UKWA, as Bingham explained, is via the portal "webarchive.co.uk", which is openly accessible. However, only a very limited amount of social media content is available to view through the site as it depends on whether the UKWA has the permission of the content owner to make it publicly available. As illustrated at the beginning of section 4.4.1, the Non-Print Legal Deposit regulation 2013 imposes certain restrictions on the way in which content from the UKWA can be accessed. The Web Archive can be browsed in its full extent only on site at one of the six legal deposit libraries (British Library, National Library of Scotland, National Library of Wales, Trinity College Dublin, University of Oxford and University of Cambridge) plus other locations such as the Kelvin Hall in Glasgow and Cardiff University. Researchers in possession of a reading pass at one of the mentioned libraries can access the reading room where the UKWA terminals are located and explore the full collection. The collection is searchable via URLs (if known) and keywords and can be also browsed by Topics or Themes available in the Special Collections section. Users can also explore the

web archive by using faceted search, for example, filtering content by year or web domain (e.g., co.uk, org.uk, ac.uk, .scot, .wales, etc).

While the location of libraries where the UKWA is accessible appears to be equitably distributed within the United Kingdom territory, with some terminals strategically located within University Libraries, the limited number of institutions providing access to the UKWA still constitutes a barrier to users who do not reside near one of those libraries, cannot physically travel for a variety of reason to a physical location, or lack adequate funding to travel to one of them. This certainly represents an obstacle for PhD students or Early Career Researchers with limited resources, and even more so for people with disability.

A further issue related to access, which is linked to provisions included in the NPLD regulation, is that – contrary to what is allowed in traditional archives – there is limited printing for items included in the UKWA collections, while downloading or any form of reproduction is not permitted (Arnold-Stratford & Ovenden, 2020). As observed by the BL Lead Curator:

> [This restriction] causes a barrier to researchers in that they can't cite that dataset […], they can't easily cite the resource they are looking at and it means that their research is not very shareable because they can't refer to that dataset.

Data citation from web archive resources is still an open problem and it is currently the subject of various studies seeking to find solutions or propose good practices to address this issue (see section 6.4).

The BL can also provide bespoke datasets, but this involves dedicated staff time and collaboration with web archive engineers to obtain the set of data the researcher is looking for, which can often be challenging to organise. However, the access limitations of the UKWA still apply to this type of service, meaning that researchers can only view the data on-site unless it is derived data.[80] Bingham explained:

> We can't pass up any raw data out unless it's derived data. One of the things that we're working on at the moment is that our technical team for the web archive is developing an API which would connect to our curation software,

[80] Ruest et al. (2022) described derivative datasets as a "transformed WARC files, with extracted items of interest. Extraction can range from the plain text of a collection to hyperlink graphs to JavaScript files, or any combination of the above" (pp. 318-319). Derived data can be generated on-demand or pre-generated for download.

> so from that somebody connected to the API would be able to download all of the metadata and would have information about curated websites.

Having access to the API would enable researchers to obtain more details about the items comprising the UKWA collections, including lists of URLs, descriptive metadata, time stamps and information about crawl frequency, and they would also be able to search derived datasets. This would represent an opportunity to facilitate the reuse of datasets. In this regard it is important to mention a recent BL blog post where Andy Jackson (2023), the UKWA Technical Lead between 2012-2023, explained how the BL web archiving team has been working with the Archive of Tomorrow project to meet their requirements and how this has resulted in the creation of an internal API. Through this API "W3ACT[81] metadata can be downloaded for entire collections, including all sub-collections and target site metadata. Authenticated W3ACT users can retrieve these full collection extracts (including unpublished collections), which are updated daily" (Jackson, 2023). According to the latest technical update in January 2023, the UKWA technical team was finalising the development of the public version of the API, which was scheduled to be released in 2023 (Jackson, 2023).

4.4.5 Long-Term Preservation

Material archived from social media and the resulting WARC files are usually treated like any other digital item and they can be subject to some of the same challenges and risks as other digital objects, such as fragility of storage media on which they are saved or replicated, technical obsolescence, changes to the digital object that may threaten its integrity and cyber threats. In terms of long-term preservation, the UKWA has adopted preservation strategies already implemented for other digital items at the British Library. The Digital Library System is the long-term digital repository developed by the British Library with the support of the other UK Legal Deposit Libraries. The Digital Library System is physically located in the British Library premises at Boston Spa in Yorkshire.

[81] The Wikipedia page created by Nicola Bingham, Lead Curator at the BL, describes the Annotation Curation Tool (W3ACT) as a tool "developed to meet the requirements of subject specialists wishing to curate web content harvested under UK Non-print Legal Deposit Legislation. W3ACT enables users to perform numerous curatorial tasks including the assignation of metadata and crawl schedules to web content, quality assurance and the ability to request permission for open access to selected websites". More information about the W3ACT can be found on GitHub: https://web.archive.org/web/20230825142818/https://github.com/ukwa/w3act

Content is replicated across nodes situated in four of the legal deposit libraries: British Library (St. Pancras), British Library (Boston Spa), National Library of Wales and National Library of Scotland. Preservation through replication involves having multiple copies of the same digital information across multiple storage environments. The Digital Library System contains all the digital content preserved by the British Library, including digitised materials, sound and visual content, e-publications and digital items archived resulting from web archiving activities.

Bingham pointed out during the interview the complexity associated with the ingestion process of WARC files into the Digital Library System. This is currently due partly to the large volume of materials included in the web archive, and partly to the backlogs of other content that are waiting to be adjusted and ingested into the main store. Indeed, Bingham noted:

> The web archive content at the moment is only partially ingested into our digital library system, the rest of it is preserved in servers that the UK web archive manages. These are still servers that are based at Boston Spa and are secure servers, content is still replicated or it's in the process to be replicated [at] the National Library of Scotland.[82]

Bingham also explained that the British Library is currently working on redeveloping the Digital Library System and once this project is completed, the UKWA will then be ingested and stored in the new system. Nevertheless, it appears to be a rather ambitious and extensive project that will require some time before it is completely finalised and fully implemented.

4.5 Future Developments

As the web and social media platforms continue to evolve and new ones emerge, the British Library is looking to find the best solutions to ensure the preservation of this resource. From a technical perspective, the British Library is planning to trial the Browsertrix[83] crawler. Browsertrix is a browser-based high-fidelity crawling tool that is able to capture and store an interactive, complex copy of websites that other tools are not able to adequately capture (LeBlanc et al., 2022). As also mentioned in section 2.6, the

[82] Replication of the UKWA to the NLS was achieved in September 2023.

[83] https://web.archive.org/web/20230918154115/https://github.com/webrecorder/browsertrix-crawler

IIPC, in collaboration with memory institutions such as the British Library, the Royal Danish Library, National Library of New Zealand and University of North Texas, and the Webrecorder.net developer Ilya Kreymer, have set out a project called "Browser-Based Crawling System For All"[84] whose aim is to support the creation of a high-fidelity crawling system driven by a full-featured user interface accessible to curators and web archivists at different institutions. This crawling system is being built on the existing Browsertrix Crawler and should provide a user-friendly interface that could complement collection activities made using Heritrix. Bingham explained:

> Browsertrix will do browser crawling and will automatically load elements that are dropped into the browser that are needed to render a web page. It will record all of those interactions but it's a little bit more automated than Webrecorder is—so we are going to be having to look at that and hopefully that will be positive for social media archiving.

The BL Lead Curator mentioned that the UKWA team is also following other interesting approaches in this area, such as those implementing the use of Social Feed Manager,[85] or APIs to harvest content from social platforms. However, Bingham pointed out that most of these tools are not suitable for a legal deposit environment, hindering the potential growth of the social media collections at the UKWA.

Conclusion

The British Library, as the main contributor to the UK Web Archive, has a legal mandate to select, archive, preserve and provide access to material made publicly available online, including websites and social media related to the United Kingdom. Since the first attempts at capturing social media, the inclusion of these platforms in web collections has challenged the UKWA's well-established archiving strategies and workflows. The combination of constraints determined by the national legal framework, ethical concerns and technical challenges has profoundly shaped the extent of the social media collection

[84] More information about the IIPC project related to the development of the Browsertrix Crawling tool can be found here:
https://web.archive.org/web/20230927030735/https://netpreserve.org/projects/browser-based-crawling/#abstract

[85] Social Feed Manager: https://web.archive.org/web/20231103003430/https://gwu-libraries.github.io/sfm-ui/

activities and the type of platforms preserved to date at the UKWA. Despite earlier positive archiving tests performed on Facebook and Flickr, the ever-changing nature of social media, unanticipated changes in platform terms and conditions conflicting with collection boundaries established by national legislation, and poor harvesting quality results have prompted the BL to exclude most of the existing platforms from its web collections until more sustainable archiving conditions are met, and to instead focus preservation efforts on a limited number of Twitter accounts.

Despite the exclusion of certain popular platforms, specific accounts that are deemed important for representativeness reasons are still recorded at a (sub)collection level to ensure that at least a trace of that information is preserved. The BL has implemented various participatory strategies aiming to mitigate representativeness concerns, including co-curation practices and crowdsourcing campaigns, which have proved to be essential to identify born-digital cultural heritage online that could adequately represent the multitude of realities existing within the national territory. Because of known technical issues faced when archiving social platforms using Heritrix, such as the inability to interact with dynamic content, the BL has explored alternative web archiving tools, with Webrecorder being one of the tested options. Although capable of higher-fidelity captures compared to traditional web crawlers, Webrecorder has proven to be rather resource-intensive. Still, archiving social media remains primarily a manual process, requiring archivists to meticulously select accounts and content to align with the parameters established by the legislation in force and which can be feasibly captured using available resources and tools.

The next chapter will delve into challenges encountered and social media archiving practices developed at the two French memory institutions involved in the preservation of the national digital cultural heritage under the French legal deposit legislation: the *Bibliothèque nationale de France* (BnF) and the *Institut national de l'Audiovisuel* (INA).

CHAPTER FIVE

Archiving Social Media Under Electronic Legal Deposit in France: A Case Study of the *Bibliothèque nationale de France* (BnF) And the *Institut national de l'Audiovisuel* (INA)

This chapter includes the second of the two case studies that aim to shed light on archiving practices and the barriers that arise while collecting content from social media platforms. The second case study, as explained in Section 1.4.2, focuses on the challenges, solutions and workflows implemented at the two institutions appointed by the French legal deposit regulation to preserve the national digital heritage (including social media), that is the *Bibliothèque nationale de France* (BnF) and the *Institut national de l'Audiovisuel* (INA). The first section of this chapter offers an overview of the web and social media archiving activities at the BnF and INA and illustrates how these two memory institutions began to develop their social media collections. This is followed by a brief summary of the legal framework in which the archiving activities of the two legal deposit institutions are embedded, discussing developments in and specific aspects of the French electronic legal deposit legislation and how these influence appraisal, preservation and access to web and social media content at the BnF and INA. The chapter also illustrates the technical framework implemented in both institutions, highlighting different approaches and tools adopted by the BnF and INA. The main portion of the chapter is dedicated to the various challenges faced by both web archiving teams when collecting and preserving material from social platforms, providing examples (where available), and illustrating solutions, workflows and practices implemented by web curators at the BnF and INA. The final section addresses potential future developments, work in progress and points of interest that need to be taken into consideration in order to allow social media collections at these two national archiving institutions to advance.

The comments and observations included in this case study are based on interviews and fieldwork carried out in Paris between 19 and 21 April 2022. Interviews were conducted with Vladimir Tybin (Head of Digital Legal Deposit) at the BnF, and with Claude Mussou (Head of Service) at INA. The fieldwork included two visits to the Bibliothèque François Mitterrand in Paris for both the BnF and INA, and one visit to the training centre in Bry-sur-Marne, near Paris, where INA's Research and Development team is located (see Sections 1.4.2 and 1.4.4). It is worth underscoring here the value of

having had the chance to visit and meet in person members of both institutions. The significance of fieldwork, particularly in the context of the French case study discussed in this chapter, resides in the insights I gained from informal conversations with various members of both INA and the BnF. These informal exchanges shed light on dynamics and workflows that a one-hour interview alone would not have allowed me to capture. Moreover, visiting their offices provided me with a deeper understanding of the extensive human resources involved in the development of each social media collection, especially when compared to other smaller national archiving initiatives. Most importantly, I had the opportunity to access at first-hand both web and social media archives through the terminals located in the Paris building, which helped me understand the impact that certain barriers and technical issues have on the preparation of the archived material for researchers. Finally, this chapter also includes information gathered through documentation available on the BnF and INA websites, previous interviews (e.g., WARCnet Papers series), blogs and papers published by their curators who actively participate in the international discourse surrounding web and social media archiving.

5.1 Overview of the Projects

In France, the responsibility to preserve digital material deemed of historical interest based on the electronic legal deposit legislation is equally distributed between the BnF and INA, respecting the continuity of their existing mandates and collections (Stirling et al., 2012).

Bibliothèque nationale de France (BnF)

First established in 1537 by François 1st, the *Bibliothèque nationale de France*'s mission, as stipulated in the 1994 decree (Décret n.94-3, 3 January 1994),[86] is to "collect, preserve, enrich and communicate the national documentary heritage".[87] According to the *Code du Patrimoine*, which incorporated in 2006 the so-called DADVSI law (see Section 5.2), the scope of the BnF was expanded to include the collection and preservation of various types of heritage material published on the Web. As the number of internet users and websites registered in France started to grow considerably in the late 1990s, action was taken to

[86] *Décret n°94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France*: https://web.archive.org/web/20240302072925/https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000545891

[87] https://web.archive.org/web/20231004134954/https://www.bnf.fr/fr/les-missions-de-la-bnf

include this means of communication and new type of publication in the national preservation strategies (Aubry, 2010).

The BnF started experimenting with web archiving around 2002. The first attempt at capturing content from the web was organised on the occasion of the 2002 presidential elections, an event that for its nation-wide significance and significant repercussions on French history and society was deemed important to be preserved for future generations. Since this first test, the BnF has covered almost 19 elections up until 2022. The special web collection preserving digital traces of the past 20 years' presidential elections on websites and social media constitutes one of the largest within the BnF's *Archives de l'Internet*. The web archive collections preserved at the BnF also include websites dating back to the early days of the development of the French web. The earliest websites from 1996, which the web curators at the BnF refer to as *incunabula* creating an interesting parallelism with the first books printed during the earliest period of typography, were acquired through an agreement with the Internet Archive,[88] one of the first initiatives to start archiving the web with the goal of preserving the whole web.

In terms of social media archiving activities, the team in charge of the *dépôt numérique* [electronic legal deposit] at the BnF began monitoring the evolution of these platforms since their first appearance around 2006. Shortly after, social media started to be included in the BnF's collection development policy, given the central role these new media were increasingly playing in French daily communications and society. The French web archive at the BnF preserves content from social platforms such as YouTube, Twitter, Instagram and Dailymotion. In May 2022, the BnF also launched its first crawl for content related to the 2022 presidential election campaign on TikTok. The BnF used to archive content from Facebook (2007-2020), but due to limits imposed by Meta Platforms Inc., crawling activities are currently suspended. Nevertheless, the web archiving team is looking into testing new archiving approaches, including capturing Facebook's mobile version (see Section 5.5).

The BnF collection development policy[89] distinguishes between *Collectes courantes* [On-going collections], which include thematic collections for each of the BnF

[88] The practice of requesting from the Internet Archive web pages from the early years of the web appears to be largely diffused among national web archiving institutions (see for example Aubry, 2010, p. 183).

[89] *Charte documentaire - Politique d'enrichissement des collections* available here: https://web.archive.org/web/20230908000733/https://multimedia-ext.bnf.fr/pdf/charte_doc_integrale.pdf

departments and regional collections, and *Collectes projet* [Event-based collections]. The latter type includes content related, for example, to special events like French election campaigns, the Olympic Games, and other cross-disciplinary collections concerning, for example, topics like environmental issues and Artificial Intelligence. The BnF has incorporated in these thematic and event-based collections audiovisual material captured from YouTube (since 2017), Dailymotion (2007-2013) and podcasts (since 2023), and social networks like Instagram (since 2020) and TikTok (since 2022).[90]

The BnF also has an "emergency collection" procedure in place for quickly capturing crises (e.g., terrorist attacks and the Covid-19 pandemic) and sites that are scheduled to be collected for a specific date (as in the case of trade fair or festival sites) or are at risk of disappearing. The latter case is exemplified by the BnF's action in 2021 to crawl content using Adobe Flash after its End of Life (EOL) announcement;[91] and the archiving effort undertaken to preserve blogs from *Le Monde*[92] or *Libération*[93] when their closure was announced respectively in 2019 and 2020, and, in collaboration with INA, still-active blogs on the *Skyblog*[94] platform before it was shut down in 2023.

In terms of frequency of harvesting, Twitter used to be crawled multiple times a day up until July 2023 (see section 5.4.3), whereas YouTube is harvested around two or

---

[90] At the time of the interview (April 2022) the first TikTok crawl was in the process of being launched by the BnF.

[91] Adobe stopped supporting Flash Player after December 2020 and blocked Flash content from running in Flash Player beginning 12 January 2021. Adobe Flash Player EOL: https://web.archive.org/web/20240226065939/https://www.adobe.com/uk/products/flashplayer/end-of-life.html

[92] *Le Monde* announcement, "La fin annoncée des blogs abonnées du Monde.fr, la fin du blog paysages sur les blogs leMonde.fr", available here: https://web.archive.org/web/20190504204009/http://cneffpaysages.blog.lemonde.fr/2019/04/14/la-fin-annoncee-des-blogs-abonnees-du-monde-fr-la-fin-du-blog-paysages-sur-les-blogs-lemonde-fr/

[93] *Libération* announcement, "Blog «Géographies en mouvement» Notre dernier post sur le blog, merci Libé", available here: https://web.archive.org/web/20230825202921/https://www.liberation.fr/debats/2020/12/01/notre-dernier-post-sur-le-blog-merci-libe_1810989/

[94] *Skyblog* is a French blogging platform, a precursor of social networks in France, hosted by the French radio station *Skyrock*. More information about the *Skyblog* collection and the collaborative effort of the BnF and INA to preserve over 12 million still-active blogs is available here: https://web.archive.org/web/20240407212055/https://www.bnf.fr/fr/la-bibliotheque-nationale-archive-les-skyblogs; https://web.archive.org/web/20230924171634/https://www.ina.fr/ina-eclaire-actu/internet-skyblog-fermeture-reseaux-sociaux-annees-2000.

three times a year. As for Instagram, the BnF launches between four and six crawls per year, with each crawl including approximately 300 Instagram accounts and covering specific topics or themes. In addition, extra captures are taken during unexpected events or crises that require the BnF to archive content outside the scheduled crawls.

The BnF's web archiving team consists of seven members, including curators responsible for the selection of content, harvesting process, preservation, access, and support for researchers. Together, they ensure the long-term preservation of significant material from the French web.

Institut national de l'audiovisuel (INA)

Created in 1975, INA's mission is to archive "French audiovisual heritage, creating content, researching, and transferring knowledge in the audiovisual and digital fields".[95] The Institute currently preserves more than 75 years of television programmes produced on public channels since the Second World War, and 85 years of radio transmissions produced both by public and private radio stations since 1930 and inherited by INA when the Institute was established. INA also preserves newsreel, which was broadcast in cinemas between 1940 and 1969, photographic collections inherited from public audiovisual companies, written documents and journals related to French broadcasters, institutions, and other professionals in the audiovisual field.

As stated on its website,[86] INA is responsible for the electronic legal deposit of material created and shared on the French web in relation to audiovisual communication, a mission that the Institute shares with the BnF. Since the collection scope of the Institute was extended to include this important source, INA has been archiving an ever-growing assortment of websites and platforms. Its collection includes websites of tv channels, programmes, TV on-demand, and other platforms such as blogs about tv or radio shows. It also archives tweets that contain links to tv shows or are related to what has been broadcast on television, such as news and events happening worldwide. As of 2023, INA's archives comprise more than 16,000 websites, plus over 15,000 accounts and 2,500 hashtags archived from various social media platforms.[96]

---

[95] See the "Our Missions" section, available at: https://web.archive.org/web/20240228222702/https://www.ina.fr/institut-national-audiovisuel/international-affairs

[96] The *Web media*: https://web.archive.org/web/20240118033437/https://www.ina.fr/institut-national-audiovisuel/collections-audiovisuelles/le-web-media

5.2 Legal Framework

The legal mandate to archive and preserve electronic publications, including digital content created and shared on the web and social media in France, has been given to the BnF and INA following the publication of the 2006 law on "Authors' Rights and Related Rights in the Information Society", known in French by the acronym DADVSI[97] (as in *Droit d'Auteur et Droit Voisins dans la Société de l'Information*). This legislation and the following implementing decrees introduced electronic legal deposit in France, clarifying the mission of the institutions already entrusted with legal deposit activities, and finally including in the BnF and INA's collection scope material published on the web.

Stirling et al. (2012) offered a detailed overview of the history of legal deposit in France. Legal deposit was first introduced on 28 December 1537 by François 1st as part of the *Ordonnance de Montpellier*. The text instituted the *Librairie royale*, a royal library that had to contain and preserve all the works "worthy to be seen" that had been and would be published, with the aim of preserving the national cultural heritage through time (Stirling et al., 2012). The French legal deposit legislation has been refined and progressively developed over the centuries in order to include different types of publications and formats. Similarly to the UK (see section 4.2), in France legal deposit was extensively used between the late 1700s and early 1900s as a means to safeguard copyright (Stirling et al., 2012). The tight bond between legal deposit of printed works and the acknowledgment of intellectual property was modified in France in 1925, when the French Copyright law, known as the *Code de la Propriété Intellectuelle*, was introduced establishing that copyright was to be considered as inherent to published works (Stirling et al., 2012). Henceforth, legal deposit became primarily linked to the preservation of cultural heritage.

The *Code du Patrimoine* [Code of Cultural Heritage] is the main text that contains the legislation related to the French cultural heritage, and, after its revision in 1992, regulates all activities related to the legal deposit. With the evolution of technologies and the increasing popularity of new media, a number of items in different formats have been added to the *Code du Patrimoine* through time. However, it was only in 2006 that the legal deposit was extended to include material published on the web. The electronic legal

[97] Loi n° 2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information (DADVSI). Available here: https://web.archive.org/web/20240117191812/https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000266350/

deposit was indeed established that year following the incorporation of Title IV (articles 39-47) of the "Authors' Rights and Related Rights in the Information Society" law (DADVSI, 2006-961) into the *Code du Patrimoine* (articles from L131-1 to L133-1). DADVSI brought the Directive 2001/29/CE[98] on the harmonisation of certain aspects of copyright and related rights in the information society into French law. For this reason, as Stirling et al. (2012) observed, it presents similarities with legislation implemented in other European countries (see Section 6.1.1).

In Sections 2.4.3 and 3.5 I discussed the importance of e-legal deposit legislation offering a broad definition of what is to be considered as "published electronic material". Given the rapid and often unforeseeable evolution of digital formats, a loose definition allows for the inclusion of any type of digital publication, encompassing all present and future developments. This allows memory institutions to start experimenting and quickly target new types of online publications as they emerge, placing them under the safeguard of national libraries and other national institutions because of their cultural significance.

The 2006 DADVSI introduced a new paragraph in article L131-2 of the *Code du Patrimoine*, adding the following description to the list of the electronic publications subject to legal deposit:

> Sont également soumis au dépôt legal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique.
> [Signs, signals, writings, images, sounds or messages of any kind communicated to the public by electronic means are also subject to legal deposit.][99]

The definition appears to be breaking digital objects down to their core elements, leaving the description as open as possible to any new technological developments consisting of *signs* or *signals* transmitted by electronic means (see also Section 6.1.2). It is thanks to this broad definition that the BnF and INA are able to collect content from social media platforms. Based on the regulation in force, legal deposit institutions have to consider two main criteria when assessing if an electronic item is in scope: it has to be made available

[98] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. Available here: https://web.archive.org/web/20240119142905/https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32001L0029

[99] In this chapter, all text in French has been translated to English using translation tools such as GoogleTranslate.

on the French Top Level Domain (e.g., .FR, .CORSICA, .PARIS); or the content has to be produced on French territory, or by a person resident in France. Identifying with certainty content that meets such territoriality criteria can be quite difficult in the context of social media because of a wide array of challenges, including its interconnectedness and the international nature of the web itself (see Section 2.5).

Moreover, the French legal deposit legislation has always had as one of its goals the intention to create collections that are as exhaustive as possible in order to preserve everything that is published on French soil (Stirling et al., 2012). However, archiving everything in the context of the web and especially social media can reveal itself as an impossible task due to the amount of information generated on the web, and the risk of collecting content that may be out of scope or could potentially endanger individuals' privacy. To prevent the latter, Article L131-2 of the *Code du Patrimoine* specifies that electronic legal deposit institutions can only archive digital material that is "communicated to the public" on the French web. This additional information helps set a limit, albeit small, on the collection perimeter excluding, for instance, emails, private messages and private areas/accounts on social media. Nevertheless, the sheer amount of content produced daily on social platforms, and the many challenges related to the comprehensive capture of such material, has led French deposit institutions to revisit the concept of "exhaustiveness" for web and social media collections. The BnF Lead Curator indeed noted that they "don't archive everything" but have made it their mission to "ensure the best representativeness of the French web". Coming to terms with the actual unfeasibility of comprehensively collecting social media content, the BnF and INA are striving to create national social media collections that are as representative as possible considering legal constraints and the vastness, fragility, and ever-changing nature of such platforms. Section 5.4.2 will explore the challenges faced and archiving strategies adopted by the BnF and INA to achieve representativeness of web collections.

While traditional legal deposit provided for publishers to deposit each and every item published in France, the *Code du Patrimoine* contains a different provision for the deposit of material published on the web. In order to facilitate the capture of an ever-increasing number of websites and other formats made publicly available online, Article L132-2-1 of the *Code du Patrimoine* permits deposit institutions to use automated

procedures to capture content.[100] Moreover, Article L132-4 specifies that authors may not prohibit the reproduction of a work when it is necessary for collection, storage and on-site consultation.[101] Allowing electronic legal deposit institutions to make copies of (copyrighted) content available on the web has been essential for the development of web and social media archives in France, as material captured from the web is indeed a representation, a reproduction, of that existing on the live web (Brügger, 2018b). The ability to make copies is therefore a necessary condition to fulfil the legal mandate to collect, preserve and make available digital material published on the web.

Article L132-4 paragraph 1 amended by 2006 DADVSI also specifies the conditions under which electronic legal deposit collections can be consulted. To comply with copyright and data protection laws, web and social media content archived by legal deposit institutions is only made available to accredited researchers on-site, including at the BnF buildings, INA's centres and other selected institutions across France (see Section 5.4.4).

5.3 Technical Framework and Frequency of Capture

At the BnF, web and social media archiving is carried out mainly using Heritrix, the web crawler created by the Internet Archive. Although the harvesting activities can be planned and overseen directly from Heritrix, the BnF digital legal deposit team has been using the NetArchiveSuite (NAS) to schedule and monitor harvests due to a more user-friendly interface. The BnF uses different tools and approaches to archive content from different social media platforms. For instance, Heritrix is used to crawl content from static webpages and social media such as Twitter and Facebook. Conversely, for platforms containing audiovisual content such as YouTube, the BnF has been using a mixed

[100] Article L132-2-1. Available here: https://web.archive.org/web/20240117185814/https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000044796411

[101] *Code du Patrimoine*, Article L132-4: 1° Consultation of the work on site by researchers duly accredited by each depository organization on individual consultation workstations, the use of which is exclusively reserved for these researchers; 2° The reproduction of a work, on any medium and by any process, when this reproduction is necessary for the collection, storage and on-site consultation under the conditions provided for in 1°. Available here: https://web.archive.org/web/20240919095728/https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006845526/2006-08-03

approach involving combined harvesting through Heritrix and Youtube-dl (see section 5.4.3). Youtube-dl[102] is an open-source, command-line programme that curators use to extract videos from YouTube.com and other video-hosting platforms. Among other features, this tool provides reports related to the persistence of audiovisual content on the chosen platform, offering information related to whether the video is still available or has been removed by its creator.

As will be illustrated in section 5.4.2, the BnF can count on an extensive network of internal and external contributors that work together to develop the French web archive collections. To facilitate the selection of URLs to be archived as part of the digital legal deposit collection based on the suggestions of its many contributors, the BnF has internally developed the *BnF Collecte du Web* (BCWeb)[103] tool. BCWeb enables contributors within and outside the BnF to participate in the selection of content by suggesting webpages, social media accounts or any other web-based content to be included in the web archive. During my visit to the BnF, curators kindly offered a virtual tour of the tool, illustrating how curators and other partners can select, manage (add, edit or deactivate) and monitor URLs to be collected through BCWeb. Within this tool, websites selected for harvesting are organised by collections, each of them assigned to specific curators who are responsible for monitoring them. Moreover, to each selected URL, curators can assign three parameters including the frequency of collection, the data budget assigned to it and the depth of crawling. Regarding data budget, curators at the BnF explained that for "budget" they mean "the number of files collected from a site" (see also section 5.4.2). An estimate of the budget that a specific website would require is usually calculated by conducting a Google search using the "site:" operator followed by the domain name of the page that is to be collected. Based on the number of URLs identified through this search, the BnF categorises the website's data budget as small (< 50,000 URLs), medium (50,000-100,000) or large (>100,000). Furthermore, BCWeb verifies the validity of a URL before initiating the crawl, preventing broken URLs from being added to the seed list. BCWeb also checks for duplicates, avoiding any risk of crawling the same URL twice. All content archived from the web and social media is saved as WARC files, which have been the recognised standard format for web archived items since 2009.

[102] https://web.archive.org/web/20231211045625/http://ytdl-org.github.io/youtube-dl/

[103] https://web.archive.org/web/20240314171034/https://collecteweb.bnf.fr/login

The BnF assigns different frequencies of crawling based on several criteria such as the frequency with which a webpage is updated, whether there is an existing archive on the website, or its risk of disappearing. Based on such criteria, the BnF has established five frequencies for items to be preserved as part of the web archive: once a year, twice a year, once a month, once a week or several times a day. Different frequencies may be assigned in particular cases, such as for news websites that are harvested once a day.

The *Institut national de l'Audiovisuel* has developed through the years its own specific tools for the collection of audiovisual content. Claude Mussou observed how in the early days of web archiving, the type of websites INA was seeking to archive were mainly composed of dynamic material, which available tools at the time were struggling to capture. To make up for the lack of suitable tools, INA developed its own in-house tools, enabling them to crawl complex and dynamic content (mainly videos) from websites at a higher frequency compared to other existing tools, while also saving storage space.

As for Twitter, INA collects data via APIs. At the time the interview and fieldwork were conducted, INA was using version 1 of the Twitter API and was planning to switch to version 2. However, following Elon Musk's acquisition of twitter.com in late-2022 and the subsequent shut down of version 1 of the Twitter API, INA has proceeded with the implementation of version 2 earlier than expected (see section 5.5).

To support selection of social media content for their archive, INA has been using a variety of curation tools which enable them to keep track of and receive alerts related to trending hashtags or topics. Among the tools mentioned during the interview are Talkwalker,[104] Hootsuite[105] and Trends24[106] (see section 5.4.2). Regarding the frequency of social media collection, INA gathers content through Twitter's Streaming API in a continuous, ongoing stream of data, although occasional glitches and interruptions may occur. As for other platforms such as YouTube or Facebook, data is collected daily via the API in order to capture the latest posts and content produced since it was last crawled.

Alongside the development of their own tools, INA has also developed its own archiving file extension, called DAFF (Digital Archiving File Format), which is compatible with the standard file format used by other web archiving institutions, such as WARC files (see section 5.4.5).

---

[104] https://web.archive.org/web/20231224034640/https://www.talkwalker.com/

[105] https://web.archive.org/web/20230904103852/https://www.hootsuite.com/

[106] https://web.archive.org/web/20230827094443/https://trends24.in/

5.4 Challenges

This section comprises an analysis of the challenges faced and solutions implemented by the respective web archiving teams at the BnF and INA to mitigate legal, curatorial, ethical and technical concerns. Interviews and informal discussions with members of the respective web archiving teams carried out during fieldwork at the BnF and INA's premises in Paris revealed the main obstacles encountered in the development of social media collections in France. These mainly stem from legal constraints imposed by social media companies on data collection and the increasing complexity of social platforms over time, requiring archiving institutions to constantly update solutions previously adopted and develop tools that can better adapt to these changes. Additionally, difficulties arise in establishing stable collaborations or at least communication channels dedicated to heritage institutions with social media companies.

5.4.1 Legal Issues

Defining the virtual borders of a national Top-Level Domain (TLD) is a common problem for many web and social media archiving institutions across the globe, including the BnF and INA (Masanès, 2006; Pennock & Beagrie, 2013). The French digital legal deposit legislation provides for the collection of websites and any "*signs, signals, writings, images sounds or messages*" transmitted through the web in the context of public communication. As in the case of the British Library in the UK (Chapter 4), the BnF and INA had to deal with the complexity of virtually delineating the boundaries of the national web. While documenting an early experience in archiving the French Web domain at the BnF, Abiteboul et al. (2002) recognised the difficulties related to providing a formal definition of what was to be considered as part of the French TLD, as the boundaries between national domains are often blurred. Identifying websites belonging to the French web for preservation purposes is usually limited to URL extensions including .FR, .JF or .CORSICA, although this criterion poses the risk of leaving out many French websites that use, for example, a .COM extension. Other selection principles include prioritising content in the French language and websites or accounts whose creators reside in the French territory. While these parameters can be helpful in identifying material to be preserved by deposit institutions, they still raise questions as, for instance, individuals on the web can communicate in languages different from their first language (Abiteboul et al., 2002).

Establishing a clear perimeter for social media content poses even greater challenges. Platforms are usually hosted on .COM domains, and their headquarters or data centres are often located in the US or countries other than France. Moreover, content created and posted on social media is often reshared by users whose geographical location may vary, and, especially when using hashtags, posts may become an integral part of broader, ongoing conversations that involve users from across the globe, using a variety of languages. To address this, a possible solution is to apply multiple criteria (e.g., language, territoriality, and relevance) in order to find a common ground that allows deposit institutions to create collections that accurately represent critical events and cultural, political and social phenomena related to France.

A major legal issue that was raised by both interviewed institutions is the limit imposed on access. By law, web and social media archives at the BnF and INA can only be accessed at specific terminals available on their premises. Although both deposit institutions can count on a wide network of more than 50 sites across France, accredited researchers are still required to physically visit those premises, posing a significant spatial and economic obstacle, particularly for people living with disabilities (see section 5.4.4). This issue became particularly evident during the first lockdown imposed at a national level during the COVID-19 health crisis. Apart from rare exceptions, material in the French web archives at INA and the BnF remained largely inaccessible as libraries were shut down until further notice. From a technical perspective, making content available online would not constitute a problem, as the infrastructure potentially already exists – as Mussou noted. However, granting broader online access to web archives, even if it is restricted using a secure authentication system, can raise concerns related to privacy, copyright and data protection. This is particularly true for social media content, which, despite precautionary measures, may still include sensitive or copyrighted material (see section 2.4.1 and 2.4.2). Restricting access to resources only to registered users and allowing consultation within authorised reading rooms through computer stations made available by libraries is intended to offer an additional layer of protection in this sense.

An additional issue for researchers seeking to engage with French web collections at the BnF and INA arises from the restrictions on reproducing web and social media archived materials. While making copies of material from the web is permitted to legal deposit institutions for the purpose of preservation, as copying is a necessary condition to enable the collection of such content (Abiteboul et al., 2002; Brügger, 2018b; Stirling

et al., 2012), making screenshots or copies of files included in web collections is limited.[107] The BnF explained that users are allowed to take screenshots for personal use only. However, if researchers want to reproduce or publish content included in the French web archives, they must request permission directly from website and social media content owners, even when the content is still available on the live web (see also section 5.4.4).

Collecting content through hashtags or from accounts that, at some point in time, decide to change their privacy settings from public to private, may pose additional legal and ethical challenges. Several studies have explored the meaning and usage of hashtags and how often this corresponds to a desire for content to be discovered and join global public discussions (A. Cui et al., 2012; Zappavigna, 2015). However, individuals who are not public figures might not be aware of or simply do not wish for their posts to be preserved for the long-term (see also Section 6.1.3). Moreover, activists and creators on social media might not want to see their content or any form of their work archived forever due to copyright, privacy and data protection concerns (Velte, 2018). While the French electronic legal deposit legislation exempts deposit institutions from the requirement to request permission from content owners prior to collection – a crucial aspect in the context of social media archiving, where seeking permission from a large number of users would be impractical and time-consuming – the BnF and INA provide individuals with the option to request the removal of content. Nevertheless, as both institutions noted, in the rare instances when this happens, content is not permanently deleted from the archive and main storage but rather hidden from public view. In any case, deposit institutions must be directly instructed by their legal department to hide contested material. Mussou (INA) added that INA usually does not remove anything from the archive unless they are specifically "instructed by the legal department or DPO [Data Protection Officer] to hide the content" in case of GDPR breaches.

Changes in privacy settings from "public" to "private" of social media accounts selected for capture exemplify the complexity of archiving items from platforms where dynamics and settings may be altered with just a few clicks. Curators at the BnF mentioned how this happened for some accounts they were archiving as part of a collection documenting the *Gilets Jaune* protests. The "Yellow Vest" movement – so called because of the yellow luminous safety vests its supporters wore – was born almost entirely on

[107] See also the regulation for "Printing from computers" on the BnF site: https://web.archive.org/web/20240117085121/https://www.bnf.fr/fr/photocopie-impression-photographie#bnf-impression-partir-des-ordinateurs

social media, initiating in November 2018 a series of demonstrations in France advocating for economic justice and asking for institutional reforms (Chrisafis, 2018). Among the material that the BnF collected from the web to preserve a trace of these protests, there were a few Facebook accounts which went private once the movement began to fall apart:

> One of the problems we faced related to the *Gilets Jaune* was that towards the end a lot of accounts moved to private. […] because of this we had a lot of issues since we are not allowed to [archive private accounts]. So, we have the beginning of the movement on Facebook but when the situation started to become more complex for the movement, they switched their accounts settings to private.

From a curatorial perspective, this understandably constitutes an issue in terms of completeness of a collection that focuses on a specific event largely documented on social media, transmitting to future generations only the beginning and the middle of a story that will probably be left with a partial (archival) epilogue. In a blog posted on the platform *Hypotheses* about the collection of web content related to the *Gilets Jaune* movement across different social media and websites, Tybin (2019) reflected on the curatorial consequences of having to stop archiving profiles due to the changing of their privacy settings. Of the over 200 seeds crawled at the start of the protests, nearly a year later, these had to be reduced to about 80 seeds due to the legislation only allowing the collection of public content, further emphasising the unstable nature of archiving similar events on social media (Tybin, 2019). Nevertheless, limiting archiving practices to public material only is essential to protect individuals' privacy.

Legal constraints and platforms' terms and conditions, especially regarding limitations on the usage of APIs for data collection, pose another major challenge for social media archiving institutions like the BnF and INA. As explained in Section 2.4.4, platforms such as Facebook and Instagram are renowned for being difficult to archive because of the limits imposed by Meta Platforms Inc., the company that owns them, on the number of accounts per institution and number of profiles that can be harvested in a set timeframe. In fact, Meta considers as a form of "data misuse" any unauthorised crawling activity that uses automated techniques to collect more than a certain amount of accounts/data per hour.[108] The rate limit imposed by Meta is one of the strategies adopted

[108] Further details on Data Misuse and strategies adopted by meta to contrast this phenomenon are available at the following link:

to combat the phenomenon and therefore the platform tends not to disclose the rate allowed, as it explicitly declared in a 2021 blogpost: "we don't want to give a roadmap to scrapers seeking to evade our defences, but one example is that we look for patterns in activity and behaviour that are typically associated with automated computer activity and stop it".[109] Thus, memory institutions like the BnF have carried out several tentative experiments collecting different amounts of accounts in order to gauge data limits. However, due to the uncertainty surrounding the results of these tests, the French archiving institution has decided not to exceed 200 accounts per crawl. In the case of Instagram, the BnF collected about 180 profiles per crawl, captured two to three times a year up to 2022, when the interview was conducted. In 2023, the BnF has successfully expanded the number of Instagram profiles collected to 300 per crawl, and the frequency increased to four to six times a year. The approach adopted for Facebook was similar. However, since the BnF has been recently blocked by Meta, any capture from Facebook has been put on hold for the foreseeable future (see Section 5.4.3).

The *Institut national de l'Audiovisuel* experienced similar issues on Facebook. Since this social platform is widely used by broadcasters, INA's curators attempted to request access to Facebook's official APIs. A member of INA's Research and Development team recounted how they tried to establish an institutional relationship with different social media platforms, including Meta. While it was difficult with Twitter to even identify the right person to contact, as most of the contact details available only pointed to the more commercial side of the business, for Facebook, INA surprisingly succeeded in initiating a conversation about being granted special access to the API:

> Three years ago, when we tried to first open an API account with Facebook – this was after the Cambridge Analytica scandal – there was already a strict process to access the API. […] After a three-month exchange to request access to do archiving for legal deposit reasons, we succeeded.

---

https://web.archive.org/web/20230601202248/https://about.fb.com/news/2021/05/scraping-by-the-numbers/

[109] Further details on rate limits and strategies adopted to combat unauthorised scraping are available at the following links:
https://web.archive.org/web/20231127151411/https://developers.facebook.com/docs/graph-api/overview/rate-limiting/ ;
https://web.archive.org/web/20231127094937/https://about.fb.com/news/2021/04/how-we-combat-scraping/

After obtaining an official Facebook API account, for which INA explicitly declared in the application form that the purpose was to archive information under electronic legal deposit, the account was soon afterwards permanently suspended for exceeding the data limit, with no chance of appeal (see Section 5.4.3).

Ideally, requesting access to APIs or signing an archiving agreement similar to the one stipulated between Twitter and the Library of Congress (see section 2.3) would be beneficial for memory institutions fulfilling their electronic legal deposit mandate. However, the lack of dedicated departments or a fast-track lane for cultural heritage institutions, coupled with the need for platforms to protect their economic interests through the sale of users' data, significantly hinder the development of national social media archives. National memory institutions find themselves having to operate in the narrow threshold existing between what is required by their legal mandate and the limitations set by platforms' terms and conditions regarding data collection.

The following section will examine the effects that the French legal framework, platforms' terms of use and institutional collection development policies have on the selection of content from social platforms, as well as the strategies implemented to ensure the representativeness of social media collections at the BnF and INA, and additional curatorial and ethical considerations that were brought to light during the interviews.

5.4.2 Selection Practices and Ethical Concerns

The national legal framework largely determines and delimits the type of material that is to be included in the web and social media archiving collections for safekeeping at both the BnF and INA. Preserving everything, all the digital memory related to France available on the internet, is not only unfeasible, as it would require an unlimited number of resources which cultural heritage institutions do not possess, but it could also lessen the value of the entire collection if appropriate archival appraisal is not performed (see Section 2.5). To this must be added, especially for social media, a series of obstacles posed by platforms' owners, the sheer amount of information published on a daily basis, and technical setbacks that make archiving these sites extremely complex. For this reason, the BnF and INA's mission focusses on ensuring the best degree of representativeness of their collections within the existing constraints. Both institutions have put in place a series of solutions, including curation tools and solid networks, that contribute to shaping collections that are as representative and inclusive as possible of the manifold layers that compose French society.

At the BnF, the electronic deposit team carries out two main annual broad crawls which aim to provide a snapshot of the whole French web domain for that year. These two broad crawls are then completed by a series of selective crawls, whose goal is to include a limited number of selected sites to enrich the material gathered through the broad crawls. The Head of the electronic legal deposit team explained how archiving activities at the BnF are "based on both national and international cooperation". Selection practices for the French web archive appear indeed to be the result of a joint effort and collaboration between the web curators that are part of the electronic legal deposit team, other internal and external collaborators, and researchers coming from different disciplines. The internal contributors include 10 coordinators (one for each department at the BnF) and more than 80 collaborators who oversee the selection of sites and social media content for their topic or area of expertise and thematic collections (e.g., Artificial Intelligence, climate change, Covid-19). As for the external contributors, this group comprises curators from the 26 libraries located in the various French regions and overseas territories. External curators have been asked to participate in the selection of content for the thematic collection about elections since 2004. The collaboration has since then been expanded to other topics of national and regional relevance such as the COVID-19 pandemic, environmental issues, and Artificial Intelligence. While all regional and overseas libraries were also invited to select content for regional collections, only four of them have participated so far: the existing collections are related to the regions of Alsace, Languedoc-Roussillon, Lorraine and Provence-Alpes-Côte d'Azur. This diverse and large group of external contributors has enabled the BnF to include in the web archive content that pertains to various areas of the French territory. These suggestions are essential to support the discoverability of local sites or social media accounts related to small, regional communities or personalities which are only known to local residents and that would probably escape from the more national focus of the main collections as these are often unknown to non-residents.

To internal and external contributors must be added research centres from across France (e.g., PACTE, SciencesPO, National Art History Institute), other institutions such as the *Bibliothèque de documentation internationale contemporaine*[110] or, for example, the Centre for Feminism Archives[111] (University of Angers). This network of contributors proposes and selects material from websites and social media platforms in order to collectively

[110] https://web.archive.org/web/20240213063418/http://www.lacontemporaine.fr/

[111] https://web.archive.org/web/20240229110356/https://bu.univ-angers.fr/CAF

generate an archive of the web that aims at comprehensively reflecting the economic, social and cultural life of France. The network is in constant expansion, and to facilitate the onboarding of potential new contributors, the electronic legal deposit department at the BnF is working on producing relevant documentation. As Tybin explained: "We have launched a process to document the work of selection [done] by contributors in order to help researchers, but also new contributors, to understand how our collections are built". Documenting archiving strategies is indeed essential to offer not only new contributors but also researchers the opportunity to get a better understanding of how to critically engage with the collections and what they can expect to find (or not) in them.[112]

As mentioned, social media content is archived as part of thematic or event-based collections, alongside other web materials such as blogs and websites. One of the curators noted that Twitter, which represents the main social media archived at the BnF as of 2023, does not have its own collection, but most of its material flows into the "Ephemeral News" collection[113] and other thematic collections (e.g., COVID-19, Presidential Elections). Among the selection strategies adopted by the electronic legal deposit department there is the call for participation launched in December 2021 to ask the public to propose material for inclusion in the Artificial Intelligence and the 2022 Presidential Election collections. The BnF explained:

> We asked the public to participate and select URLs that they think should be added to the collections. For AI, we received 25 profiles suggestions and 80 URLs. It was a good result. And I think it would also help in terms of representativeness.

Although the numbers of sites suggested may seem small compared to the sheer amount of material produced daily on the web and social platforms, as was noted during the interview, the campaign proved to be a positive experience, so much so that it has been replicated for different topics. Moreover, on the BnF website it is possible to suggest

---

[112] Insights about the making of some of their collections are available on the BnF's blog *Web Corpora*: https://web.archive.org/web/20240502150044/https://webcorpora.hypotheses.org/category/preserver-la-memoire-du-web/la-fabrique-des-collections

[113] The collection *Actualité Éphémère* [ephemeral news] is described by the BnF as a collection that "concerns phenomena which are relayed on social networks (Twitter, Facebook, etc.), blogs or sites which we may fear will have a limited lifespan". [translated with GoogleTranslate] https://web.archive.org/web/20240205171918/https://www.bnf.fr/fr/decouvrir-les-collections-darchives-web-de-la-bnf#bnf-les-collections-presse-et-actualit-s

URLs for collection at any time by submitting the provided online form.[114] Although no explicit limit has been stated, the wording of the description at the top of the page suggests that the focus remains on websites rather than social media profiles. Nevertheless, the intricate system of collaboration implemented by the BnF and its partners, as well as crowdsourcing activities like the ones described, can help to shape social media collections, aiming at representing the plurality of the voices, histories and culture of the many different territories, local entities and minority groups in France.

The approach implemented by INA is different. The type of cultural heritage material they are called to preserve includes anything related to TV and radio personalities, channels, and programmes as well as users' interactions generated around TV shows or radio transmissions. INA is also interested in preserving French video channels and on-demand TV services, and new types of digital formats made available for streaming or download on the Internet, such as podcasts. As part of their varied archive, INA seeks to capture reactions to TV programmes, news, and big events on social media. During the interview, Mussou offered a detailed overview of their social media archiving activities, describing the process and tools they use to refine the selection of content. At first, curators at INA decided to expand the scope of the existing web collection by capturing social media accounts associated with archived sites. However, it soon became evident that only a small number of these websites were actually linked to social media accounts, prompting INA to explore new paths for the selection of in-scope material. INA curators mentioned some experiments with keyword searches but the outcomes were often unreliable and not exactly relevant to the topic they were looking to capture. For example, Mussou explained that when trying to search for material about a popular French TV drama titled *Plus belle la vie*[115] using this approach, they would obtain disparate and overly generic results. Due to the difficulties of identifying relevant content to be captured, as in the case described, INA has started to introduce in their workflow a series of curation tools to support what is mostly manual selection of social media material.

Among the tools used, Mussou mentioned Talkwalker, which works as a search engine where curators can set alerts for specific keywords such as "audiovisuel";

[114] https://web.archive.org/web/20231027095214/https://www.bnf.fr/fr/signaler-un-site-web-la-bnf

[115] *Plus Belle la Vie* is a popular French tv soap opera that narrates the everyday life of people living in a fictitious neighbourhood called Le Mistral, in Marseille. More information about the tv show can be found here: https://web.archive.org/web/20240106032350/https://www.tf1.fr/tf1/plus-belle-la-vie-pblv

Hootsuite, a social media feed manager on which curators can monitor, for instance, Twitter or Facebook timelines; and Trends24 which curators use for following and receiving alerts about trending hashtags on Twitter worldwide. These and other tools enable INA curators to complement manual selection, which is mostly based on the team's broad knowledge about anything related to TV and radio shows and main events reported in the news happening in or related to France.

The archiving advantages generated from both the BnF's and INA's selection approaches need, however, to be balanced with the technical and legal complexity of archiving material from social platforms. Due to the limitations imposed by social media platforms described in section 5.4.1, combined with annually established data budgets and the resources required to manually select material of interest, the BnF has restricted not only the number of accounts to be collected from sites like Instagram, but also reduced its collection frequency to two or three times a year. The BnF noted:

> For [Instagram] accounts the contributors select a list of profiles they want to archive, and we select 180 accounts starting from the list provided by the contributors. And if we cannot crawl all the accounts suggested, we will crawl them in the next crawls.

Even though the number of profiles collected and the frequency has been increased in 2023 (see Section 5.4.1), such constraints represent a significant element to consider when assessing the representativeness of social media collections, requiring contributors and curators to pay extra care when selecting material as only a small number of accounts can actually be harvested. Moreover, even when archiving institutions manage to archive content from Facebook or Instagram, the harvested data may present several technical issues (see Section 5.4.3), leading the BnF to focus their resources on sites that have a good-quality capture rate. Similarly, a limit has been set on the number of accounts to be collected for YouTube and TikTok. For the latter, during the first archiving test conducted in the spring of 2022, curators established a maximum of around 150 accounts and hashtags related to candidates and supporters involved in the election campaign (see also Faye et al., 2024). This was done to ensure the smooth running of the harvesting activities and reduce the chance of being blocked. As for YouTube, the BnF has imposed limits in order to remain within a designated annual data budget, which according to Faye et al. (2024) is set between 3 and 5TB. Tybin explained that at the beginning of each calendar year, the BnF allocates a storage budget for the electronic legal deposit department, covering content collected over the course of twelve months. For YouTube,

the BnF's collection development policy stipulates that the collection on this platform is to be carried out at the level of the channel taken in its entirety and not individual videos (see also, Faye et al., 2024, p.192). One of the BnF curators explained that this decision stems from the need to be selective with the data archived due to limited data budget, and to ensure representativeness of collections. By selecting only channels that are of interest in their entirety, the BnF can ensure a balanced and representative collection of content, covering a wide range of themes without exceeding their storage limits. These restrictions, coupled with the increasing quality and, consequently, the size of video files, have required the BnF curators to meticulously plan each crawl: "Before launching a crawl, we estimate the budget of the channels we want to crawl and if we reach the limit, we remove some channels and crawl them instead next time." Having to collect entire YouTube channels instead of single videos means that the selection of YouTube channels must be considered carefully, as the channel as a whole has to be deemed of cultural interest in order to be included in the seed list.

As for the *Institut national de l'Audiovisuel*, the web archiving team has also been facing challenges related to limitations imposed by platforms, specifically on the quantity of data harvested via Twitter's APIs. For example, web archivists at INA explained that through Twitter's API, they can retrieve about 3200 of the latest tweets and retweets – in reverse chronological order – published by a single account (Faye et al., 2024). This means that for a selected account the archiving team would potentially not be able to collect it comprehensively if more than 3200 tweets/retweets had been posted since the profile was created (Twitter, 2023a). A similar issue has been encountered when archiving tweets related to a specific hashtag. When following hashtags, the Twitter Search API can only collect tweets a few seconds after being published and can also only retrieve tweets no older than seven days (Faye et al., 2024). Since hashtags can go viral all of a sudden after remaining dormant for a long period of time, the existing temporal limitations put in place by Twitter risk hindering the complete capture of the history of social media phenomena. Moreover, for hashtags collected through the Twitter Filtered Stream API, which provides access to a real-time stream of public posts related to a keyword query, one of INA's web archivists observed that "With the [Filtered Stream] API you can only collect 50 tweets per second, […] so if we are following the hashtag #ukraine or #covid we [can] only collect the first 50 tweets published in that second" (see also, Faye et al., 2024, p.7).

This limit also constitutes an additional issue when trying to capture, for instance, the worldwide reaction to events like the 2015 terrorist attack in Paris. As Mussou noted:

> One of our main collections in the Twitter archive is related to the 2015 terrorist attack in Paris and we started collecting as soon as it happened following hashtags. This was an issue because actually lots of the hashtags were not in French, like #prayforparis […] and for these hashtags there were too many tweets to collect.

In the attempt to collect as many tweets as possible related to the #prayforparis hashtag, INA had to face the fact that, because of the above-mentioned rate limit of 50 tweets per second, they would never be able to capture all of them, consequently producing gaps in the information preserved. It is worth pointing out, however, that there is and will always be a discrepancy between the sheer amount of what is produced daily on social platforms and the actual amount of content archived and preserved by memory institutions such as the BnF and INA based on the tools and resources available. Moreover, this discrepancy is only destined to increase as the amount of interactions produced online has been growing exponentially over time and the institutional capacity to archive them is not able to follow this trend. Nevertheless, it is important to underline that the mission of the electronic legal deposit (as opposed to the traditional one) is not to archive everything that is published on the web but to ensure the best representation possible of the French websphere.

Moreover, according to Statista.com (2023), among the platforms with the most active number of users in France in 2023 are Facebook, YouTube and Instagram, with Twitter ranking sixth. This data appears to somewhat match the degree to which platforms are archived at both the BnF and INA as part of their electronic legal deposit collections, although adjustments have been made necessary due to the scope of each institute and the previously mentioned set of legal and technical obstacles.

Finally, a common issue reported by both the BnF and INA is the changes that social media platforms bring to their sites. The everchanging nature of social media results in archiving institutions having to adapt rather frequently their workflows and tools' settings to capture new features or other modifications implemented (Faye et al., 2024). Both institutions are working on producing documentation that will keep track of the evolution of barriers, solutions and archiving strategies adopted over time. Making this information openly available may be beneficial not only to future curators or contributors, but most of all to all those web archive users who seek to understand how collections have been developed and how this may potentially affect the analysis of the corpora they wish to work on.

5.4.3 Technical Obstacles

As mentioned in Section 5.3, the BnF and INA have been using two different technical frameworks when it comes to social media archiving. The BnF has been using predominantly Heritrix for the collection of content from the different types of social platform they currently archive. Heritrix is, however, a tool initially designed to respond to archiving needs that were typical of static webpages and therefore it is unable to properly capture and interact with dynamic elements on the page, such as scrolling pages or playing videos (Faye et al., 2024). This is exemplified by the problem experienced by the BnF when harvesting Twitter through Heritrix. As explained during the interview, the BnF collects Twitter accounts or selected hashtags twice a day. Despite specific configurations and external scripts used to harvest social media effectively, Heritrix still has issues in interacting with some of the dynamic elements on social media platforms. As a result, the content captured is inevitably limited to the tweets displayed at the top of the feed, due to the web crawler's inability to scroll down and prompt the loading of older posts:

> […] We only get the top 20 tweets, so if it is a hot topic with many tweets per day, [Heritrix] would be able to capture only a fraction. So, for topics or hashtags there is a lot of content that is not archived because of this limitation of 18/20 tweets [per crawl] (Tybin, Interview).

The same happens for accounts of public figures included in the seed list. If one of the selected personalities tends to publish more than twenty tweets a day, the chances are that the content captured is incomplete.

Section 5.4.2 highlighted how legal and curatorial challenges prompted the BnF to restrict collection activities on Instagram, to the point of necessitating manual selection and harvesting of content. From a technical perspective, manual harvesting became necessary as the BnF is unable to archive content directly from Instagram and therefore has to rely on a free online tool called Picuki.[116] This tool functions as a sort of search engine and allows the user to view content from public profiles without having to log into the main platform. Most importantly, it displays Instagram photos or Reels (short video clips of up to 60 seconds) along with their caption. The latter feature is essential, as when harvesting Instagram profiles or hashtags using Heritrix, the web crawler would only capture the grid of images and the information contained in the description box below

[116] https://web.archive.org/web/20231204020658/https://www.picuki.com/

the user's profile picture, and not the caption associated with each post. This limitation becomes particularly significant in research contexts where the analysis of captions plays a crucial role, such as understanding user sentiments, contextualising images, or tracing the evolution of narrative themes and societal phenomena over time (Blackwood, 2021; Phillips et al., 2022). In order to mitigate this issue, the BnF has introduced Picuki into its workflow: once BnF curators together with their network of contributors have selected relevant Instagram accounts or hashtags for collection, web archivists manually search each username and hashtag on Picuki, and then crawl the URLs generated on Picuki via Heritrix. This and other workflows are discussed and revised collectively at the BnF during annual workshops hosted within the library to identify appropriate technical approaches for the collection of specific platforms. For example, Instagram was the topic for the 2020 workshop, while TikTok was the subject of the one held in 2022.

The BnF offers several interesting examples of the variety of tools and approaches that can be used in combination with more traditional web crawlers like Heritrix to supplement the capture of missing information, capture dynamic content that crawlers are often unable to collect or circumvent specific constraints. In particular, the BnF has configured Heritrix with several parameters, including a four-hour limit for the overall duration of downloaded videos per crawl, which would prevent the library from comprehensively collecting YouTube channels. For this reason, curators have established a specific workflow for the capture of YouTube content, which begins with the creation of a seed list of relevant channels through BCweb, the management tool developed by the BnF (see Section 5.3). In this regard, since YouTube channels can have more than one type of URL pointing to the channel homepage, the BnF specified that they only archive channels through the "channel ID" URL bearing the following syntax: "youtube.com/channel/[ID number]". In fact, YouTube distinguishes between three types of URL on its Help page:[117] the standard URL based on the unique channel ID; the handle URL which is the one beginning with the symbol "@" and separated from the root "youtube.com" by a back slash; the last category includes customised URLs (which are not available anymore) and legacy username URLs. Once the correct ID-based URLs have been added to the seed list, the BnF proceeds with the collection of content, which is articulated in two phases. First, archivists crawl the webpage using Heritrix to capture the structure of the page and metadata; and then they use the command-line programme

[117] https://web.archive.org/web/20230721045225/https://support.google.com/youtube/answer/6180214?hl=en-GB

"YouTube-dl" to download videos from the selected YouTube channels, also recovering some type of technical and descriptive metadata, such as title, date and duration (see also Faye et al., 2024; see also Section 5.3). The implementation of this two-step approach has enabled the BnF to mitigate some of the mentioned technical limitations that characterise both Heritrix and YouTube-dl (e.g., script needing to be frequently updated to adapt to platform's changes; issues with capturing metadata for larger channels). Besides, one of the curators noted that the information captured through YouTube-dl regarding the size of videos files that are included in a specific channel has proved to be particularly useful when evaluating and planning the allocation of the data budget for the year.

Preserving the "look and feel" of platforms over time is essential for understanding their evolution and the context in which the archived material was once embedded. However, it is not always possible to capture and accurately replay every single aspect and feature of each site across devices (see Section 5.4.4). When preparing content for access, the BnF developers described their efforts to recreate the complex structure of YouTube pages, placing all the components and data acquired via Heritrix and YouTube-dl back into their respective positions on the page. However, keeping up with platforms' continuous interface changes and the removal or addition of new features, represents a great technical challenge to preserving social media material. Representatives from both French deposit institutions reported issues in this sense during the interviews. Yet, I only fully grasped the extent of the problem when curators at the BnF quickly went through a long document used internally to keep track of all the changes they had been implementing, adapting, and adding over the years in order to be able to capture and replay social media platforms. The list of settings they had expressly for YouTube was particularly impressive. Among the examples illustrated during my visit at the Bibliothèque François-Mitterrand, the one that perhaps made obvious the process that institutions go through when trying to rapidly adjust to platforms' evolution and changes in their terms of use is related to the "disable_polymer" parameter. Curators explained that when reconstructing a YouTube page in the Wayback machine the electronic legal deposit team had to disable the dynamic layout of YouTube for them to be able to replay the archived content using the structure of the legacy website version (see section 5.4.4.):

> In 2020 we had to disable the new dynamic layout of YouTube […] and had to add the parameter "disable_polymer" [to the Wayback Machine] so that we could use the old source code. But in 2022, YouTube does not allow the use

> of that parameter anymore. We had to change the way content is displayed once it has been collected as it did not display correctly as it did before.

When in 2022 YouTube launched its new User Interface, permanently eliminating the legacy version, the developers at the BnF had to find a new solution to play back and provide access to the information collected after that date. As the BnF curators explained, although the "disable_polymer" command line did not work anymore, the BnF was still able to collect single elements from YouTube pages, but not the structure they were embedded in. For this reason, the developers had to recreate the structure of the YouTube page from scratch and, afterwards, piece together each element collected on the reconstructed page in the Wayback machine. Indeed, one of the BnF curators observed: "It looks like YouTube, but the source code of the page it's not the original one, as it has been recreated by the BnF developers. It has been reconstructed." These kinds of changes generated at a platform level, can cause issues not only from an archival, technical perspective, but also for web archive users. Bucher & Helmond (2017) illustrated how any change brought to websites transforms the way users interact and engage with platform activities, such as liking and sharing. Thus, it is important that such platform evolutions are captured and reproduced in web and social media archives, especially those that aim at transmitting the look and feel of a platform alongside information. As social platforms continue to evolve at a rate that appears to be constantly increasing, social media archiving institutions are forced to adapt as quickly as the available resources, both human and financial, allow. Yet, even when a solution to a problem is found, no matter how rapidly, a new feature or tiny change brought to a platform risks posing a new obstacle in a non-stop race against time to safeguard traces of this ephemeral content.

As for INA, developers at the training centre in Bry-sur-Marne explained that one of the main issues they had to face in relation to collection through APIs was not, however, specifically related to the evolution of the API itself, despite its being subject to change (see also Faye et al., 2024). As they noted, the Twitter APIs have remained more or less stable over the last few years, with approximately two major changes in the timespan of 10 years. Rather, their concerns revolved primarily around conditions of access, particularly the difficulty in setting up accounts to access the information, which, as they emphasised at the time of the interview, was becoming increasingly complex. This appeared to be particularly evident when they initiated the application for the Twitter Academic API, only for it to be suddenly deprecated in 2023 (see section 5.3 and 2.4.4). While INA is one of the very few social media archiving institutions that have succeeded

in securing access to the Twitter Academic API, they described the numerous questions – "more than the previous year", they underlined - that had to be answered in order to prove their status and obtain an account. Securing an account, however, does not appear to provide any additional assurance to the API user, especially to archiving institutions. In the event that Twitter API accounts are blocked or permanently suspended for exceeding data limits (see section 5.4.2), archiving institutions often have to wait days before being able to obtain a new one and resume collection, losing in the meantime the chance of capturing material that could potentially have already disappeared.

In this regard, INA's archivists have pointed out some additional technical issues encountered when using this method of collection. On some occasions, for example, they experienced outages with the Twitter Real Time API:

> We have for some days, gaps in the data, because sometimes for two hours the API was down, and we only realised that after two hours - or we didn't have access to the network for some reason.

Technical glitches, such as the one described, can result in gaps in the collection of information streams, making it difficult to assess what is missing.

Whether using web crawlers or APIs, the examples provided in this chapter illustrated the manifold challenges and constraints posed by platforms that institutions face when attempting to archive social media. The ideal scenario for memory institutions would involve platforms offering support and granting permission to archive content from their sites for preservation purposes. Yet, no mention is made in any of the platforms' terms of use about harvesting for long-term preservation purposes by memory institutions operating under legal deposit legislation. This leaves archiving initiatives to deal with legal and technical obstacles as they strive to meet their obligations with the resources and tools available.

5.4.4 Accessing the French Web and Social Media Collections at INA and the BnF

The BnF and INA both provide access to their web and social media collection onsite to researchers who possess a Research Pass.[118] Access to the electronic legal deposit collections comprising social media items is restricted due to a complex set of legal

[118] https://web.archive.org/web/20240131204430/https://www.bnf.fr/fr/bibliotheque-de-recherche#bnf-conditions-d-acc-s

constraints aimed at safeguarding, for instance, copyright and privacy. Although limited to onsite consultation only, the BnF and INA can count on a dense network of BnF sites, INA centres and regional libraries located across the whole French territory. In total, the *Archives de l'Internet* collections are accessible at more than 26 locations[119] (Figure 9), including a couple of overseas libraries and archives.[120]

FIGURE 9: Where to access the Archives de l'Internet (source: BnF website)

INA's *WebMédia* collection is accessible at numerous locations across France (Figure 10), with some INA centres offering "Expert Consultation" alongside autonomous consultation stations.[121] "Expert Consultation" includes technical support from INA's team based at some of its centres and tools for the analysis of data.

[119] The BnF's website, "Where can I consult the BnF Internet archives?". Available here: https://web.archive.org/web/20231017225440/https://umap.openstreetmap.fr/fr/map/ou-consulter-les-archives-de-linternet-de-la-bnf_73737#6/46.362/3.450

[120] Access overseas is available at the *Bibliothèque départementale de La Réunion*, the *Archives de la Martinique - Fort-de-France* and the *Archives Départementales de la Guadeloupe – Gourbeyre*.

[121] https://web.archive.org/web/20231115040419/http://www.inatheque.fr/consultation.html

FIGURE 10: Where to access INA's Webmédia collection (source: INA website)

Despite the rich network of points of access available on the French territory, barriers to accessing web and social media collections persist, for example, for individuals who might not be able (e.g. people living with disabilities) or have the resources to travel to specific locations.

At the *François-Mitterrand* building, which is one of the BnF and INA sites that I was able to visit while carrying out interviews, numerous computer stations are made available to accredited researchers in the reading rooms positioned at the garden level,[122] at the base of the four-tower structure that make up the BnF building.

During my fieldwork I was granted permission to take photos of the workstations through which researchers can access either the BnF or INA web collections. The BnF's web archive collections are searchable through terminals like the one pictured in Figure 11, which are available in any of the lower ground reading rooms.

[122] The BnF's Floor plan – Rez-de-Jardin, available here: https://web.archive.org/web/20231018173259/https://www.bnf.fr/sites/default/files/2019-02/plan_rdj_bis.pdf

FIGURE 11: Computer workstations (access to BnF's web archive collections)

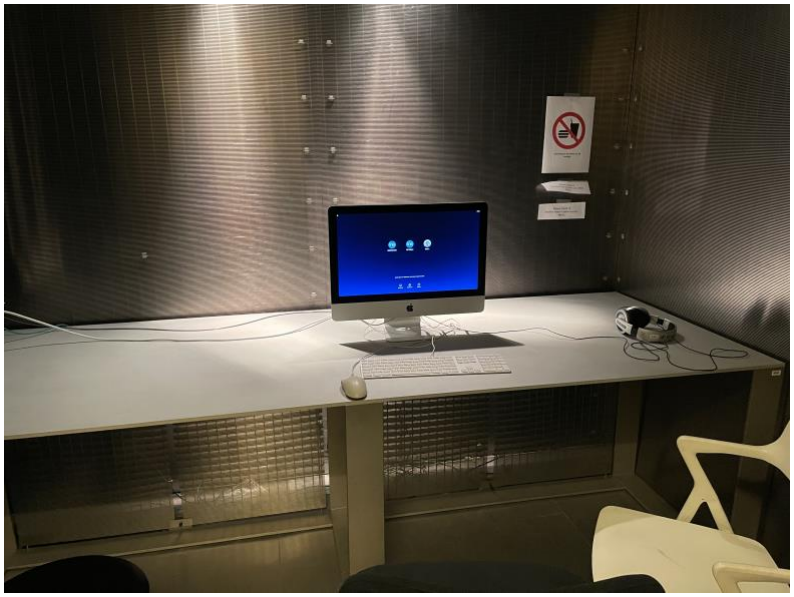

FIGURE 12: Audiovisual station (access to INA's WebMédia collections)

In the Inathèque, located in reading room P on the garden level, it is possible to browse the *WebMédia* via terminals placed within individual cubicles and in various audiovisual stations (Figure 12). These stations are small, soundproof rooms designed for researchers to view and listen to the material archived since there is no audio output available on the reading room computers.

As mentioned, only accredited researchers with a Research Pass are granted access to the these reading rooms. While this restriction adds an extra layer of security, it also

limits access to the wider public. Research Passes are issued upon the payment of a fee, which, although reduced for certain categories (e.g., the elderly, people with disabilities and high school students), might still pose an additional barrier to PhD students and researchers with limited budgets.

In terms of the way content preserved as part of the electronic legal deposit collections is made available for consultation (e.g., interface, searchability, data available, visualisation), the BnF and INA have adopted two completely different approaches, according to their different missions, methods of collection adopted, and challenges encountered.

At the BnF, the electronic legal deposit department has favoured the preservation of the context in which content archived from a wide array of platforms was embedded. The BnF's goal is to recreate the appearance and the user experience of browsing content on a certain platform within its original navigation context, in a manner that closely mimics what users would see and experience on the live web:

> Web pages are reconstructed as close as possible to the original one, with possible navigation within and between sites. […] We know that for researchers/users it's important to have the context, to see what [the platform design] looked like at a specific time.

Recreating the context to some extent is important for future researchers who may not have the opportunity to experience the act of browsing social media platforms first-hand, as we do now. However, the archived page represents only a snapshot in time of something that is not "live", dynamic and interconnected anymore (Brügger, 2018a). While platforms' appearance and navigation can be replicated to a certain degree, there are often limitations imposed by collection development policies, legal or technical constraints. For example, when archiving video material from the web, the BnF has deactivated the capture of advertisement and promotional banners that often appear at the beginning or during video playback on YouTube. Since this can be of interest for researchers, the missing elements have been promptly flagged in a PDF document available on the BnF website which describes the collection and some of the curatorial decisions made during its development, including missing ads and promotional banners associated with the channels archived.[123] The documents describing collections and the

[123] https://web.archive.org/web/20240623100105/https://www.bnf.fr/sites/default/files/2021-03/videos_parcours.pdf

so called "Guided Tours" (more on these below) also come with a list of website, account or channel URLs included in the web archive, providing crucial information to researchers before visiting the BnF premises. The BnF also collects CSS stylesheets for some of the platforms and websites archived, so that they can provide additional information or display in their customised version of the Wayback machine the archived material as it appeared on the live web at different points in time.

Conversely, INA has decided to focus on the data rather than providing a high-fidelity reconstruction of the original interface. While it can be pointed out that by adopting this approach, users may miss information about the original context, by focussing on the data alone and developing in-house their own interface called *WebMédia*, developers at INA have been able to provide various options for users to search, visualise, and analyse the data collected. Perhaps the most interesting and original examples in this sense can be found under the *WebMédia* section called "Lab", where users can browse some of the experiments INA's developers are working on based on requests and suggestions made by researchers. Among these experimental means to play back or offer new data interpretation, there is the *SocialTV (fr)*. At the time of my visit, in this section there were three videos available: an investigative report about plastic; a political debate related to the 2017 Presidential elections; and, lastly, a video regarding the terrorist attack at the Charlie Hebdo headquarters. The peculiarity of these videos is that, when a user clicks on one of them, they are directed to a page where the video is displayed on the left-hand side, while on the right, a Twitter timeline shows a stream of tweets using the dedicated hashtag for the selected TV programme. The Head of INA's services explained that "the two streams are synchronised and if you play the video, you can see the [Twitter] feed moving alongside and synchronising with the TV show". Below the video player, there is also a graph showing the fluctuation of engagement on Twitter during the broadcast. By clicking on any part of the graph, the video and the Twitter feed jump to the selected time. Due to the high volume of tweets posted simultaneously, however, the stream of tweets often scrolls down rapidly, attempting to synchronize with the video, making it difficult if not impossible to read any of the comments posted by the audience. The only way the user can read all the tweets published at a particular time, especially for sections of the video that registered higher levels of engagement, is to pause the video. Nevertheless, the SocialTV experiment, as Mussou commented, "shows how much social media is connected to the main media". Through the SocialTV, INA manages not only to capture reactions and public engagement on social media in real time, but also to

preserve a trace of the collective experience of viewers commenting live on Twitter during tv programmes.

In terms of searchability, the web and social media content can be retrieved from the *Archives de l'Internet* at the BnF using URLs. Although a full-text search option was under development at the time the interview was conducted, so far, users of the French web archive need to know the exact URL in order to find content through the search toolbar. To partially mitigate this problem, the BnF has created a series of Guided Tours[124] for collections such as the one titled *Des Vidéos À La Chaîne, Un État Des Lieux*[125] [Video Channels, An Overview] which focuses on YouTube channels, or the one dedicated to the COVID-19 pandemic, facilitating discovery of the archived material while also providing a starting point for users approaching the web archive for the first time. In addition, lists of URLs archived as part of the BnF thematic collections, including social media, are available as datasets on the data.gouv.fr website.[126]

Privacy and copyright concerns are some of the reasons behind the decision to restrict reproduction of material from the web archive to only limited forms (see section 5.4.1). The BnF Head of the electronic legal deposit department reported that there have been discussions with the BnF reproduction service in order to start offering a screenshots on demand service, although it would still potentially require single researchers to request permission directly from owners of the material they wish to publish. However, as discussed in section 2.4.1, identifying and reaching out to content creators on social media can be time consuming as often content is shared and reshared multiple times, across platforms, making it difficult to find the owner of specific content.

The BnF offers support to researchers using the web archive through various means. Apart from the already mentioned series of Guided Tours, which include the description and list of subjects comprised in these curated research paths, the BnF can provide seed lists and also offer help with data extraction and visualisation through the BnF DataLab service. For example, one researcher studying the 2003 internet

[124] The full list of Guided Tours available can be found at this link: https://web.archive.org/web/20240314121207/https://www.bnf.fr/fr/parcours-guides-archives-de-linternet

[125] The documentation and description of this guided tour can be found here: https://web.archive.org/web/20231004050318/https://www.bnf.fr/sites/default/files/2021-03/videos_parcours.pdf

[126] https://web.archive.org/web/20240418120240/https://www.data.gouv.fr/fr/datasets/collectes-thematiques-du-web-par-la-bnf/

phenomenon called "Harlem Shake"[127] reached out to the electronic legal deposit Department seeking assistance regarding a dataset they were working on:

> Usually, we first send the requested data to researchers, and it can be a very long list of URLs. However, researchers often don't know [where] to start from that long list of URLs. So, data engineers provide some data visualisations and a description of all the treatment they have done with the data (Tybin, Interview).

Data visualisation can indeed offer researchers an essential starting point for the elaboration of different interpretations about provided data, although the BnF noted that statistical inaccuracies may occur due to the different frequency with which content is crawled from different social platforms (e.g., twice a day) compared to websites (e.g., once a day for newspapers or only during annual crawls). This emphasises the importance of providing researchers with detailed information regarding the criteria used to archive content from social media, thus ensuring they have all the necessary elements to interpret the data.

As for INA's *WebMédia*, all web and social media content is fully indexed, and users can perform a full-text search using the search bar at the top of the page. Researchers can also filter content by type of data, hashtag, mentions, and language, although the accuracy of the language analysis operated by Twitter is not always reliable (Pla & Hurtado, 2017). There is also an option to include or exclude retweets and quoted posts. Moreover, for the data available through the "Twitter archive" section, INA offers a series of data visualisations using the metadata harvested from Twitter. These visualisations can be useful for gaining a basic understanding of the results and for conducting sentiment analysis. In fact, users can choose from the Dashboard section whether to view the top 10 hashtags, mentions, usernames and emoji used in the tweets included in the search results. Among the data visualisations made available on *WebMédia* there is also an option to see the top ten links and images embedded in the tweets. However, these two options had to be temporarily deactivated as they require longer to process, significantly slowing down INA's web archive browser. By expanding the "Stats" container, the user can benefit from a few more graphs showing percentages about the distribution of retweets, quoted tweets, tweets per country, language, certified accounts and also a distribution of

[127] Further information about the Harlem Shake meme can be found at the following link: https://web.archive.org/web/20230205014437/https://knowyourmeme.com/memes/harlem-shake

results per city. Regarding the latter, Mussou noted, however, that similarly to the language metadata, this may not be completely reliable "because a lot of users do not have the geolocalisation switched on".

Finally, INA's video archive includes content from YouTube, Twitter, Facebook and Dailymotion, for which they collect a wide variety of metadata along with the video file itself. This allows the user to perform a full-text search of all the metadata available. During the interview, it was also noted that INA does not collect engagement data such as the number of likes, views or how many times a video has been reshared. Although the Institute recognised that this information could be valuable to researchers, the decision was made because keeping this information up to date for the large number of videos INA preserves (more that 26 millions), would have required an even larger amount of resources and time.

5.4.5 Long-Term Preservation

For long-term preservation of the digital cultural heritage archived, the BnF has developed its digital repository, the Scalable Preservation and Archiving Repository (SPAR) system, which involves the combined use of two types of storage element: disk storage for files that need to be accessed, and tape for long-term storage (Fauduet & Peyrard, 2010). In addition to guaranteeing the long-term safeguarding of the websites and social media content archived under electronic legal deposit, the BnF seeks to ensure data interoperability across institutions. In fact, everything that is harvested from the web as part of the *Archive de l'Internet* is currently stored as Web ARChive (WARC) files, the recognised standard file format used by most national web archives to aggregate all the information related to web crawls. Although the WARC file was first specified in 2009, the BnF only adopted this format in 2014. As for videos, these are all downloaded and preserved as .mp4 files.

The *Institut national de l'Audiovisuel* has opted for a different type of file called DAFF (Digital Archiving File format), which has been developed by INA's engineers to meet the specific needs of the Institute. As video files are quite heavy – and with the quality of videos and devices constantly increasing – engineers created a format that would enable INA to save valuable archiving disk space. Essentially, the DAFF format archives separately each piece of content in a record, that is metadata (e.g., URL, download date, etc.) and the original HTML content (Lobbé, 2018; Pehlivan et al., 2021). This file format is interoperable with existing file standards and can be easily converted to and from

WARC files. As for long-term preservation, INA launched in 1999 a Backup and Digitisation Plan for their audiovisual heritage, which includes replication and storage of its collections in digital form across its three data centres.

5.5 Future Developments and Perspectives

Both the BnF and INA are closely monitoring and adapting to the evolution of the web and social media as new features are introduced, and new platforms emerge. At the BnF, curators and engineers of the electronic legal deposit department are looking into finding solutions that could enable them to start archiving content from Facebook again. As mentioned in section 5.4.1, the strict limits imposed by Meta resulted in the web archiving team being unable to crawl data. The BnF is currently experimenting with archiving Facebook from the mobile app. Although these tests are still in their quite early stages, the first few results appear to be promising: the BnF was able to crawl the Facebook home page, showing only four or five latest posts, and it appeared to be displaying correctly once replayed, but they were not able to capture each post page individually. Once fine-tuned, this approach could enable the BnF not only to resume the capture of this platform but also to preserve Facebook as billions of users experience it every day.

Finally, the BnF is part of the *Réseau de Partenaires pour l'analyse et l'exploration de données numériques* [Network of Partners for the Analysis and Exploration of Digital Data] (ResPaDon) project,[128] whose aim is to develop and diversify the use made by researchers of web archives kept at the BnF. To achieve this goal, they are seeking to implement remote access to the *Archives de l'Internet* from a number of universities across France. This will be supported by sharing relevant documentation, data and text mining activities as well as training for students and researchers on how to use web archives. It is hoped that the positive outcome of this experiment will be to spark a change in the law, allowing for less restrictive regulation of access for web archives created under electronic legal deposit.

Regarding potential future development for INA's web and social media archive and the *WebMédia* interface, Mussou commented:

[128] Further information about the *Réseau de Partenaires pour l'analyse et l'exploration de données numériques* (ResPaDon) project can be found here: https://web.archive.org/web/20230603115109/https://respadon.hypotheses.org/les-membres/a-propos

> We are very lucky because we have a team of dedicated engineers who monitor the evolution of websites and update the access interface in accordance, and they also experiment [with tools and features] that researchers might need.

During my visit to INA's training centre at Bry-sur-Marne, one of the engineers gave a quick demonstration of the new interface they were working on at the time. Among the novelties and improvements, they were looking into implementing, the engineers would like to add a higher degree of fidelity for the archived web and social media content, and to also allocate more space in the interface to images and videos. Moreover, they were looking into reorganising and expanding the filters section, enabling users to filter search results more comprehensively. They were also planning to introduce a new feature to allow users to search for tweets that share the same picture: the embedded image has to have, however, the same URL in order to produce results, as this feature does not work in the same way as an image recognition software. Furthermore, the new interface was being built to support a wider range of platforms than Twitter or YouTube, paving the way for a potential expansion of the type of social media archived at INA. But, as mentioned, this is still a work in progress.

Lastly, following Elon Musk's acquisition of Twitter in 2022 (see section 2.4.4), INA had to switch, earlier than planned, to a new version of the Twitter API. In contrast to the BnF where this event has not created any major issue due to the archiving methods and tools they use, Twitter's acquisition and the subsequent changes in the API's terms of use have generated new technical and curatorial issues for INA. In response to a follow up email I sent to INA in March 2023 to understand the impact that this event had on archiving initiatives participating in this study (see Section 1.4.5), Mussou noted:

> Following the decision to shut access to the API V1, we have implemented access to API V2 which limits to 10 million the number of tweets we can collect per month (we used to collect between 20 to 50 million tweets per month).

The new rate limitation has led INA's curators to the decision to stop using the Twitter Timeline API, and no longer collect retweets. They also decided to rely on Twitter's automatic language detection tool to try and circumscribe the collection to tweets published in French, restricting the collection boundaries even further.

Unfortunately, shifts such as those determined by change of platforms' ownership and terms of use, are not new to the history of social media platforms (Le Follic & Chouleur, 2016; Thomson & Kilbride, 2015; van der Vlist et al., 2022). Bruns (2019), for

example, discussed at length the implications of what the author defined as an "APIcalypse" following the 2016 Cambridge Analytica scandal and the unannounced, drastic changes implemented by Facebook to API access and web crawlers rate limits. The consequences of those restrictions still reverberate in the words and examples mentioned by both the BnF and INA, as they are still unable to effectively archive Meta platforms at scale. Nevertheless, it is important to note here that the follow-up email occurred prior to the drastic changes implemented by Elon Musk in 2023 and the establishment of a paid tier system for access to data, the effects of which are still to be fully measured (see section 6.1.4). Still, since the summer of 2023, INA is no longer able to collect content from Twitter using the deprecated APIs.

Circling back to future developments in the social media archive, INA stated in the follow-up that they have no plans to initiate any capturing activities on Mastodon, the platform to which many Twitter users appeared to be fleeing in the early days of Musk's take over, commenting that "not many people are actually publishing on the platform even though accounts have been opened". Moreover, Mussou felt the need to reiterate how:

> As a heritage institution, we are still struggling to gain access to other social networks, notably Meta's accounts and publications on their platforms.

Although the changes brought to the access to and use of the Twitter APIs has generated new challenges that add to the existing ones, memory institutions such as INA appear to perceive the difficulties encountered in capturing and establishing collaborations with social media platforms like those owned by Meta as one of the main obstacles to social media archiving.

Conclusion

The BnF and INA share the mandate to archive, preserve and provide access to the digital cultural heritage published on the French websphere, including websites and social media platforms. While web archiving practices have consolidated over almost 20 years of activities, the development of social media collections related to France has brought to light new challenges. The complexity of the national legal framework, data protection and the intersection with restrictions imposed by social media platforms heavily shape social media collections at both electronic legal deposit institutions. The BnF and INA have developed strategies such as networks of contributors (BnF) and the combined use of

different tools (INA) to identify heritage materials to be archived in order to mitigate biases and ensure representativeness of collections on a national scale. However, ethical concerns coupled with a variety of technical challenges, mostly concerning lack of high-fidelity captures, limitations on access to data, and the rapid evolution of platforms' interfaces and terms of use, represent further impediments to social media collections. Thus, memory institutions find themselves constantly striving to keep up with new solutions that, once implemented, are often already obsolete.

Although both French institutions are working toward fulfilling the same legal mandate, the distinct priorities reflected in the BnF and INA's mission statements (section 5.1) seem to partially explain the adoption of different approaches in terms of harvesting (Heritrix vs. APIs) and access (high-fidelity vs. raw data). Nevertheless, as Faye et al. (2024) emphasised, the implementation of different collection and consultation methods has enabled both French institutions to capture materials that, despite some expected overlap, largely complement each other, capturing a plurality of voices and perspectives related to specific events or phenomena. This combined effort offers researchers a diverse range of sources and research approaches, as exemplified by the tools and support offered by both the BnF and INA Labs, potentially leading them to the exploration of new research paths and increasing their engagement with social media collections.

The case study of the BnF and INA sheds light on the varied nature of social media archiving practices – even within the same country and legal framework – and on how the concept of "best practices" in this context may vary from institution to institution, depending on their mission and user requirements. The next chapter will further explore some of the themes and challenges highlighted in the two case studies, offering a comparative analysis of the practices, obstacles and opportunities that emerged through interviews with twelve memory institutions at different stages of development of their social media archiving initiatives.

CHAPTER SIX

Comparative Analysis of Social Media Archiving Initiatives: Challenges, Practices, and Opportunities

This Chapter aims to offer a transnational comparative analysis of the key challenges and aspects that emerged from interviews and fieldwork carried out with twelve memory institutions at various stages of their archiving projects (see Section 1.4.4 and 1.4.5). It will consider how national legal frameworks, platforms' terms of use, and archiving institutions' preservation strategies and decision-making processes may be shaping existing social media collections and affecting the potential development of new, emerging initiatives. The following discussion is based on data gathered through interviews and fieldwork, and documentation produced by institutions and openly available on their official websites or kindly shared by the interviewees, such as collection development policies, reports, journal articles and blogposts. As described in Section 1.4.5, semi-structured interviews were conducted with twelve archiving initiatives between April 2022 and September 2022. A series of follow-up emails were also exchanged with interviewees in March 2023, to capture their reaction to Twitter's acquisition and subsequent changes implemented between 2022-2023. A complete list of the interviewees is included in Appendix D.

6.1 Legal Challenges and Ethical Concerns

This section begins by exploring how the combination of different layers of national legal frameworks influences social media archiving activities. While in most countries the national electronic legal deposit legislation enables institutions to archive the national web domain, such regulation also imposes specific limitations, affecting for instance the granularity of collections and both access to and re-use of material archived (Ogden & Maemura, 2021; Winters, 2020). Drawing from the analysis of clauses defining born-digital heritage materials that fall under national e-legal deposit legislation, I highlight the impact that these may have on the development of web and social media archiving initiatives. This is followed by a discussion about additional obstacles stemming from data protection and copyright law, illustrating the diverse approaches taken by various institutions in addressing these challenges. Lastly, social media companies impose several

legal constraints on access to data and consequently on archiving practices. In particular, the latest Twitter "APIcalypse", borrowing the term coined by Bruns (2019) to describe the severe restrictions implemented by Facebook after the Cambridge Analytica scandal (Isaak & Hanna, 2018; see Section 2.4.4), offers a good example about the power that social media companies exert on access to information and the state of uncertainty that social media archiving initiatives face in navigating the rapidly evolving landscape of these platforms and their policies.

6.1.1 National Legal Frameworks

National legal frameworks significantly shape the development of social media archiving initiatives, whether they are long-term endeavours, pilot projects or are exploring potential pathways for new undertakings. Among the institutions interviewed, many cited national legal frameworks as a primary concern, highlighting how outdated legislation or inadequately worded clauses often lead to an impasse that can affect the very existence of social media archiving initiatives and the nature of their collections (see Section 6.1.2). Moreover, national legal frameworks establish the boundaries within which cultural heritage materials can be lawfully archived, as well as how archived resources can be used and reused (Cunnea et al., 2020; Stirling et al., 2012; Winters & Prescott, 2019).

In Section 3.5, I illustrated the impact of the absence of specific legal provisions, such as electronic legal deposit legislation, which explicitly include or, at least, allow space for the inclusion of material published on the web and social platforms. The lack of legal clarity can constitute a significant barrier to the establishment of social media archiving initiatives at an institutional level. For instance, in Italy, national deposit institutions are still waiting for approval of the regulation provided for by art. 37 D.P.R. 252/2006, which would finally introduce the digital legal deposit of "documents disseminated through the web" deemed culturally significant and publicly available (see Section 3.5.1). Conversely, in Lithuania, the government is contemplating the inclusion in the *Dokumentų ir archyvų įstatymas* [Law of Documents and Archives][129] of provisions concerning information shared on the web. However, an official timeframe has not been set yet to amend the existing regulation (see Section 3.5).

Provisional discussions about the introduction of electronic legal deposit legislation appear to be a determining factor in motivating national memory institutions to start thinking about or planning the development of a social media archiving initiative.

[129] All translations in this chapter have been translated using the Google Translate tool.

The effectiveness of such discussions is often amplified and supported through pilot projects carried out by national memory institutions to demonstrate the feasibility of the archiving endeavour, while also highlighting any potential challenges and safeguards that forthcoming regulation will need to address and establish. Many are the web archiving efforts that have started with often small, test cases undertaken by national libraries or archives prior to the introduction of a specific legal remit (Koerbin, 2017; Webster, 2017). Institutions like the National Széchényi Library (NSZL) in Hungary and the Royal Library of Belgium (KBR) have adopted similar approaches to test the viability and sustainability of social media archiving. Both initiatives serve as good examples of how short-term or pilot projects can help make the case for a change in the existing national legal deposit legislation. In particular, the two-year project BESOCIAL[130] was launched in 2020 at KBR with the aim of "creating a sustainable strategy for archiving social media in Belgium […], also making the whole concept of born digital data part of our cultural national heritage" (Geeraert et al., Interview). Friedel Geeraert, one of the curators involved in the project and Expert in Web Archiving at KBR, explained that while BESOCIAL was coming to an end at the time the interview for this study was carried out, the efforts made in demonstrating the feasibility and value of a social media archiving initiative played an essential role in convincing the director of the KBR about the need to establish a national social media archive. However, securing resources and a legal mandate were identified as essential prerequisites for its realisation.

In fact, Belgium does have legal deposit legislation concerning electronic materials, but it only provides for copies of digital works to be deposited on a physical carrier by publishers or authors (Geeraert et al., Interview). The traditional deposit system used for both paper and digital publications, such as newspapers and journal articles, could not be applied to the complex and dynamic environment of social media platforms. Therefore, the existing law had to be updated for the KBR to directly harvest content from social media as part of their legal deposit mandate. In this regard, the BESOCIAL researcher clarified that while "the law has been changed in the meantime to also include electronic publications on non-material carriers, […] it still needs to go into the Royal Decree" which had not been published yet at the time of the interview (Geeraert et al., Interview; see also Van Camp, 2020). The Royal Decree will enable the practical implementation of the new law. It is important to underscore the active role that the KBR,

[130] https://web.archive.org/web/20240104174025/https://www.kbr.be/en/projects/besocial/

in collaboration with the Belgian Policy Office (BELSPO),[131] had in revising the legal deposit legislation and the royal decrees stipulating the mission of the Royal Library itself to implement crucial changes that could finally allow for the inclusion of web and social media material (Van Camp, 2020).

Similarly, the web and social media archiving project established at the National Széchényi Library (NSZL) in Hungary began as an experimental web archiving effort while waiting for the legislation to be approved by the government. However, only after the introduction of the new law in 2021 was the NSZL able to start planning and developing a long-term strategy for the safeguarding of information shared on the national TLD and web-based material concerning Hungary (Geeraert & Németh, 2020; Németh, 2020; Németh et al., Interview).

Some of the longest-standing web and social media archives also began with small web archiving pilot collections – usually on a permissions basis – long before the introduction of national e-legal deposit legislation. For instance, the BL created a few special web collections to capture events like the 2005 UK General Election and the July 2005 terrorist bombings in London. These early tests provided the BL with the opportunity to advocate for the value and the fragility of the information shared on the web, paving the way for consultations and the entering into force of the NPLD legislation in April 2013 (DPC, n.d.-b).

The introduction of national electronic legal deposit legislation appears to be essential in consolidating the various early web archiving experiments developed over the years. This is exemplified by the Danish *Netarkivet*, the *Bibliothèque nationale de France*, the *Institut national de l'Audiovisuel* (see Section 5.2), and the British Library (see Section 4.2), where updated legal deposit legislation was introduced respectively in 2005, 2006, and 2013. When the mandatory archiving law passed in 2004 and came into effect in 2005 in Denmark, the *Netarkivet* at KB became one of the first institutions to receive the legal mandate to preserve the content made publicly available on the national web domain and other material published on the web by or about Danish people, later allowing for the inclusion of social media platforms (Laursen & Møldrup-Dalum, 2016). Receiving a legal mandate represents a crucial step not only to officially preserve web and social media, thus recognising their cultural value, but also for obtaining indispensable resources to develop further and sustain in the long-term such complex archiving effort.

[131] https://web.archive.org/web/20240204200719/https://www.belspo.be/belspo/brain2-be/index_en.stm

Nevertheless, archiving activities involving the capture of social media platforms are not solely dependent on the existence of legal deposit legislation. National legal frameworks often provide for other types of archiving institutions to preserve information produced by central government and its agencies. This is the case of the UK National Archives (TNA) and Archives New Zealand (ANZ). The Digital Preservation Analyst at Archives New Zealand explained that the ANZ is "legally obliged to capture all of the transactions between government agencies and the public as they are all public records" (Ng, Interview), thus including also communications occurring on social media platforms between said agencies and the public. Analogously, The National Archives of the United Kingdom (TNA) states on the UK Government Web Archive (UKGWA) website that its mission is to "capture, preserve and make accessible UK central government information published on the web".[132] These mandates clearly exempt these institutions from the obligations of legal deposit institutions. The aim of the latter is to capture a snapshot of the born-digital cultural heritage disseminated across, but not limited to, the national web domain as a whole. Conversely, memory institutions such as TNA or ANZ are required to preserve records produced by government agencies and public figures (e.g. civil servants) in order to adhere to a two-fold mandate: safeguarding a record of government activities and promoting transparency by making archived material, including content published on social media, freely available to the public (Espley et al., 2014). In fact, the UKGWA, for instance, is not only fully accessible through its portal but it is also open to search engines (Espley et al., 2014). In contrast, legal deposit institutions are often subject to stricter regulation, particularly concerning access to the archived material (see Section 6.4).

It is worth noting here, however, that the UKGWA constitutes a somewhat different experience from the ANZ. The ANZ reportedly mainly preserves what is periodically transmitted to it by single governmental departments or agencies (Ng, Interview). Preparing social media content to be transmitted to Archives New Zealand is often an added task to an already busy schedule for many departments and therefore limited, until 2022, to sporadic screenshots (often stored as PDFs) of web and social media material that is deemed particularly noteworthy. Conversely, while the UKGWA is obviously part of TNA, it appears to operate separately from the main archive. Public records – no matter the format – are generally transferred by government and other

[132] https://web.archive.org/web/20240428164459/https://www.nationalarchives.gov.uk/webarchive/

organisations subject to the Public Records Act[133] to TNA for permanent preservation. Information generated on social media or websites does not appear instead among the types of born-digital material that are subject to transfer to TNA.[134] While this can be ascribed to technical constraints and the need of specialised resources to crawl sites that transferring bodies might not possess, it clearly separates the two archiving activities, despite being part of the same institution. The UKGWA, in fact, directly captures and preserves websites and some social media accounts of central government, collecting snapshots of them throughout their lifespan.[135]

An alternative to the two legal scenarios described above – both offering a solid legal ground to start preserving consistently and for the long-term social media platforms – is the Portuguese web archiving initiative. The manager of the service, Daniel Gomes, explained that Arquivo.pt[136] started its archiving activities around 2007, building on a previous and successful pilot project led in 2001 by a group of researchers from the Faculty of Sciences at the University of Lisbon. Unlike other countries located in the European area, the Portuguese web archive cannot count on any electronic legal deposit legislation or other similar regulation, since Arquivo.pt is neither part of a national library nor a national archive. It is, however, embedded in the Portuguese national research and education landscape but not linked, for example, to any university library, therefore setting this archiving initiative apart from any other web archive described so far. Nonetheless, Arquivo.pt has received from the Portuguese government the mandate to capture, preserve and provide access to the Portuguese digital heritage published on the web. Far from being a limitation in this case, the absence of a specific piece of legislation that would regulate both scope and access has allowed Arquivo.pt to archive a large variety of born-digital information, including social media sites, as long as they are publicly available on the web and are related to and of interest to the Portuguese community. Moreover, thanks to this special mandate, Arquivo.pt can provide open access to the archived web material through its online portal, with no restrictions to physical premises or geographical borders (see also Section 6.1.3). However, Gomes (2017) clarified how such free access has to be

[133] https://web.archive.org/web/20240409082136/https://www.legislation.gov.uk/ukpga/Eliz2/6-7/51

[134] See "What are born-digital records":
https://web.archive.org/web/20240404224211/https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/what-are-born-digital-records/

[135] https://web.archive.org/web/20240117042759/https://www.nationalarchives.gov.uk/webarchive/about/

[136] https://web.archive.org/web/20240424100505/https://arquivo.pt/?l=en

provided in a way that respects authors and their activities, establishing, for example, specific guidelines to protect rights-holders and imposing an embargo in order not to compete with the live sites (Gomes, 2017; Vlassenroot et al., 2019).

Another case that is worth mentioning in this context is the Library of Congress (LoC), a research library that officially serves the United States Congress (Zimmer, 2015). The US legal framework does not envisage a specific legislation like the electronic legal deposit regulation introduced in other areas of the globe. However, as the national library of the United States, LoC has been capturing content from social media as part of its pre-existing web archive in the fulfilment of its mission to preserve the history of the American people, although this has not been without challenges (see Section 6.3).

As illustrated in this section, receiving a mandate – whether through the introduction of specific legislation or directly from the government – appears to be the *conditio sine qua non* to officially collect and preserve public web and social media content at an institutional level, thus recognising their cultural value. Moreover, such a mandate is essential for obtaining the indispensable resources needed to further develop and sustain in the long-term such complex archiving efforts.

6.1.2 A Matter of Definition: e-Legal Deposit Legislation and the Scope of Social Media Collections

The mere existence of a piece of legislation that provides national archiving institutions with a legal mandate to archive content published on media other than paper, in the interest of preserving digital traces of a certain country, does not ensure the ability of national deposit institutions to automatically start archiving material from the web and, specifically, from social media. Digital technologies have evolved and continue to advance quite rapidly, while governments and lawmakers often fail to prioritise addressing challenges linked to the variety and increasing complexity of the digital world and its preservation in the long-term. Moreover, the manner in which existing legislation is phrased can pose additional obstacles to the preservation of the web and social media platforms (Gooding & Terras, 2020a). A strict or too broad definition of the type of born-digital material that is to be preserved could generate uncertainty among web archivists, hindering the preservation of culturally relevant content shared on the web (see for example, Gooding & Terras, 2020; Winters, 2020).

As illustrated in Section 3.5.1, the case of Italy offers an interesting example about how the introduction of a far too specific definition in the 2005 Digital Administration Code regarding what qualifies as "digital document" can be identified as one of the

contributing factors to the absence of a national web and social media archiving initiative at scale, particularly concerning the public sector. In fact, Italian legislators appear not to have adequately considered the wide array of existing types of digital records that may constitute an expression of activities or communication between the public and government agencies, and their potential future evolution. This narrow definition has indeed precluded national archiving institutions, specifically the *Archivio Centrale dello Stato*[137] or any other deposit libraries, from systematically including, for example, government websites or official social media accounts in a broader national preservation strategy (Storti, 2023).

An analysis of the electronic legal deposit legislation related to interviewed deposit institutions revealed the tendency of these legal texts to opt for broad definitions of the type of content that is to be considered in scope, to allow space for interpretation and remain open to future technological developments without requiring constant updates of the legislation. For example, when the Danish Act on legal Deposit of Published Material (Act no. 1439 of 22/12/2004) passed in 2004, Denmark gathered into a single act all regulations about preservation of works published in Denmark, regardless of format (Larsen, 2005). In particular, chapter III article 8 of the aforementioned 2004 Danish Act calls for the submission of all "Danish material published in electronic communication networks",[138] further specifying that "Danish material" meant content that "is published from internet domains and the like that are specifically assigned to Denmark, or that is published from other internet domains and the like and is aimed at an audience in Denmark".[139] Interestingly, the 2004 legislation also prescribes the ways in which authors or creators should submit the material, such as submitting copies or facilitating access to content by providing access, for example, to material behind paywalls or any other information necessary to allow the deposit institutions to fulfil their mandate. To clarify, this pertains solely to public material, such as online newspapers, and does not include private content protected by password on social platforms. Like other legislation in the European area, the 2004 Danish Act makes use of the word "published" [140], which can

[137] The *Archivio Centrale dello Stato* [Central National Archive] is the archiving institution designated by the *Codice dei Beni Culturali* [Cultural Heritage Code] to safeguard and ensure access to documents produced in the fulfilment of the central government's activities.

[138] ACT no. 1439 of 22/12/2004, article 8. [Text translated using Google Translate] Available here: https://web.archive.org/web/20240224042201/https://www.retsinformation.dk/eli/lta/2004/1439#K3

[139] ibidem

[140] See definition provided in Brügger (2017), p.186: "Publication=making available to the public".

here be interpreted in the sense that the material must be made accessible to the public in electronic "communication networks", meaning any systems or structures that facilitate communication or exchange of information among groups of people. Despite the discussions and important ethical concerns surrounding the meaning of "published", especially in the context of social media, and the degree of users' awareness on the matter discussed in Section 2.4.2, the definition provided in the 2004 Danish Act clearly allows for the inclusion of social media platforms in the *Netarkivet* collection scope (Larsen, 2005; Schostag & Fønss-Jørgensen, 2012).

In the UK, the Legal Deposit (Non-Print Works) legislation defines the digital material that the six deposit institutions are required to archive and preserve as a "work published in a medium other than print […] that is published on line"[141] but expressly excluding material that consists only of sound and/or video recording. Again, such a broad description has allowed the British Library and its partners to include content from some social media platforms as part of their pre-existing web archive collections. However, the explicit exclusion of material primarily consisting of sound or video recordings has led to the BL's inability to collect content from platforms like YouTube or TikTok, or any future platforms that are predominantly based on these types of media (see Section 4.4.1).

Probably the broadest definition among those examined for the purposes of this study is the one formulated by the French lawmakers. Using a definition that appears to be inspired by De Saussure's semiotic theory of structuralism (Saussure, 2006), the French electronic legal deposit legislation goes into the structural elements of online communication. It includes among the material to be preserved under electronic legal deposit any kind of "*signs*, *signals*, texts, images, sounds or messages" publicly transmitted through the Internet. Such an extremely comprehensive definition has been essential for the BnF and INA to add to their web collections content collected from a wide variety of social media platforms, also leaving sufficient space for any other type of communication format that may be developed in the future. These *signs* and *signals* must however be made publicly available on the web and be published in or be related to France (see Section 5.2).

[141] The Legal Deposit Libraries (Non-Print Works) regulations 2013, 13(1). Available here: https://web.archive.org/web/20240506154952/https://www.legislation.gov.uk/uksi/2013/777/regulation/13/made?view=plain

Seemingly inspired by the above, in Luxembourg the digital legal deposit legislation introduced with the Grand-Ducal regulation of 6 November 2009,[142] and modified by the Grand-Ducal regulation of 21 December 2017, identifies among the works published on the national territory and subject to legal deposit at the *Bibliothéque national du Luxembourg* (BnL), all those "publications without material support made available to the public through an electronic network […] as well as all *signs*, *signals*, writings, images, sounds or messages of any kind". Although the BnL is still experimenting with social media archiving, the legislation at hand clearly provides for collection at scale of public content shared on social sites and any future evolution of the sort.

By contrast with the above, the legal framework (comprising the Public Records Act 1958[143] and the Copyright Designs and Patents Act 1988[144]) under the terms of which the UK National Archives operates instead, has allowed the UKGWA to start capturing government websites and social media with no need to modify the existing legislation (Espley et al., 2014; Winters, 2020). In particular, as Espley et al. (2014) underlined in their article tracing the early social media archiving approaches at the UKGWA, the Public Records Act 1958 "purposefully negates to apply a format definition to the record and suggests that the keeper take action to ensure the preservation of any and all records" (p.34). Moreover, the Public Records Act further clarified that by "record" is to be intended "not only written records but records conveying information by *any other means* whatsoever" (Public Records Act 1958, s.10; see also Espley et al., 2014; Winters, 2020).

It appears crucial then that future legislation addressing the preservation of digital items and the web should keep in consideration the unpredictability of their evolution and forge sufficiently broad definitions. Policymakers need to set boundaries within which national archiving institutions can lawfully operate and archive material published on the web in order to safeguard the rights of private citizens as well as rights-holders' commercial interests (Muir, 2020). However, lawmakers and consultants involved in the drafting of, for instance, updated national legal deposit regulations must choose words

---

[142] *Règlement grand-ducal du 6 novembre 2009 relatif au dépôt legal, modifié par le Règlement grand-ducal du 21 décembre 2017*. The full-text of the regulation is available here: https://web.archive.org/web/20240903135904/https://legilux.public.lu/eli/etat/leg/rgd/2009/11/06/n8/jo

[143] https://web.archive.org/web/20240409082136/https://www.legislation.gov.uk/ukpga/Eliz2/6-7/51

[144]https://web.archive.org/web/20240303183100/https://www.legislation.gov.uk/ukpga/1988/48/contents

wisely as any word used to define born-digital objects in scope has the power to include or exclude digital artefacts from web archive collections and define the digital memory that deposit institutions will be able to transmit to future generations.

Among the common threads of challenges emerging from the analysis of existing electronic legal deposit legislation is the fact that the digital material must pertain to a territoriality criterion. In Sections 4.2 and 5.2, it has been noted that for websites, provenance is usually determined by verifying whether the site belongs to the country code top-level domain (ccTLD), including its sub-domains, assigned to a certain country, or whether the owner of the site resides in that same country. However, when it comes to social media platforms, which usually adopt a .COM extension (e.g., twitter.com, facebook.com), ascertaining the geographical location of content or the place of residence of its author can pose significant challenges when harvesting content at scale, especially when the content has been shared and re-shared several times (Bingham & Byrne, 2021). Consequently, while having a legal mandate can facilitate the development of social media archiving initiatives – provided that the definition formulated in the legislation is broad enough to allow the inclusion of such content – this legislation often comes with a series of constraints that heavily shape the resulting collections, generating gaps in the preserved digital culture.

Almost all of the institutions I had the chance to interview mentioned how abiding by a territoriality criterion represents one of the first challenges that archiving institutions have to face when starting to plan the development of social media collections. Ben Els from the BnL observed how from "a legal point of view [archiving social media] is quite tricky" (Els, Interview). Els mentioned the mandate the BnL has for publications made available on the web in the territory of Luxembourg, expressing doubts about the meaning of such criteria when applied to social media and how this "grey zone" is often a source of concern among web archivists. This has also been confirmed by other long-standing institutions, such the BL, INA and the BnF. Nevertheless, these institutions have established their own additional selection parameters and solutions matured over ten years of web archiving activities, enabling them to sift through and identify relevant social media content within the boundaries set by their respective legal frameworks (see Section 6.2).

Still, limiting collection to only one country's domain or geographical borders risks excluding the transnational aspect of events and conversations unfolding on social platforms, which is a uniquely valuable trait (Huc-Hepher & Wells, 2021; Schafer et al., 2019). As Bingham and Byrne (2021) noted, the legal constraints placed on web and social

media collections inevitably generate gaps, as conversations among users from across the globe are partially captured. It is worth noting that these gaps are the result of a combination of legal, curatorial and technical factors, with legal constraints – including deposit regulations, data protection laws and social media company policies – playing a major role (see Section 6.1.4).

6.1.3 "A Grey Area": Balancing Data Protection Laws, Copyright Concerns and the Right to Information.

National legal deposit legislation is just one variable in the complex equation that shapes social media collections. Because of the sheer amount of personal information that can potentially be found on social sites, archiving institutions must carefully consider the content they appraise for preservation to respect the privacy of individuals and third parties whose information may be included, for instance, within public figures profiles (Windon & Youngblood, 2024). For most archiving institutions interviewed this constitutes a grey area to navigate due to the difficulties in drawing a solid line between public and private spheres on social media for many users. For example, the UK National Archives (TNA), which preserves government-related social media profiles in the context of the Freedom of Information Act 2000 (FOIA),[145] noted that for certain public figures they had to face the uncertainty of how to handle profiles that share a mix of public and personal content. In this regard, Claire Newing, web archivist at TNA, commented:

> We capture Twitter's channels of secretaries of state, but we don't publish those at the moment, there's a question over […] the fact that they are tweeting about things as a secretary of states, but there are also other tweets which they've created as a private citizen.

Although the user may publish this content on social media with no privacy restrictions attached, the blurred line between "public" and "private" often generates legal and ethical dilemmas that even legal teams may struggle to disentangle (see Section 2.4.2). This uncertainty is the cause of many concerns, especially among memory institutions that are still in the planning phase, as they are unsure about how to properly handle this kind of situation. Problematic examples like the one mentioned above seem to arise particularly during the capture of local or national elections. Memory institutions often use these

[145]https://web.archive.org/web/20240302170644/https://www.legislation.gov.uk/ukpga/2000/36/contents

recurring political occasions as test-events to experiment with and assess the feasibility of social media archiving efforts, given the resonance they have across the whole country or region and the wide engagement registered among the population. For example, Ben Els from the BnL explained that while collecting material concerning a local election, there were some cases in which certain candidates would have on their social media accounts, especially on Facebook, an intertwined mixture of posts concerning their private life on one side, and public, elections-related content, on the other: "In many cases we had a mix of vacation photos, family photos and maybe some pictures to say: 'here I am putting up posters for the campaign'." (Els, Interview). Similar cases may cause concerns due to the uncertainty over whether third parties have granted permission to share those images publicly.

Some legal deposit institutions such as the BnF, INA, the BL or the *Netarkivet* are exempted from requesting permission to collect publicly available content. However, no specific provisions exist in the current legislation on how to deal with such a mixture of personal and public content. Moreover, as data protection laws classify political opinions and/or affiliations as "special category data", web archivists must exercise extra caution with content shared, for example, in relation to election campaigns. Collection development policies at single institutions could introduce additional ethical guidance for similar instances. Still, manually assessing and filtering every single item across hundreds of social media accounts would obviously require extra time and resources that many institutions simply do not have, potentially leading to further scaling down existing archiving initiatives.

Moreover, due to the large, automated scale at which most archiving institutions operate, and despite the efforts made by web archivists to carefully select only publicly available content for collection, a small, occasional amount of content shared on social media by private citizens can end up in web and social media collections. As mentioned, in the context of data protection law, national archiving institutions can benefit from specific exceptions for archiving purposes in the public interest (Michel, 2021). Nevertheless, to guarantee individuals' rights and freedoms, national web archiving institutions have implemented specific takedown policies. This means that the interested party can raise a complaint and ask, in exceptional circumstances, for the material to no longer appear in the archive. Some institutions offer the opportunity to send a notification by simply filling out an online form which is often available on the institution's portal (e.g., UKGWA), or by sending the takedown request via email (e.g., the BnF). As observed

by some of the institutions interviewed, such as the BL and the *Netarkivet*, the number of complaints and takedown requests have been quite low so far, with many archives yet to receive requests related specifically to social media content. While this can be seen as a positive outcome resulting from a combination of strict legal provisions and meticulous selection practices by web archivists, it cannot be excluded that the lack of awareness among the public about the existence of such collections might also have contributed to the limited number of takedown requests.

Furthermore, the act of "taking down content" from web archives has more to do with hiding it from the public view than actually deleting it from storage. As the Senior Digital Collection Specialist at the Library of Congress noted: "We honour take down requests, but we always keep the content so we can keep the integrity of the work. But we put it on the restricted list on the Wayback level" (Lyon et al., Interview). Likewise, the BnF mentioned how, despite not being required by law to do so, they honour the GDPR's right to erasure through their takedown policy, blacklisting the content that data subjects ask to be removed. Furthermore, INA explained that no material is hidden from the archive unless they are instructed to do so by their legal department (see Section 5.4.1). In fact, when receiving takedown requests, these are usually examined by each institution's legal teams to assess their legitimacy before proceeding with hiding content from access.

The takedown procedure has been the subject of an interesting discussion at KBR. One of the researchers working on the BESOCIAL project recounted how the legal advice they were given while developing the procedure was to always question whether the takedown request was related to providing access to the material archived or was instead a complaint about preservation:

> The legal advice that we obtained they recommended us for each case to look at two different aspects: is it a complaint about providing access, or is it a complaint about preservation? For preservation the right to information will prevail more frequently over the right to privacy, whereas for access it would be the other way around, where people would have better grounds to argue that the right of privacy was not guaranteed. The legal analysis also showed that we could argue that we should not delete the content indefinitely, but rather that we should not make it visible anymore. So, it isn't a request for deleting things, but more of not providing access to that kind of content anymore. (Geeraert et al., Interview)

The distinction between the right to capture public content from social media for cultural heritage preservation purposes and providing access to it is indeed crucial from a legal perspective. It is worth noting, however, that none of the interviewed institutions has specified for how long the web archived material will remain hidden or if it will ever be made accessible again in the future.

Copyright is another aspect that archiving institutions must consider when collecting content from social platforms. This includes not only user-generated material but also the platform's logo, the design, and the code captured on each site alongside the content (Hockx-Yu, 2014). Most national electronic legal deposit legislation analysed for the purposes of this study seems to provide for specific exceptions to national copyright regulations. In particular, exceptions concern the ability granted to archiving institutions to copy content in scope under electronic legal deposit legislation for preservation purposes (Gooding & Terras, 2020a). As is also explained in Section 5.2, the act of archiving material from the web, whether websites or dynamic platforms, entails producing a copy of the original item made available on the live web (Brügger, 2018b; Chambers et al., 2021; Sepetjan & Graff, 2011). Thus, the copyright exception described above has been essential to allow web and social media archiving activities in most deposit institutions, such as the BL, INA and the BnF. Furthermore, exceptions to copyright law for research or educational purposes may offer opportunities for the development of test collections. For example, in Belgium this specific copyright exception has played an essential role in allowing the BESOCIAL project at KBR to build a sample collection and advocate for the importance of preserving social media. However, the material collected as part of the BESOCIAL project may not be reusable at the end of the project, meaning that the born-digital cultural heritage gathered in this context will likely not be included in any future collection created after the introduction of the e-legal deposit legislation.

In the intersection of copyright and data protection laws, an additional challenge may arise for smaller archiving initiatives that do not benefit from e-legal deposit provisions exempting them from requesting permission from the rights-holders. Requesting permission to archive websites from their rightful owner has always been considered a difficult endeavour to fulfil, even in the late 1990s when the number of websites was considerably lower compared to now (Grotke, 2017; Koerbin, 2021). In the context of social media, seeking the authorisation to harvest content from single users appears even more burdensome given the billions of users active daily and the millions of

images, comments and status updates generated per minute.[146] Often e-legal deposit legislation exempts deposit institutions from having to request permission to archive copyrighted material. In France, rights-holders cannot object to the copying of their material under legal deposit as it is necessary for the collection and preservation of content (Chambers et al., 2021; Sepetjan & Graff, 2011). Likewise, deposit institutions in the United Kingdom, Luxembourg and Denmark do not require any authorisation from the rights-holder for archiving in scope material from social platforms. In Hungary, the recently implemented deposit legislation has followed this same path already traced by most deposit legislation in other European countries.

On the opposite side of the spectrum sit memory institutions that do not benefit from any legal deposit mandate and do not copy copyrighted web material for research purposes, and therefore need to find their own approach to lawfully archiving content from social media within the complex set of boundaries established by national legal frameworks. An interesting example in this context is the case of the Museum of London (MoL), which has been preserving the social and urban history of London and Londoners over time through the collection of more than seven million objects.[147] In particular, the Museum seeks to collect stories and memories from living people. Thus, when an important sports event like the Olympics took place in the UK's capital in 2012, curators at the MoL sought the unique opportunity to capture, among other memories of the event, Londoners' reactions and interactions on social media, specifically on Twitter. The MoL Curator, Foteini Aravani, explained that the decision to carry out their first web collecting experiment on Twitter was mainly due to its openness and the ability to capture "the opinions and the voices and the stories of Londoners". Since this was their first attempt at capturing social media, the Curator noted how they wanted to approach this new, unexplored territory for the MoL in the safest way possible from a legal perspective. In fact, some of the main concerns they had to face when planning the collection were related to data protection and copyright, from both users and platforms. To mitigate these, the MoL decided to request permission from a selected number of users to collect

[146] Stats provided on Statista.com concerning the amount of content uploaded by users per minute on social media platforms shows that, for example, in December 2023 users posted around 360,000 tweets on Twitter, and liked more than 4,000,000 posts on Facebook. More information available here: https://web.archive.org/web/20240822190615/https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

[147]https://web.archive.org/web/20240117054237/https://www.museumoflondon.org.uk/collections/about-our-collections/what-we-collect

their tweets about the 2012 London Olympics. Through an open call organised in collaboration with Westminster University, participants were invited to apply as one of the Twitter users who would tweet using hashtags (e.g., #citizencurator) and be captured as part of the London Olympics collection. The shortlisted 18 participants were asked to sign an agreement with the MoL stating that they would tweet between five to 10 tweets per day about the Olympics. Moreover, the use of the hashtag #citizencurator would further indicate participants' permission for the MoL to collect the tweet in which the said hashtag was included, and as Aravani remarked "this was for [the MoL] a way to deal with copyright".

This permission-based approach was made necessary due to the Museum of London not being able to benefit from the same copyright exemptions as legal deposit institutions. Aravani explained the complexities of navigating multiple layers of copyright, involving both users and the platform. In fact, the MoL decided to also contact and request permission from Twitter Inc. itself for the use of its logo and platform design:

> The second layer of copyright sits with Twitter itself because they own the copyright of the logo, the design of the platform and […] that was one of the reasons why we got in touch with Twitter: to get permission to collect the screenshots having their logo on there as well (Aravani, Interview).

While recruiting participants and seeking permission from Twitter were certainly challenging, Aravani noted that the main obstacle resided in the significant time and effort needed to set up both phases, and in particular the prolonged wait for a response from the social platform. Nevertheless, this approach was feasible for a small collection effort focusing on a limited number of items, especially considering that the tweets collected were published by users who were aware and conscious of their tweets being collected. Seeking permission and establishing copyright agreements on a larger scale would be impossible to implement, as it would require an even larger number of resources and time.

In light of the MoL example, having a legal mandate providing for ad hoc exemptions, thus facilitating archiving activities for legal deposit institutions, appears to be crucial for the development of social media archiving initiatives at a national level. Moreover, the efforts to which the MoL went to ensure social media data was lawfully collected, strictly abiding by copyright and data protection provisions, might explain why not many other institutions in the GLAM sector other than national libraries and archives have, so far, either included or planned to include collections or exhibitions involving large amounts of user generated content from social media. Moreover, in the past decades,

undertakings like the Mass-Observation Project (MOP) showed how relying on permission and voluntary participation can lead to biases and undermine the overall representativeness of the material collected. Similarly to the social media collection at MoL, MOP's aim was to capture the perspective of "ordinary" people on various aspects of contemporary social life, although on a much larger scale. Pollen (2013) described in length the debates and concerns raised by MOP in terms of representativeness and the "self-selection" of contributors. Similar objections could be moved to MoL, as recruiting participants prior to the creation of content may filter the spontaneity out of conversations and reactions to events unfolding on social media. However, as Aravani clarified, the 2012 Olympics collection was "an experimental project" to test how MoL could practically collect tweets. Therefore, building on an argument made about MOP (Sheridan, 1937 through Pollen, 2013), individuals selected for MoL's first social media collection test could somewhat be considered as "participants in" an experiment, selected specifically to reflect on a particular event. In fact, the approach changed for subsequent collections. For instance, for the "Going Viral"[148] collection dedicated to Londoners' experience of the first COVID-19 lockdown, curators identified individual tweets for acquisition after they had been published (see also, Section 6.4). This time, they collected spontaneous reactions to the event, securing permission from the rights-holders before including them in the collection. Documenting appraisal and acquisition choices as well as workflows is essential – particularly in similar cases – to fully understand collections development and their potential biases.

As illustrated so far, national electronic legal deposit legislation facilitates preservation of web and social media materials. However, this enabling legal framework also sets restrictions that severely affect access to the resulting collections (Schafer & Winters, 2021; Winters, 2020). Despite being essential to protect individuals' rights, the interaction between the different layers of legal constraints set by national legal frameworks is a primary aspect that archiving institutions looking into developing their own social media collections must consider, especially those still in the planning phase. As Fien Messen, one of the researchers involved in the BESOCIAL project at KBR, noted "the whole legal aspect is quite complicated, […] as we have to find a balance between the right to information and the right to privacy — and that's [often] really a case-by-case

[148] "Museum of London acquires 'viral' Tweets for Collecting COVID". https://web.archive.org/web/20240515093309/https://www.museumoflondon.org.uk/news-room/press-releases/museum-london-acquires-tweets

process" (Geeraert et al., Interview). Most of the e-legal deposit legislation examined for the purpose of this study permits access to web and social media collections only within the premises of legal deposit institutions, usually to researchers who possess a reader card at the archiving institution, and through specific consultation terminals made available in the reading rooms.

In France, for example, lawmakers included in the e-legal deposit legislation an exception to the "right of communication to the public" (Chambers et al., 2021, p.18). Article L132-4 of the *Code du Patrimoine* states in fact that the copyright holder may not prohibit deposit institutions such as the BnF and INA from providing access to archived content (namely web and social media collections) to duly accredited researchers on the condition that this is to be provided only onsite.[149] Moreover, as mentioned in Section 5.4.4, limited reproduction is allowed for content archived under the e-legal deposit legislation, especially for material included in the French web archive. If researchers want to publish material from the web archive in its original form as part of their research, they must however obtain permission directly from the rights-holders. The complexity derived from the way social media platforms work, the reuse of content and the stratified layers of ownership, can make finding the rightful owner and requesting their permission quite challenging. Moreover, even when the request is not ignored by the owner, the time that passes between the request and the actual granting of permission can be significant often making the researcher question whether they really need to cite specific content in its original form (Ahmed et al., 2017).

Almost all of the legal deposit institutions that took part in this study are only allowed to provide access within their premises, with some being able to place some terminals in regional or local libraries (see for example Section 5.4.4). The decision to include such strict provisions in terms of access is rooted in the necessity to safeguard the interests of rights-holders, as well as to protect citizens' privacy and their information. In some cases, such as with the BL or the BnL, access to the reading rooms containing the terminals for accessing the web archive is granted simply by possessing a Reader Pass or library card (see Section 4.4.4). Other institutions instead require researchers to go through a longer process to obtain access to their web collections. In Denmark, for example, the KB only grants access to the *Netarkivet* to researchers or PhD students who apply for it, as the web archive can contain sensitive personal data, thus extra care needs

[149]https://web.archive.org/web/20240903135905/https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006845526?etatTexte=VIGUEUR

to be applied when providing access to its content.[150] The application process requires the researcher to complete a form in which they have to provide detailed information about the project and the names of other people involved in the project. If other co-investigators need to access data from the *Netarkivet* for the same project, they will need to complete a new application, as the KB can only provide individual access to the archived web collections.

Prior to my fieldwork at the KB in September 2024, I personally applied for access. The application process is rather easy: from the *Netarkivet* website,[151] I downloaded the "Applicant Declaration" form, which is available in both Danish and English under the "Research Access" section.[152] Then, I completed the form, describing my current PhD project and providing personal information (e.g., name, surname, email address), phone number (necessary to receive One Time Passwords), and the institution I was affiliated with. Regarding the latter, it is important to note that the KB states on its website that access is provided only to researchers or PhD students "affiliated with a Danish research institution". In my case, I was able to apply because I was going to briefly "reside" at the KB for the time of my fieldwork. Moreover, the application form asks the researcher to specify the timeframe of the project – I declared one week corresponding to the length of my fieldwork – which will determine for how long the KB will provide access to the web archive collections. The turnaround time for receiving a response is about three working days. Based on this account, access to the *Netarkivet* may appear to be more restrictive compared to other European institutions, especially due to the fact that researchers must be affiliated with a Danish research institution, and master's students are only granted access in special circumstances. However, these initial restrictions are balanced by a more permissive mode of access. Once the request for access is approved, researchers receive two separate emails including their username and password that, coupled with a two-factor authentication (OTPs are sent to the phone number provided at the time of the application), grants access to the web collections without restrictions in terms of location. Researchers can indeed access web archived data at home, without

[150] More information about the application process to access the *Netarkivet* is available here: https://web.archive.org/web/20240225221854/https://www.kb.dk/find-materiale/samlinger/netarkivet/forskningsadgang

[151] https://web.archive.org/web/20240117034741/https://www.kb.dk/en/find-materials/collections/netarkivet

[152] https://web.archive.org/web/20240117215033/https://www.kb.dk/en/find-materials/collections/netarkivet/research-access

having to travel to one of the KB sites located in Copenhagen and Aarhus. The extensive freedom of collection granted to the KB by the Danish legal deposit legislation in gathering content publicly available on the web, along with the possibility of inadvertently harvesting sensitive personal data during broad and selective crawls, explain the implementation of such a strict (yet in other respects permissive) access policy.

The COVID-19 pandemic, and the subsequent series of lockdowns imposed by governments across the globe to tackle the spread of the virus, have made patent how restrictions on access pose severe challenges to researchers and anyone interested in consulting the data preserved in national web archives, with almost all existing web archive collections created under legal deposit being completely inaccessible to users until restrictions were lifted. Sections 4.4.1 and 5.4.1 discussed how the recent COVID-19 pandemic affected access to the UK Web Archive, and the BnF and INA's web collections, highlighting the need for allowing wider access to web archives, albeit in a controlled manner to prevent any potential misuse of sensitive data. Initiatives such as the *ResPaDon* project[153] in France (see Section 5.5) or the recently established Legal Deposit User Forum[154] in the UK (see Section 4.5) are working towards understanding users' needs and finding ways to improve access conditions to web archive collections, for example by implementing terminals in university libraries, in the first case, or amplifying users' voices and feedback, in the latter, to bring some improvement to the existing access system (see Section 6.4).

Although necessary for privacy and data protection, restrictions on access, and specifically those concerning onsite consultation only, have profound repercussions on the engagement with and development of web and social media archives. As Lise Jaillant (2022) once duly noted, "archives are meant to be used, not locked away" (p.551). For web and social media archives in particular this is indeed a crucial point. As most archived content remains available on the live web, institutions encounter difficulties in conveying the importance for researchers of physically visiting their premises, often requiring travel to different cities, to access archived web material via provided terminals. Archiving initiatives have made remarkable progress in the past twenty years in terms of raising awareness of the existence of web archives among various communities of practice (Schafer & Winters, 2021). Specifically for social media, institutions like the BnF, INA

[153] https://web.archive.org/web/20240515101418/https://www.museumoflondon.org.uk/news-room/press-releases/museum-london-acquires-tweets

[154] https://web.archive.org/web/20230922080748/https://libguides.tcd.ie/legal-deposit-user-forum

and the KB are aware of a few researchers working on their archived material from social platforms like YouTube and Twitter (see also Fay et al., 2024; Vlassenroot et al., 2021). However, restrictions on access and low engagement with web-archived resources, especially among academic researchers, place a heavy burden on archiving initiatives. These initiatives may indeed find it difficult to advocate for additional funding needed to scale up and support the increasing number of resources required for web archiving activities over time. Particularly emblematic in this context is the case of the delayed restoration of access to the UKWA following the BL's cyber-attack in October 2023. Driven by the need to restore access to the most frequently used services first, the BL has been deferring the date for a potential restoration of access to the web archive collections, prioritising other services instead. Initially postponed to July 2024, the BL stated in a blog published in August 2024 that the restoration "won't for the time being include the UK Web Archive", attributing the decision to technical issues that required different solutions than those implemented for the other NPLD collections (British Library, 2024). Nevertheless, greater awareness and usage of the UKWA, particularly among students, might have prompted at least the provision of a more specific timeframe for restoring the service.

Furthermore, results from an extensive survey carried out by members of the Web ARChive studies network (WARCnet)[155] to identify skills, tools and knowledge required and challenges encountered across diverse communities of practices in web archives research, brought attention to the understudied socio-economic aspect (Healy et al., 2022). Available resources to travel to a certain library can influence the engagement with and usage of web and social media archives, especially among students and early career researchers who may have limited or no funding to cover multiple trips to consult material over the course of their research, where remote access is not an available option (see also Byrne et al., 2023). Participants in the Web Archives Research, Skills and Tools survey (WARST) also pointed out that restrictions on how archived content is accessed constitute an added barrier to research using web archives, specifically for researchers working on transnational collaborative projects, due to the diverse national legislation in force in each country (Healy et al., 2022). In this regard, initiatives such as the WARCnet network, carried out between 2020 and 2023 (see Section 1.2), aimed to address these challenges, promoting high-quality national and transnational research into transnational

[155] https://web.archive.org/web/20240104222259/https://cc.au.dk/en/warcnet/about

events. A further analysis on discoverability, replay issues related to access to and engagement with social media collections will be discussed in Section 6.4.

6.1.4 "You Shall Not… Archive": Social Media Platforms and the Constraints Imposed on Social Media Archiving Initiatives

Social media companies impose an additional layer of legal constraints that affect data access, limiting the quantity and frequency of information that can be accessed within a set timeframe. No matter the method used to scrape data (e.g., Heritrix, APIs) or the reason (e.g. research, preservation of cultural heritage material), social media companies have applied increasingly strict policies over the years to limit the amount of data that can be collected from their platform, especially after events like the Cambridge Analytica scandal (see Section 2.4.4; Isaak & Hanna, 2018; Schneble et al., 2018). While these restrictions are justified as a way to protect users' privacy, they primarily serve to protect the social media companies' legitimate business interests, allowing them to sell valuable information, such as user behaviour data, to the highest bidder (Helmond et al., 2015; Puschmann, 2019; van der Vlist et al., 2022). The introduction of a certain degree of control over data access is also necessary to prevent capture (and misuse) of potentially sensitive information by unauthorised third parties. However, these same restrictions add a further layer of limitation to the already complex labyrinth of constraints governing social media collection activities at an institutional level. Understanding how limitations imposed by social media companies affect the development of social media collections is important, as these limitations, combined with other technical challenges (see Section 6.3), strongly determine the amount of data that can actually be collected from each platform, partly explaining why some platforms are generally more archived than others.

Facebook, for example, is one of the platforms that appears to be causing the most trouble among social media archiving initiatives. Despite being a popular social site counting over 3 million monthly active users in 2023 according to Statista.com (Statista.com, 2023a), Facebook is only the second most mentioned social media platform being archived or in planning to be archived by the archiving initiatives that completed the survey (see Section 3.3). Moreover, in terms of quantity of data collected, the number of archived Facebook profiles appears to be limited compared to the large amount of information archived from Twitter. On the one hand, this can be attributed to the nature of the platform itself as being predominantly considered as a sort of personal diary where users tend to share information with a closed groups of "Friends" or followers (Burkell

et al., 2014; Good, 2012; Sinn et al., 2013), thus falling outside the scope of many institutions as the content is not made publicly available. On the other hand, archiving institutions have been struggling for some time now because of the inability to access the data they are required to collect – as in the case of e-legal deposit institutions – or to collect data to the extent they would like without risking being locked out of the platform. Many memory institutions have experienced being blocked by Meta at least once, either because they exceeded the amount of requests per hour or because Facebook detected behaviour patterns that could be attributed to web crawlers. The ongoing issues faced by archiving institutions have prompted web archiving teams to strictly limit the number of public Facebook accounts they archive or to exclude the platform completely from capture. For example, the BnF has decided to momentarily pause collection of this site until better archiving solutions are found (see Sections 5.4.1), whereas the BL has excluded it from its collection development policy apart from rare occasions (see Sections 4.4.2). Concerns about access to information on Facebook have also been raised by other institutions such as the KB, KBR, and BnL.

To address these limitations, some interviewees reported attempting to reach out to Meta. For instance, TNA mentioned several difficulties in establishing any form of communication with Facebook. Nonetheless, even in the rare instances where institutions succeed in establishing a communication channel with social media companies, this does not necessarily provide any additional assistance or safeguards for the institution. This is clearly exemplified by INA's experience described in Section 5.4.1, where obtaining an official account for the purpose of "cultural heritage preservation" did not offer any guarantees, as Facebook proceeded to close their account for exceeding rate limits only a few months later. The same problem in terms of rate limits has been registered by institutions trying to collect content from Instagram. Researchers in the BESOCIAL project reported that one of the primary struggles they encountered in attempting to archive content from Instagram was the risk of being blocked due to the numerous accounts and hashtags they had identified for preservation and attempted to collect. This has led curators to the decision of severely limiting the number of Instagram accounts to be included in their collection.

In terms of access to data specifically for research purposes, Twitter has hitherto demonstrated a larger openness compared to other platforms, providing different levels of access to their historical data. A well-known example demonstrating Twitter's recognition of the unique value of the information shared on its microblogging site by

billions of people is the agreement signed with the Library of Congress in 2010 to establish a Twitter archive, which unfortunately has remained an isolated case in the social media archiving landscape, at least up until the time this thesis was written (see Section 2.3; Bruns, 2018; Fondren & Menard McCune, 2018). Moreover, in 2020 Twitter introduced the so called "Twitter Academic"[156] API, a new tool that was immediately met with great interest by scholars because of the potential that dedicated access, including full-archive search, would mean for the research community. Given the greater amount of data that this type of access would provide to successful applicants, a few archiving institutions among those taken into consideration in this study applied for access to the Twitter Academic API. Of these institutions, all described the long process required to fill in the application form, which included extensive questions about the project for which they needed access to the Twitter Academic API. After numerous email exchanges with Twitter, most of them saw their request denied. The only exception was a project submitted by one of INA's web archiving team members who commented: "We managed to get one of these accounts but it's not easy, because you have to prove that you are a researcher and have a project". Moreover, INA noted that Twitter wanted to ensure that the applicant was a researcher with sufficient technical skills to use the API (see Section 5.4.3). The high rate of rejections for applications submitted by cultural heritage institutions may suggest that Twitter perceives researchers as having different objectives than archiving institutions, providing the former with a specific level of access to their full archive, while leaving the latter to juggle restrictions and technical limitations with no direct support from the platform they are essentially contributing to preserve in the long-term.

However, the acquisition of Twitter by Elon Musk in November 2022 drastically changed all the positive prospects that archiving institutions anticipated achieving with this type of API (Paul & Milmo, 2022). In fact, in February 2023 the new administration announced with a tweet that Twitter would no longer support free access to the Twitter API (Twitter Dev, 2023a), substituting the old access system with a set of paid tiers ranging from "Basic" to "Enterprise" (Twitter, 2024). Although the backlash generated by the announcement brought the introduction of an additional "Free" tier, the type of

[156] The page related to the Twitter Academic access does not exist anymore, but a web archived copy is available through the Wayback Machine: https://web.archive.org/web/20230524034313/https://developer.twitter.com/en/products/twitter-api/academic-research/application-info

access provided is much more limited than the amount of data made available through the Twitter Academic API.

The events following the acquisition of Twitter, and the implementation of a paid tier system, make apparent how social media archiving is dependent on and heavily shaped by conditions and constraints imposed by social media companies, which often change unexpectedly and in unpredictable ways. This specific event is particularly significant for the social media archiving landscape, as Twitter was the most archived social platform until 2023, according to the survey results discussed in Chapter 3. For this reason, I followed up via email[157] with some of the social media archiving initiatives I had previously interviewed to capture their reactions to the news of Twitter's acquisition, the apparently short-lived risk of seeing one of the most researched platforms disappear in the aftermath of the mass-layoffs at the end of 2022, and the immediate consequences of the launch of paid tiers for accessing the Twitter API. Rather than the prospect of seeing Twitter suddenly disappear, institutions appeared to be mostly concerned about the economic implications of paid access, as this would mean allocating more resources to cover the additional expenses for access to the new API version. When I received responses to the follow-up emails between the end of February and the beginning of March 2023, the pricing and specifics for each tier had either not yet been published or had just been released. Thus, some comments from the consulted institutions were largely based on information available on the Twitter Developer Platform or circulating among communities of practice. Section 5.4.6 described the effects of and adjustments that INA had to implement following the changes imposed by Twitter on access through the new pricing scheme. While institutions harvesting Twitter via Heritrix were less impacted than those relying solely on the official Twitter API, the announcement sparked many discussions within institutions and internationally among members of the IIPC regarding the potential influence this new system could have on preservation and research communities. Bingham commented that the BL and the other UK deposit libraries were "concerned with preserving as much as we can of Twitter that is in scope (accounts of public figures and official accounts of organisations) given this instability",[158] despite the general feeling that Twitter was not going to disappear, as things started to settle down a few months after the acquisition.

[157] Follow-up emails were sent out in January 2023, see also Section 1.4.5.

[158] Nicola Bingham, email received on 15 March 2023.

Anders Klindt Myrvoll, Web Curator at the *Netarkivet*, reiterated the concerns regarding the difficulty of getting access to the Twitter Academic API and the negative outcome their application had. He expressed his doubts about the new pricing system, noting that it did not make the whole process any easier for archiving institutions, and added that they would like "some kind of archival access from Twitter,"[159] despite being aware that the chances of this happening in the near future are unfortunately quite low. On 30 March 2023, Twitter announced through their @TwitterDev (now @XDevelopers) profile that they were looking into new solutions for providing access to academics, encouraging them to use, in the meantime, one of the tiers available (Twitter Dev, 2023b). Yet, at the time of writing no access specifically created for preservation purposes has been made available nor has any plan to implement anything similar ever been announced.

While I will not go into the details of how the Twitter APIs work as it goes beyond the scope of this thesis, I would like to point out how even the highest level of access, the "Enterprise" tier, which is the only one to date that allows access to the Full-Archive search API,[160] that is the full corpus of data produced on the platform since its creation in 2006, has its limits. In fact, as stated in the "Search API: Enterprise" overview page: "Each customer has a defined rate limit for their search endpoint. The default per-minute rate limit for Full-Archive search is 120 requests per minute, for an average of 2 queries per second (QPS)" (X Developer Platform, 2023). In an interview conducted by Schafer et al. (2020) for the WARCnet papers series about collections dedicated to the COVID-19 pandemic, Jérôme Thièvre, INA's former Manager of the Digital Legal Deposit department and now Head of Web Archiving Services since 2023, recounted that INA was able to collect over 120 million tweets related to COVID-19 since the start of the pandemic. Thièvre specified, however, that this number only referred to the tweets that INA managed to collect considering the rate limits set for the real-time API they were using at the time, which limited collection to 50 tweets per second (Schafer et al., 2020). Therefore, it is evident that the changes implemented in 2023 considerably reduced the

[159] Anders Klindt Myrvoll, email received on 16 March 2023.

[160] Since September 2023, X-Twitter has added a "Pro" tier that sits in between the "Basic" and "Enterprise" one in terms of price and rate limits. More details here: https://web.archive.org/web/20240518154805/https://developer.x.com/en/docs/twitter-api/getting-started/about-twitter-api#v2-access-leve

rate limit of tweets per month, significantly impacting the collection capacity of institutions using the Twitter APIs (see also Section 5.4.6).

LinkedIn represents another interesting case in this regard. Despite not being among the most archived sites, LinkedIn constitutes an important platform to preserve, particularly for material related to public figures or companies. The main limitation that emerged from the interviews is related to the fact that LinkedIn does not allow access to public content unless a user is logged into the platform, which might not be allowed by law for some institutions. Moreover, Daniel Gomes explained that Arquivo.pt used to archive content from LinkedIn but, in recent years, this has no longer been possible because of the limits the company started to impose on access to data. Such restrictions were seemingly implemented after the platform was acquired by Microsoft in 2016. Citing a commitment to user privacy, LinkedIn Corp. began enforcing stricter rules for web crawling around 2017, mandating permission to crawl data generated on its site (LinkedIn, 2017). The LinkedIn user agreement (2022) clearly states under the "Don'ts" section that by joining the platform the user agrees to not "develop, support or use software, devices, scripts, robots or any other means or processes (including crawlers, browser plugins and add-ons or any other technology) to scrape the Services or otherwise copy profiles and other data from the Services".[161] Particularly emblematic in this regard appears to be the legal dispute *hiQ Labs Inc. Vs LinkedIn Corporation*, which stemmed from LinkedIn's attempt to block hiQ Labs from scraping public profile data. After six years of litigation, the case led to a significant ruling on the legality of web scraping and data access rights in April 2022.[162] Despite the positive ruling in favour of hiQ Labs, the U.S. District Court for the Northern District of California ruled in November 2022 that hiQ Labs had breached LinkedIn's user agreement by scraping and through the creation of fake accounts (Neuburger, 2022). The litigation ended with a settlement agreement between the two parties whose terms have not been disclosed. While the case demonstrated that web scraping is legal in certain instances, it also indicated that social media companies may sue third parties for breach of contract to stop scraping activities.

[161] LinkedIn user agreement, Section 8.2, article 2. Available here: https://web.archive.org/web/20240518221005/https://www.linkedin.com/legal/user-agreement#dos

[162] HIQ LABS, INC. V. LINKEDIN CORPORATION, No. 17-16783 (9th Cir. 2022). https://web.archive.org/web/20240503223218/https://law.justia.com/cases/federal/appellate-courts/ca9/17-16783/17-16783-2022-04-18.html

The limitations imposed by LinkedIn on web crawling activities have caused many issues for web archiving initiatives, as documented on the Internet Archive's support page: "LinkedIn is blocking crawls of select personal and organizational profile pages. Web crawls cannot archive pages that return the HTTP 999 status code currently" (Archive-It Help Center, 2023a). It is because of these limitations that archiving institutions like TNA had to reject archiving requests for this platform on multiple occasions:

> We aren't able to get [LinkedIn] at all and so there was an organisation that I was talking to – what we call short term bodies – that is an organisation that's up for a purpose and then disappears after a while, […] they put quite a lot of stuff on LinkedIn but I had to tell them "sorry we can't, we're not able to capture LinkedIn".

The restrictions imposed by social media companies have significant repercussions for the ability of national archiving institutions to fulfil their mandate to preserve digital cultural heritage on social media. It is also important to note that platforms tend to limit the number of accounts users can create to access the official APIs to one per institution or project. Coupled with the rate limits already imposed on data access, this makes archiving material on a national scale nearly impossible, especially when archiving time-sensitive events. For this reason, some institutions may be compelled to create more than one account to effectively preserve significant content on social media, slightly pushing the boundaries of what is allowed by existing policies. Navigating all these factors clearly represents a great challenge for memory institutions when it comes to planning, selecting and effectively collecting such ephemeral and fragile born-digital material, which ultimately affect the shape of resulting collections.

6.2 Selection Practices

National legal frameworks and the additional limitations set by social media companies trace the primary boundaries within which institutions involved to different degrees in social media archiving activities can select material to be preserved as part of their collections. While these multi-layered legal constraints and various technical challenges (see Section 6.3) serve, in a certain sense, as an initial culling of the myriad content generated daily on social platforms, they also pose significant challenges to social media archiving initiatives. As web archivists try to untangle the intricate network of

conversations shared across national borders on social sites and select only publicly available material relevant to specific countries, they must consider the increasing risk of generating gaps in the information archived, raising concerns about the representativeness of social media collections. The following sections will explore the main issues raised by interviewed institutions concerning selection practices for social media. It will reflect on how national legal frameworks, social media terms and conditions, and technical issues affect the development and granularity of collections, illustrating approaches implemented by archiving initiatives to ensure representativeness at a national level.

6.2.1 Shaping National Social Media Collections: Selection Challenges and Representativeness Concerns

Social media has unlocked the opportunity for users to participate in an ongoing global conversation, whose interlocutors can join in at any point in time and from anywhere in the world (Simon, 2012; van Dijck, 2011, 2015). Moreover, social media platforms enable the creation of a network of reactions to events on a global scale, mixing responses from individuals with diverse cultural backgrounds, and amplifying voices that for centuries have often been silenced or not properly included in mainstream archives (Barrowcliffe, 2021; Bergis et al., 2018).

Over the years, there has been increasing recognition of the need to provide space for marginalised groups of society whose stories had been filtered for a long time through the lenses of their colonisers, leading institutions to recognise the need to revise and "decolonise" traditional archive collections and their descriptions (Ghaddar & Caswell, 2019). With web and social media archives, curators are offered the unique chance and the hard task to shape collections that aim at reflecting the diversity and the full granularity of a specific country's society and culture. However, while the goal appears to be clear, archiving institutions must face the complex reality of social media data, the manifold technical challenges, legal constraints, and limited resources available in the public cultural heritage sector. These factors inevitably influence the texture and shape of the resulting social media collections. This is especially true considering that small, resource-constrained web archiving teams have to select content from the vast cauldron of information generated on social platforms.

Before proceeding with the analysis of the challenges occurring in the selection phase, it is worth pointing out that most interviewed institutions tend to not distinguish between websites and social media collections. Social media content is usually added to existing thematic collections or collections created in response to events like terrorist

attacks or health emergencies.[163] The tendency to aggregate content from websites and social media platforms, despite the unique traits that distinguish the two types of materials, can be attributed to technical approaches, such as the method of collection. Chambers et al. (2021) observed that social media material harvested through web crawlers and preserved in WARC files is generally treated as any other web material and thus incorporated into existing web collections alongside websites; whereas content collected via APIs is often preserved as a separate collection. This distinction is clearly illustrated in the French case study (see Chapter 5). The BnF, which archives both websites and social media mainly through web crawlers, aggregates them by topic without differentiation between the two sources. Conversely, INA has chosen to collect data from social platforms via APIs and distinguishes between the two types of material. Specifically, tweets are maintained as a distinct collection known as the "Twitter Archive", which occupies a dedicated section on *WebMédia*, INA's web archive access interface.

The frequency of collection usually depends on the type of platform from which the content is captured, how frequently accounts are updated, the risk of content disappearing, and whether it is related to a specific topic or event. For instance, Twitter is often collected more than once a day (especially when archived via APIs), once a day, or weekly. In contrast, other platforms such as Facebook and Instagram, which are more problematic to archive efficiently due to legal and technical constraints, are typically collected between once a month and once a year.

When it comes to selection of social media content, as mentioned, national legal frameworks have a profound influence on the type of material selected for preservation, drawing a first limit on the vast amount of data from which archiving institutions may select content. National e-legal deposit legislation tends to circumscribe the radius of action of deposit institutions to works made publicly available by individuals or organisations who reside in or share content related to the country in scope (see Section 6.1). One of the primary concerns that emerged from the interviews was related to the difficulty of clearly identifying national borders on something so internationally interconnected as content generated on social platforms. As discussed in Section 6.1.1,

[163] See for example the series of paper published in the context of the WARCnet Network project about collection activities at various national institutions concerning web material about the COVID-19 pandemic. The full list of papers is available here: https://web.archive.org/web/20240211102313/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports

the territoriality criterion is often the cause of many concerns among web curators, as they must carefully assess the provenance of content among the sheer amount of data generated on social media. However, assessing provenance of content on social media can be laborious and not always reliable. For example, relying on geolocalisation data provided on social platforms or other data suggesting a certain location may not be accurate and can be subject to a high error rate (Burton et al., 2012; Graham et al., 2014). For this reason, archiving institutions have mostly opted to hand-pick accounts of public figures or organisations for which they can determine with a high degree of certainty their pertinence to the country in scope. This is the approach that institutions such as the BL (see Section 4.4.2) and the BnF (See Section 5.4.2) have adopted to identify content pertaining to the territoriality principle established by national legal deposit regulation.

To sift through the sheer amount of content available on social platforms, national memory institutions must apply additional criteria beyond those provided by national legal frameworks to further refine the scope of archivable content. For example, some institutions mentioned the use of criteria such as "language" to help them identify relevant social media accounts. While this approach may be suitable for countries where national languages are spoken by a limited number of people within a specific area, it can be counterproductive for refining in-scope material in countries where the official language(s) is also spoken widely outside national borders. This is exemplified by the English language which is often used as a common means of communication among speakers worldwide. For institutions such as the LoC or the BL located in English-speaking countries, it would be too difficult, if not impossible, to ascertain the provenance of each post on social media that is written in English. Moreover, selecting content based solely on the official language(s) of a certain territory would risk overlooking the diverse kaleidoscope of content generated by minority communities speaking different languages (Huc-Hepher & Wells, 2021).

In the BESOCIAL project, researchers posited the use of official languages as one of the potential additional selection criteria to identify content for preservation as part of Belgium's online cultural heritage at the KBR. However, the language criterion soon proved unsuitable for such archiving endeavour, as Belgium shares its three official languages – French, Dutch and German – with other countries, including France, Netherlands and Germany. One of the researchers of the BESOCIAL project indeed noted: "We were quite surprised about how difficult it is to get hashtags and accounts that are really linked to Belgium especially hashtags, especially since we share languages with

France, with Netherlands — language isn't really a criterion we can use" (Geeraert et al. Interview).

Hashtags are another difficult element to appraise when archiving social media. As observed in Section 5.4.2, hashtags serve as a means to mark topics on social media, facilitating content discovery as well as participation in online conversations about events and specific themes (Alfano et al., 2022; Bruns & Burgess, 2011; A. Cui et al., 2012). However, archiving social media posts through hashtags can lead to the capture of content that goes beyond the countrywide scope, and consequently outside the perimeter set by the national electronic legal deposit legislation. This is the reason why the BL, for instance, has chosen to limit the capture of hashtags to only exceptional occasions, where the pertinence to the United Kingdom is evident, as in the case of the hashtag #Brexit (Bingham, Interview; see Section 4.4.2). Moreover, hashtags may sometimes be in languages different from those spoken in the country of the archiving institution. This is the case of the viral hashtag #prayforparis created following the events of the 2015 Terrorist Attack, which included messages of support from all over the world and in many different languages (see Section 5.4.2). Both the BnF and INA decided to archive this hashtag due to its relevance to their mandate and as a significant testimony of what occurred during those terrible moments and immediately thereafter (Tybin, Interview; Mussou, Interview).

Similarly, the KBR tried to identify hashtags clearly linked to Belgium by searching for those including the two-letter code for Belgium (BE) or Belgian city names. But, as Geeraert noted, apart from rare examples in which there was a clear reference to national events, identifying hashtags is "not a straight science and it's very approximate" (Geeraert et al., Interview). Using curation tools such as those described by INA in Section 5.4.2 (e.g., Hootsuite, Trends24) may facilitate the identification of trending hashtags in a certain country. Nevertheless, due to the challenges in verifying the origin of a large volume of content aggregated by hashtags, and the resources needed to sift through and archive the vast amount of associated data, memory institutions often tend to limit the number of hashtags they collect.

Interestingly, some institutions also considered identifying relevant social media accounts by searching for flags placed next to individuals' names or handles, as indicative of their connection to a specific country. The use of flag emojis in online communication has received little attention, with only a few studies investigating their meaning in social media discussions. For example, Kariryaa et al. (2022) examined the use of flag emojis in

more than 600 thousand tweets shared by political parties and members of the parliament in Germany and United States. Results showed that in both countries the national flag emoji was frequently used in the context of political communication, especially during events of national importance such as Independence day, or national elections (Kariryaa et al., 2022). Moreover, Kariryaa et al. observed that the usage of the flag emoji depends on political and cultural context, with right-wing parties tending to use it more frequently than left-wing parties. In particular, the flag emoji appeared to be used with a higher frequency in Germany by far-right political parties. Similar trends were found in the US, where the Republican party and its members seemed to make more extensive use of the national flag emoji compared to Democrats (Kariryaa et al., 2022). Likewise, in Italy, the use of the Italian tricolour flag emoji next to Twitter handles often indicates support for the right-wing or even far-right parties. Conversely, supporters of Italian left-wing parties tend to use the European Union flag emoji, which may or may not be juxtaposed with the national flag. Exceptions to this trend obviously occur, as in the case of the Catalan flag emoji, which seems to be less associated with right-wing parties (Cubells Pastor, 2020; Medero & Maestre, 2023). Nevertheless, the uncertainty surrounding the usage of national flag emojis and the possible association with extremisms or nationalist movements makes this potential criterion one that should be considered thoroughly in all its nuances and meanings. The decision to select accounts or posts based on any of the Unicode flags used on social media could indeed have important consequences on the development of national collections and equal representation of different political trends and views.

Still, identifying all the public figures and organisations pertaining to one country represents a great challenge for archiving institutions, both those with long-standing web archiving experience, and even more so for those that are testing or planning to preserve social sites. An interesting concern in this sense was raised by Ben Els, Digital Curator at the BnL, which is testing social media archiving activities with small sample collections. Like many other archiving institutions, the BnL has started experimenting with events such as national or local elections because of their historical, societal, and political significance. For websites, the BnL usually gets a full list of URLs comprising the .LU domain from the national registrar, ensuring a comprehensive capture of the whole Luxembourg national web domain. Conversely, "for social media [this] is impossible" (Els, Interview). The sheer amount of content generated and made publicly available daily on various platforms makes discovering relevant content rather challenging. This issue appears to be particularly evident, for instance, when institutions need to capture accounts

of candidates in national and local elections who do not have a clear presence on social media or have recently converted their private accounts into public spaces to promote their candidacy. In this regard, Els described the difficulty of tracking down candidates that had a less prominent presence on social platforms: "It was a bit difficult to research all the people: these were local elections, so they were in some cases less prominent: […] they are not always clearly present as a political figure on social media" (Els, Interview).

The same discoverability issue has been raised by other memory institutions such as the KB, which has reported problems in locating social media accounts for less known candidates. Anders Klindt Myrvoll, Web Curator at the KB, explained that one of the reasons behind this challenge, in their specific case, could be attributed to several people often sharing the same name in Denmark, making it difficult to identify the exact account, especially if there is little activity on said profile regarding the elections. While efforts to locate accounts by searching for political party logos and other elements associated with election campaigns have produced positive results, Klindt Myrvoll noted that receiving a comprehensive list of candidates, complete with the correct social media handles for each candidate, would be essential to allow the national web archive to fulfil their mandate and capture future national or local elections more accurately. Likewise, Daniel Gomes from Arquivo.pt pointed out that in Portugal official lists of candidates often do not include adequate information about social media accounts. While most politicians have a well-established online presence, as seeking engagement on social media among target communities has proved to be essential to support the electoral campaigns (Bruns et al., 2021; Cogburn & Espinoza-Vasquez, 2014), Gomes explained that in Portugal many candidates appear to create Facebook pages in concomitance with their candidacy and then abandon them quickly (or delete them) soon after the end of the campaign.

In addition, Gomes noted that keeping track over time of all these profiles has proved to be particularly challenging, especially in the context of smaller communities and villages. Portugal has indeed a high number of small boroughs, each one periodically hosting local election campaigns. The lack of a comprehensive list containing information about each candidate, which could be easily collected at the time of the submission of the electoral roll, exposes national web and social archiving initiatives such as Arquivo.pt and the KB to the risk of missing valuable information. To mitigate this issue, the Portuguese web archiving team has been compiling an internal list of candidates for both national and local elections that occurred in the past few years, including information related to any website or social media profile connected to their political activities (Gomes,

Interview). While this list aids web archivists when compiling seed lists for the collection of in scope material, it also represents an important resource for present and future researchers interested, for instance, in political matters.

As mentioned in Section 6.1, a further concern for social media archiving initiatives stems from the mixture of private and public content, particularly when public figures share personal information, leading to complex legal and ethical challenges that memory institutions have yet to resolve. Expanding on the privacy and data protection matter, web archivists, such as the BL (Section 4.4.2) and the BnF (Section 5.4.1), have shared concerns about the inability to capture culturally significant material due to content being accessible only to a restricted number of members of a certain page or group. In this regard, the BnL provided an interesting example while discussing the impact of the COVID-19 pandemic on web and social media archiving activities. Ben Els mentioned that during the first lockdown in Luxembourg, numerous solidarity groups were created on Facebook to support, for instance, expats who, due to language barriers, might have had issues in fully understanding news and information provided through the national main channels in one of the country's official languages. Beyond the emotional and societal value of these episodes of community support, these groups represent a unique and invaluable source documenting the events and sentiments during the first weeks, if not months, of uncertainty following the national closure of commercial activities and travel restrictions due to the health crisis. Among the myriad support groups created online during that time, Els described how even smaller communities, such as local neighbourhoods, began to form closed Facebook groups where neighbours would check on each other, and younger people would offer to go grocery shopping for the elderly living nearby. Despite these episodes being worth preserving and of great interest to researchers, the BnL recognised that because of privacy and data protection concerns such material could not be included in their web collection:

> I don't know what we can archive in these groups… there are very intimate questions, and these people wouldn't know that what they had shared would be archived. So, we have to say, this is not going to be in the public collection – we cannot archive it, but it is the thing we are asked about the most.

While these necessary curation decisions are made to protect the privacy of group members, they unfortunately create significant gaps in the resulting web and social media archives, raising concerns about the overall representativeness of national collections.

The uncertainty surrounding whether to archive, or more generally, how to treat content shared within closed groups on social media – especially when their closed status is necessary to protect the visibility of minority groups or activists with a history of persecution (Bergis et al., 2018a) – can particularly affect representation of minority groups in social media collections. In the midst of the 2020 pandemic, for example, migrant and ethnic minorities sought on social media that community support that could no longer be provided in person in local community centres (Goldsmith et al., 2022b). Despite the risk of being exposed to forms of misinformation while browsing content on their feed, Goldsmith et al. (2022) observed that social media platforms, particularly Facebook pages or groups related to religious and local communities, had been essential to bridge language barriers, facilitating the dissemination of official, accurate information about a wide range of matters, including vaccination campaigns. Moreover, in a study analysing the Internet presence and performance of European minority languages media, Ferré-Pavia et al. (2018) concluded that in the last decade minority groups have increased exponentially their online presence, with a fair number of them exclusively active on social media, with no associated websites. While this is most likely due to the scarcity of resources and infrastructures available to these groups for maintaining websites over time (Ferré-Pavia et al., 2018), the aforementioned findings appear to be particularly noteworthy if considered in relation to web and social media archiving practices. As several minority groups on social media cannot be archived due to legal constraints, inevitably escaping the net cast by memory institutions, what will remain of their online presence if no website can be alternatively archived as a trace of their existence and activities?

In this regard, Els recounted how a call for suggesting websites relevant to the COVID-19 pandemic (see Section 6.2.2) sparked a larger reflection on the challenges of discovering small communities on social media, as many of them are closed behind the walls of private groups. Moreover, a quick search on Facebook for religious groups in Luxembourg revealed numerous communities whose online presence is circumscribed to the Meta platform. Els indeed commented:

> There's no equivalent to the social media presence on regular websites or in other media. There are no other traces [about them] on newspapers or brochures books. Thus, they should be archived.

On a similar note, Friedel Geeraert from the BESOCIAL project in Belgium observed:

> I tried to look into it and most of them actually had a closed Facebook page. So, in order to approach minority groups, you would need to find a way in somehow.

Although beyond the scope of this thesis, it is important to highlight the need for further work to engage with these communities, understand their concerns, and develop strategies that involve them more closely in the preservation of their stories (see Sections 6.2.2, and 8.2). Still, the discoverability of these small communities on social media appears to be a significant issue for memory institutions at all stages of development of a social media archive. Moreover, discoverability issues combined with the legal constraints that prevent many national archiving organisations from capturing these realities make the endeavour of creating collections that accurately represent the diverse and stratified structure of society even more daunting.

Although it has been pointed out on many occasions over the course of this study how the vast amount of content produced daily on social media makes archiving everything that has been (or will be) published practically impossible, considering the time and resources this gigantic endeavour would require. While legal frameworks and collection development policies implemented by each institution help to broadly delimit the radius of actions within which web archivists can select social media content, it is ultimately the web archivists who have to appraise and select material that is worth archiving based on the historical, societal and political value this content may have in supporting future researchers to understand important events and our present times. For these reasons, the pursuit of completeness is often replaced by a broader goal of developing collections that aim to represent the many faces of individual countries. This is, for example, the approach that the BnF has adopted, as pointed out in Section 5.4.2, aiming to ensure the representativeness of all French realities rather than exhaustiveness in the *Archive de l'Internet*.

6.2.2 Different Approaches, One Goal: Mitigating Concerns of Representativeness in Social Media Collections through Practices of Co-curation and Participation

In post-modernist archival theory the concept of "archival representation" has been extensively discussed (Charlton, 2017; Jimerson, 2006; Kaplan, 2002). It has been identified as a socially constructed practice in constant evolution, firmly embedded in archival processes (Yakel, 2003), where the record is considered as "a surrogate of some transaction, activity, or event" (Charlton, 2017, p.2). Elizabeth Kaplan (2002) further

examined the concept of "representation" in archives by linking archival practices to anthropology, highlighting how "both [disciplines] are concerned with representations — of people, of cultures, of events, and ultimately of history and of memory […]" (p.211), and noting that "both exercise *power* in the creation and usage of records […], of information" (p.211). It is indeed the acknowledgment of the power involved in selecting certain records over others that has raised concerns about the potential biases existing in mainstream narratives (T. Cook & Schwartz, 2002; Harris, 2002; Schwartz & Cook, 2002). Collecting practices have often led to documenting only one side of history, thus silencing marginalised groups due to structural power dynamics (Schwartz & Cook, 2002). For this reason, Jimerson (2006) underscored the need to recognise the biases that inevitably shape collections, and emphasised memory institutions' commitment to "ensuring that our records document the lives and experiences of all groups in society" (p.30), not just part of it.

As mentioned in Section 2.5, web and social media collections are not immune to the biases and power structures that are intrinsic to archiving and preservation practices. Moreover, the rigid legal frameworks within which social media collections are developed, especially those created under e-legal deposit legislation, coupled with inequalities of access to social media for certain groups of society (Lutz, 2022; see also Section 2.2.1), can impose additional constraints on achieving representativeness of the material archived (see Section 4.4). Despite these constraints, it remains crucial that social media collections at an institutional level reflect, as much as possible, the kaleidoscope of stories that social platforms have enabled individuals to share, allowing them to make their voices heard and to document from their standpoint significant events related to their own histories. However, identifying marginalised groups or stories related to specific regions in the national territory can prove challenging due to the sheer amount of material generated on social platforms (C. Cui et al., 2023).

To mitigate structural biases, bridge gaps and help address known issues of representativeness, many of the archiving initiatives interviewed have implemented participatory archiving strategies designed to make the most of the limited budgets, staff and time available (Pendergrass et al., 2019). It is important to note that the participatory approaches illustrated in this section have not been developed specifically for the enrichment of social media collections but have been tested over time and seamlessly applied to complement the collection strategies of most web archiving initiatives. As mentioned above, social media content is often included in existing web archive

collections, resulting in an equally minimal distinction, in the context of participatory practices, between social platforms and websites.

Crowdsourcing is a popular approach that has been widely used by web archiving initiatives to improve the granularity and representativeness of existing collections (Bingham & Byrne, 2021; Schafer & Winters, 2021). Crowdsourcing involves the "contributions from a large online community to undertake a specific task, create content, or gather ideas" (Terras, 2015, p.421). In the context of web and social media archiving, this may entail targeted campaigns promoted to gather suggestions from the public for web-based material on specific themes, or it can take the form of an open, permanent call to propose URLs that should be preserved, according to the scope of the collection (Schafer & Winters, 2021; Terras, 2015). For example, at the beginning of the BESOCIAL project at the KBR, researchers decided to start by defining the concept of Belgian cultural heritage online. Because of the breadth of the scope, the BESOCIAL project team decided to first identify a few main themes. This top-down approach allowed them to compile an initial seed list organised around topics like "the COVID-19, museums, festivals, heritage collections, food, minority groups and experts' communities in Belgium" (Geeraert et al., Interview). However, since BESOCIAL was developed as a pilot project seeking to investigate the feasibility and sustainability of social media archiving activities in Belgium, after the first round of collection, the researchers decided to test different bottom-up approaches to selection. In particular, they directly asked the Belgian public to suggest social media content – including text-based material, accounts or hashtags – that should be preserved as part of the online national heritage collection. Fien Messen from the BESOCIAL project described how they sent out an open invitation for suggesting content addressed to the whole Belgian public:

> We started up a crowdsourcing campaign in order to fill the gap of representativeness and that was a success, I think, also in promoting the whole social media archiving concept within Belgium. We received around 800 suggestions by the Belgian public.

While researchers in the project recognised that more work needed to be done in future projects to capture material that could fairly represent all of the minority communities existing in Belgium, overall, the crowdsourcing call appeared to be successful as the number of responses received demonstrated. Not only did this bottom-up approach enable BESOCIAL web archivists to discover new online material to be preserved, shedding light on minority groups, entities and stories that might have escaped those

initial top-down rounds of collection, but it also helped raise awareness about the social media archiving initiative among the wider public. Essential in this sense was the promotional campaign developed in collaboration with the KBR Communication team, which was circulated on the KBR website and through various news media. This included participating in "The World Today", a programme aired on the Belgian Radio1,[164] and an article on VTR's website, which is the national public-service broadcaster for the Flemish Community of Belgium (Mechant et al., 2022, pp. 17–18).

Some web archives have dedicated pages on their portals where users can easily nominate websites for preservation (e.g., the BnL, the BnF, the BL, Arquivo.pt). These online forms seem however to be mostly focused on receiving website suggestions rather than social media content: the prompts often refer to websites or webpages rather than expressly to social media. Nevertheless, recent campaigns promoted between 2020 and 2021 on the occasion of the COVID-19 pandemic have encouraged users also to suggest public social media content, significant hashtags and posts that should be included in the national web archive. Section 5.4.2 described how the *Bibliothèque national de France* launched a few similar campaigns on their website through which the French public could nominate webpages and social media content about specific thematic collections, such as Artificial Intelligence and the 2022 presidential election, obtaining high engagement and relevant content. While participation in these crowdsourcing campaigns can be fluctuating, any suggestion can still be of great value as it may uncover underrepresented themes and help identify minority communities that may have slipped through the mesh of the net used by memory institutions to sift through relevant content. It is important to note, however, that the effectiveness of such campaigns in terms of representativeness is closely linked to the way these are promoted and through which communities they are circulated, as a wide diffusion among target groups could be crucial to obtain suggestions that can truly enrich the content that is already the object of preservation. In particular, crowdsourcing campaigns aimed at reaching a higher level of representativeness of collections should also consider the fact the engaging in such participatory activities may appeal to certain sociodemographic groups of society, excluding others due to a wide range of digital inequalities, such as access to the internet, digital skills or adequate

[164] The article and an excerpt from the radio interview is available here: https://web.archive.org/web/20240430094329/https://www.vrt.be/vrtnws/nl/2021/10/28/online-erfgoed/

hardware (Bonacchi et al., 2019). In this regard, the "Whose Knowledge"[165] initiative has provided some important evidence of how crowdsourcing initiatives (e.g., Wikipedia) can privilege specific areas of the globe and certain groups in the production of knowledge, requiring memory institutions to consider and be aware of potential imbalances among the results collected through such campaigns.

Fostering collaboration both within and outside archiving institutions, as well as involving researchers-curators in the selection phase, are further interesting approaches used alongside crowdsourcing campaigns. Some of the archiving initiatives interviewed described how the selection process for web and social media content is supported by a network of collaborators, including internal departments and external contributors from other institutions. For example, the network established at the BnF, which is detailed together with its numerous contributors in Section 6.4.2, facilitated the inclusion of regional stories, initiatives, and public figures that characterise the vast French territory across the various web collections. This content might otherwise have been difficult to discover for curators located, for example, in Paris, due to their specific geographical relevance.

There are other institutions like the BL that have established, instead, collaborations with curators who are representatives of minority groups, and thus are more likely to have connections with members of said communities and to bring into collections a privileged perspective on what aspects and stories should be added to web archives (Brewis et al., 2021; see also Section 4.4.2). Although co-curated collections might still be the subject of disputes when members of those same communities disagree with certain selection choices, such discussions can produce positive outcomes. In fact, after receiving criticism regarding the inclusion of transphobic or homophobic material in the "LGBTQ+ Lives Online" collection (see Section 4.4.2), the UKWA enriched their collection descriptions to highlight curation choices and critically, potentially controversial aspects. Improving transparency about collection approaches, as well as acknowledging potential gaps and limitations in specific collection documentation, holds significant value for both the institution and (future) researchers.

[165] More information about "Whose Knowledge" and issues related to public online knowledge production is available here:
https://web.archive.org/web/20240523132508/https://whoseknowledge.org/issues/public-online%20knowledge/

Furthermore, regarding the broad spectrum of issues that come with documenting the lives of marginalised or vulnerable groups, Bergis et al. (2018) emphasised the need to always try to respect the rights of content creators from these communities. In this regard, the "Social Humans"[166] project created as part of the "Documenting the Now" project has proposed an interesting solution: two sets of labels – one for content creators (SH-C label) and one for archives (SH-A label) – to address ethical issues in re-using social media data generated by vulnerable communities. The project is still in the testing phase, and while its wide implementation could take a long time, these labels have the potential to offer a means for creators to express their wishes regarding whether they want to see their data used and how; and for archives, to add an additional layer of information to posts and/or whole datasets, including warnings about harmful content or flagging accounts that appear to be bots. Similarly, the "Archive of Tomorrow" project, mentioned in Section 4.4.2, has also investigated methods to ethically present sensitive or harmful content, particularly in collections containing public health information, offering recommendations about adding specific content warnings at a sub-collection level. Last but not least, it is worth noting that ensuring greater diversity within archiving teams can also provide more varied perspectives and a heightened awareness of issues specific to the diverse communities that make up the complex tapestry of society.

The participatory approaches mentioned in this section offer a unique opportunity to discover and portray the multifaceted reality of communities existing in a certain country, supporting institutions in their attempt to capture a representative sample of perspectives and experiences about significant events or themes that might be of interest to the public and, especially, future generations of researchers. While the impact of participatory approaches on the actual degree of representativeness in social media collections still needs to be assessed, social media remains an invaluable, unprecedented source of information generated by diverse groups of individuals from different backgrounds and locations sharing their unique experiences about daily life or events. Preserving a representative record of their interactions on these platforms will significantly enhance future generations' knowledge and understanding of our present time.

[166] https://web.archive.org/web/20240724053729/https://www.docnow.io/social-humans/about.html

6.3 Collecting Methods: Web Crawlers, APIs and Technical Challenges

The analysis of the data gathered through the interviews conducted with a wide array of institutions, located in different countries, with diverse legal frameworks and amount of resources allocated to social media archiving activities, has revealed that, apart from some similarities in the tools used, each institution employs a unique combination of methods for capturing social platforms for preservation purposes[167]. Some institutions, such as TNA, have decided to outsource the technical side of social media collection to external companies like MirrorWeb.[168] Others, like INA, have been developing instead ad hoc tools in-house to support the institution's specific requirements. The chosen methods inevitably vary depending on the platform being collected and, more importantly, are tailored to the types and quantities of resources (and developers) available to each institution. As highlighted in a blogpost discussing social media archiving at the UK Web Archive, Byrne (2017) pointed out how archiving social media is inherently different from collecting traditional websites, due to the presence of many dynamic elements that existing tools struggle to capture and the frequency with which both single accounts and the interface of the platforms are updated.

From a technical perspective, social media archiving often involves a long list of trial and error, where the suitability and quality of capture concerning existing tools appear to be highly influenced by the frequent changes implemented by the social platforms, such as new functionalities or interface updates. Additionally, social media companies often try to prevent the capture of data from their sites by imposing undisclosed limits on unauthorised crawling activities (see Section 6.1.3). These restrictions are primarily intended to safeguard their business interests, prompting users to pay for access to their official APIs to obtain the desired information.

Since many of the participating cultural heritage institutions have started archiving social media as a consequent extension of a pre-existing web archiving effort, it is not surprising that one of the most used tools to archive social sites is a web crawler like Heritrix (see Section 2.6). The Internet Archive's Heritrix has been widely used for over two decades to archive websites at scale and, when social media began to appear, archiving institutions started experimenting with the capture of these sites using the tools at hand. While Heritrix manages to archive at a satisfactory level traditional websites, it encounters several issues with more complex, dynamic sites such as social media platforms. Section

[167] A full list of tools mentioned in this thesis can be found in Appendix E.

[168] https://web.archive.org/web/20240801131437/https://www.mirrorweb.com/

4.4.3 discussed some of the problems that the British Library has been facing while archiving (or attempting to archive) social media platforms, such as Twitter or Facebook. In particular, Bingham pointed out how Heritrix is not a high-fidelity crawler and how its inability to properly interact with dynamic content on social sites, such as "read more" buttons, for example, represents an important aspect to consider, especially when evaluating the comprehensiveness and fidelity of the material collected using this tool. Analogously, curators at the BnF expressed concerns about the representativeness and completeness of content archived via Heritrix from Twitter. The limitation of capturing only the top 20 tweets visible on the page may result in gaps when targeted accounts generate a high volume of daily tweets (see Section 5.4.3). Nevertheless, Heritrix remains one of the most widely used tools in the context of social media archiving. Its popularity stems not only from its international adoption but also from its ability to provide a system that helps institutions in managing and scheduling web harvesting as well as checking for duplicates.

For archiving initiatives that use Heritrix via the paid subscription service "Archive-it",[169] such as BnL, one of the main problems is the data limit imposed by their subscription tier. Started in 2005 by the Internet Archive, the Archive-it service allows organisations to harvest, index and provide access to web archived collections. As a usage-based subscription service, institutions establish at the start of the calendar year how much data they plan to archive. The data budget determines the amount of data an institution can archive from the web in a year (e.g. 256GB, 1TB), and it resets to zero at the beginning of each new subscription year. Moreover, data that is unsaved, deleted, or expired from text crawls is not counted towards the budget (Archive-It, 2024). A predetermined data budget requires institutions to carefully plan and closely monitor their crawls to ensure they remain within the allocated limit. Although this can be increased at any point in time during the year, institutions tend to stay within the agreed data budget, particularly when the subscription tier must be pre-approved by library or archive managing boards. This limitation poses a significant challenge, as unexpected events that need to be archived with high priority could require institutions to reduce capture frequency or pause ongoing crawls to allocate space for an emergency collection. Moreover, the BnL pointed out that some of the experiments they run to archive material from social platforms such as Facebook tended to collect a larger quantity of data compared to traditional websites (Els,

[169] See Section 2.3; more information is available here:
https://web.archive.org/web/20240323115608/https://archive-it.org/learn-more/

Interview). In fact, Archive-it guidelines for capturing social media suggest setting a higher data limit for this type of crawls, in order to get more complete captures (Archive-It Help Center, 2023b). However, applying larger seed-level data limits can become problematic for institutions that have a contract with Archive-It and pay a subscription fee based on a set amount of data, especially because, as Ben Els pointed out, "even though it costs more, and it takes up more data budget, we saw that the capture[s] are still very unreliable".

Frequent issues encountered by social media archiving initiatives when performing collection experiments, for example on Facebook, using Heritrix range from missing elements such as images or videos, to a general poor-quality capture of the platform's layout. Moreover, some initiatives are even blocked at the very start, with web crawlers only managing to capture Facebook's login page. As a result, many of the institutions interviewed chose to either pause collection or entirely stop collecting any material from problematic platforms like Facebook (e.g., the BL and the BnF). In this regard, it is worth mentioning the case of LoC. Section 2.3 illustrated the convoluted events surrounding the LoC's agreement with Twitter Inc. for the preservation of the Twitter historical archive, including content from the platform's creation to 2010. In the years after the 2017 blogpost where the library announced it would archive only selected content from Twitter (Library of Congress, 2017), the Library continued to harvest social media less and less due to the numerous technical issues encountered, until a decision was taken in 2020 to develop a separate project to the one concerning websites. The project would investigate solely the preservation of social platforms, as one of the LoC Digital Collection Specialists explained in the interview carried out in April 2022:

> We took all the social media out of the crawl about two years ago and we just put a halt to all social media, and instead we started this project to research what it would take for us to archive social media, especially Twitter, Facebook and Instagram.

LoC's decision to explore alternatives to web crawlers, which had been their primary method until the establishment of the new social media-focused project, underscores the need to develop specific strategies that take into account the different nature and characteristics of these platforms. LoC's team has so far experimented with various APIs and tools (e.g., Twarc, YouTube-dl). However, at the time of the interview, the first official crawl was still in the planning phase, and no update has been released to date.

The Archive-it Help-Center page "Social media and other platforms status"[170] offers an insightful summary of known technical issues surrounding the capture of platforms with Heritrix and replay with Wayback (see Section 2.3). A rapid browse of the provided information reveals that only a few of the social platforms that can be archived using the tools included in the Archive-It service have no known issues. In Chapter 3, results from the survey revealed that some of the most archived social media platforms are Twitter, Facebook, Instagram, and YouTube. According to the Archive-it Help Center status page, however, all of the mentioned platforms present a series of issues, ranging from incomplete captures to platforms preventing crawlers from accessing certain types of account. In some cases, Archive-It even discourages the capture of certain sites. This is the case with Twitter. Following the recent changes implemented by the new owner since late-2022, Archive-It has gone from flagging a few issues mostly related to replay in March 2023[171] to advising Heritrix users to pause crawling activities on Twitter, due to changes in content visibility and reading limits.[172] In fact, in July 2023, Elon Musk announced the introduction of restrictions on the number of viewable posts for both verified and unverified Twitter accounts, in an attempt "to address extreme levels of data scraping & system manipulation"(Musk, 2023). These rate limits were initially presented as a temporary solution but have since become permanent (Roush, 2023).

Rather problematic is also the collection of Instagram using Heritrix, as Meta is blocking captures of most organisations and public profile pages (Archive-It Help Center, 2023a). Due to its popularity among users worldwide (Statista.com, 2023b), archiving institutions are keen to preserve Instagram, despite the challenges in capturing it. The reason can be found in the numerous limitations imposed by Meta in terms of access to information and for its complex and dynamic layout with which tools like Heritrix are unable to interact. Nevertheless, some workarounds have been identified. For example, the BnF is successfully collecting Instagram profiles using a combination of Heritrix and

[170] https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

[171] Archive-It Help Center "Social media and other platforms status", web archived on 25/03/2023. https://web.archive.org/web/20230325144024/https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

[172] Archive-It Help Center "Social media and other platforms status", web archived on 30/11/2023. https://web.archive.org/web/20231130180611/https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

Picuki,[173] an online tool that allows the exploration, viewing, and sharing of publicly available Instagram content. As illustrated in Section 5.4.3, the BnF uses Heritrix to crawl the URL generated after retrieving a specific hashtag or account through Picuki.

An alternative, high-fidelity tool to archive social media platforms is Rhizome's Webrecorder. Webrecorder is a web archiving service that involves generating a recording of a webpage that the user browses, recording every interaction. This allows the capture of social media content and any interaction or dynamic content with a higher degree of fidelity than web crawlers like Heritrix, helping also to preserve a trace of how social platforms functioned as well as the user experience to some degree. There are a few initiatives that have tested the Webrecorder tool alongside other collection approaches for the purpose of web and social media preservation. For example, the Library of Congress recounted that before 2020 they experimented with Webrecorder but deemed it not "an optimal solution" for the scale of their web collection (Thomas et al., Interview). Nevertheless, Digital Preservation Specialists at the LoC involved in the new project focusing solely on social media explained that they were planning new tests with the Webrecorder tool at the time of the interview.

Among the institutions that tested Webrecorder, several noted that capturing social media accounts or events with this tool involves a significant amount of time and dedicated curators. Experiments carried out by the BL (Bingham et al., 2020; see also Section 4.4.3, TNA (UKGWA Team, 2020), and independent researchers (Hawes, 2020) have documented the laborious process required to meticulously prepare ahead of the capture, even when using the autopilot function. In particular, Hawes (2020) noted some difficulties in capturing Instagram Stories and user tags on posts when using the Webrecorder autopilot. This required her to supplement the automated collection with manual captures, which necessitated a careful planning of the sequence of interactions performed and elements to be recorded. In 2020, TNA has also used Webrecorder to capture frequently updated resources like the COVID-19 arcgis.com cases dashboard, observing however that a number of significant resources relating to COVID-19 were impossible to capture regularly because of the time required to manually record all those interactions (UKGWA Team, 2020). Webrecorder certainly represents a useful method to capture websites and platforms with peculiar features that cannot be otherwise archived or to preserve a trace of browsing experience for future generations. However, the time and resources required to collect social media using this tool do not appear to be

[173] https://web.archive.org/web/20240401042437/https://www.picuki.com/

compatible with archiving practices at scale. To mitigate the issue, Ilya Kramer, developer and creator of Webrecorder, has been working with the IIPC to develop a browser-based web crawler called "Browsertrix" which aims to complement collection activities made using Heritrix. Some institutions like the BL have planned to start testing it with social media, hoping to achieve better results than with traditional web crawlers (see Section 5.5).

The examples provided in this section demonstrate how hybrid social media archiving approaches, combining automated tools like Heritrix with more manual or semi-automated tools such as Webrecorder, along with inventive procedures involving curation tools or command-line programmes like YouTube-dl (see Section 5.4.3), have enabled institutions to circumvent technical and legal limitations posed by social platforms. Despite being laborious and resource-intensive to implement, these approaches offer good solutions to the complexities of archiving ever-changing and dynamic content on social media.

A different method to archive data from social media is to use Application Programming Interfaces (APIs). Thomson (2016) fittingly described APIs as a sort of "backdoor into a social media platform […] to call raw data directly from the platform" (p.7), providing access to rich metadata that is not visible on social platforms (Pehlivan et al., 2021). The type of data and metadata that can be obtained through APIs varies from social site to social site, as well as how far back in time data can be retrieved. Earlier in this chapter (section 6.1.4) I discussed at length the implications of the changes and constraints imposed by social media platforms such as Facebook and Twitter on data access, illustrating their impact on the interviewed institutions. Nonetheless, it is worth noting here some additional opportunities and limitations related to harvesting data through APIs for preservation purposes emerged during the interviews. Moreover, I would like to point out that among the interviewees, the number of institutions making use of platforms' official APIs – predominantly to access Twitter data – appear rather small compared to those using traditional web crawlers alone or in combination with other tools.

One of the issues raised by institutions collecting information via APIs is the often complex process to obtain access to it. Section 5.4.3 described the hurdles experienced by INA to apply for a Facebook and Twitter official API account as a cultural heritage institution and how obtaining one did not offer any additional assistance or safeguards for the institution. Similarly, LoC pointed out that one of the main blocks concerning the

collection of Facebook material came from Meta due to their application process and how they hand out developer access (Thomas et al. Interview).

Following the announcement of the introduction of the Twitter Academic API in 2020, various social media archiving initiatives applied to gain access to it. This API, before being deprecated in 2023, represented a potential optimal solution for many archiving institutions as it granted access to the platform's full historical archive (Twitter Inc., 2021; see also Section 6.1.3). However, only one of the institutions interviewed (INA) managed to secure access to it (see Section 5.4.3). Conversely, others, including the KBR, LoC and the KB, saw their applications being rejected despite going through a long assessment process which involved many questions and stages to ascertain the nature of the project for which they required access to that specific API. LoC recounted that "Twitter was very interested in […] how we were going to use the data, what we were going to be reusing and displaying" (Thomas et al., Interview), especially because they were applying as a cultural heritage institution. As also mentioned in Section 6.1.4, there is no specific route established by any of the existing social media platforms specifically for cultural heritage institutions seeking to preserve material related to their country, despite most of these institutions having a legal mandate to do so.

On top of this, the paid tier system implemented by X/Twitter in June 2023 adds further obstacles for many web archiving institutions. With already limited budgets to support web preservation efforts in the long-term, many institutions would not be able to afford the expensive fees required for the "Pro" or "Enterprise" tiers, which are the only options providing access to historical data. For this reason, initiatives like INA, TNA and the *Netarkivet* that were collecting Twitter data via its official APIs have since then stopped archiving this platform for the time being due to the lack of adequate free or reasonably affordable access options that would allow them to sustainably archive content at scale.

Some concerns also emerged during the interviews regarding the opaqueness surrounding data sampling when archiving social media through APIs. Although most archiving initiatives do not aim for exhaustiveness when it comes to social media collections, it is essential for them to understand how sampling mechanisms work and how representative the information gathered through APIs is, so that this information can be properly documented (Littman et al., 2018; Pehlivan et al., 2021). Driscoll & Walker (2014) argued that APIs have indeed to be regarded as a "black box", where the lack of transparency in how sampling of data is conducted may introduce biases into the portion of data archived. This opaqueness may open memory institutions to the risk of

preserving potentially unbalanced views and opinions regarding certain topics, particularly, in election campaign collections where it is essential to preserve a balanced plethora of opinions from all the parties involved. For this reason, understanding APIs sampling mechanisms is essential to provide researchers using web archived resources with a comprehensive knowledge of the potential biases and gaps introduced in the collections.

Some of the social media archiving initiatives interviewed pointed out alternative API-based tools to harvest information when access to the official API is not available, such as Social Feed Manager[174] and Instaloader.[175] For example, the BESOCIAL team at the KBR experimented with Social Feed Manager for the capture of content from Twitter. Social Feed Manager is an open-source tool developed by researchers and library staff at the George Washington University to collect social media, including Twitter and Sina Weibo (Social Feed Manager, 2018). This tool allows institutions to collect data from social media via APIs, while complying with platforms terms and conditions (Acker & Kreisberg, 2020). When the BESOCIAL team's application for access to the Twitter Academic API was rejected, they explored alternative methods to harvest social media. After a careful consideration of the open source and commercial tools available on the market, which was documented in detail in their WP1 Report (Chambers et al., 2021), they decided to incorporate Social Feed Manager in their workflow to archive posts and hashtags from Twitter. The main problem they experienced with this tool was related to the limit imposed on the number of hashtags that could be added per each collection. Fien Messen explained that they were able to create collections including a maximum of ten hashtags, requiring them to manually create new collections for every ten hashtags they selected for harvesting, which resulted in a rather time-consuming task (Geeraert and al., Interview).

Archiving institutions like the KBR and LoC have managed to archive public content from Instagram using Instaloader, a command line tool that allows the downloading of pictures and Reels, including the associated caption and metadata. LoC, for example, decided to test Instaloader after halting collection of Meta platforms through its official APIs. However, Digital Preservation Specialists at LoC explained that testing this tool in a scalable manner was challenging as "it runs into Instagram timeout policies pretty quickly, so you can only access a certain number of accounts for a certain number

[174] https://web.archive.org/web/20240528092738/https://gwu-libraries.github.io/sfm-ui/about/

[175] https://web.archive.org/web/20240807055139/https://instaloader.github.io/

of minutes, otherwise you get timed out for 24 hours. So, it's a pretty slow method, but it works well" (Thomas et al., Interview). The main issue with this type of method appears indeed not to be related to specific technical issues with this tool but rather to the limitations imposed by Meta. In this regard, the KBR commented that they tried to resolve this issue by using scripts that would mimic the way humans browse content to avoid triggering platforms' security measures against automated web scraping. However, the KBR found this method was not sustainable for their project (Geeraert et al. Interview). Moreover, Fien Messen (KBR) underscored that because of the possibility of being blocked by the platform, they had to limit the number of accounts and hashtags they initially planned to harvest. Although Instaloader keeps track of requests and sends a notification whenever the rate limit is about to be reached, institutions using this tool have experienced instances where their accounts were blocked from the platform.

An interesting tool worth mentioning, although not cited by any of the archiving institutions participating in this study, is *Zeeschuimer*.[176] This browser extension, successfully used in the collaborative web preservation project Saving Ukrainian Cultural Heritage Online (SUCHO),[177] allows users to collect data directly from the platform interface. However, *Zeeschuimer*, as defined on its GitHub page, is designed primarily for researchers. Its suitability for a national cultural heritage institution, particularly in the context of e-legal deposit legislation, remains to be verified. As Bingham observed (see Section 4.5), there are not many tools that are compatible with electronic legal deposit contexts, thus restricting the pool of methods that most social media archiving institutions can draw technical solutions from.

Finally, it is important to highlight the crucial role played by developers and other specialized technical profiles in social media archiving initiatives. Having full-time developers dedicated to web archiving projects is not the norm, as was underscored by Geeraert at the end of their interview. She noted that one of the key lessons from the BESOCIAL project, which they planned to implement in the new KBR-based project called BelgicaWeb,[178] was the necessity of hiring a developer to provide more specific technical support for the team. Geeraert pointed out the numerous challenges in attracting

[176] https://web.archive.org/web/20240108140527/https://github.com/digitalmethodsinitiative/zeeschuimer

[177] https://web.archive.org/web/20240430013245/https://www.sucho.org/

[178] Belgicaweb is a BELSPO funded project which aims to make Belgium's born-digital heritage accessible and Findable, Accessible, Interoperable and Reusable (FAIR).
https://web.archive.org/web/20240523232033/https://www.kbr.be/en/projects/belgicaweb/

ICT specialists to a federal research institution due to wages not being competitive enough compared to the tech industry. Web and social media archiving teams are often rather small, ranging from a couple of curators to fewer than a dozen members, including developers who are sometimes shared with other departments. At the *Netarkivet*, for instance, the only full-time role in the web archiving teams is the Web Curator, whereas developers and other tech professionals are involved part-time in numerous other projects across the KB, often leaving insufficient time to dedicate to specific research support activities, such as creating bespoke datasets for researchers (see Section 6.4). The diversity of national legal frameworks and technical expertise, as well as the resources available to support social media archiving efforts in the long term at individual institutions, are at the root of the variety of methods and combinations of tools described in this section. While there is no "one-size-fits-all" solution at present, testing new approaches and developing new tools through national and international collaboration among institutions, such as the efforts made by IIPC members in partnership with Webrecorder, represent a promising path for advancing social media archiving practices.

6.4 Access to Social Media Collections and Public Engagement

Web and social media archives preserve content generated online to safeguard a trace of cultural heritage online and to ensure that this remains accessible through time to the public. Scholarship has explored the numerous challenges related to providing access to web archives, particularly those subject to restrictions imposed by national legal deposit legislation and how these affect their use for research purposes (e.g., Bell et al., 2022; Gomes, 2017; Thomson & Kilbride, 2015; Winters, 2020). In this section, I will focus on issues and specific aspects related to providing access to archived social media mentioned by web archivists during the interviews. It is worth pointing out that providing access to archived social media resources often shares similar issues with archived websites, especially when these originate from limitations imposed by national legal frameworks (Gooding & Terras, 2020b; Winters, 2020). Such restrictions can significantly impact the overall engagement with web collections, potentially placing web and social media archives in a vicious cycle of low funding, and consequently limiting future opportunities to enhance their service. This section will begin with a summary of the type of access available at participant institutions, some of the challenges they face – such as increasing

public engagement and discoverability of social media content – and additional access services provided to users, including data extraction and preliminary data analysis.

When planning to use social media collections created under national legal deposit legislation, researchers must consider the prospect of having to sustain additional costs and allocate time in order to travel to specific buildings where access to the archived web material is provided (Healy et al., 2022). This constitutes a major obstacle particularly for early career researchers and people with disabilities who may not have resources and/or the physical ability to travel to a specific building. In most of the institutions interviewed, users must be accredited researchers and thus possess a reader card, as in the case of the BnF and INA (see Section 5.4.4) or go through a specific application process providing details of their research project to be granted on-site and remote access to the web archive, such as for the *Netarkivet* at KB (see Section 6.1.3). Others, like BL (see Section 4.4.4), may provide limited online access through their portal, allowing users to browse collections, although only a small portion of their archive can be viewed outside their premises unless BL has been granted permission from the rights-holder — which is rarely the case for social media content (see Section 4.4.1). A small number of institutions, thanks to their specific legal mandates to either preserve a record of the Government's online presence (TNA) or web material in the interest of their country (Arquivo.pt), are exceptions to this general trend and provide unrestricted access to their collections.

Restrictions on access strongly impact the level of usage of web and social media collections. In fact, while studies concerning web archives with online unrestricted access, such as the UKGWA, revealed that they appear to be well used (Bell et al., 2022), those with on-site access only remain largely unexplored (Masanès et al., 2021). Web archivists across all the institutions interviewed underscored the need to improve access to collections. This sentiment has grown particularly strong during the COVID-19 lockdown which saw entire collections, especially web archives, remain inaccessible and unused. Therefore, extending access to other buildings and, especially, opening to remote access for accredited researchers could be a key way to improve awareness and overall engagement with archived social media resources, especially given the low rate of requests received to date to consult this relatively new material.

Moreover, unlike traditional archives or other legal deposit collections, there is often very limited copying or downloading from web and social media collections, with researchers being mostly unable to take home copies of the consulted material, due to limitations imposed by national legal frameworks (see Section 6.1.3). In addition, the

survey conducted by Healy et al. (2022) on the skills and challenges related to the use of web archives, highlighted that another obstacle to research engagement with web collections is the lack of standards or guidelines on how to cite material from web archives. For example, the "Persistent Web Identifier" (PWID) has been recently developed with the aim of supporting references to material in web archives with restricted access by providing detailed information, including the web archive where the reference was found and validated, thus eliminating any possibility of ambiguity (Jurik & Zierau, 2017; Zierau, 2022). However, the widespread adoption of such a persistent identifier across different web archives and in researchers' data management practices may take a long time. Important steps towards a standardised way to cite web archived resources have been made with the inclusion of web archives in the ISO 690:2021(en)[179], which offers guidelines for the citation of web pages that are only available through web archives. Nevertheless, although crucial in facilitating the use and reuse of archived material for research purposes, citing archived web and social media resources from legal deposit web collections or archives with restricted access remains a significant challenge, particularly in terms of ensuring the transparency of the research methods employed (Healy et al., 2022, p.106).

Some of the longest-standing web archiving initiatives (e.g., the BL, the BnF, INA, the BnL) are actively working to improve access to and engagement with web collections, particularly by seeking to increase the number of access points at satellite institutions. A good example in this sense is the *ResPaDon* project mentioned in Section 5.4.6. Although France already has one of the densest networks of institutions providing access to the French web archive, the BnF has been supporting the *ResPaDon* project to further improve access by installing terminals at additional locations across the French territory, especially at university libraries (Désos-Warnier, 2023). Providing access on the premises of university libraries has been identified as a potential way to raise awareness about web and social media collections and their usage for research among students and researchers. A similar approach has been adopted by the *Netarkivet* which is accessible both remotely and from physical terminals located at the KB's main building in Copenhagen and its other branch in Aarhus, which also functions as Aarhus University Library. Providing access to the web archive, along with other services offered by the university library,

[179] ISO 690:2021(en) Information and documentation — Guidelines for bibliographic references and citations to information resources. Available here: https://web.archive.org/web/20240830153047/https://www.iso.org/obp/ui/#iso:std:iso:690:ed-4:v1:en

creates opportunities for students to browse web and social media collections and potentially make use of them in the course of their studies and beyond.

The improvement of the service and expansion of the number of access locations have also been considered by the six UK legal deposit libraries. While providing wider access to the UKWA would require a revision of the law – a process that will necessarily take time, given that more than ten years passed between the first draft of the NPLD legislation and its enactment – the UK deposit libraries have shown an overall concrete willingness to improve their service by forming a Legal Deposit User Forum (LDUF) in 2023. The LDUF aims to gather feedback from users of the legal deposit collection, including the UKWA, to influence the design and development of the service.[180] However, due to the difficulties BL faced following a ransomware attack that has affected most collections including the UKWA since October 2023,[181] the work of the user forum is still in its early stages.

Raising awareness among academic circles remains a priority, especially in terms of developing reproducible research frameworks that can serve as starting points for scholars interested in investigating topics such as the history of the web and social platforms (Byrne et al., 2023). Nonetheless, it is equally important to foster visibility and raise awareness of the importance and role of web archives among the wider public. Increasing engagement with web and social media collections could strengthen the case at higher levels for broadening access to the preserved web and social media material. The problem of access and awareness of web archived collections has been extensively discussed both within individual archiving institutions and international networks like the IIPC. Els (BnL) noted that together with other members of the IIPC they were exploring ways to open collections to the wider public without the constraint of physically visiting the building, stating that "all web archives would like to have more visitors and more requests to use the web archive". This resonates even more for archived social media material, which is still not widely archived due to the many legal, ethical, and technical constraints that hinder the ability to scale up collection efforts and, therefore, constitutes

[180] More information about the Legal Deposit User Forum is available here: https://web.archive.org/web/20240731152248/https://www.lib.cam.ac.uk/news/if-you-use-our-legal-deposit-collections-we-need-your-help

[181] Details of the cyber-attack can be found here: https://web.archive.org/web/20240628204851/https://en.wikipedia.org/wiki/British_Library_cyberattack

a relatively small portion of the existing collections compared to archived websites. The limited number of accounts captured as well as restrictions on access and (re)use of archived resources, can results in researchers preferring to independently harvest their own datasets rather than using what is available at national institutions.

Participatory approaches described in Section 6.2.2 may help increase public engagement with social media collections, especially among minority groups and local communities by directly involving them in the preservation process of their identity and history (Terras, 2015). Crowdsourcing campaigns, such as those organised by the BnF, the BnL, the BL or the KBR (see section 6.2.2), offer significant opportunities to enhance public outreach and raise awareness about web and social media collections, underscoring their societal and cultural value beyond academia. In this regard, an interesting example comes from an exhibition prepared by the Museum of London (MoL) in the summer of 2022. The Museum of London offers an itinerary through the history of London and its inhabitants, showcasing items from prehistoric times to the present day. As part of their attempt to preserve traces of the lived experience of Londoners regarding significant events (see also Section 6.1.3), curators of the museum captured content from social media on the occasion of the Olympic Games in 2007 and the COVID-19 pandemic in 2020. One of the challenges that Foteini Aravani, Digital Collections Curator at the MoL, had to face once the content had been acquired was how to engagingly display and make these peculiar born-digital objects available to the public. For the exhibition titled "Into the Twitterverse", which I visited in July 2022, Aravani and a PhD student tested four different methods of displaying content collected from Twitter. This included displaying the text as "digital rain", taking inspiration from the shower of green computer code falling from above of the famous sci-fi film franchise "The Matrix"; placing it on an interactive map of London which the user could explore through a touchscreen; or showcasing it as printed and framed pictures, like artworks in a gallery (Figure 13).

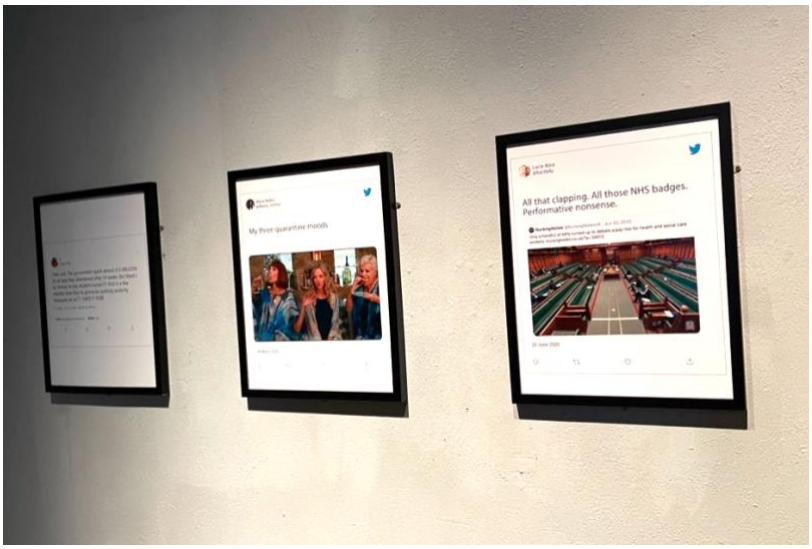

FIGURE 13: "Printed Screenshots", "Into the Twitterverse" Exhibition, Museum of London, July 2022.

Although the exhibition was only a short-term experiment to gather insights on visitors' reactions, it was an interesting experience that stimulated the public to perceive tweets in a different way, beyond their original context. By decontextualizing Twitter posts and displaying them as objects within a museum collection, MoL encouraged visitors to reflect on the meaning and significance of what are often perceived merely as posts, highlighting instead their value as museum-worthy artifacts. Similar experiences can represent opportunities to raise public awareness about the cultural value of content shared on social media, especially when associated with an effective publicity campaign aimed at engaging a broad spectrum of society.

When it comes to providing access to and replaying archived material, social media presents additional challenges compared to websites due to its dynamic nature and constantly changing interfaces (Thomson, 2016). Depending on their mission and scope, social media archiving institutions either opt for preserving the "look and feel" of social platforms (e.g., the BL, the BnF, the BnL), or the raw data (e.g., INA, KB, KBR). Preserving the full experience of using and interacting with other individuals on social media will never be possible, as even high-fidelity tools like Webrecorder often capture content through the perspective of and limited to the itinerary established by the curator (Hawes, 2020). Moreover, unexpected changes in platforms' terms of use, interfaces and access to APIs, as illustrated in the previous section, can affect the institutions' ability to correctly replay the harvested data. For example, the BnF faced issues replaying YouTube content due to the deprecation of the legacy interface, which required them to manually reconstruct the structure of the page to properly replay the archived material (see Section

5.4.3). Besides, several initiatives do not archive information related to interactions, such as the number of "likes", shares, or reactions (e.g., emoji reactions on Facebook), as keeping this data up to date would be too resource intensive (Newing, Interview). For other material, such as comments and retweets, most institutions have decided to exclude them from capture due to privacy and data protection concerns, consequently preserving only one side of conversations taking place on social media (Pennock & Beagrie, 2013). Advertisements either embedded in archived websites or shown alongside social media content are also often excluded from capture (see Section 5.4.4). Given the interest that interactions, comments, and other elements may have for present and, especially, future researchers, it is essential for institutions to document and flag missing items and gaps so that users can critically engage with collections. Several web archives participating in this study have mentioned the work they have been doing to compile additional documentation which will be essential for researchers to understand curation choices or technical constraints. For instance, Klindt-Myrvoll mentioned that the *Netarkivet* team at the KB is currently working on assembling specific documentation about their web and social media collections, which will be made available through the access interface.

Archiving initiatives including LoC, the BnL and Arquivo.pt, reported replaying issues for platforms like Facebook and Instagram, where changes in the terms of use (see Section 6.1.4) have led to restrictions for web crawlers, resulting in institutions being able to capture only "404 error" pages or the platform's login pages. In addition, difficulties have also been observed in replicating the infinite scrolling feeds found on the live TikTok platform for accounts with multiple videos archived using Archive-It's web crawlers. In some cases, according to the Archive-It Help Center (2023), pages with multiple TikTok videos can only replay the first video, while the remaining ones can only be watched by opening them in a new browser tab. Geeraert also commented about the issues in reproducing the experience of certain features on TikTok, especially the "For You" page, which is heavily personalised and dependent on individual users' interactions with the content.

Social media archiving institutions tend to replay and provide access to content using access platforms such as the Internet Archive's Wayback Machine whose interface can be personalised based on the institution (e.g., the BnF, the BnL, Arquivo.pt). Other archiving initiatives like the UKWA and the *Netarkivet* use SolrWayback instead.[182] SolrWayback is a search interface developed by the KB as a web application similar to the

[182] https://web.archive.org/web/20240414035457/https://github.com/netarchivesuite/solrwayback/

Wayback Machine, that enables users to browse historical WARC files. Conversely, at TNA, access to archived social media channels, which is separate from archived websites, is provided through a custom interface created by MirrorWeb,[183] the company to which TNA outsources the capture of in scope online material. Finally, there are some institutions that have opted for developing their own replaying interface from scratch as in the case of INA (see Section 5.4.4) or the KBR. Over the duration of the two-year project, the latter was only able to produce a sample interface by reusing a tool previously developed at the University of Louvain (Geeraert et al., Interview). The BESOCIAL project tested some features they wanted to potentially implement in a final version of the access platform, which would include some analysis tools like charts, word clouds, and hashtags occurrences. The concrete development of a user-friendly access platform and API that would enable access at data level for born-digital heritage preserved at the KBR, will be explored in the new, BELSPO funded project called "BelgicaWeb", which is set to begin in the second half of 2024.

Archiving institutions offer several entry points for users to access and explore archived social media material, including URL or keyword search, topic-based browsing, derived data, or seed lists. It is worth reiterating that most initiatives do not distinguish between archived websites and social media accounts (see Section 6.2), as both are often aggregated into the same thematic collections and thus searchable through the same interface. Conversely, institutions like TNA and INA provide distinct access to their web and social media archives, requiring the user to select a specific section on the portal to browse or search materials from archived social platforms. For example, the UKGWA portal accessible through TNA's website offers the ability to either search the full archive of social media channels preserved, or select between YouTube, Twitter and Flickr channels to browse only material related to a specific platform (Figure 14).

[183] https://web.archive.org/web/20240710091550/https://www.mirrorweb.com/

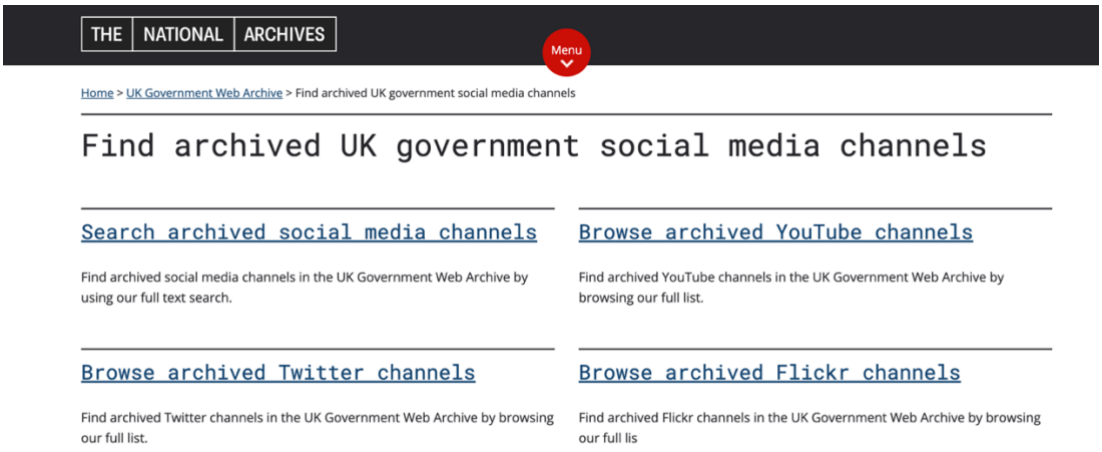

FIGURE 14: Overview of access points offered on the UKGWA's Social Media Archive section (June 2024).

Almost all social media collections accessible either online or on-site at the archiving initiatives interviewed are full-text searchable using one or more keywords. Users can also perform URL searches, but they must know the exact URL of the archived source. This can be rather challenging when, as described in Section 5.4.3, institutions like the BnF may decide to archive only a specific type of URL among the different types available on platforms like YouTube for the same video. Therefore, it is essential for institutions to provide detailed information about the types of searches that can be performed and values that can be used to refine queries through the main search bar (e.g., Boolean terms, quote marks). Search results can be further refined by using the Advanced Search or the facets usually located on the left-hand side of the screen. Some institutions, like TNA, offer users the ability to filter by platform, year, channel, or include/exclude a given keyword. Moreover, in Section 5.5 I described some of the new filters that INA was going to implement in the new *WebMédia* interface, such as excluding retweets and mentions or identifying in INA's Twitter Archive section images with the same URLs. Image search is a feature that other archiving initiatives have been testing. In particular, the *Netarkivet* allows users to search for pictures related to specific topics in certain areas of Denmark, based on the location metadata collected as part of the data. Similarly, users of Arquivo.pt can browse images related to queried keywords.

Some web and social media archives also provide basic data visualisation such as graphs and N-grams, either through the replay interface or on demand. INA offers various types of data visualisation, including lists of top ten hashtags or frequent emojis used in the tweets retrieved in relation to a query, as well as unique ways of showcasing Twitter feeds, as in the case of the SocialTV (see Section 5.4.4). The *Netarkivet* also

provides some data visualisations (e.g., visualization of search result by domain, linkgraphs, domain stats, Ngram and wordclouds) which can be exported and downloaded in different formats directly from the access interface (Lauridsen, 2021).

Thematic collections are also good access points to explore web archives especially when users do not have a clear idea about the topic they want to research (see Section 4.4.2). However, since thematic collections often gather both websites and social platforms, isolating material captured from the latter can sometimes be difficult, as well as identifying which, among the many web collections available, include social media material. This information is often not clearly conveyed on institutional websites. Many only mention the inclusion of social media in the web archive without specifying the types of platforms archived or the collections in which users can find them. In this regard, the "Guided Tours" made available on the BnF website offer a suitable solution to this issue, providing detailed lists of URLs related to the material associated with a specific topic (see Section 5.4.4).

Equally useful appears in this context the seed lists made publicly available by some institutions such as the BnL, the BnF and NZSL on their websites. These lists of collected URLs allow researchers to get an idea of the material contained in the web archive prior to their visit. Making seed lists or derived data available online, or at least shareable upon request, facilitates engagement with the collections, especially for research concerning transnational events (Aasman et al., 2021). For instance, studies conducted within the WARCnet Network provided opportunities to combine and experiment with derived data and metadata from various web archives, enabling the development of a shared, transnational corpus across European countries and offering new perspectives on how to use web archived resources (Aasman et al., 2021, 2022). Derived data represents an optimal option for those web and social media archives that for legal reasons cannot provide bulk access to primary data. This is the case with the UKWA, which provides access to several derived datasets, API services and a few tools through the "UKWA Open Data" GitHub repository.[184]

Some institutions also offer support for the capture of bespoke datasets. At the BL, for example, this service is limited by the fact that data can only be viewed within the institution's premises as the UKWA can't share any data unless it is a derived dataset (see Section 4.4.4). The Netarkivet can provide on-demand datasets to researchers, but depending on the type and size of the dataset, the library may charge a fee to compensate

[184] https://web.archive.org/web/20231221073205/https://data.webarchive.org.uk/opendata/

for the additional hours of work. In fact, engineers at the KB in Copenhagen can only dedicate a limited amount of their time to creating bespoke datasets upon researchers' requests, as they work part-time as members of the web archiving team. During the fieldwork carried out at the KB in September 2022, I had the opportunity to join a meeting between the *Netarkivet* team and a researcher seeking to collect data from Twitter through the library. Participating in this meeting provided a unique insight into the dynamics and the work done by the web archiving team, highlighting their expertise and the limited resources (especially human resources) dedicated to web and social media archiving initiatives. This is especially notable in the longest-standing institutions like the *Netarkivet*, which can count on a relatively large team of experts, making the challenges faced by smaller teams and emerging initiatives even more evident.

In the vicious cycle that sees social media archiving initiatives running on often insufficient resources, constantly finding a balance between legal and technical constraints, improving access to collections while ensuring they are safeguarded in the long-term remains a crucial point. This further underscores the need to raise more awareness about the existence and value of web and social media collections within and beyond academia.

6.5 Long-Term Preservation

In the course of this study, it has been remarked several times how social media is a highly ephemeral content, whose characteristics make it challenging to capture and replay for access, so much so that the Digital Preservation Coalition included social media platforms in multiple categories in its list of "Endangered Digital Species".[185] Although a small but steady number of national legal deposit organisations and independent NGO projects are archiving or looking into archiving social platforms, the long-term preservation of archived social media material still requires further development.

Responses to the interviews showed that institutions have, overall, adopted for web and social media archiving similar digital preservation strategies to those implemented for other digitised and born-digital items preserved in the respective memory institutions' main storage systems. Strategies involve, for instance, the creation

[185] https://web.archive.org/web/20240218121656/https://www.dpconline.org/digipres/champion-digital-preservation/bit-list/social-media

of multiple copies stored safely across different nodes in separate, distant locations, combined with tape storage for long term preservation. As described in Section 5.4.5, the BnF mentioned how web archived content is preserved in its "Scalable Preservation and Archiving Repository", a trusted digital repository that combines disk preservation for access to the archived resources and tape for long-term preservation. Content archived as part of the UKWA, including social media, is kept in its main digital library store, which is physically located in a separate location to the main BL building in London. Moreover, archived data is also replicated across nodes located in each of the six legal deposit libraries.

However, it appears that, while all institutions consider long-term preservation as a primary concern, not many social media archiving initiatives have compiled written documentation detailing digital preservation procedures specifically developed for archived social media content. This trend seem to confirm the point raised by Vlassenroot et al. (2021), who observed a general lack of common understanding about the meaning of "digital preservation procedures" and preservation formats among web archiving institutions. In fact, when asked to describe long-term preservation challenges or practices for social media, participants in this study mostly mentioned the use of the Web ARChive (WARC) file format and the recently released Web Archive Collection Zipped (WACZ),[186] which serves as a package format for WARCs. Despite being originally developed as a storage format and not specifically created for long-term preservation (Vlassenroot et al., 2021), the WARC file format appears to be largely intended also as a standard to preserve in the long-term content and related information crawled from social platforms, or collected through social media APIs (Pehlivan et al., 2021). An exception to this trend comes from INA which instead uses the Digital Archiving File Format (DAFF), a file extension developed, as Mussou explained, to support the special requirements of the *Institut national de l'Audiovisuel* and the type of born-digital material they archive which mainly includes videos, and text with embedded video files (see Section 5.4.5). It needs to be pointed out that the lack of specific information related to long-term preservation policies for social media may be related to the fact that interviews were mostly conducted with curators whose focus is not predominantly on digital preservation strategies as this task often falls under the responsibility of a separate, specific team. Nevertheless, it demonstrates that there is a need to conduct further research and raise more awareness

[186] https://web.archive.org/web/20240522182322/https://webrecorder.net/2021/01/18/wacz-format-1-0.html

about the potential issues surrounding the long-term preservation of archived social media material.

The long-term preservation of social media content is not limited to strategies of bit preservation and accessibility through time, but also involves the safeguarding of other information that will be essential for future generations to understand the context and user experience. In this regard, it is worth noting how social media is very often accessed by users through their mobile devices, which provides an experience of navigation and of content fruition essentially different from the desktop version. Yet, archiving initiatives often tend to collect the latter versions rather than the mobile one. Despite memory institutions' efforts to replicate social media navigation and interfaces as closely as possible to the live web, capturing aspects such as community experience and the immediacy of communication is inevitably challenging, as these dynamics are difficult to encapsulate in static snapshots of what was originally in flux. Moreover, the frequency with which social media platforms modify functionalities and layouts, and the constant emerging of new platforms, pose additional challenges to the preservation of social media content. Archiving institutions must constantly adjust the settings of their capturing tools or seek new collection methods to match these changes. This is demonstrated by the example of the BnF's experience with YouTube in Section 5.4.3 and the massive document that curators use to record all the tests and changes implemented over time. Tools like Webrecorder can certainly support the preservation of user experience and context, although, as mentioned in Section 6.3, the resulting material often portrays a mediated experience of the user's journey planned by a curator. Nevertheless, Webrecorder does not represent a sustainable and scalable approach neither in the short nor in the long term, as pointed out by many interviewees. While the sustainable preservation of archived social media with the resources currently available remains a pressing problem across memory institutions worldwide, more research is certainly needed to develop specific strategies to ensure the long-term preservation of the content (and context) archived from these platforms.

Conclusion

This chapter has offered a comparative analysis of the challenges encountered and practices implemented at twelve archiving initiatives at different stages of the development of their social media collections. Using relevant examples provided by web archivists about various aspects of social media archiving, this chapter illustrated how the

unique nature of social platforms poses new obstacles, setting this material aside from traditional websites and thus requiring specifically developed strategies for effectively preserving them.

The cross-national comparative study highlighted the ongoing conflict between the empowering aspects of national legal frameworks and the limitations they impose on archiving initiatives, especially those operating under national legal deposit legislation. Receiving a legal mandate is an essential factor for consolidating and scaling up social media collections. Yet, unclear or too strict definitions of what material should be archived, along with necessary limitations such as adherence to territoriality criteria or the public availability of the accounts/posts to be archived, generate concerns among web archivists. Safeguarding the privacy and copyright of individuals and rights-holders, albeit important, generates additional restrictions on the type of material that can be preserved and, even more so, on the modalities by which this material can be accessed, used, and re-used. Limitations imposed by social media companies in the attempt to protect their main source of revenue add a further layer of constraints to an already complex archiving endeavour. Moreover, the unexpected and continuous changes implemented in platforms' terms of use, along with the ever-changing nature of social sites and the lack of specific communication channels for assisting cultural heritage institutions in preserving these platforms, make archiving social media particularly difficult and in certain cases (e.g., Facebook) impossible.

Selection of content is heavily influenced by legal constraints, with some of the selection criteria previously applied to websites, such as identifying in-scope materials through site extensions included in national TLDs, proving to be ineffective for social platforms. Moreover, due to the sheer amount of content generated daily on these sites and the inversely proportional amount of resources available to sustain the archiving effort, memory institutions abandon the ideal goal of "completeness" for "representativeness". However, legal, technical and resource constraints expose institutions to the risk of introducing gaps and imbalances into collections, raising questions about the actual degree of representativeness of social media collections. To mitigate such concerns, web archivists have integrated participatory practices into their workflow (e.g. crowdsourcing, co-curation, and networks of contributors) which aim to ensure an effective representation of stories and voices from all strata of society.

Web archivists use a wide array of tools, ranging from web crawlers to APIs, often combining different tools and approaches to gather snapshots of events or conversations

on social media. Some of the technical issues that emerged from this study, such as difficulties in capturing dynamic content, are known to web archiving but are often exacerbated by the volume, ephemerality, and constant evolution of social platforms. Sudden changes implemented in the platforms' terms of use often cause issues that, when left unsolved for too long, lead institutions to set certain platforms aside until better solutions for archiving them are found. In the meantime, important traces of present events are lost forever.

Documenting curatorial choices, legal, and technical challenges is essential for providing researchers with crucial information to understand how collections have been developed and to critically approach them. While access to archived social media material is generally limited to on-site terminals for legal reasons, web archivists are seeking to expand access across national territories and improve discoverability of content. However, more needs to be done to raise awareness and improve engagement with web and social media collections within and, especially, beyond academia. Increased usage of these collections may help web archivists to advocate for the necessary funding to scale-up collection activities and sustain the long-term preservation of archived social media materials.

CHAPTER SEVEN

Guidelines for The Development of Social Media Archives

The diverse range of practices and solutions adopted by the participating institutions demonstrate that the development of best practices and standards for social media preservation at an institutional level is still a work in progress. Nevertheless, archiving initiatives have come a long way in the capturing of social sites, with the 2020s health crisis and armed conflicts being propelling forces. The comparative analysis of workflows and approaches adopted at memory institutions was essential in identifying shared positive outcomes and key takeaways matured through the many trial-and-error archiving attempts carried out by web archivists. Building on the international comparative analysis provided in Chapter Six, this chapter will offer a set of guidelines and recommendations to support the development of future social media archiving initiatives, not only at a national level but also among a wider range of organisations, including areas of the GLAM sector other than libraries and archives.

7.1 Guidelines and Recommendations

Before illustrating the guidelines for the development of social media archives, it is important to stress again that the aim of this research is by no means to criticise any of the approaches illustrated in the previous chapters. Based on the analysis and reflections that emerged from a combination of literature review, examination of documentation (e.g., reports, blogs) and practices implemented by various institutions, the following recommendations seek to guide future archiving initiatives through the complex and sometimes obscure maze of concerns and challenges related to social media preservation, generating a critical reflection on potential archiving approaches and sparking further discussions.

1. ***Social media archiving initiatives should document in detail information related to the legal and technical context, selection choices and curation approaches.*** In Chapter 6, it has been highlighted on several occasions how the interconnectedness of legal, ethical, and technical obstacles and concerns has a profound influence on the archiving strategies adopted and significantly shapes the

development of social media collections. Producing proper documentation intended to explain critical points of the legal framework within which the social media materials are collected, describing selection and archiving strategies adopted and pointing out technical issues that may have affected capture and replay of certain content could offer researchers crucial information about reasons behind the inclusion or exclusion of content. Moreover, following the example set by BL, where certain types of material are impossible to collect although essential to properly record an event, web archivists and curators should record the missing item(s) and the reasons for its unobtainability. Such documentation would enable researchers to get a better understanding of how collections have been built, and about any potential gaps that should be taken into consideration when studying content from web and social media collections (see also Recommendation no.8). Producing detailed documentation about social media collections is not only essential to good collection management but also ensures transparency about decision making processes.

The involvement of volunteers or external funding could be sought where limited time and resources are available to support documenting information about the collections.

2. ***Collection development strategies should carefully consider and document any criteria that may have ambiguous relevance or significance***. As illustrated in Section 6.2.2 some selection criteria need to be considered in relation to the cultural context of each country. For example, using official languages or specific pictograms largely used in online communications may assume different meanings in various contexts (e.g., the case of the flag emoji and its political meaning in some countries). Moreover, the interconnected reality of social media discussions, and their tendency to blur the space defined through geographical borders, pose a challenge to many countries who share the same language and may risk exceeding the territoriality criterion imposed by national electronic legal deposit legislation. Nevertheless, archiving institutions should consider the potential value of ambiguous criteria or exceptions that may apply to the institution's collection development policy. In some cases (e.g., the COVID-19 pandemic), the relevance of hashtags or content generated beyond borders can offer a significant contribution to portraying the breadth of events that have a global resonance and impact. Moreover, it can also offer an

opportunity to connect national collections across borders, facilitating studies of transnational events.

3. ***The feasibility and effectiveness of archiving workflows should be tested through small-scale sample collections on significant events, although risks of inaccurate results should be considered***. There is a tacit trend to test the feasibility of social media archiving efforts on events of national significance, such as general elections. Despite the risk of incurring hiccups and glitches that may jeopardise the quality of the results, evaluating archiving strategies or the effectiveness of tools on occasions of national relevance may offer an opportunity for institutions to request additional funding to support an archiving test related to important events for the whole country. Moreover, capturing for instance the progression of general election campaigns up until voting day allows institutions to collect information within a circumscribed timeframe, making the archiving effort certainly demanding, yet manageable. Moreover, choosing a recurring event happening in a defined span of time would allow curators to plan the details of the test in advance. Nevertheless, the risk of not properly capturing a record of such major events should be factored in. Any issues or obstacles that emerge during the test phase should be accurately documented for later assessment.

4. ***Participatory strategies should be implemented to support the development and enrichment of inclusive, representative social media collections. Targeted promotion campaigns can amplify their efficacy***. The examples provided in Chapters 4 and 5 regarding additional collection strategies aimed at mitigating potential gaps and representativeness concerns regarding web and social media collections demonstrated the value of collaboration across departments of the same institution as well as with external contributors. Crowdsourcing campaigns inviting the public to suggest websites and social media content that should be included in the national web archive have proved to be efficient and cost-effective. Moreover, in most cases participatory approaches have achieved a two-fold result: raising awareness among the public about the existence and value of web and social media archives and improving the rate of engagement with these collections among the wider public. It is worth noting however that these campaigns should be promoted especially among minoritised groups or seek collaboration with community archives

or activist groups, thus enabling the enrichment and improvement of the representativeness of collections. Where available, institutions' public engagement and communication teams could support the development of efficient promotional strategies involving, for example, exposure on institutional social media channels and in national and local newspapers, as well as a more granular approach including workshops or activities at local libraries (see Section 6.2.3).

5. ***Attention should be paid to potential mental health repercussions related to the selection and preservation of traumatic or harmful content in social media collections***. Since most thematic collections including social media content often revolve around traumatic events, such as terrorist attacks, armed conflicts or health crises, and can include hate speech as well as historical and contemporary discrimination, it is essential to draw attention to the potential mental health implications that may affect the work of web archivists and curators, or any other persons (e.g., volunteers) involved in the selection, capture or quality assurance of potentially distressing material. Although this aspect did not emerge from the interviews, it is important to dedicate one of the recommendations to the need for developing and implementing adequate safeguards in this regard. There is an increasing corpus of studies that have begun exploring secondary trauma in the context of archiving institutions (Regehr et al., 2023; Sloan et al., 2019; Williamson et al., 2020) and how to process emotions that may emerge when working closely with such material in a positive way. Results from on-going projects such as "Protecting the Investigator in Traumatic Research Areas" (PETRA)[187], will shed light on formulating best practice guidelines to support researchers and professionals dealing with content where secondary trauma is a possibility.

6. ***Technical frameworks should be evaluated based on resources and expertise available at the archiving institution***. National legal frameworks, individual institutions' mandates and priorities (e.g., "look and feel" Vs raw data), as well as resources available to sustain the archiving efforts in the long term, influence the type

[187] PETRA is one of the Community Interest Groups in the UK-Ireland Digital Humanities Association. More information about the project can be found here: https://web.archive.org/web/20231204005529/https://digitalhumanities-uk-ie.org/community-interest-groups/protecting-the-investigator-in-traumatic-research-areas/

of tools adopted and methods used to archive social platforms. Therefore, the evaluation of collection methods should ensure that the chosen technical framework aligns with what the institution can realistically support and manage. Web crawlers may best suit institutions that want to focus on preserving a high-fidelity snapshot of social platforms and/or those that already have familiarity with this method of collection because of previous web archiving activities. Conversely, collection through APIs offers a good solution for the capture of raw data and rich metadata. Nevertheless, as highlighted in Section 6.3, both methods come with a set of obstacles that must be carefully assessed and weighted against the available budget and the scope of collections, as changes implemented at platform level may affect the ability to collect content and the overall quality of results. If it is not possible to develop the necessary expertise internally or attract professionals with competitive salaries to directly support the technical side of social media harvesting at scale, institutions may need to explore alternative methods for archiving social media. These include subscribing to services such as Archive-It, which provides sufficient training and some technical support, both in terms of harvesting and replaying; or outsourcing the capture of the selected material to private companies (e.g., TNA using MirrorWeb). It should be considered, however, that these solutions often require additional funding to cover expensive fees and may involve limits on the budget of data archivable over a set timeframe.

7. ***Access to social media collections should consider the peculiarities of the material, potentially providing specific features, tools and additional entry points to facilitate discovery***. Providing and ensuring access to web and social media collections is an essential step in the long-term preservation of these resources. Depending on the restrictions imposed by national legal frameworks on access and resources available at an institutional level, access to social media collections should be facilitated by user-friendly interfaces. The current trend sees archiving institutions aggregating with no distinction both websites and social media, mostly due to technical factors and preferences expressed by some users. Nevertheless, implementing a series of facets developed specifically to help users refine their social media related query as much as possible would benefit a wider variety of potential users interested in certain platforms or communication dynamics (e.g., retweets, mentions). This will become increasingly important as the amount of social media

content archived potentially grows over time. Moreover, access to social media collections should be facilitated through a variety of entry points such as the BnF's "Guided Tours" (see Section 5.4.4), or innovative and engaging ways to display content like INA's SocialTV (see Section 5.4.4), or thematic collections. Users, especially those approaching social media collections for the first time, can benefit from multiple points of access. In addition, features such as data visualisations, Ngrams, or image search tools can offer an initial understanding of the web archive data retrieved.

8. ***Access interfaces should provide quick access to collection documentation and other information that can help researchers to critically approach and query archived social media data***. As mentioned, documenting collection development strategies as well as the legal and technical context is an important aspect of the management and preservation of collections. Moreover, such documentation is essential for researchers approaching archived social media resources, as it provides insights into the multiple factors and choices that shaped the collection. For this reason, documentation regarding social media collections should be made available online on the website of the archiving institution and/or through the access interface. Information about the collection should also be accompanied by a guide including tips for exploring social media resources, a description of the types of search and filters available, and limitations regarding the results obtained (e.g., relevance to the query or ranking of results).

9. ***User forums, including researchers and users from diverse backgrounds, should be established early in the development of social media collections to understand user requirements and gather valuable insights for improving the collections and access to them.*** Seeking engagement and collaborating with potential users and researchers working with social media data would provide essential insights into the type of platforms, material and aspects related to the context of social media that should be considered by archiving teams in the selection stage. This could be even more effective if user forums or focus groups are established from the beginning of social media archiving activities, contributing to the development of collections and the access interface. Examples such as those provided in Section 6.4 regarding the Legal Deposit User Forum at BL or the

ResPaDon project in France demonstrated the importance of user testing and feedback to (re)evaluate aspects of the service access and identify areas of improvement. In addition, involving a diverse range of potential users other than researchers (e.g., students, journalists, teachers) with no previous experience with archived web and social media material could help archiving initiatives understand the type of information and training resources that first-time users might need to approach these collections. Involving a varied group of users could also support the development of strategies to engage with and raise awareness of this resource's value among additional user communities.

10. ***Social media archiving initiatives should engage with and establish international networks of collaboration***. Interviews highlighted the importance of establishing networks of collaboration with initiatives that are involved in similar archiving projects, both nationally and internationally. Sharing lessons learned and solutions to common issues is essential to support the advancing of social media archiving. In this regard, initiatives such as the IIPC[188] or the Belgian PROMISE[189] project demonstrated the importance of creating spaces in which institutions at different stages of their projects as well as researchers/end users can share their knowledge, creating opportunities to discuss challenges and work together towards finding solutions or developing new, shared tools. The latter is crucial in a context where unexpected changes in social media terms of use can drastically hinder the progress attained, and immediate actions need to be taken to rapidly find solutions. Collaborating in the development of tools and guidelines may foster, with time and widespread implementation, the establishment of international standards. Also, in an ever-evolving landscape like that of social media platforms, where data constituting a shared memory and cultural heritage (see Section 2.2.1) is gatekept by a few companies, international networks could advocate for the implementation of clear long-term preservation plans within social platforms, and which should include the establishment of closer collaboration with cultural heritage institutions.

11. ***Dedicated long-term preservation strategies should be developed and updated regularly to adapt to the evolution of the social media landscape***. The

[188] https://web.archive.org/web/20240111024120/https://netpreserve.org/about-us/

[189] https://web.archive.org/web/20240117214854/https://www.kbr.be/en/projects/promise-project/

long-term preservation of archived social media within memory institutions is often assimilated into that of other born-digital items. While this may be effective, there is clearly a need for establishing specific strategies to ensure that this unique type of material is properly preserved. These strategies should take into consideration the rapidly changing nature of social media platforms, requiring institutions to quickly adjust archiving and preservation approaches to the adoption of new tools and collecting methods. In this context, it could be useful to introduce in the workflow a "living" document which is not only regularly updated at every stage of the archived social media material preservation journey but is also promptly reviewed when modifications to archiving strategies are triggered, for instance by changes implemented at a platform level. These revisions should also be promptly shared within the archiving team. The documentation should include the description of the development, implementation and use of tools, the file formats adopted, and steps taken to preserve and maintain the integrity of and access to such material in the long term. This information will be essential to understand, for instance, the evolution of the archiving methods, acquisition strategies, and file formats adopted to store the archived material, or metadata management. Documenting processes specifically developed for the archived social media content will ensure that the preservation procedures are carried out consistently over time (DPC, n.d.-a).

7.2 Additional Recommendations

Social media archiving initiatives at any stage of collection development should work closely with the institution's legal team (if available) or seek advice from experts to address legal issues such as those described in Section 6.1.3. For example, in case of uncertainty surrounding the capture of potentially sensitive data or handling a mix of private and public content, consulting legal experts would be essential to resolve similar impasses. Obviously, the ability to request external consultation depends on the available resources of individual initiatives. Nevertheless, decisions made regarding legal concerns should be properly documented for future reference (see Recommendation no.1). Finally, organising periodic workshops within individual institutions that would involve professionals from across the various departments, such as those described in Section 5.4.3, could bring new perspectives to unsolved issues supporting the complex process of finding solutions or elaborating new approaches for problems concerning specific platforms.

Conclusion

The guidelines and recommendations presented in this chapter are designed to support the development of social media archives and collections, serving as a starting point for memory institutions new to this archiving endeavour, while also encouraging further discussions and reflections among those with established experience in archiving social platforms. While stemming from the analysis of obstacles, practices and solutions that emerged from a study primarily involving memory institutions operating at a national level (e.g. national archives, deposit libraries), these guidelines intend to offer practical recommendations to a wider variety of institutions, including private institutions, university archives or research-based archiving initiatives, and extending beyond areas of the GLAM sector other than libraries and archives.

Considering the rapid changes that characterise social media platforms, these guidelines will likely need be frequently revised, with new recommendations added to reflect the changing landscape. To support ongoing revisions and ensure broad dissemination within communities of practice, I propose making these guidelines publicly accessible as a living document (e.g., a shared Google document or a Wikipedia page). This format would enable web archivists and other professionals involved in the preservation of social media to contribute valuable input, offer suggestions and share insights, helping to collaboratively shape common strategies to archive social platforms over time.

CHAPTER EIGHT
Conclusion and Future Work

This thesis investigated the long-standing issues and latest concerns faced by social media archiving initiatives at various stages of development of their collections, offering several new insights into the social media archiving landscape. Positioned at the crossroads of born-digital archiving, digital preservation, archival science, web archiving studies, cultural heritage studies, media studies and digital humanities, this thesis expands knowledge of the specific challenges that social media platforms pose to archiving institutions compared to other born-digital materials, such as websites. More broadly, as born-digital cultural heritage generated on social media becomes increasingly central to digital humanities and archival studies, this thesis advances understanding of this important resource and offers a roadmap for ensuring its long-term accessibility for future research across digital culture, media studies, and other fields, including political science, public health, and social science studies. In particular, this research makes a valuable contribution to web archive studies and internet history by offering comprehensive insights into the diverse practices of social media archiving. Through web archivists' first-hand experiences, it illuminates how and why certain content is archived while other content is not, offering opportunities for critical reflection when engaging with social media collections.

This thesis demonstrates how legal frameworks and limitations imposed by single platforms heavily influence curatorial practices and consequently the ability of memory institutions to preserve social media content, often determining the type of platforms and to what extent material can be effectively and sustainably archived from these sites. By capturing the immediate reactions and consequences of the latest "API-calypse" following the 2022 Twitter's acquisition on archiving institutions, this thesis also demonstrates how the ever-changing nature of social media and their economy deeply affect the advancement of social media archiving.

Moreover, this thesis demonstrates, through a cross-national comparative analysis of selection and curation practices, how these processes are shaped by legal, cultural, and technical contexts. It highlights differences between social media and traditional websites, demonstrating, for example, that although both born-digital materials are often part of the same collections, social media requires the applications of specific selection criteria. Furthermore, given the power that national memory institutions hold in preserving

narratives and the memory of events that will be transmitted to future generations (Jimerson, 2006; Schwartz & Cook, 2002), this research examines whether collections are developed with consideration for the multitude of societal groups and communities within society, emphasising the critical importance of ensuring the representativeness of collections. Consequently, it explores participatory archiving practices implemented by various memory institutions, illuminating the multiple benefits of strategies such as co-curation and crowdsourcing. These approaches not only help address concerns regarding the representation of the multiple layers of society but may also improve awareness and engagement with web and social media archives beyond academia.

To ensure transparency and critical engagement with social media collections, the thesis encourages archiving initiatives to produce and maintain up to date documentation of decision-making processes as well as the context within which a social media collection is developed. It further highlights the need to compile documentation specifically designed to ensure the long-term preservation, integrity, and access of archived social media materials.

In addition, this thesis examines the social media archiving landscape, identifying the types of institutions that are archiving or planning to archive social media platforms, the specific platforms being preserved, and whether these archiving initiatives are extensions of previous web archiving efforts. It explores the reasons behind the absence of social media archiving initiatives in certain regions, attributing the concentration of such activities in the Global North to geopolitical and power dynamics. Additionally, it identifies some of the underlying factors hindering the development of social media archiving initiatives even within Global North countries.

Lastly, this thesis seeks to advance social media archiving practices by offering guidelines and recommendations that can be taken as a starting point by any institution aiming to develop a social media collection. The numerous, detailed examples presented in the two case studies (Chapters Four and Five), and the comparative analysis (Chapter Six), provide practical insights into the challenges web archivists encounter when archiving social media and the solutions implemented to mitigate such obstacles.

In the following final sections, I will review the main arguments and findings of this thesis, retracing the methodology, reasoning and discussions from each chapters. This chapter will conclude with some final reflections and suggestions for future research.

8.1 Thesis Findings

Several conclusions have emerged from the survey analysis and the cross-national comparative study. First, I have argued that, although stemming from web archiving preservation practices, and often embedded in the exact same legal framework and sharing similar concerns, social media archiving has revealed itself as something increasingly moving on a parallel road to traditional website preservation, posing specific legal, curatorial, technical and accessibility challenges. In Chapter Two, I underscored the cultural value of social media platforms and the discussions taking place on these sites, and how this has been further reinforced by the central role these sites played during health and political crises, supporting for example the exchange of critical information and documenting armed conflicts (e.g., the 2022 Russia-Ukraine war). Despite being an integral part of the World Wide Web, social media platforms have increasingly become separate ecosystems with unique characteristics, specific communication dynamics, and technical aspects that increasingly distinguish them from traditional websites (Helmond & Vlist, 2019). Moreover, because of the fragility that social media platforms share with other virtual environments in the digital world, and the ephemerality that characterises the content generated on these platforms by billions of users worldwide (SalahEldeen & Nelson, 2012), I emphasised the need for immediate measures to appropriately preserve this important resource, which is and will continue to be essential for understanding the 21st century. Like novel monks "copying" material from the web and social platforms against the often foretold but as yet unrealised "Digital Dark Age" (Kuny, 1998), memory institutions worldwide have started collecting some content from social media in order to preserve fragments of it for future generations despite the difficulties encountered on many fronts. In fact, while web archiving has matured over the past quarter of a century, consolidating shared practices and standards across practitioners worldwide (Ben-David, 2021), social media archiving is relatively new and institutions are still figuring out appropriate workflows, often adapting archiving strategies and standards from web archiving (Thomson, 2016).

In the spirit of framing the state of the art, Chapter Three mapped out social media archiving initiatives, gathering essential information about their stage of development and their scope. The survey shed light on emerging patterns related to the phenomenon of social media archiving, drawing critical attention to the uneven geographical distribution of the archiving initiatives. Pointing out the influence of economic and technical divides as well as power dynamics that characterise the opposition

Global North/Global South, I underlined the gaps that such dynamics produce in the preservation of the global collective memory generated on social media. Moreover, Chapter Three discussed the issues surrounding the discoverability of smaller, independent social media archiving initiatives flourishing outside national archiving institutions, which are often more difficult to identify through a simple web search. While the number of social media archiving initiatives is certainly larger than the group captured in this study, due to the time constraints and limitations identified in Section 1.5, results illuminated the diverse range of organisations – beyond libraries and archives – seeking to preserve this important material because of its cultural value and its widespread use. Moreover, the survey and desk research results revealed that between 2017 and 2023 there has been a rise in initiatives independent of any previous web archiving endeavour, often emerging from the need to record crises in real-time. These "archives of crisis" (Bingham et al., 2024) are exemplified by initiatives like the "Telegram Archive of the War" or the collection assembled in the aftermath of the 2017 Manchester Arena terrorist attack. The analysis of the survey results also highlights that the most archived platforms do not necessarily reflect the popularity of individual platforms at a national level. Social media archiving initiatives have tended to focus their efforts on Twitter given its relative openness and the "publicness" of the conversations taking place on this platform. However, existing usage reports about the popularity and number of active users of social sites (e.g., Statista.com, 2023), show that Twitter is not as widely used as other sites like Facebook, Instagram or TikTok. Similar discrepancies raise concerns about the potential gaps in the preservation of the collective memory generated on social media, and the representativeness of the complex multitude of voices and communities worldwide. This is particularly concerning when several social platforms are, for example, only used in areas of the Global South (e.g., WeChat) where there might be no archiving initiative preserving or planning to preserve them. I argue that autonomous, non-mainstream social media archiving initiatives, especially those operating outside of a legal deposit mandate and thus devoid of any collection limit concerning geographical borders, or those led by activist groups could represent an opportunity to safeguard stories from marginalised communities and events related to Global South countries. However, rather than relying on "benevolent" (Lor & Britz, 2004) archiving activities led by Global North initiatives, which risk being patronising, efforts should focus on empowering the Global South to archive its online cultural heritage. Nevertheless, it is important to note that the challenges inherent in social media archiving, even for long-standing web archiving institutions,

could potentially exacerbate existing Global South-North web archiving dynamics and barriers (Colin-Arce et al., 2023).

As discussed in Section 4.5, the absence of web and social media archiving initiatives in some of the Global North countries appears to be related to obstacles including concerns about selection criteria and technical issues. But it is the lack of specific legislation that clearly regulates or assigns a mandate for the collection and preservation of material published on the web at an institutional level that has emerged as one of the prime reasons that hinder the development of initiatives of the sort. To support this statement, I used the case of Italy, where delays in the approval of updated legal deposit legislation that openly states the requirement for national libraries to archive web and social media, slow politic dynamics, and limited resources have particularly affected the launch of any archiving activities concerning social sites. Expanding on this point, Chapters Four and Five offered, by contrast, two in-depth case studies concerning social media archiving at national archiving institutions in the context of electronic legal deposit, respectively in the United Kingdom and France. The two case studies explored the national legal context in which these archiving initiatives are embedded, demonstrating the importance of benefiting from electronic legal deposit legislation that has been updated to allow for the inclusion of web-based material and scale up archiving efforts. However, the two case studies also shed light on the complex set of constraints imposed by those same enabling national legal frameworks, social platforms and technical aspects, illustrating how these influence the development of social media collections at well-established web archiving institutions. Drawing from UK and French examples, I noted the plurality of social media archiving approaches and practices — even within the same legal context, as the case of BnF and INA demonstrated. In fact, despite sharing the same legal mandate, the two French institutions implemented different collection development strategies and methods. These choices appear to mirror their institutional mission statements (see Section 5.1), technical expertise and resources available, generating two diverse yet complementary collections covering a broad range of web-based cultural heritage materials (Faye et al., 2024).

The transnational comparative analysis presented in Chapter Six shed light on the extent to which memory institutions are currently able to archive social media and the intricate set of challenges that inevitably shape the collections that researchers are and will be able to study. National legal frameworks are the primary factor that greatly influences the development of social media archiving initiatives. While most social media archives

have been established at national deposit libraries, the chapter highlights alternative archiving scenarios operating outside electronic legal deposit legislation. These include institutions preserving governmental records (e.g., TNA) or research-based initiatives appointed by the government to preserve national cultural heritage on the web (e.g., Arquivo.pt). Building on the observations made in Chapter Three, I underscore the importance of receiving a legal mandate. Obsolete legislation or inadequate wording in existing deposit regulations are among the reasons behind the absence of national social media archiving initiatives in certain countries (e.g., Italy), potentially leading to extensive gaps in preserving the collective memory generated on social platforms. Collection experiments such as those conducted by the BESOCIAL project at the KBR or the British Library during the 2005 UK General Election, have been crucial in demonstrating the feasibility of web and social media archiving activities, emphasising the value of web-based material. Similar small-scale tests can support memory institutions in advocating for the implementation of appropriate national legislation, enabling them to consistently preserve or scale-up the preservation of these platforms. Furthermore, Chapter Six highlighted the tendency in electronic legal deposit legislation in force in the countries of interviewed institutions to loosely define the type of born-digital material in scope. Recognising the power that regulatory language has in shaping collections – and thereby the memory of present times that will be preserved – I emphasise the importance of forthcoming legislation taking into account the unpredictability of the evolution of the web. This requires moulding definitions that are sufficiently broad to allow for the inclusion of future born-digital formats that may hold cultural value. It is important to reiterate here the complexity of legal frameworks, and how their enabling power is often counterbalanced by restrictions that severely limit the scope of collections and access to them (Schafer & Winters, 2021). Having a legal mandate to archive social sites is not an end goal, but rather a preliminary step necessary to initiate the actual archiving endeavour. I discussed how the need for lawmakers to set clear boundaries for collecting web-based information under electronic legal deposit – usually identified with geographical borders and related country code top-level domains – clashes with the interconnected nature of the conversations generated on social media. This has required institutions to narrow the scope of collections, often opting for sampling approaches to keep the capture of these sites manageable over time. Moreover, while national archiving institutions are exempt from certain copyright and data protection obligations when archiving for the purpose of preserving national cultural heritage, these still pose significant obstacles for other

organisations in the GLAM sector or independent projects. The case of the Museum of London, and its approach to mitigate privacy and copyright concerns, demonstrates the complexity of the effort and time required to collect social media material for institutions operating outside of an e-legal deposit context.

Besides, the limitations imposed by social media companies on access to and reuse of data generated on their platforms to protect their business interests affect the granularity of collections and, more broadly, the sustainability of archiving efforts at an institutional level. Repeated poor results in capturing content from platforms or sudden changes in platforms' terms of use, like those implemented following Twitter's acquisition by Elon Musk in 2022 or the Cambridge Analytica scandal in 2016, heavily influence the ability of archiving initiatives to capture material for preservation purposes. I highlight how the attempts to contact social media platforms, whether to apply for access to specific APIs (e.g., Twitter Academic API) or request permission to archive, have often faced the absence of specific communication channels for cultural heritage institutions. Even in the few instances where memory institutions succeeded in applying for access to official APIs, as in the case of INA with Facebook, no additional support is provided to archiving institutions, underscoring social media companies' lack of interest in addressing the specific needs of memory organisations.

Legal constraints deeply influence selection and curatorial decisions, often determining the type and extent of social platforms archived. I draw attention to the incompatibility of some selection criteria previously adopted for websites, such as those concerning the use of country code top-level domains to establish the pertinence to territoriality criteria. Difficulties in ascertaining the provenance of content on social media, combined with legal concerns and technical issues, often result in institutions having to manually select material on social platforms, which requires the long-term investment of time and resources that most institutions do not have. As a result, the amount of social media content that can realistically be manually assessed by curators, and thus archived, is much smaller compared to the overall volume of websites archived within the same crawls. I highlight how the pursuit of comprehensively preserving social platforms is often replaced by a more feasible intent to create collections that are representative of the many faces of one country. Building on post-modern archival theory, I draw attention to Kaplan's (2002) definition of "representation" and the power that archiving institutions exercise through the act of selecting certain records rather than others. Acknowledging the biases that inevitably shape archival collections and the power

dynamics that have led institutions for centuries to preserve mainstream histories, silencing marginalised groups (Jimerson, 2006; Schwartz & Cook, 2002), I advocate for appropriately representing minority communities and the stories social platforms have enabled them to share in national social media collections. This is essential to accurately portray and transmit significant information about present times to future generations. In this regard, I shed light on the participatory practices implemented by archiving initiatives to mitigate structural biases and address representativeness concerns. Participatory approaches, such as crowdsourcing campaigns prompting the public to suggest URLs for preservation or co-curation practices involving a network of collaborators from specific communities and local areas, provide an opportunity to enrich collections and ensure the preservation of a representative sample of voices from various societal groups. Moreover, I underscored the added benefits of ensuring diversity within archiving teams to widen perspectives about themes and issues related to the varied tapestry of communities existing in a certain country.

In addition to legal and curatorial challenges, Chapter Six also analysed the technical frameworks implemented at various archiving institutions, illustrating the unique combination of tools each institution has adopted for capturing social platforms, depending on the resources and technical expertise available. For instance, the BnF has been using Heritrix in combination with Picuki to capture Instagram in order to mitigate issues and rate limits often encountered when capturing this site through web crawlers. Challenges emerging from the interviews are mostly linked to the use of tools that have not been developed specifically for the collection of highly dynamic content such as that shared on social sites, or to the limits imposed by social media policies on access to data. It is important to note that social media platforms have often been added to the scope of many memory institutions as a natural extension of preexisting web archiving activities, and thus often captured using tools at hand, such as web crawlers like Heritrix, which were initially developed for website collection. I described the several technical issues that have emerged from the use of this type of method to acquire social media sites, which range from the inability to interact with and collect dynamic content to poor-quality in capturing the structure of platforms, resulting in missing several items and features that are essential to properly preserve social media sites. Conversely, institutions using official APIs have raised concerns about the unpredictable changes often implemented by social platforms. Additional concerns stem from the lack of transparency on sampling of data as well as on the difficult and time-consuming application process to request access to

official APIs. While alternatives to the two main collection methods exist, including Webrecorder and third-party tools such as Social Feed Manager, these are still subject to a series of limitations and technical issues that may lead to loss of important information about the context or the user experience on social media. For this reason, it is essential for technical frameworks to be thoroughly documented as these shape the content and information included in social media collections. Moreover, I argue for this information to be made available to researchers to facilitate a critical study of the archived social media material.

Chapter Six discussed the problem of access and the generalised lack of awareness of web and social media collections. Apart from a few notable exceptions like the KB, TNA and Arquivo.pt, legal frameworks tend to restrict access to web and social media archive collections to the institution's premises. Building on existing studies discussing issues stemming from restrictions on access to web and social media archives (Gomes, 2017; Healy et al., 2022; Thomson & Kilbride, 2015; Winters, 2020), I consider the socio-economic impact of these restrictions on the level of usage of social media collections. I maintain that participatory approaches as well as experiences such as the "Into the Twitterverse" exhibition at the MoL can offer opportunities to raise awareness about the values and existence of web and social media archives among the wider public, specifically among minority groups. I also examined the variety of entry points offered at interviewed institutions, highlighting good practices aimed at mitigating access restrictions, such as making seedlists available to the public for download, and providing bespoke or derived datasets. Nevertheless, I noted the importance of providing users with detailed instructions and information on how to effectively search for social media archived material through the access interfaces. This is crucial as most often archived social media are included with no distinction in web archive collections and the facets available on access interfaces are sometimes not sufficiently tailored to the specific needs of research concerning social platforms. Thus, I argue for the need to provide ad hoc gateways and/or more detailed facets designed around the peculiarities of archived social platforms, enabling users to efficiently filter social media materials from websites according to their research goals.

Lastly, I draw attention to the need to establish specific preservation strategies for social media material that goes beyond preservation activities developed for any born-digital material or the use of WARC/WACZ file format standards. I conclude by underscoring the importance of producing detailed documentation and keeping it up to

date. This should include information about the original context in which the archived material was embedded, as well as curatorial choices, decision-making processes, legal and technical contexts, and preservation strategies. Such documentation is essential to safeguard not only the long-term preservation of the collected material, but especially to ensure the transmission of critical information needed to understand archived social media and the genesis of collections.

Overall, this thesis has highlighted the complexity of the legal, ethical and technical factors and challenges that intervene in shaping collections development strategies and the histories that future researchers will be able to study. It also shed light on the variety of practices and solutions adopted by institutions, depending on the resources invested in social media archiving efforts as well as on the cultural background of each country, which is mirrored in some of the selection criteria adopted, themes and topics identified by archivists and the wider public through participatory approaches. It is important to note that the difficulties in finding a "one-size-fits-all" approach, coupled with the ever-changing social media landscape, have profound consequences on the potential establishment of common international standards specifically developed for the preservation of social platforms. For this reason, I suggested guidelines and recommendations that are based on what can be considered "good practices" rather than "best practices". The guidelines proposed in Chapter 7 underscore the importance of documentation, international collaboration and the need to consider the unique nature of social platforms, which sets them apart from traditional websites. I emphasise how appropriately recording information about the development of social media collections is essential for future researchers to critically engage with the archived material. Moreover, when potential biases, gaps, missing elements, and the contextual information about social platforms are properly highlighted and reviewed over time, this documentation will ensure the accurate long-term preservation of archived social media content.

Lastly, by recording issues and collecting strategies adopted in the first decade of social media archiving, this thesis has contributed to capturing a trace of the direct experiences, voices and memories of web archivists who have been dealing first-hand with the complex endeavour of preserving the ephemeral, providing a glimpse on the early history of the development of social media collections.

8.2 Looking Ahead: The Future of Social Media Archiving and Research Opportunities.

In the course of this study the social media landscape and archiving initiatives have known profound transformations. The pandemic and the various political crises have not only favoured the rise of new platforms like TikTok, but also prompted many institutions or independent initiatives to start collecting content from social platforms. Conversely, the acquisition of Twitter in 2022 and the subsequent changes to the platform, branding and terms of use have caused significant upheaval for many archiving institutions, most of which had to halt the preservation of Twitter altogether. Consequently, the continued evolution of platforms and the related archiving activities offer several additional research opportunities to be explored in the future.

The first path for research has emerged from the analysis of the geographical location of social media archives (see Chapter 3). Time constraints imposed by the postgraduate programme, coupled with the discoverability issues concerning social media archiving initiatives, have limited the number of institutions that could be identified. The number of social media archiving efforts is certainly greater than those described in this study, especially in relation to independent projects that may be more difficult to pinpoint due to their operating outside the main web archiving networks used in the dissemination strategy. Thus, further research should focus on identifying additional existing projects. It would also be valuable to create a resource where individual initiatives, regardless of their size, can report their existence. Such a resource would also benefit researchers who seek to use social media in their research but might not be aware of existing social media collections. A possible outcome in this sense would entail creating a Wikipedia page – following the model of the one already existing for web archiving initiatives (Gomes et al., 2010) – where institutions could add their name, general information about their collection, and the type of platforms archived.

This thesis has compared and analysed the manifold challenges to social media archiving arising from a combination of legal constraints, technical issues, ethical and selection concerns. This is not to say, however, that these are the only challenges that could emerge from the development and preservation of archived social media collections. There are certainly other issues and concerns specific to other institutions or smaller archiving initiatives that could not be covered in this study. For example, one possible research route could involve investigating the reasons and challenges hindering the development of social media collections in institutions located in the Global South.

In addition, building on the concerns related to the representativeness of collections at a national level, especially regarding marginalised communities, further research is needed to understand the extent (if any) to which community archives are capturing social sites. Research in this sense could potentially foster the establishment of networks of collaboration between community archives, independent projects, and national legal deposit institutions.

Moreover, even though one of the aims of this study was to understand the variety of challenges concerning the preservation of social media sites, the information collected revealed a sort of state of imbalance towards legal and selection issues rather than the more technical aspects and long-term preservation strategies. One of the reasons for this imbalance is that the web archivists interviewed were predominantly involved in the curatorial side of social media archiving rather than the technical side. In fact, the main points regarding technical issues that emerged from the interviews primarily focused on the tools used and replay issues rather than delving into proper technical details. Further research focusing on the technology and expertise needed for preserving this specific born-digital material would be beneficial for institutions looking into developing social media archives. In particular, investigating potential options or following the development of alternative methods of acquisition to the now-deprecated official Twitter API may offer a window onto testing practices and evaluation processes that can support institutions in moving past the impasse created by the new restrictive terms of use introduced after the 2022 Twitter acquisition. In terms of long-term preservation, as mentioned, no specific strategies have been established for archived social media beyond the use of WARC/WACZ file formats and other precautions implemented for other born-digital material. This is a matter worth exploring further, especially in light of the overall lack of long-term preservation strategies implemented by social media companies. This is a significant concern, as the fragility of the online environment places the vast amount of valuable information generated on social platforms at the mercy of organisations that have often shown apparent indifference to the fate of user-generated content. This was demonstrated once again, for example, by the glitch that deleted photos uploaded to Twitter between 2011 and 2014 (Novak, 2023).

One of the rare virtuous exceptions in this context is a long-term preservation project established by Flickr in 2022, named the "Flickr Foundation", which aims to keep Flickr photos visible for a century. They are in the process of developing a 100-year plan aimed at creating "an accessible social and technical infrastructure to protect [Flickr's]

invaluable collection for future generations".[190] Particularly interesting appears to be their idea of a "Data Lifeboat", which they describe as a sort of "archival sliver that is updated if things change at flickr.com",[191] and which would include a copy of individuals' presence on Flickr, including photos and their metadata to understand their networked structure (Flickr Foundation, n.d.). The call for ensuring continuity of access to content created by users has also been accepted by Wordpress.com, one of the most popular publishing platforms. In an announcement made on 25 August 2023, WordPress Founder Matt Mullenweg introduced a 100-year plan whose goal is to "secure online legacy for a lifetime,"[192] providing users with a series of solutions aimed at safeguarding their online legacy. Interestingly, the 100-year plan includes the automatic submission of users' sites to the Internet Archive (if it is public), ensuring that a snapshot of the website is saved over time. While it is still unclear whether users will be willing to pay for such a plan – which appears to involve a one-time payment of $38,000 (Wordpress, 2024) – the impact and implications that similar initiatives might have on the continued preservation of online content for future generations could be significant. Moreover, it could be essential in raising awareness about this critical issue among other platforms and technology companies, like ripples in the water.

This thesis has also highlighted how in many instances no distinction is currently being made between archived websites and social media, resulting in both being aggregated within the same collections. Whilst recognising the fact that in some cases the amount of social media material collected appears limited compared to two decades of archived websites and thus may not justify the creation of separate collections, it can be argued that in terms of access this conglomeration can add further limitations. Future work could assess access interfaces aiming to understand whether and how these have evolved to accommodate queries related to social sites. Moreover, as emphasised in one of the proposed recommendations (see Recommendation no.9), there is a need to further understand how archived social media materials are being or could be used by researchers, supporting institutions in the improvement of their services. Facilitating engagement with

[190] https://web.archive.org/web/20240628002644/https://www.flickr.org/announcing-the-flickr-foundation/

[191] https://web.archive.org/web/20230311104151/https://www.flickr.org/programs/content-mobility/data-lifeboat/

[192] https://web.archive.org/web/20231216090256/https://wordpress.com/100-year/

these collections could attract vital funding, helping to scale up and sustain archiving efforts over time.

Looking ahead, it is worth drawing attention to the potential opportunities that the introduction of the European Digital Services Act (DSA)[193] in 2023 may present for access to data on social platforms. The DSA is an unprecedented piece of internet legislation that, among other issues, seeks to tackle disinformation and enforce more transparency from online giants such as Meta Platforms Inc., Twitter/X, and TikTok. In response to the coming into force of the DSA, companies like Meta have announced important changes to their policies which include additional transparency, accountability, and user empowerment (Clegg, 2023). In this spirit, for example, Meta introduced cards that provide more information about ranking systems for Feed, Reels and Stories in addition to the "Why Am I Seeing This?" feature. However, the most interesting announcement in terms of access to data concerns the release of two new tools for researchers: The Meta Content Library and API tools. In a blogpost shared in 2022, Meta stated that through these new tools researchers would be able to "search, explore, and filter the publicly available content on a graphical User Interface (UI) or through a programmatic API" (Clegg, 2023), providing access to content across Facebook and Instagram. While the benefits of these two new tools still need to be investigated, their release could represent an important step forward for the preservation of publicly available content on both Meta's platforms. Still, no specific access points to data for cultural heritage institutions have been taken into consideration to date.

As the social media archiving landscape keeps evolving, institutions will also need to reflect on how to sustainably develop and maintain these born-digital collections. Sustainability concerns extend beyond the lack of financial and human resources for long-term preservation to encompass the increasing environmental impact of maintaining the vast amount of information captured on the web and social sites (Kilbride, 2024). Considering the impending climate emergency (United Nations, 2020), memory organisations have begun to interrogate themselves on the environmental costs that preserving and retaining all these born-digital materials entail and how to potentially reduce their carbon footprint. Discussions surrounding archives and the environment, such as those reported by Vavassori et al. (2023), highlighted the need to raise awareness

[193] regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). Available at: https://eur-lex.europa.eu/eli/reg/2022/2065/oj

about the carbon footprint of digital archives, including web archives. However, this need appears to clash with the additional financial costs associated with reducing storage and deduplicating material during web crawling (Vavassori et al., 2023). The environmental impact of developing, scaling up and maintaining web and social media archives is a problem that archiving institutions must critically address in the near future, requiring the introduction of more environmentally responsible tools and archiving practices to sustainably ensure the preservation of these invaluable resources.

APPENDIX A

Survey Questionnaire

Survey - Social Media Archiving Projects

6. What is your current country of residence? *

7. What is the name of your cultural heritage/memory institution? *

8. What kind of cultural heritage/memory institution do you work for? *

- ◯ Library
- ◯ Archive
- ◯ Museum
- ◯ Gallery
- ◯ Private institution
- ◯ Other

9. If you selected OTHER, please specify below

10. What is your position in the institution?

11. Does your cultural heritage/memory institution archive social media or have plans to archive social media? *

- ◯ Yes
- ◯ No

12. If your answer was NO, why not?

13. Please select the answer that best describes the status of your social media archiving project *

- ◯ Long-term project
- ◯ Pilot/ fixed-term project
- ◯ Still in the planning phase

14. When did your institution start archiving social media? *

Please indicate the year

15. Is your social media archiving project part of a wider web archiving initiative? *

Please select your answer

- ◯ yes
- ◯ no

16. What type of social media does your institution archive? *

Please select all that apply

- ☐ Twitter
- ☐ Facebook
- ☐ Instagram
- ☐ TikTok
- ☐ WhatsApp
- ☐ LinkedIn
- ☐ Snapchat
- ☐ Other

17. If you selected OTHER, please specify below

18. Please give details below of any other social media archiving projects (no matter the size of the collection) that you know of and would like to suggest for this research, or any other information you think would be relevant for this project.

(e.g. community-led/activists social media archives, universities, museums, etc. ...)

19. Please enter your institutional email if you are happy to be contacted by the researcher for further clarification of your responses and/or for a potential follow-up interview

APPENDIX B

Questions and Key to Survey Follow-ups

Follow-up questions:

- What are the main obstacles to the development of a social media archive in your institution? Does your national legal framework include provisions regarding the preservation of content published on the web, and specifically social media?
- Have you ever received requests from users/members of the public about whether you preserve this type of material?
- Are there any other issues that you would like to address or comments about the prospect of developing a collection at [name institution]?

List of follow-up interviews via email with institutions that have no plans to archive social media in the near future:

BUJOKAS, DARIUS – Chief specialist of the Department of Management and Use of Documents and Archives at the National Archives of Lithuania (Interviewed via email, 25 January 2023).

STORTI, CHIARA – Manager for the *Magazzini Digitali* project and web archiving activities, National Central Library of Florence (Interviewed via email, 23 January 2023).

APPENDIX C

General Interview Questions[194]

- Can you briefly present yourself and your role at [*name institution*]?
- When and why did your institution start to systematically archive social media? What is the scope of the collection and how frequently do you harvest social media data?
- How did you develop the selection criteria for your social media collections?
- Have you taken into consideration concerns about inclusivity and diversity?
- What are the main legal issues that you had to work on?
- What happens if someone requests to delete content collected by your institution? How do you manage deleted content?
- Can you tell me about any technical issues you have dealt with?
- What kind of tools does your institution use to harvest social media content? Did you observe any issues or challenges when using [tool/API]? How did you deal with them?
- Is your collection openly accessible? If so, how do researchers access the archived social media collection and what are the most frequent issues related to access?
- Has the covid-19 pandemic had any impact on providing access to the collection? How did [*name institution*] respond to this issue?
- In the survey, you stated that your social media archiving initiative is/is not related to any previous web archiving project. How do you think this may have influenced the development of your project?
- Is there any other concern or challenge you face when archiving social media content that you would like to address?

Bonus question for pilot/fixed term projects

- What are the plans for after the end of your pilot/fixed-term project?

[194] These questions were adapted based on the type of the project (long-term, pilot/fixed term, or planning phase) and institution. Follow-up questions were asked to clarify certain aspects including the development of special collections, and specific harvesting methods used or developed.

Interview Questions

[*Questions for institutions that are still in the planning phase*]

- What is the scope of your institution? What is the context in which the social media collection would be included?
- In the survey, you declared that your institution is still in the planning phase. What do you consider the main issues for the development of your social media collection?
- How does the legal framework of your country affect the development of this initiative?
- Has the covid-19 pandemic influenced the development of your project in any way? (Has it sped the process up or slowed it down?)
- Do you have a rough idea of what the scope of your collection will most likely be?
- Have you taken into consideration concerns about inclusivity and diversity? Will your selection policy address these issues?
- What tools do you plan to use to harvest social media?
- What type of access would you be providing to users?
- Are there any other issues that you would like to address or comments about the development of a social media archive at [*name institution*]?

APPENDIX D

List of Interviewees

[Alphabetical Order]

ARAVANI, FOTEINI – Curator of Contemporary Collections at the Museum of London (Interviewed via Zoom, 28 June 2022).

BINGHAM, NICOLA – Lead Curator of Web Archives, UK Web Archive at the British Library (Interviewed via Zoom, 23 June 2022).

ELS, BEN – Digital Curator, *Bibliothèque nationale du Luxembourg* (Interviewed via Zoom, 12 April 2022).

GEERAERT, FRIEDEL – Expert in Web Archiving at the Royal Library of Belgium, and MESSENS, FIEN – Researcher on the BESOCIAL project (Interviewed via Zoom, 13 July 2022).

GOMES, DANIEL – Head of Arquivo.pt (Interviewed via Zoom, 23 June 2022).

KLINDT MYRVOLL, ANDERS – Programme Manager, *Netarkivet* at the Royal Danish Library (KB) (Interviewed in Copenhagen, 26 September 2022).

MUSSOU, CLAUDE – Head of Service, *Institute national de l'Audiovisuel* (Interviewed in Paris, 19 and 21 April 2022).

NEMETH, MARTÓN – Web Archivist, and DROTOS, LASZLO – Coordinator of the Web archiving Department, National Széchényi Library (Interviewed via Zoom, 25 April 2022)

NEWING, CLAIRE – Web Archivist, The National Archives (Interviewed via Zoom, 30 June 2022).

NG, JOSHUA – Digital Preservation Analyst, Archives New Zealand (Interviewed via Zoom, 10 May 2022).

THOMAS, GRACE – Senior Digital Collection Specialist, and LYON, MEGHAN – Digital Collection Specialist, Library of Congress (Interviewed via Zoom, 2022).

TYBIN, VLADIMIR – Head of Digital Legal Deposit, *Bibliothèque nationale de France* (Interviewed in Paris, 19-20 April 2022).

APPENDIX E

Lists of Tools

Archive-It – https://archive-it.org/

Browsertrix – https://browsertrix.com/

Brozzler – https://github.com/internetarchive/brozzler

Heritrix – https://github.com/internetarchive/heritrix3

Hootsuite – https://www.hootsuite.com/

Instaloader – https://instaloader.github.io/

Picuki – https://www.picuki.com/

Social Feed Manager – https://gwu-libraries.github.io/sfm-ui/

Talkwalker – https://www.talkwalker.com/

Trends24 – https://trends24.in/

Twarc – https://twarc-project.readthedocs.io/en/latest/

Webrecorder [Conifer] – https://conifer.rhizome.org/

YouTube-dl – https://github.com/ytdl-org/youtube-dl

Zeeschuimer – https://github.com/digitalmethodsinitiative/zeeschuimer

Bibliography

AASMAN, S., Bingham, N., Brügger, N., De Wild, K., Gebeil, S., & Schafer, V. (2021). Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections. *WARCnet Paper, Aarhus*. https://web.archive.org/web/20240104222307/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf

AASMAN, S., Clavert, F., de Wild, K., Gebeil, S., Brügger, N., Schafer, V., & Sirajzade, J. (2022). *Studying Women and the COVID-19 Crisis through the IIPC Coronavirus Collection*. https://web.archive.org/web/20231127152255/https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-covid-19-crisis-through-the-iipc-coronavirus-collection/

ABITEBOUL, S., Cobena, G., Masanes, J., & Sedrati, G. (2002). A first experience in archiving the French Web. *Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002 Rome, Italy, September 16–18, 2002 Proceedings 6*, 1–15. https://doi.org/10.1007/3-540-45747-X_1

ACKER, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

AHMED, W., Bath, P. A., & Demartini, G. (2017). Using Twitter as a data source: An overview of ethical, legal, and methodological challenges. *The Ethics of Online Research*. https://doi.org/10.1108/S2398-601820180000002004

ALFANO, M., Reimann, R., Quintana, I. O., Chan, A., Cheong, M., & Klein, C. (2022). The Affiliative Use of Emoji and Hashtags in the Black Lives Matter Movement in Twitter. *Social Science Computer Review*. https://doi.org/10.1177/08944393221131928

ALLEGREZZA, S. (2019). The Future of Our Personal Digital Memories: It's Time To Start Thinking About It. *Atlanti+*, *29*(1), 55–65.

ALLYN, B. (2022, March 14). *Telegram is the app of choice in the war in Ukraine despite experts' privacy concerns: NPR*. https://web.archive.org/web/20221125172220/https://www.npr.org/2022/03/14/1086483703/telegram-ukraine-war-russia

ARCHIBALD, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*, 1609406919874596. https://doi.org/10.1177/1609406919874596

ARCHIVE-IT. (2022, September 9). *What is Brozzler?* Archive-It Help Center. https://web.archive.org/web/20240106171334/https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler-

ARCHIVE-IT. (2024, February 6). *How to monitor your data budget*. Archive-It Help Center. https://support.archive-it.org/hc/en-us/articles/208000096-How-to-monitor-your-data-budget

ARCHIVE-IT HELP CENTER. (2023a, April 22). *Social media and other platforms status*. https://web.archive.org/web/20230422005647/https://support.archive-it.org/hc/en-us/articles/9897233696148-Status-of-monitored-platforms

ARCHIVE-IT HELP CENTER. (2023b, November 28). *Archiving Facebook – Scoping Guidance*. Archive-It Help Center. https://web.archive.org/web/20231128091201/https://support.archive-it.org/hc/en-us/articles/208333113

ARNOLD-STRATFORD, L., & Ovenden, R. (2020). UK non-print legal deposit: From regulations to review. In *Electronic Legal Deposit: Shaping the library collections of the future* (Facet Publishing, Vol. 1, p. 1).

ARP, C. (2019). *Archival Basics: A Practical Manual for Working with Historical Collections*. Rowman & Littlefield.

ARVIDSON, A., Persson, K., & Mannerheim, J. (2000). *The Kulturarw3 Project—The Royal Swedish Web Archiw3e—An Example of 'Complete' Collection of Web Pages*. https://web.archive.org/web/20240202103808/https://eric.ed.gov/?id=ED450729

AUBRY, S. (2010). Introducing web archives as a new library service: The experience of the national library of France. *Liber Quarterly*, *20*(2), 179–199.

BAILEY, S., & Thompson, D. (2006). UKWAC: Building the UK's First Public Web Archive. *D-Lib Magazine*, *12*(1). https://doi.org/10.1045/january2006-thompson

BARROWCLIFFE, R. (2021). Closing the narrative gap: Social media as a tool to reconcile institutional archival narratives with Indigenous counter-narratives. *Archives and Manuscripts*, *49*(3), 151–166. https://doi.org/10.1080/01576895.2021.1883074

BELL, M., Storrar, T., & Winters, J. (2022). Web Archives and the Problem of Access: Prototyping a Researcher Dashboard for the UK Government Web Archive. In L. Jaillant (Ed.), *Archives, Access and Artificial Intelligence* (p. 61). Bielefeld University Press.

BEN-DAVID, A. (2021). Critical web archive research. In *The Past Web* (pp. 181–188). Springer.

BEN-DAVID, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, *2*(1–2), 179–201. https://doi.org/10.1080/24701475.2018.1455412

BERCOVICI, J. (2010). Who Coined 'Social Media'? Web Pioneers Compete for Credit—Forbes. *Forbes*. https://web.archive.org/web/20240228120932/https://www.forbes.com/sites/jeffbercovici/2010/12/09/who-coined-social-media-web-pioneers-compete-for-credit/

BERGIS, J., Summers, E., & Mitchell, V. (2018). *Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations*. Documenting The Now (white paper). https://web.archive.org/web/20230708195547/https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

BERTOT, J. C., Jaeger, P. T., Munson, S., & Glaisyer, T. (2010). Social media technology and government transparency. *Computer*, *43*(11), 53–59. https://doi.org/10.1109/MC.2010.325

BINDER, M. (2019). *MySpace lost 12 years of music and photos, leaving a sizable gap in social network history*. Mashable. https://web.archive.org/web/20240304040955/https://mashable.com/article/myspace-data-loss/

BINGHAM, N., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web

archive. *Big Data & Society*, *8*(1), 2053951721990409. https://doi.org/10.1177/2053951721990409

BINGHAM, N., Byrne, H., Lelkes-Rarugal, C., & Rossi, G. C. (2020, May 29). Using Webrecorder to archive UK political party leaders' social media after the UK General Election 2019. *UK Web Archive Blog*. https://web.archive.org/web/20230530013406/https://blogs.bl.uk/webarchive/2020/05/using-webrecorder-to-archive-uk-political-party-leaders-social-media-after-the-uk-general-election-2.html

BINGHAM, N., Schafer, V., Winters, J., & Ben-David, A. (2024). Conclusion: A Highly transformative age for web archives. In S. Gebeil & J.-C. Peyssard (Eds.), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*. 10.36253/979-12-215-0413-2.29

BLACKWOOD, R. (2021). *Vernacular Mythologies: Instagram, Starbucks and Meaning-Making by Non-Elites at Paris Orly Airport* (1). *1*, Article 1. https://doi.org/10.3828/mlo.v0i0.361

BONACCHI, C., Bevan, A., Keinan-Schoonbaert, A., Pett, D., & Wexler, J. (2019). Participation in heritage crowdsourcing. *Museum Management and Curatorship*, *34*(2), 166–182. https://doi.org/10.1080/09647775.2018.1559080

BONACCHI, C., & Krzyzanska, M. (2019). Digital heritage research re-theorised: Ontologies and epistemologies in a world of big data. *International Journal of Heritage Studies*, *25*(12), 1235–1247.

BRAGG, M., & Hanna, K. (2013). Web Archiving Life Cycle Model. *Archive-It Blog*. https://web.archive.org/web/20240223133910/https://archive-it.org/learn-more/publications/web-archiving-life-cycle-model/

BREWIS, G., Ellis Paine, A., Hardill, I., Lindsey, R., & Macmillan, R. (2021). Co-curation: Archival interventions and voluntary sector records. *Area*, *n/a*(n/a). https://doi.org/10.1111/area.12768

BRIGHT, P. (2012). Europe proposes a 'right to be forgotten'. *Ars Technica*. Available at: https://arstechnica.com/tech-policy/2012/01/eu-proposes-a-right-to-be-forgotten/

BRINKMANN, S. (2014). Unstructured and semi-structured interviewing. *The Oxford Handbook of Qualitative Research*, 277–299.

BRITISH LIBRARY. (2024). *Restoring our services – 30 August 2024 update*. https://blogs.bl.uk/living-knowledge/2024/08/restoring-our-services-30-august-update.html

BROCK, A. (2012). From the blackhand side: Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media*, *56*(4), 529–549. https://doi.org/10.1080/08838151.2012.732147

BRÜGGER, N. (2012). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research/Historische Sozialforschung*, 102–117.

BRÜGGER, N. (2015). A brief history of Facebook as a media text: The development of an empty structure. *First Monday*. https://doi.org/10.5210/fm.v20i5.5423

BRÜGGER, N. (2017a). Chapter 23—Webraries and Web Archives – The Web Between Public and Private. In D. Baker & W. Evans (Eds.), *The End of Wisdom?* (pp. 185–190). Chandos Publishing. https://doi.org/10.1016/B978-0-08-100142-4.00023-3

BRÜGGER, N. (2017b). Web History and Social media. In J. Burgess, A. E. Marwick, & T. Poell (Eds.), *The SAGE Handbook of Social Media* (pp. 196–212). SAGE Publications. http://ebookcentral.proquest.com/lib/ulondon/detail.action?docID=5151795

BRÜGGER, N. (2018a). A brief Outline of Temporalities of the Web Online and in Web Archives. In V. Schafer (Ed.), *Temps et temporalités du Web* (pp. 57–74). Presses universitaires de Paris Nanterre. https://doi.org/10.4000/books.pupo.6073

BRÜGGER, N. (2018b). *The Archived Web: Doing History in the Digital Age*. MIT Press.

BRÜGGER, N., & Milligan, I. (2018). *The SAGE Handbook of Web History*. SAGE.

BRÜGGER, N., & Schroeder, R. (2017). *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press. 10.14324/111.9781911307563

BRUNS, A. (2018, January 10). *The Library of Congress Twitter Archive: A Failure of Historic Proportions*. Medium. https://web.archive.org/web/20220507025042/https://medium.com/dmrc-at-

large/the-library-of-congress-twitter-archive-a-failure-of-historic-proportions-6dc1c3bc9e2c

BRUNS, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

BRUNS, A., Angus, D., & Graham, T. (2021). Twitter campaigning strategies in Australian federal elections 2013–2019. *Social Media+ Society*, *7*(4), 20563051211063462. https://doi.org/10.1177/20563051211063462

BRUNS, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, 1–9.

BRUNS, A., & Weller, K. (2016). Twitter as a first draft of the present: And the challenges of preserving it for the future. *Proceedings of the 8th ACM Conference on Web Science*, 183–189. https://doi.org/10.1145/2908131.2908174

BUCHER, T., & Helmond, A. (2017). *The affordances of social media platforms*.

BURGESS, J., & Bruns, A. (2012). Twitter Archives and the Challenges of 'Big Social Data' for Media and Communication Research. *M/C Journal*, *15*(5), Article 5. HTTPS://DOI.ORG/10.5204/MCJ.561

Burgess, M. (2016, January 18). Friends Reunited to close after 15 years. *Wired UK*. https://web.archive.org/web/20220704021554/https://www.wired.co.uk/article/friends-reunited-closed

BURKELL, J., Fortier, A., Wong, L. (Lola) Y. C., & Simpson, J. L. (2014). Facebook: Public space, or private space? *Information, Communication & Society*, *17*(8), 974–985. https://doi.org/10.1080/1369118X.2013.870591

BURKEY, B. (2020). Repertoires of Remembering: A Conceptual Approach for Studying Memory Practices in the Digital Ecosystem. *Journal of Communication Inquiry*, *44*(2), 178–197. https://doi.org/10.1177/0196859919852080

BURTON, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). ' Right time, right place' health communication on Twitter: Value and accuracy of location information. *Journal of Medical Internet Research*, *14*(6), e156. https://doi.org/10.2196/jmir.2121

BYRNE, H. (2017, April 18). The Challenges of Web Archiving Social Media. *UK Web Archive Blog*. https://web.archive.org/web/20230204212951/https://blogs.bl.uk/webarchive/2017/04/the-challenges-of-web-archiving-social-media.html

BYRNE, H., Cannelli, B., Noguera, C., Kurzmeier, M., & De Wild, K. (2023). *Looking ahead: After web (archives)?* https://web.archive.org/web/20240216223623/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Looking_ahead.pdf

CADAVID, J. A. P. (2014). Copyright Challenges of Legal Deposit and Web Archiving in the National Library of Singapore. *Alexandria*, *25*(1–2), 1–19. https://doi.org/10.7227/ALX.0017

CASWELL, M., Migoni, A. A., Geraci, N., & Cifor, M. (2017). 'To be able to imagine otherwise': Community archives and the importance of representation. *Archives and Records*, *38*(1), 5–26. https://doi.org/10.1080/23257962.2016.1260445

CHAMBERS, S., Birkholz, J., Geeraert, F., Pranger, J., Messens, F., Lieber, S., Mechant, P., Michel, A., & Vlassenroot, E. (2021). *BESOCIAL: final report WorkPackage1 an international review of social media archiving initiatives*. 91.

CHARLTON, T. (2017). The treachery of archives: Representation, power, and the urgency for self-reflexivity in archival arrangement and description. *The iJournal: Student Journal of the University of Toronto's Faculty of Information*, *3*(1). https://theijournal.ca/index.php/ijournal/article/view/28894

CHOKSHI, N. (2019, March 19). Myspace, Once the King of Social Networks, Lost Years of Data From Its Heyday. *The New York Times*. https://web.archive.org/web/20240410050708/https://www.nytimes.com/2019/03/19/business/myspace-user-data.html

CHOWDHURY, G. G. (2002). Digital Divide: How can digital libraries bridge the gap? *International Conference on Asian Digital Libraries*, 379–391.

CHRISAFIS. (2018). *Who are the gilets jaunes and what do they want? | France | The Guardian*. The Guardian. https://web.archive.org/web/20230407204833/https://www.theguardian.com/world/2018/dec/03/who-are-the-gilets-jaunes-and-what-do-they-want

CLEGG, N. (2023, August 22). *New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect | Meta*. https://web.archive.org/web/20240712234453/https://about.fb.com/news/2023/08/new-features-and-additional-transparency-measures-as-the-digital-services-act-comes-into-effect/

COGBURN, D. L., & Espinoza-Vasquez, F. K. (2014). From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign. In *Strategy, Money and Technology in the 2008 Presidential Election* (pp. 282–306). Routledge.

COLIN-ARCE, A., Fernández-Quintanilla, S., Benítez-Pérez, V., García-Monroy, A., & Rogel-Salazar, R. (2023). *Web Archiving en español:Barriers to Accessing and Using Web Archives in Latin America*. https://www.youtube.com/watch?v=plQURfARGBc

COOK, L. (2015). The Right to Be Forgotten: A Step in the Right Direction for Cyberspace Law and Policy. *Journal of Law, Technology, & the Internet*, *6*(1), 121.

COOK, T. (2011). 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists*, *32*(2), 173–189. https://doi.org/10.1080/00379816.2011.619688

COOK, T., & Schwartz, J. M. (2002). Archives, records, and power: From (postmodern) theory to (archival) performance. *Archival Science*, *2*, 171–185. https://doi.org/10.1007/BF02435620

CORMODE, G., & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday*. https://doi.org/10.5210/fm.v13i6.2125

COSTA, M., Gomes, D., & Silva, M. J. (2017). The evolution of web archiving. *International Journal on Digital Libraries*, *18*(3), 191–205. https://doi.org/10.1007/s00799-016-0171-9

COX, R., & Samuels, H. (1988). The archivist's first responsibility: A research agenda to improve the identification and retention of records of enduring value. *The American Archivist*, *51*(1–2), 28–42. https://doi.org/10.17723/aarc.51.1-2.gkw6742413344ug8

CRESWELL, J. W., author. (2022). *Research design: Qualitative, quantitative, and mixed method approaches / John W. Creswell.* (1452226105). SAGE.

CUBELLS PASTOR, A. (2020). *The visual rhetoric of right-wing populism: An analysis of Vox's visual communication on Instagram*. Malmö universitet/Kultur och samhälle. https://web.archive.org/web/20240527101848/https://www.diva-portal.org/smash/get/diva2:1482209/FULLTEXT01.pdf

CUI, A., Zhang, M., Liu, Y., Ma, S., & Zhang, K. (2012). Discover breaking events with popular hashtags in twitter. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1794–1798. https://doi.org/10.1145/2396761.2398519

CUI, C., Pinfield, S., Cox, A., & Hopfgartner, F. (2023). Participatory Web Archiving: Multifaceted Challenges. In I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, & R. D. Frank (Eds.), *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity* (pp. 79–87). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28035-1_7

CUNNEA, P., Hamilton, G., Hawley, G., & Saunderson, F. (2020). The Influence of Legal Deposit Legislation on the Digital Collections of the National Library of Scotland. In M. Terras & P. Gooding (Eds.), *Electronic Legal Deposit: Shaping the library of the Future*. Facet Publishing.

DAVIS, C. (2014). Archiving the Web: A Case Study from the University of Victoria. *The Code4Lib Journal*, *26*. https://web.archive.org/web/20240205125400/https://journal.code4lib.org/articles/10015

DAY, M. (2006). The Long-Term Preservation of Web Content. In J. Masanés (Ed.), *Web Archiving* (pp. 177–199). Springer. https://doi.org/10.1007/978-3-540-46332-0_8

DERRIDA, J. (2002). Archive Fever. In C. Hamilton, V. Harris, J. Taylor, M. Pickover, G. Reid, & R. Saleh (Eds.), *Refiguring the Archive* (pp. 38–38). Springer Netherlands. https://doi.org/10.1007/978-94-010-0570-8_4

DÉSOS-WARNIER, C. (2023). *Faire réseau autour des archives du web: Bilan et perspectives du projet ResPaDon*. https://web.archive.org/web/20240213171344/https://bbf.enssib.fr/tour-d-horizon/faire-reseau-autour-des-archives-du-web-bilan-et-perspectives-du-projet-respadon_71102

DIJCK, J. van. (2013). The Culture of Connectivity: A Critical History of Social Media. In *The Culture of Connectivity*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199970773.001.0001/acprof-9780199970773

DPC. (n.d.-a). *The What, Why and Who of Documentation—Digital Preservation Coalition*. Retrieved 6 July 2024, from https://web.archive.org/web/20240523001047/https://www.dpconline.org/digipres/implement-digipres/digital-preservation-documentation-guide/digital-preservation-documentation-what-why-who

DPC. (n.d.-b). *UK Web Archive: Celebrating 15 years—Digital Preservation Coalition*. Retrieved 24 January 2024, from https://web.archive.org/web/20240124153207/https://www.dpconline.org/events/digital-preservation-awards/dpa2020-ukwa

DRISCOLL, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication*, *8*, 20.

DRUBIN, D. G., & Kellogg, D. R. (2012). English as the universal language of science: Opportunities and challenges. *Molecular Biology of the Cell*, *23*(8), 1399–1399. https://doi.org/10.1091/mbc.e12-02-0108

DUPONT, H. (1999). Legal Deposit in Denmark—The New Law and Electronic Products. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *9*(2), Article 2. https://doi.org/10.18352/lq.7539

ELALDI, S., & Yerliyurt, N. S. (2017). The Efficacy of Drama in Field Experience: A Qualitative Study Using MAXQDA. *Journal of Education and Learning*, *6*(1), 10–26. https://doi.org/10.5539/jel.v6n1p10

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*, *25*(1–2), 31–50. https://doi.org/10.7227/ALX.0019

EVANS, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*.

FARRELL-BANKS, D. (2020). 1215 in 280 Characters: Talking about Magna Carta on Twitter. In *European Heritage, Dialogue and Digital Practices* (pp. 86–106).

Routledge. https://primo.getty.edu/primo-explore/fulldisplay?vid=GRI&docid=GETTY_ALMA51197623930001551&context=L

FAUDUET, L., & Peyrard, S. (2010). A Data-First Preservation Strategy: Data Management In SPAR. *iPRES*. https://phaidra.univie.ac.at/o:185415

FAYE, A., Thièvre, J., & Schafer, V. (2024). Le temps des plateformes: Enjeux, différences et complémentarité de l'archivage des médias sociaux numériques à la Bibliothèque nationale de France et à l'Institut national de l'audiovisuel. *978-2-86000-390-2*. https://orbilu.uni.lu/handle/10993/60404

FERRÉ-PAVIA, C., Zabaleta, I., Gutierrez, A., Fernandez-Astobiza, I., & Xamardo, N. (2018). Internet and social media in European minority languages: Analysis of the digitalization process. *International Journal of Communication*, *12*, 22.

FIELDING, N. G. (2012). Triangulation and mixed methods designs: Data integration with new research technologies. *Journal of Mixed Methods Research*, *6*(2), 124–136. https://psycnet.apa.org/doi/10.1177/1558689812437101

FLICKR FOUNDATION. (n.d.). Data Lifeboat. *Flickr Foundation*. Retrieved 18 July 2024, from https://web.archive.org/web/20230311104151/https://www.flickr.org/programs/content-mobility/data-lifeboat/

FLORINI, S. (2014). Tweets, Tweeps, and Signifyin' Communication and Cultural Performance on "Black Twitter". *Television & New Media*, *15*(3), 223–237. https://doi.org/10.1177/1527476413480247

FONDREN, E., & Menard McCune, M. (2018). Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive. *Preservation, Digital Technology & Culture*, *47*(2), 33–44. https://doi.org/10.1515/pdtc-2018-0011

GARDE-HANSEN, J. (2009). MyMemories?: Personal Digital Archive Fever and Facebook. In J. Garde-Hansen, A. Hoskins, & A. Reading (Eds.), *Save As … Digital Memories*. Palgrave Macmillan. https://doi.org/10.1057/9780230239418_8

GARG, K., Jayanetti, H. R., Alam, S., Weigle, M. C., & Nelson, M. L. (2023). Challenges in replaying archived Twitter pages. *International Journal on Digital Libraries*. https://doi.org/10.1007/s00799-023-00379-w

GEERAERT, F., & Németh, M. (2020). Exploring special web archives collections related to COVID-19: The case of the National Library of the National Széchényi Library in Hungary. *WARCnet Papers*. https://web.archive.org/web/20230902214709/https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_Hungary.pdf

GHADDAR, J. J., & Caswell, M. (2019). "To go beyond": Towards a decolonial archival praxis. *Archival Science*, *19*(2), 71–85. https://doi.org/10.1007/s10502-019-09311-1

GIBBY, R., & Brazier, C. (2012). Observations on the development of non-print legal deposit in the UK. *Library Review*, *61*(5), 362–377. https://doi.org/10.1108/00242531211280487

GIZZI, M. C., & Rädiker, S. (2021). *The Practice of Qualitative Data Analysis: Research Examples Using MAXQDA*. BoD–Books on Demand. 10.36192/978-3-948768058

GOLDSMITH, L. P., Rowland-Pomp, M., Hanson, K., Deal, A., Crawshaw, A. F., Hayward, S. E., Knights, F., Carter, J., Ahmad, A., Razai, M., Vandrevala, T., & Hargreaves, S. (2022a). Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: A systematic review. *BMJ Open*, *12*(11), Article 11. https://doi.org/10.1136/bmjopen-2022-061896

GOLDSMITH, L. P., Rowland-Pomp, M., Hanson, K., Deal, A., Crawshaw, A. F., Hayward, S. E., Knights, F., Carter, J., Ahmad, A., Razai, M., Vandrevala, T., & Hargreaves, S. (2022b). Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: A systematic review. *BMJ Open*, *12*(11), e061896. https://doi.org/10.1136/bmjopen-2022-061896

GOMES, D. (2017). Web preservation demands access. *Digital Preservation Coalition - Blog*. https://web.archive.org/web/20240507162150/https://www.dpconline.org/blog/wdpd/web-preservation-demands-access

GOMES, D., Miranda, J., & Costa, M. (2010). *A survey on web archiving initiatives – sobre.arquivo.pt*. Foundation for National Scientific Computing (FCCN).

https://web.archive.org/web/20240528052109/https://sobre.arquivo.pt/pt/a-survey-on-web-archiving-initiatives/

GOMES, D., Miranda, J., & Costa, M. (2011). A Survey on Web Archiving Initiatives. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 408–420). Springer Berlin Heidelberg.

GOOD, K. D. (2012). From scrapbook to Facebook: A history of personal media assemblage and archives. *New Media & Society*, *15*(4). http://journals.sagepub.com/doi/abs/10.1177/1461444812458432

GOODING, P., & Terras, M. (2020a). 'An ark to save learning from deluge'? Reconceptualising legal deposit after the digital turn. In *Electronic Legal Deposit: Shaping the Library Collections of the Future* (pp. 203–228). facet publishing.

GOODING, P., & Terras, M. (2020b). *Electronic Legal Deposit: Shaping the library collections of the future*. Facet Publishing.

GOODING, P., Terras, M., & Berube, L. (2019). *Towards user-centric evaluation of UK non-print legal deposit: A digital library futures white paper*.

GRAHAM, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568–578. https://doi.org/10.1080/00330124.2014.907699

GRAY, L. M., Wong-Wylie, G., Rempel, G. R., & Cook, K. (2020). Expanding qualitative research interviewing strategies: Zoom video communications. *The Qualitative Report*, *25*(5), 1292–1301. https://doi.org/10.46743/2160-3715/2020.4212

GROTKE, A. (2017). *Getting Started in Web Archiving*.

GUIDANCE on the Legal Deposit Libraries (Non-Print Works) Regulations 2013. (2013). https://web.archive.org/web/20230803193640/https://www.gov.uk/government/publications/guidance-on-the-legal-deposit-libraries-non-print-works-regulations-2013

HABELSBERGER, B. E., & Bhansing, P. V. (2021). Art Galleries in Transformation: Is COVID-19 Driving Digitisation? *Arts*, *10*(3), 48. https://doi.org/10.3390/arts10030048

HALCOMB, E. J., & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary? *Applied Nursing Research*, 6. https://doi.org/10.1016/j.apnr.2005.06.001

HAM, F. G. (1993). Selecting and appraising archives and manuscripts. *Society of American Archivists*. Selecting and appraising archives and manuscripts

HARRIS, V. (2002). The archival sliver: Power, memory, and archives in South Africa. *Archival Science*, *2*(1), 63–86. https://doi.org/10.1007/BF02435631

HAWES, A. (Director). (2020). *Archiving 1418-Now using Rhizome's Webrecorder: Observations and reflections* [Video recording]. https://www.youtube.com/watch?v=YnRBza0y18Q

HEALY, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. https://doi.org/DOI 10.17605/OSF.IO/VF7GT

HELMOND, A., Rogers, R., & ASCA (FGw). (2015). *The web as platform: Data flows in social media* [Urn:nbn:nl:ui:29-1.485895]. https://dare.uva.nl/personal/pure/en/publications/the-web-as-platform-data-flows-in-social-media(a990b0ba-8010-4cec-b5c8-ba9f3f4bceb3).html

HELMOND, A., & Vlist, F. van der. (2019). Social Media and Platform Historiography: Challenges and Opportunities. *TMG Journal for Media History*, *22*(1), Article 1. https://doi.org/10.18146/tmg.434

HENNINGER, M., & Scifleet, P. (2016). How are the new documents of social networks shaping our cultural memory. *Journal of Documentation*, *72*(2). https://www.emeraldinsight.com/doi/abs/10.1108/JD-06-2015-0069

HERN, A. (2019, March 18). Myspace loses all content uploaded before 2016. *The Guardian*. https://www.theguardian.com/technology/2019/mar/18/myspace-loses-all-content-uploaded-before-2016

HOCKX-YU, H. (2014). *Archiving Social Media in the Context of Non-print Legal Deposit*. IFLA WLIC 2014, Lyon, France. http://library.ifla.org/999/

HOCKX-YU, H. (2015, August). *Meeting the Challenges of Preserving the UK Web*. Helen Hockx's Blog: Things I Cannot Say in 140 Characters.

https://web.archive.org/web/20230322230919/https://hhockx.wordpress.com/2015/08/11/meeting-the-challenges-of-preserving-the-uk-web/

HOLMBERG, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, *101*(2), 1027–1042. https://doi.org/10.1007/s11192-014-1229-3

HU, J. (2020, August 3). *The Second Act of Social-Media Activism*. The New Yorker. https://web.archive.org/web/20230209035450/https://www.newyorker.com/culture/cultural-comment/the-second-act-of-social-media-activism

HUC-HEPHER, S., & Wells, N. (2021). Exploring Online Diasporas: London's French and Latin American Communities in the UK Web Archive. In *The past web: Exploring web archives* (pp. 189–201). Springer.

ICA. (2011). *Universal Declaration on Archives | International Council on Archives*. https://www.ica.org/en/universal-declaration-archives

IIPC. (n.d.). Browser-based crawling system for all. *IIPC*. Retrieved 12 February 2024, from https://web.archive.org/web/20230323170429/https://netpreserve.org/projects/browser-based-crawling/

IIPC. (2017, March 15). *Strategic Plan 2016 – 2017*. https://web.archive.org/web/20170315022037/http://netpreserve.org/sites/default/files/IIPC-Strategic-Plan-2016-2017.pdf

ISAAK, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, *51*(8), 56–59.

JACKSON, A. (2023, January 16). UK Web Archive Technical Update—Winter 2022. *UK Web Archive Blog*. https://web.archive.org/web/20230301124541/https://britishlibrary.typepad.co.uk/webarchive/

JACOBSEN, T., Punzalan, R. L., & Hedstrom, M. L. (2013). Invoking "collective memory": Mapping the emergence of a concept in archival science. *Archival Science*, *13*(2–3), 217–251. https://doi.org/10.1007/s10502-013-9199-4

JAILLANT, L. (2022). More Data, Less Process: A User-Centered Approach to Email and Born-Digital Archives. *The American Archivist*, *85*(2), 533–555. https://doi.org/10.17723/2327-9702-85.2.533

JEFFREY, S. (2012). A new Digital Dark Age? Collaborative web tools, social media and long-term preservation. *World Archaeology*, *44*(4), 553–570. https://doi.org/10.1080/00438243.2012.737579

JENNER, B. M., & Myers, K. C. (2019). Intimacy, rapport, and exceptional disclosure: A comparison of in-person and mediated interview contexts. *International Journal of Social Research Methodology*, *22*(2), 165–177. https://doi.org/10.1080/13645579.2018.1512694

JIMERSON, R. (2006). Embracing the Power of Archives. *The American Archivist*, *69*(1), 19–32. https://doi.org/10.17723/aarc.69.1.r0p75n2084055418

JUDGMENT IN CASE C-131/12 Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González (Court of Justice of the European Union 13 May 2014). https://curia.europa.eu/juris/document/document_print.jsf?doclang=EN&text=&pageIndex=0&part=1&mode=DOC&docid=152065&occ=first&dir=&cid=667631

JURIK, B., & Zierau, E. (2017). *Data management of web archive research data*. https://sas-space.sas.ac.uk/9675/

KAPLAN, E. (2002). 'Many paths to partial truths': Archives, anthropology, and the power of representation. *Archival Science*, *2*, 209–220. https://doi.org/10.1007/BF02435622

KARIRYAA, A., Rundé, S., Heuer, H., Jungherr, A., & Schöning, J. (2022). The Role of Flag Emoji in Online Political Communication. *Social Science Computer Review*, *40*(2), 367–387. https://doi.org/10.1177/0894439320909085

KENDALL, T. (2021). From Binge-Watching to Binge-Scrolling: TikTok and the Rhythms of #LockdownLife. *Film Quarterly*, *75*(1), 41–46. https://doi.org/10.1525/fq.2021.75.1.41

KERN, F. G. (2018). The trials and tribulations of applied triangulation: Weighing different data sources. *Journal of Mixed Methods Research*, *12*(2), 166–181. https://doi.org/10.1177/1558689816651032

KETELAAR, E. (2017). Archival turns and returns: Studies of the archive. In A. Gilliand, S. Mckemmish, & A. Jau (Eds.), *Research in the archival multiverse* (pp. 228–268). Monash University Publishing.

KILBRIDE, W. (2024, March 21). *The Anthropocene Remembered: Digital Memory After the Climate Crisis - Digital Preservation Coalition*. https://web.archive.org/web/20230526094342/https://www.dpconline.org/blog/the-anthropocene-remembered-digital-memory-after-the-climate-crisis

KLIMAN-SILVER, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. *Proceedings of the 2015 Internet Measurement Conference*, 121–127. https://doi.org/10.1145/2815675.2815714

KOERBIN, P. (2017). Revisiting the world wide web as artefact: Case studies in archiving small data for the National Library of Australia's PANDORA archive. *Web*, *25*, 191–206.

KOERBIN, P. (2021). National web archiving in Australia: Representing the comprehensive. In *The past web: Exploring web archives* (pp. 23–32). Springer.

KOŠČÍK, M., & Myška, M. (2019). Copyright law challenges of preservation of" born-digital" digital content as cultural heritage. *European Journal of Law and Technology*, *10*(1). https://web.archive.org/web/20240919144950/https://ejlt.org/index.php/ejlt/article/view/664

KROMBHOLZ, K., Merkl, D., & Weippl, E. (2012). Fake identities in social media: A case study on the sustainability of the Facebook business model. *Journal of Service Science Research*, *4*(2), 175–212. https://doi.org/10.1007/s12927-012-0008-z

KUCKARTZ, A. M., & Kuckartz, U. (2002). *Qualitative text analysis with MAXQDA*. Fundación Centro de Estudios Andaluces.

KUNY, T. (1998). *The digital dark ages? Challenges in the preservation of electronic information*. Undefined. https://api.semanticscholar.org/CorpusID:128073783

LARSEN, S. (2005). Preserving the Digital Heritage: New Legal Deposit Act in Denmark. *Alexandria*, *17*(2), 81–87. https://doi.org/10.1177/095574900501700204

LAURIDSEN, J. (2021, February 25). *SolrWayback 4.0 release! What's it all about?* https://web.archive.org/web/20240523053351/https://netpreserveblog.wordpress.com/2021/02/25/solrwayback-4-0-release-whats-it-all-about/

LAURSEN, D., & Møldrup-Dalum, P. (2016). Netarkivet 10 år. *Revy*, *39*(2), 6–9.

LE FOLLIC, A., & Chouleur, M. (2016). La collecte des médias sociaux, un enjeu pour la constitution des collections de dépôt légal du web à la Bibliothèque nationale de France. In *Pérenniser l'éphémère. Archivage et médias sociaux* (pp. 109–124).

LEBLANC, Z., Janco, A., Wermer-Colan, A., Dombrowski, Q., Kijas, A., Majstorovic, S., Strong, D., & Peaslee, E. (2022). A Conversation with the Organizers of Saving Ukrainian Cultural Heritage Online (SUCHO). *Journal of Library Outreach and Engagement*, *2*(1), Article 1. https://doi.org/10.21900/j.jloe.v2i1.969

LEETARU, K. (2015). *Why It's So Important To Understand What's In Our Web Archives*. Forbes. https://web.archive.org/web/20210224013158/http://www.forbes.com/sites/kalevleetaru/2015/11/25/why-its-so-important-to-understand-whats-in-our-web-archives/%5C

LIBRARIES, U. L. D. (2023). *Archive of Tomorrow: Capturing public health discourse in the UK Web Archive* (British Library) [Report]. UK Legal Deposit Libraries. https://bl.iro.bl.uk/concern/reports/965c8aa9-f782-40f1-8046-ed5d893e13fc

LIBRARY OF CONGRESS. (2013, January 4). Update on the Twitter Archive at the Library of Congress [Blog]. *Library of Congress*. https://web.archive.org/web/20210511123014///blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

LIBRARY OF CONGRESS. (2017, December 26). Update on the Twitter Archive at the Library of Congress [Blog]. *Library of Congress*. https://web.archive.org/web/20210509124919///blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/

LIEBER, S., Van Assche, D., Chambers, S., Messens, F., Geeraert, F., Birkholz, J. M., & Dimou, A. (2021). BESOCIAL: A Sustainable Knowledge Graph-Based Workflow for Social Media Archiving. *SEMANTICS2021, the 17th International Conference on Semantic Systems*, 198–212.

LINKEDIN. (2017). *LinkedIn Crawling Terms and Conditions*. https://web.archive.org/web/20240121094137/https://www.linkedin.com/legal/crawling-terms

LINKEDIN. (2022). *User Agreement*. https://web.archive.org/web/20240518221005/https://www.linkedin.com/legal/user-agreement

LITTMAN, J. (2019, January 8). *Twitter's Developer Policies for Researchers, Archivists, and Librarians*. Medium. https://web.archive.org/web/20210424204948/https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2

LITTMAN, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., Vij, R., & Wrubel, L. (2018). API-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, *19*(1), 21–38. https://doi.org/10.1007/s00799-016-0201-7

LOBBÉ, Q. (2018). Revealing historical events out of web archives. *International Conference on Theory and Practice of Digital Libraries*, 312–316.

LOBE, B., Morgan, D., & Hoffman, K. A. (2020). Qualitative Data Collection in an Era of Social Distancing. *International Journal of Qualitative Methods*, *19*, 1609406920937875. https://doi.org/10.1177/1609406920937875

LOMBORG, S. (2012). Researching Communicative Practice: Web Archiving in Qualitative Social Media Research. *Journal of Technology in Human Services*, *30*(3–4), 219–231. https://doi.org/10.1080/15228835.2012.744719

LOMBORG, S., & Bechmann, A. (n.d.). Using APIs for Data Collection on Social Media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

LOR, P., & Britz, J. J. (2004). A moral perspective on South–North web archiving. *Journal of Information Science*, *30*(6), 540–549. https://doi.org/10.1177/0165551504047925

LOUGEN, C. (2009). *The Sage Encyclopedia of Qualitative Research Methods Lisa M. Given*. *49*(1), 101.

LUTZ, C. (2022). Inequalities in social media use and their implications for digital methods research. *The SAGE Handbook of Social Media Research Methods*, 679–690.

MACDONALD, B., & Walker, R. (1975). Case-study and the social philosophy of educational research. *Cambridge Journal of Education*, *5*(1), 2–11. https://doi.org/10.1080/0305764750050101

MAEMURA, E. (2023). *Sorting URLs out: Seeing the web through infrastructural inversion of archival crawling.* https://www.tandfonline.com/doi/epdf/10.1080/24701475.2023.2258697?needAccess=true

MAEMURA, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, *69*(10), 1223–1233. https://doi.org/10.1002/asi.24048

MAKHORTYKH, M., & Sydorova, M. (2017). Social media and visual framing of the conflict in Eastern Ukraine. *Media, War & Conflict*, *10*(3), 359–381. https://doi.org/10.1177/1750635217702539

MARSHALL, C. C., & Shipman, F. M. (2012). On the institutional archiving of social media. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. https://doi.org/10.1145/2232817.2232819

MARSHALL, C. C., & Shipman, F. M. (2015). Exploring the ownership and persistent value of Facebook content. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 712–723.

MARSHALL, C. C., & Shipman, F. M. (2017). Who Owns the Social Web? *Communications of the ACM*, *60*(5), 52–61. https://doi.org/10.1145/2996181

MASANÈS, J. (2006). Web archiving: Issues and methods. In *Web archiving* (pp. 1–53). Springer.

MASANÈS, J., Major, D., & Gomes, D. (2021). The Past Web: A Look into the Future. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 285–291). Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_22

MECHANT, P., Birkholz, J. M., Messens, F., Michel, A., Rolin, E., Vandepontseele, S., Vlassenroot, E., Watrin, P., & Budulan, D. (2022). *Towards a sustainable social media archiving strategy for Belgium*. https://web.archive.org/web/20230907070126/https://www.belspo.be/belspo/brain2-be/projects/FinalReports/BESOCIAL_FinRep.pdf

MEDERO, R. S., & Maestre, R. L. (2023). Should You Put an Emoticon on Your Flag? How Subliminal Visual Stimuli Can Change Political Opinions. *Nationalities Papers*, 1–20. https://doi.org/10.1017/nps.2023.62

MESSENS, F., Birkholz, J. M., Chambers, S., Geeraert, F., Michel, A., Mechant, P., Vlassenroot, E., Lieber, S., Dimou, A., & Watrin, P. (2021). BESOCIAL– Towards a sustainable strategy for archiving and preserving social media in Belgium. *Digital Humanities Benelux 2021 Conference*. https://biblio.ugent.be/publication/8712394

MICHEL, A. (2021). Web Archiving in the Public Interest from a Data Protection Perspective. In *Deep diving into data protection: 1979-2019: Celebrating 40 years of research on privacy data protection at the CRIDS* (pp. 181–200). Larcier.

MILAN, S. (2015). When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Social Media+ Society*, *1*(2), 2056305115622481. https://doi.org/10.1177/2056305115622481

MILLIGAN, I. (2015, July 14). *Web Archive Legal Deposit: A Double-Edged Sword*. IAN MILLIGAN Digital History, Web Archives, and Contemporary History. https://web.archive.org/web/20190412051710/http://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/

MILLIGAN, I. (2018, March 27). *Ethics and the Archived Web Presentation: "The Ethics of Studying GeoCities"*. Ian Milligan. https://web.archive.org/web/20211101122805/https://ianmilli.wordpress.com/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/

MILLIGAN, I., Ruest, N., & Lin, J. (2016). Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16*, 107–110. https://doi.org/10.1145/2910896.2910913

MIRRORWEB. (2018, April 25). *What are the implications of GDPR for digital archiving?* MirrorWeb. https://web.archive.org/web/20240215213052/https://www.mirrorweb.com/blog/what-are-the-implications-of-gdpr-for-digital-archiving

MOHR, G., Kunze, J., & Stack, M. (2008). *The WARC File Format 1.0 (ISO 28500)*. https://escholarship.org/uc/item/9nh616wd

MORAN-ELLIS, J., Alexander, V. D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., & Thomas, H. (2006). Triangulation and integration: Processes, claims and implications. *Qualitative Research*, *6*(1), 45–59. https://doi.org/10.1177/1468794106058870

MUIR, A. (2020). Publishers, legal deposit and the changing publishing environment. *Electronic Legal Deposit: Shaping the Library Collections of the Future*, *1*, 121.

MUSK, E. [@elonmusk]. (2023, July 1). *To address extreme levels of data scraping & system manipulation, we've applied the following temporary limits: - Verified accounts are limited to reading 6000 posts/day—Unverified accounts to 600 posts/day—New unverified accounts to 300/day* [Tweet]. Twitter. https://web.archive.org/web/20230705030327/https://twitter.com/elonmusk/status/1675187969420828672

NASCIMENTO, L. da S., & Steinbruch, F. K. (2019). "The interviews were transcribed", but how? Reflections on management research. *RAUSP Management Journal*, *54*, 413–429. https://www.emerald.com/insight/content/doi/10.1108/RAUSP-05-2019-0092/full/html

NATOW, R. S. (2020). The use of triangulation in qualitative studies employing elite interviews. *Qualitative Research*, *20*(2), 160–173. https://doi.org/10.1177/1468794119830077

NÉMETH, M. (2020, June 26). *From pilot to portal: A year of web archiving in Hungary*. https://web.archive.org/web/20230922080739/https://netpreserveblog.wordpress.com/2020/06/26/a-year-of-web-archiving-in-hungary/

NEUBURGER, J. D. (2022, December 8). *hiQ and LinkedIn Reach Settlement in Data Scraping Lawsuit*. https://web.archive.org/web/20240205032458/https://natlawreview.com/article/hiq-and-linkedin-reach-proposed-settlement-landmark-scraping-case

NEWING, C. (2020, November). *The UK Government Social Media Archive – now bigger, more comprehensive and searchable—Digital Preservation Coalition*. https://web.archive.org/web/20210409232615/https://www.dpconline.org/blog/wdpd/blog-claire-newing-wdpd

NORTH CAROLINA STATE GOVERNMENT WEB SITE ARCHIVES AND ACCESS PROGRAM. (n.d.). *About*. Retrieved 17 May 2021, from https://web.archive.org/web/20231128042853/https://webarchives.ncdcr.gov/about.html

NOVAK, M. (2023, August 19). *Twitter Deletes All User Photos And Links From 2011-2014*. Forbes. https://web.archive.org/web/20230824215538/https://www.forbes.com/sites/mattnovak/2023/08/19/twitter-deletes-all-user-photos-and-links-from-2011-2014/?sh=6a8beea878fe

ODOM, W., Zimmerman, J., & Forlizzi, J. (2010). Virtual Possessions. In K. Halskov & M. G. Petersen (Eds.), *Proceedings of DIS 10 Designing Interactive Systems (DIS),* (pp. 368–371).

ODOM, W., Zimmerman, J., & Forlizzi, J. (2011). Teenagers and their virtual possessions: Design opportunities and issues. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1491–1500.

OGDEN, J. (2020). *Saving the Web: Facets of Web Archiving in Everyday Practice* [Phd, University of Southampton]. https://eprints.soton.ac.uk/447624/

OGDEN, J. (2021). "Everything on the internet can be saved": Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*, 1–20. https://doi.org/10.1080/24701475.2021.1985835

OGDEN, J., Halford, S., & Carr, L. (2017). Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. *Proceedings of the 2017 ACM on Web Science Conference*, 299–308. https://doi.org/10.1145/3091478.3091506

OGDEN, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-021-00032-5

OLIFFE, J. L., Kelly, M. T., Gonzalez Montaner, G., & Yu Ko, W. F. (2021). Zoom Interviews: Benefits and Concessions. *International Journal of Qualitative Methods*, *20*, 16094069211053522. https://doi.org/10.1177/16094069211053522

O'REILLY, T. (2005). *What Is Web 2.0*. https://web.archive.org/web/20240518071636/https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

ORTNER, C., Sinner, P., & Jadin, T. (2019). The History of Online Social Media. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 372–384). https://doi.org/10.4135/9781526470546.n25

PALMBERGER, M., & Gingrich, A. (2022). Qualitative Comparative Practices: DImension, Cases and Strategies. In pages 94-108, *The SAGE Handbook of Qualitative Data Analysis*. SAGE Publications Ltd. https://doi.org/10.4135/9781446282243

PANKHURST, S. (2016, January 15). *FriendsReunited—The sunset of an era*. Medium. https://web.archive.org/web/20210507111355/https://medium.com/@liife/friendsreunited-the-sunset-of-an-era-3e5b2ea7bb11

PAUL, K., & Milmo, D. (2022). Elon Musk completes Twitter takeover and 'fires top executives'. *The Guardian*. https://web.archive.org/web/20230206170109/https://www.theguardian.com/technology/2022/oct/27/elon-musk-completes-twitter-takeover

PEHLIVAN, Z., Thièvre, J., & Drugeon, T. (2021). Archiving Social Media: The Case of Twitter. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 43–56). Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5

PENDERGRASS, K. L., Sampson, W., Walsh, T., & Alagna, L. (2019). Toward Environmentally Sustainable Digital Preservation. *The American Archivist*, *82*(1), 165–206. https://doi.org/10.17723/0360-9081-82.1.165

PENNOCK, M., & Beagrie, N. (2013). *Web-Archiving* (DPC Technology Watch Report). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.5280&rep=rep1&type=pdf

PETERSON, A. (2015, September 21). *French regulators tell Google to hide 'right to be forgotten' removals on all sites—The Washington Post*. The Washington Post.

https://web.archive.org/web/20220710231246/https://www.washingtonpost.com/news/the-switch/wp/2015/09/21/french-regulators-tell-google-to-hide-right-to-be-forgotten-removals-on-all-sites/

PHILLIPS, A. A., Walsh, C. R., Grayson, K. A., Penney, C. E., & Husain, F. (2022). Diversifying Representations of Female Scientists on Social Media: A Case Study From the Women Doing Science Instagram. *Social Media + Society*, *8*(3), 20563051221113068. https://doi.org/10.1177/20563051221113068

PIETROBRUNO, S. (2013). YouTube and the social archiving of intangible heritage. *New Media & Society*, *15*(8). https://doi.org/10.1177/1461444812469598

Pla, F., & Hurtado, L.-F. (2017). Language identification of multilingual posts from Twitter: A case study. *Knowledge and Information Systems*, *51*(3), 965–989. https://doi.org/10.1007/s10115-016-0997-x

POLLEN, A. (2013). Research Methodology in Mass Observation Past and Present:'Scientifically, about as valuable as a chimpanzee's tea party at the zoo'? *History Workshop Journal*, *75*(1), 213–235. https://doi.org/10.1093/hwj/dbs040

PUSCHMANN, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

RAAD, E., Al Bouna, B., & Chbeir, R. (2016). *Preventing sensitive relationships disclosure for better social media preservation. 15*(2), 173–194. https://doi.org/10.1007/s10207-015-0278-9

RAYMOND, M. (2010). How Tweet It Is!: Library Acquires Entire Twitter Archive. *Library of Congress Blog*. https://web.archive.org/web/20210517000008/https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/

REGEHR, C., Duff, W., Ho, J., Sato, C., & Aton, H. (2023). Emotional responses in archival work. *Archival Science*. https://doi.org/10.1007/s10502-023-09419-5

RICHARDSON, A. (2020). The Coming Archival Crisis: How Ephemeral Video Disappears Protest Journalism and Threatens Newsreels of Tomorrow. *Digital Journalism*, *8*(10), 1338–1346. https://doi.org/10.1080/21670811.2020.1841568

RINGEL, S., & Davidson, R. (2020). Proactive ephemerality: How journalists use automated and manual tweet deletion to minimize risk and its consequences for social media as a public archive. *New Media & Society*, 1461444820972389. https://doi.org/10.1177/1461444820972389

RITCHIE, J., Lewis, J., Nicholls, C. M., & Ormston, R. (2013). *Qualitative research practice: A guide for social science students and researchers*. SAGE Publications.

ROBINSON, M. (2018). Talking of heritage: The past in conversation. In *The Routledge Handbook of Language and Superdiversity*. Routledge.

ROCKEMBACH, M. (2017). Inequalities in digital memory: Ethical and geographical aspects of web archiving. *The International Review of Information Ethics*, *26*.

ROSEN, J. (2012). The right to be Forgotten. *Stanford Law Review Online*, *64*, 82–88.

ROSSI, G. C., Pyke, T., Pope, J., Skains, R. L., & Wisdom, S. (2022). The New Media Writing Prize Special Collection. *Electronic British Library Journal*, *2022*. https://bl.iro.bl.uk/

ROUSH, T. (2023). *Elon Musk Announces 'Temporary' Reading Limit On Twitter Amid Outage*. Forbes. https://web.archive.org/web/20240525174846/https://www.forbes.com/sites/tylerroush/2023/07/01/elon-musk-announces-temporary-reading-limit-for-unverified-twitter-accounts-amid-outage/?sh=444e6ee075fa

RUEST, N., Fritz, S., & Milligan, I. (2022). Creating order from the mess: Web archive derivative datasets and notebooks. *Archives and Records*, *43*(3), 316–331. https://doi.org/10.1080/23257962.2022.2100336

RUNNACLES, J. (2014, October 8). *The Role of Social Media in the Hong Kong Protests*. Diplomatic Courier. https://web.archive.org/web/20220706045747/https://www.diplomaticourier.com/posts/the-role-of-social-media-in-the-hong-kong-protests

SALAHELDEEN, H. M., & Nelson, M. L. (2012). Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and Practice of Digital Libraries* (pp. 125–137). Springer. https://doi.org/10.1007/978-3-642-33290-6_14

SAUSSURE, F. de. (2006). *Writings in General Linguistics*. Oxford University Press.

SCHAFER, V., Musiani, F., & Borelli, M. (2016). Web archiving, governance and STS. *French Journal of Media Research*, *6*, 1–23.

SCHAFER, V., Thièvre, J., & Banckemane, B. (2020). Exploring special web archives collections related to COVID-19: The case of INA. *WARCnet Papers*.

SCHAFER, V., Truc, G., Badouard, R., Castex, L., & Musiani, F. (2019). Paris and Nice terrorist attacks: Exploring Twitter and web archives. *Media, War & Conflict*, *12*(2), Article 2. https://doi.org/10.1177/1750635219839382

SCHAFER, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-021-00037-0

SCHENSUL, S. L., Schensul, J. J., LeCompte, M. D., & Ph.D, M. D. L., M. A. (1999). *Essential Ethnographic Methods: Observations, Interviews, and Questionnaires*. Rowman Altamira.

SCHNEBLE, C. O., Elger, B. S., & Shaw, D. (2018). The Cambridge Analytica affair and Internet-mediated research. *EMBO Reports*, *19*(8), e46579. https://doi.org/10.15252/embr.201846579

SCHOSTAG, S., & Fønss-Jørgensen, E. (2012). *Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective*. *41*(3–4), 110–120. https://doi.org/10.1515/mir-2012-0018

SCHWARTZ, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival Science*, *2*(1), 1–19. https://doi.org/10.1007/BF02435628

SEPETJAN, S., & Graff, E. (2011). Le dépôt légal en France. *Les Cahiers de Propriété Intellectuelle*, *1*(23), 169–186.

SEYFI, M. (2017). The Relationship Between Autobiographical Memory and Social Media: Sharing Childhood Photographs on Social Media. *Global Media Journal*, *8*(15), 57–70.

SHANE-SIMPSOn, C., Manago, A., Gaggi, N., & Gillespie-Lynch, K. (2018). Why do college students prefer Facebook, Twitter, or Instagram? Site affordances, tensions between privacy and self-expression, and implications for social capital. *Computers in Human Behavior*, *86*, 276–288.

SHERIDAN, D. (1937). Ordinary hardworking folk? *Volunteer Writers in Mass-Observation*, *50*, 1981–1991.

SHIOZAKI, R. (2021). Future uncertainties for preserving tweets: Peoples' perceptions in Japan. *Journal of Librarianship and Information Science*, 09610006211037392. https://doi.org/10.1177/09610006211037392

SIBONA, C., & Walczak, S. (2012). Purposive Sampling on Twitter: A Case Study. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 3510–3519. https://doi.org/10.1109/HICSS.2012.493

SIMON, R. I. (2012). Remembering together. In *Heritage and Social Media:Understanding Heritage in a Participatory Culture* (pp. 89–106). Routledge.

SIMONS, Helen. (2009). *Case study research in practice*. SAGE Publications Ltd. https://doi.org/10.4135/9781446268322

SINN, D., Syn, S. Y., & Syn, S. Y. (2013). Personal Documentation on a social network site: Facebook, a collection of moments from your life? *Archival Science*, *14*(2), 95–124. https://doi.org/10.1007/s10502-013-9208-7

SLOAN, K., Vanderfluit, J., & Douglas, J. (2019). Not 'Just My Problem to Handle': Emerging Themes on Secondary Trauma and Archivists. *Journal of Contemporary Archival Studies*, *6*(20). https://elischolar.library.yale.edu/jcas/vol6/iss1/20

SMITH, L. (2006). Uses of Heritage. In C. Smith (Ed.), *Encyclopedia of Global Archaeology* (pp. 10969–10974). Springer International Publishing. https://doi.org/10.1007/978-3-030-30018-0_1937

SMITH, W. (1997). *The National Library of Australia's Pandora Project*. *47*(3), 169–179. https://doi.org/10.1515/libr.1997.47.3.169

SOCIAL FEED MANAGER. (2018, May 23). *About*. Social Feed Manager. https://web.archive.org/web/20240528092738/https://gwu-libraries.github.io/sfm-ui/about/

STATISTA.COM. (2023a). *Biggest social media platforms 2023 | Statista*. https://web.archive.org/web/20231210153436/https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

STATISTA.COM. (2023b). *Monthly Active Users by Social Media Platform (in millions)*. https://web.archive.org/web/20231210153436/https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

STATISTA.COM. (2024). *Réseaux sociaux les plus utilisés en France 2023*. Statista. https://web.archive.org/web/20240316154409/https://fr.statista.com/statistiques/491792/france-reseaux-sociaux-messageries-instantanees-penetration/

STEEL, B., Parker, S., & Ruths, D. (2023). *The Invasion of Ukraine Viewed through TikTok: A Dataset* (arXiv:2301.08305; Issue arXiv:2301.08305). arXiv. https://doi.org/10.48550/arXiv.2301.08305

STEWART, E. (2019, August 20). *How China used Facebook, Twitter, and YouTube to spread disinformation about the Hong Kong protests*. Vox. https://web.archive.org/web/20210304180659/https://www.vox.com/recode/2019/8/20/20813660/china-facebook-twitter-hong-kong-protests-social-media

STIRLING, P., Illien, G., Sanz, P., & Sepetjan, S. (2012). The state of e-legal deposit in France: Looking back at five years of putting new legislation into practice and envisioning the future. *IFLA Journal*, *38*(1), 5–24. https://doi.org/10.1177/0340035211435323

STORRAR, T. (2014, May 8). Archiving social media [Text]. *The National Archives Blog*. https://blog.nationalarchives.gov.uk/archiving-social-media/

STORTI, C. (2023). 'Resource not found': Cultural institutions, interinstitutional cooperation and collaborative projects for web heritage preservation. *JLIS.It*, *14*(2), Article 2. https://doi.org/10.36253/jlis.it-533

TAGG, C., Hu, R., Lyons, A., & Simpson, J. (2016). Heritage and social media in superdiverse cities: Personalised, networked and multimodal. *Working Papers in Translanguaging and Translation*, 35.

TERRAS, M. (2015). Crowdsourcing in the digital humanities. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 420–438). Wiley Online Library.

THE NATIONAL ARCHIVES. (2018, June). *Preservation Policy*. https://www.nationalarchives.gov.uk/documents/preservation-policy-june-2018.pdf

THOMSON, S. D. (2016). *Preserving Social Media* (16–01; DPC Technology Watch Report). https://www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file

THOMSON, S. D., & Kilbride, W. (2015). Preserving social media: The problem of access. *New Review of Information Networking*, *20*(1–2), 261–275. https://doi.org/10.1080/13614576.2015.1114842

TREEM, J. W., Dailey, S. L., Pierce, C. S., & Biffl, D. (2016). What We Are Talking About When We Talk About Social Media: A Framework for Study. *Sociology Compass*, *10*(9), 768–784. https://doi.org/10.1111/soc4.12404

TSUDA, Y. (1998). Critical studies on the dominance of English and the implications for international communication. *Nichibunken Japan Review*, 219–236.

TWITTER. (2017, November 3). *Developer Policy – Twitter Developers*. https://web.archive.org/web/20200105002041/https://developer.twitter.com/en/developer-terms/policy

TWITTER. (2020, March 10). *Developer Agreement and Policy – Twitter Developers*. https://web.archive.org/web/20200418215448/https://developer.twitter.com/en/developer-terms/agreement-and-policy

TWITTER. (2021a, January 26). *Introducing the new Academic Research product track—Announcements*. Twitter Developers. https://twittercommunity.com/t/introducing-the-new-academic-research-product-track/148632

TWITTER. (2021b, May). *Twitter Terms of Service*. https://web.archive.org/web/20210531152032/https://twitter.com/en/tos

TWITTER. (2022, September 4). *Twitter Terms of Service*. https://web.archive.org/web/20220904022235/https://twitter.com/en/tos

TWITTER. (2023a, August 18). *Timelines standard v1.1 to v2 migration guide | Docs | Twitter Developer Platform*. https://web.archive.org/web/20230818152238/https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/migrate/standard-to-twitter-api-v2

TWITTER. (2023b, December 7). *Display requirements – Twitter Developers | Twitter Developer Platform*.

https://web.archive.org/web/20231207104654/https://developer.twitter.com/en/developer-terms/display-requirements

TWITTER. (2024). *Twitter API Documentation—X APIv2*. X Developer Platform. https://web.archive.org/web/20240613081752/https://developer.x.com/en/docs/twitter-api

TWITTER. (n.d.). *The Twitter rules: Safety, privacy, authenticity, and more*. https://web.archive.org/web/20210616151805/https://help.twitter.com/en/rules-and-policies/twitter-rules

TWITTER DEV, [@TwitterDev]. (2023a, February 2). *Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead*  *[Tweet]*. Twitter. https://web.archive.org/web/20230224053552/https://twitter.com/twitterdev/status/1621026986784337922

TWITTER DEV, [@TwitterDev]. (2023b, March 29). *For Academia, we are looking at new ways to continue serving this community. In the meantime Free, Basic and Enterprise tiers are available for academics. Stay tuned to @TwitterDev to learn more. [Tweet]*. https://web.archive.org/web/20230606074402/https://twitter.com/TwitterDev/status/1641222788911624192

TYBIN, V. (2019). Les Gilets jaunes sous l'œil du dépôt légal numérique [Billet]. *Web Corpora*. https://web.archive.org/web/20230328051802/https://webcorpora.hypotheses.org/750

UK DATA FORUM. (2013). *UK Strategy for Data Resources for Social and Economic Research*. https://esrc.ukri.org/files/research/uk-strategy-for-data-resources-for-social-and-economic-research/

UKGWA Team. (2020). *Capturing the UK Government Response to the Coronavirus (COVID-19) Pandemic at The National Archives UK - Digital Preservation Coalition*. Digital Preservation Coalition - Blog. https://web.archive.org/web/20240111114050/https://www.dpconline.org/blog/series-wa-coronavirus-newing-3

UNESCO. (2003a). *Charter on the Preservation of Digital Heritage*. UNESCO. https://en.unesco.org/about-us/legal-affairs/charter-preservation-digital-heritage

UNESCO. (2003b). *Convention for the Safeguarding of the Intangible Cultural Heritage—UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000132540

UNESCO. (2003c). *Guidelines for the preservation of digital heritage*. https://unesdoc.unesco.org/ark:/48223/pf0000130071

UNITED NATIONS. (2020). *The Climate Crisis – A Race We Can Win*. United Nations; United Nations. https://web.archive.org/web/20240707001731/https://www.un.org/en/un75/climate-crisis-race-we-can-win

VAN CAMP, A. (2020). How to assess and improve the coverage of the Legal Deposit collection in Belgium? *In Monte Artium*, *13*, 71–103. https://doi.org/10.1484/J.IMA.5.122155

VAN DER VLIST, F. N., Helmond, A., Burkhardt, M., & Seitz, T. (2022). API Governance: The Case of Facebook's Evolution. *Social Media + Society*, *8*(2), 20563051221086228. https://doi.org/10.1177/20563051221086228

VAN DIJCK, J. (2011). Flickr and the culture of connectivity: Sharing views, experiences, memories. *Memory Studies*, *4*(4), 401–415. https://doi.org/10.1177/1750698010385215

VAN DIJCK, J. (2015). After Connectivity: The Era of Connectication. *Social Media + Society*, *1*(1), 2056305115578873. https://doi.org/10.1177/2056305115578873

VAVASSORI, V., Goudarouli, E., & Winters, J. (2023). *Archives and the environment report*. OSF. https://doi.org/10.31219/osf.io/pkbg5

VELTE, A. (2018). Ethical Challenges and Current Practices in Activist Social Media Archives. *The American Archivist*, *81*, 112–134. https://doi.org/10.17723/0360-9081-81.1.112

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. https://doi.org/10.1007/s42803-019-00007-7

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*. https://doi.org/10.1007/s42803-021-00036-1

WEBBER, J. (2020, March). *15 Years of the UK Web Archive—The Early Years*. https://web.archive.org/web/20231130101934/https://blogs.bl.uk/webarchive/2020/03/15-years-of-the-uk-web-archive.html

WEBSTER, P. (2015, March 20). How fast does the web change and decay? Some evidence. *Web Archives for Historians*. https://webarchivehistorians.org/2015/03/20/how-fast-does-the-web-change-and-decay-some-evidence/

WEBSTER, P. (2017). Chapter Eleven: Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger (Ed.), *Web 25. Histories from 25 Years of the World Wide Web* (pp. 179–190). Peter Lang. https://www.peterlang.com/view/9781433140655/chapter-012.xhtml

WELLER, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society: An Introduction* (Vol. 89). Peter Lang Publishing.

WERF, T. van der, & Werf, B. van der. (2014). *The paradox of selection in the digital age*. IFLA WLIC 2014, Lyon, France. https://library.ifla.org/id/eprint/1042/

WILLIAMSON, E., Gregory, A., Abrahams, H., Aghtaie, N., Walker, S.-J., & Hester, M. (2020). Secondary Trauma: Emotional Safety in Sensitive Research. *Journal of Academic Ethics*, *18*(1), 55–70. https://doi.org/10.1007/s10805-019-09348-y

WINDON, K., & Youngblood, J. (2024). Privacy Considerations in Archival Practice and Research. In *Human Privacy in Virtual and Physical Worlds: Multidisciplinary Perspectives* (pp. 205–234). Springer Nature Switzerland Cham.

WINTERS, J. (2017). *Will history survive the digital age?* HistoryExtra. https://web.archive.org/web/20210417025255/https://www.historyextra.com/period/20th-century/will-history-survive-the-digital-age/

WINTERS, J. (2020). Giving with one click, taking with the other: E-legal deposit, web archives and researcher access. In M. Terras & P. Gooding (Eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future*. Facet Publishing.

WINTERS, J., & Prescott, A. (2019). Negotiating the born-digital: A problem of search. *Archives and Manuscripts*, *47*(3), 391–403. https://doi.org/10.1080/01576895.2019.1640753

WOLK, R. M. (2004). The effects of English language dominance of the Internet and the digital divide. *2004 International Symposium on Technology and Society (IEEE Cat. No.04CH37548)*, 174–178. https://doi.org/10.1109/ISTAS.2004.1314348

WONG, A., Ho, S., Olusanya, O., Antonini, M. V., & Lyness, D. (2020). The use of social media and online communications in times of pandemic COVID-19. *Journal of the Intensive Care Society*, 1751143720966280. https://doi.org/10.1177/1751143720966280

WORDPRESS. (2024, September 5). Introducing the 100-Year Plan: Secure Your Online Legacy for a Century – WordPress.com News. *Wordpress.Com*. https://web.archive.org/web/20240905200456/https://wordpress.com/blog/2023/08/25/introducing-the-100-year-plan/

X DEVELOPER PLATFORM. (2023, December 7). *Search API: Enterprise*. https://web.archive.org/web/20231207233407/https://developer.twitter.com/en/docs/twitter-api/enterprise/search-api/overview

YAKEL, E. (2003). Archival representation. *Archival Science*, *3*, 1–25. https://doi.org/10.1007/BF02438926

YIN, R. K. (1994). Discovering the future of the case study. Method in evaluation research. *Evaluation Practice*, *15*(3), 283–290.

ZAPPAVIGNA, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, *25*(3), 274–291. https://doi.org/10.1080/10350330.2014.996948

ZHAO, X., & Lindley, S. (2014, May 26). Curation through Use: Understanding the personal value of Social Media. *Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems (CHI 2014)*. https://www.microsoft.com/en-us/research/publication/curation-through-use-understanding-the-personal-value-of-social-media/

ZHAO, X., Salehi, N., Naranjit, S., Alwaalan, S., Voida, S., & Cosley, D. (2013). The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 1–10). Association for Computing Machinery. https://doi.org/10.1145/2470654.2470656

ZIERAU, E. (2022). URN Namespace Registration for Persistent Web IDentifiers (PWID). *Internet Assigned Numbers Authority (IANA)*, *2022-11–18*. https://www.iana.org/assignments/urn-formal/pwid

ZIMMER, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*, *20*(7). https://web.archive.org/web/20201201055921/http://firstmonday.org/article/view/5619/4653

ZOOM INC. (n.d.). *Security | Zoom*. Retrieved 24 March 2022, from https://web.archive.org/web/20220324180945/https://explore.zoom.us/en/trust/security/