Making social media archives: Limitations and archiving practices in the development of representative social media collections

Beatrice Cannelli

Abstract: Social media has become an important digital space where individuals can participate in ongoing global discussions and document instances of historical events. Social media offers marginalized communities a means to express their identities, voice their concerns, and tell their stories. Archiving institutions have started to include social media in their collections because of its enduring value. However, constraints set by legal and technical frameworks and limited resources available at single institutions can influence the overall representativeness of content archived on social sites. This chapter explores the impact these constraints have on the development of representative social media collections and illustrate participatory approaches that can help to mitigate concerns.

Keywords: social media archiving, representativeness of collections, participatory archive.

Social media has become an important digital space where individuals can participate in ongoing global discussions, offering at the same time a platform for sharing and documenting instances of historical events, as health and political crises in the early 2020s have demonstrated (Simon 2012; van Dijck 2011). Moreover, social media provides an opportunity for marginalized communities to express their identities, voice their concerns, and tell their stories (Bergis et al. 2018). The cultural value and historical relevance of social media content has been widely recognized (Henninger and Scifleet 2016; Pietrobruno 2013), leading cultural heritage institutions worldwide to include the material generated on these sites in their preservation strategies in order to ensure its safeguard and accessibility in the long term (Bingham and Byrne 2021; Fondren and Menard McCune 2018; Schafer and Winters 2021; Storrar 2014). Social media archiving initiatives have the ability to preserve fragments of our (online) present, passing down to future generations of researchers key information to understand the 21st century. For this reason, it is essential that the plurality of voices emerged on social platforms is adequately reflected in the resulting archive collections. However, developing social media archives has proved to be a difficult endeavor under many points of view. Although social media archiving inherits some of the challenges identified over more than 25 years of web archiving activities, web curators have been dealing with a series of new technical, ethical, and legal issues that are specific to social media sites and have been limiting the scale of collection, thus potentially influencing the granularity and representativeness of archived

Beatrice Cannelli, University College London, United Kingdom, beatrice.cannelli@postgrad.sas.ac.uk, 0000-0002-8645-9503

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Beatrice Cannelli, *Making social media archives: Limitations and archiving practices in the development of representative social media collections*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.08, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 57-75, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

social media collections (Thomson 2016). Also, appraising and selecting content out of the sheer amount of information generated daily on social platforms requires time and resources that web archiving teams often do not possess.

Scholarly literature has discussed concerns of comprehensiveness and representation in web collections focusing on archiving strategies, the influence of socio-technical infrastructures, and cultural perspectives (Bingham and Byrne 2021; Brügger 2018; Hegarty 2022; Maemura 2023), calling for a critical approach to web archives (Ben-David 2021). However, few have addressed the questions surrounding the representativeness of social media archived material (Chambers et al. 2021; Schafer et al. 2019). The unique dynamics that regulate social platforms, the ephemerality of content, and the distinctive curatorial challenges that archiving this born-digital material pose to collecting institutions, call for further examination of the limitations and practices related to the development of representative social media collections at a national level. Drawing from interviews[1] and fieldwork conducted as part of wider, ongoing PhD research investigating the challenges and opportunities related to the development of social media archives, this chapter explores the factors that may impact the degree of representativeness of social media collections and the actions taken by existing social media archiving initiatives to mitigate these concerns.

In the first section, I will delineate the context in which social media archives are embedded, drawing attention to platforms' representation, geopolitical dynamics, and archival narrative disparities. I will consider the popularity of certain platforms and how this is not always mirrored in the collections developed by existing web and social media archiving initiatives in the Global North. In the second section, I will discuss how the need to preserve this important historical resource often collides with the numerous constraints imposed by national legal frameworks, social media policies, technical challenges, and inadequate resources necessary to guarantee the long-term sustainability of archiving efforts at an institutional level, setting the stage for representation concerns, biases, and narrative gaps in national social media collections. In the final section of this chapter, I will identify some of the steps taken by existing web and social media archiving initiatives to mitigate representativeness and inclusivity concerns. I will conclude by arguing that the use of crowdsourcing strategies and participatory approaches are examples of good practices that can not only sustain the development of more comprehensive collections, but also offer a

[1] Semi-structured interviews were conducted between April and September 2022 with twelve web archivists at national archiving institutions currently archiving or planning to archive social media. Insights and examples that emerged during the interviews are referenced in the footnotes.

unique opportunity to raise awareness of the existence of social media archives and their cultural significance across diverse layers of society.

1. Questions about inequalities and representation in the social media archiving landscape

A rich body of post-modernist archival literature has discussed the meaning of representation in the archives specifically with regard to selection practices, highlighting the potential repercussions that archival choices may have on the history told through the cultural heritage thus preserved (Caswell et al. 2017; Yakel 2003). In particular, concerns have been raised about biases existing in mainstream narratives and how collecting practices have frequently led to documenting one side of history, silencing groups of people placed at the margin of society because of structural power dynamics (Harris 2002; Jimerson 2006; Schwartz and Cook 2002).

With the advent of social media platforms, plus a diffused democratization of mobile devices and access to the internet, many of those marginalized voices have found multiple virtual spaces where they could make themselves heard. As Barrowcliffe (2021) noted, social media offered minority groups a means to convey counter-narratives, documenting, from their standpoint, critical events related to their own history, as these unfold on both a national and a global scale. For this reason, archiving social media represents an unprecedented opportunity for memory institutions to preserve historical traces of the present that have the potential to portray the multi-leveled landscape of voices prompting or joining conversations on these sites. A kaleidoscope of stories coming from communities that have often been underrepresented or misrepresented in mainstream media and repositories.

However, while social media has been amplifying certain protests, movements, and events contributing to the online unfolding of viral phenomena such as those expressed through the hashtags #BlackLivesMatter or #MeToo, it still mirrors societal and geographical inequalities existing offline, if not exacerbating some of those differences (Jackson 2020; Lutz 2022). The roots of these inequalities, as Lutz (2022) explained, are to be found in the different layers of social media divide, which involves among other factors the uneven distribution of access to not only mobile internet, for example, but also to the plethora of existing social platforms that may differ from country to country, with subsequent repercussions on the empowerment of certain marginalized groups rather than others (Lutz 2022).

The social media divide stemming from geopolitical dynamics, as well as other aspects that will be discussed below, appears to have heavily influenced the geographical distribution and development of social media

archiving initiatives. While numerous web archiving initiatives have emerged at various national memory institutions, consolidating over the past twenty years techniques and collection strategies to safeguard national Top-Level Domains (TLDs), the preservation of social media is still finding its pace and space, with only a few countries consistently archiving this born-digital material. As I reported in a blogpost recently published on the International Internet Preservation Coalition (IIPC) blog, the preservation of social media material appears to be fundamentally located in Global North countries, especially in North America, Europe, and Oceania (Cannelli 2022). This uneven distribution has raised questions about the potential gaps in the overall preservation of the collective memory generated on social platforms (Cannelli 2022). The reasons behind these discrepancies include geopolitical factors and challenges that are still unresolved, which makes this material particularly difficult to collect and provide successful access to (Bruns and Weller 2016; Pehlivan et al. 2021; Thomson 2016). As emerged from the aforementioned preliminary study, imbalances are also to be found in the type of social media that are currently being preserved by Global North cultural heritage institutions. Among the most archived platforms to date there is Twitter (officially rebranded as X in April 2023), followed by Facebook and Instagram; conversely, sites such as YouTube and TikTok, which has become very popular in the past couple of years, only appear at the very bottom of the list (Cannelli 2022). However, if these data are compared to the list of social media sites counting the highest number of users in the past couple of years, some discrepancies emerge between the platforms that are being archived and the ones that users across different countries engage in. In fact, the high number of users active on Meta Inc. platforms confirms the interest of most archiving initiatives in taking steps to preserve these sites (Statista.com 2023). Contrastingly, YouTube, which counted over 2.5 billion monthly active users as of October 2023, seems to be infrequently included in social media collections for various reasons. On the opposite side of the spectrum sits Twitter, which is largely archived in North American and European institutions, despite the number of users active on this platform being considerably lower than on other, more popular social sites (Statista.com 2023). These trends generate concerns regarding the representativeness of the social media cultural heritage that will be passed down to future generations, also highlighting the need to create positive conditions that could ease the barriers that prevent the development of social media archiving initiatives, especially in countries of the Global South (Colin-Arce et al. 2023).

Moreover, the combination of these imbalances consequently raises questions regarding the actual granularity of social media content collected on a national scale. In the following section, I will offer an overview of the factors that may affect the type of social media platforms archived as well

as the level of representativeness of national social media collections, examining restrictions set by legal frameworks, technical aspects, and available resources.

2. Factors influencing representativeness of social media collections

Web and social media archives play a central role in shaping the image that future generations will be able to remember and study about present events. A complex set of elements intervening in the development of national social media collections should be taken into consideration, as these have a profound impact on the overall structure, gaps, and narratives preserved.

The making of traditional archives involves a series of selection and appraisal procedures that can only be applied to a certain extent to social media, due to its unique characteristics. Reflecting on web archiving practices, Masanès et al. (2021) observed how "archiving the 'whole' Web is not attainable, due to resources and time limitations, as well as its de facto infinite generative nature." Preserving social media appears to some extent even more challenging than traditional websites, owing to its highly ephemeral, dynamic nature and the sheer volume of content generated each second on an ever-growing number of platforms. For this reason, instead of striving to achieve an impossible and unnecessary level of comprehensiveness, archiving institutions aim at providing the best representation possible of events and discussions on social sites (Masanès et al. 2021).

One aspect to consider when it comes to archived social media content is that many national archiving institutions rarely distinguish between websites and social media, creating collections that indiscriminately include both types of artifacts. While this is relatively justified by the fact that social media is indeed part of the web, it is undeniable that the latter has evolved into a separate phenomenon. Unlike websites that are collected through multiple approaches combining broad-scope, annual, and several selective crawls, only a rather small selection of social media accounts or hashtags is included in existing web collections, organized around specific themes or events, and often captured in the context of emergency collection campaigns to document unexpected crisis (Schafer et al. 2019). Although archival practices and collection development policies may vary between institutions, there are several factors that affect almost all social media archiving initiatives and may have a profound effect on the granularity of collections.

2.1 Legal constraints

National legal frameworks, including digital legal deposit legislation and policies established by social media companies to regulate the use and reuse of content shared on their own platforms, are among the legal constraints that may impact the degree of representativeness of social media collections.

National legal frameworks have a significant influence on the content preserved as part of social media collections at national memory institutions, especially for those operating under digital legal deposit legislation. In an overview of existing non-print legal deposit legislation offered in a report compiled by researchers involved in the BESOCIAL project (Chambers et al. 2021), it emerged how the minimum common denominator of most of these regulations is that they define born-digital content as that which is related to or published within national borders. On the one hand, while this criterion coincides with the sovereignty that a government possesses over affairs within a territorial or geographical area, on the other hand it fulfills the need to preserve the digital history and cultural heritage of specific countries. This parameter implicates, however, geographical boundaries that tend to blur in the context of the World Wide Web and particularly social media. Because of the international interconnectedness that characterizes the web and even more so social media interactions, it is extremely difficult for web archivists to disentangle billions of threads of discussions and ascertain with absolute certainty content provenance on social sites. In a recent article discussing the archival strategies implemented in the development of the UK Web Archive, Bingham and Byrne (2021) reported the uncertainty surrounding the process of identifying content on social media that originates on national soil. As they explained, establishing the boundaries of the national web domain for websites is facilitated to a certain extent by the selection of sites bearing domain extensions assigned to the national TLD. Conversely, social media platforms are mostly hosted on .com domains and thus located outside the country in scope (Bingham and Byrne 2021). Moreover, assessing provenance of content shared on social platforms can be laborious and not always reliable. For example, relying on geolocalization data available on these sites has proved to be a challenging task as geographical information can be subject to high error rates and inaccurate (Graham et al. 2014). For this reason, archiving institutions have mostly resolved to hand-picking accounts of organizations or public figures for which provenance or pertinence to the country in scope can be determined with confidence. Similar archiving approaches, however, contribute to the formation of inevitable gaps in the collections, which, in some cases, cannot be filled due

to the low persistence, high ephemerality, and constant evolution of social media content (Richardson 2021; Ringel and Davidson 2020).

Digital legal deposit provisions and data protection legislation place another layer of restrictions on selection criteria. In order to safeguard individuals' privacy, the collection of born-digital material under the governing law is usually limited to content that is made publicly available on social sites. Such limitation, however, often leads to the exclusion of content that might be in scope but accessible only upon authentication. Particularly affected by this is the capture of platforms like Facebook, where users tend to share information among a selected group of 'friends' or among closed groups of people (Sinn et al. 2013). Especially in the latter case, the constraints imposed by existing regulations, although necessary to protect the users' privacy, can lead to the formation of important gaps in the cultural heritage preserved. For example, displaced or marginalized groups appear inclined to share information and communicate with members of their own community within private Facebook groups (Goldsmith et al. 2022; Good 2012). In order to capture this content, archiving institutions would be required to log into the platform or be invited to join said groups, which might not be authorized by national digital legal deposit legislation. This clearly constitutes a problem in terms of representativeness of collections as many of these communities are often only present on social media and have no website that could be archived instead (Ferré-Pavia et al. 2018). Besides, as observed by web archivists at the Luxembourg Web Archive[2], there are some additional dilemmas that come into play when trying to preserve social media, such as problems with online discoverability of a variety of small realities spread across the national territory, or concerns emerging from public figures' accounts that share personal information alongside public communication (Schafer and Els 2020).

In this already complicated panorama, social media policies add another layer of legal constraints that heavily affect the granularity of information collected on their platforms, particularly concerning access to data. Social media companies impose strict limitations, for example, on the quantity and frequency with which information can be captured within a set time frame. That, coupled with other technical challenges, may explain in part why many institutions archive certain platforms (e.g. X-Twitter) more than others. As mentioned in the previous section, Facebook appears among the most archived social platforms. However, when looking closely at the amount of Facebook materials included in existing collections, it becomes evident how some institutions only archive a limited number of relevant

[2] Ben Els (National Library of Luxembourg), interviewed via Zoom by Beatrice Cannelli, 12 April 2022.

profiles on this platform. Due to the numerous restrictions set by Meta Inc. on harvesting, many institutions have seen their accounts periodically blocked when exceeding the set rate limit. Web archivists' reduced ability to regularly capture information without having to worry about accounts being suspended or restricted, consistently affects Facebook's preservation, and specifically impacts all those organizations, communities, and public figures active only on this site.

Furthermore, platform acquisitions from third parties can lead to changes in social media policies, making sites more difficult if not impossible to archive. For example, problems surrounding the capture of platforms like LinkedIn can be connected to the implementation of stricter rules for web crawling following Microsoft's acquisition in 2016. The LinkedIn User Agreement explicitly states that users are forbidden to scrape data from the site using third party software[3]. Although it is not among the most archived social platforms, LinkedIn is still relevant to some institutions such as the UK National Archives that are left unable to capture potentially relevant content for their UK Government Web Archive because of the restrictions in place[4]. Similarly, the most recent acquisition that has had a major impact on existing social media archives and whose repercussions are yet to be fully assessed, especially for institutions collecting through the platform's official application programming interfaces (APIs), is the one concerning X-Twitter. The takeover in October 2022 of Twitter by Tesla Inc. CEO, Elon Musk (Clayton & Hoskins 2022), led to a series of changes that culminated—from a social media archiving perspective—in the upheaval of the Twitter API access system known until that point. The new leadership decided to end free access to its APIs, including the much-praised Academic Twitter API, in favor of a paid tiers system that, to date, does not include any access specifically designed for research or preservation purposes. According to information made available on the X Developer Platform[5], the only tier available that provides free access to data via the Twitter API v2 comes with several limitations in terms of the total amount of data that can be retrieved per month; whereas the tier that offers the highest level of access, including a full-archive search, requires the payment of a fee that many archiving institutions with already limited budgets will not be able to sustain both in the short- and long-term.

---

[3] LinkedIn, Prohibited software and extensions: https://web.archive.org/web/20231116062416/https://www.linkedin.com/help/linkedin/answer/a1341387

[4] Claire Newing (The UK National Archives), interviewed via Zoom by Beatrice Cannelli, 30 June 2022.

[5] Further information about access to the Twitter API v2 can be found at the following link: https://web.archive.org/web/20231208180144/https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api#v2-access-leve

Representativeness of social media collections can also be influenced by technical challenges encountered while using different archiving techniques. As there are no standard approaches to collecting and preserving social media material, archiving initiatives use different methods to capture information on social sites. The decision of which method to adopt is often based on requirements (e.g. preservation in the context of legal deposit), resources, and expertise available at single institutions. These methods include the use of Application Programming Interfaces (APIs) to access data on social media sites and traditional web crawlers (e.g. Heritrix).

As clarified at the beginning of this chapter, most archiving institutions do not aim for an exhaustive archiving of content on social media, but rather a representative snapshot of the discussions and digital cultural heritage generated on these sites. However, the restrictions applied by social media platforms through policies and terms of use pose several technical challenges to the development of representative collections. Pehlivan et al. (2021) provided a comprehensive overview of the archival challenges related to data collection via APIs, discussing the restriction rules imposed specifically by Twitter. Among the different types of Twitter APIs described, the authors pointed out how the Sample API allowed institutions to collect 1% of all public tweets selected randomly in real time, which did not include historical tweets as it was not Twitter's intention for the Search API to focus on exhaustiveness but rather on relevance to the chosen keywords (Pehlivan et al. 2021). Although archiving institutions might not aim for completeness when it comes to social media collections, it is important for them to understand how sampling mechanisms work and how representative those random samples are of the whole data available, so that this can be accurately documented. Numerous studies have shed light on the biases existing in sampling mechanisms (González-Bailón et al. 2014; Tromble et al. 2017; Wu et al. 2020), but only a handful of them have taken into consideration potential repercussions on institutional archiving and long-term preservation (Acker and Kreisberg 2020; Littman et al. 2018; Pehlivan et al. 2021). The opaqueness of criteria concerning sampling and the changes applied to algorithms that can occur at any time without making API users aware, can influence the representativeness of content collected using this method (Hino and Fahey 2019). The main risk lies in portraying and preserving potentially unbalanced perspectives on historical events and culture in the long term. For example, this can be particularly problematic in the case of controversial topics or election campaigns where the amount of social media content has increased significantly in the past decade and for which it is essential to preserve the different opinions of all the parties involved.

Additional challenges to shaping representative social media collections arise from the use of archiving tools, such as web crawlers, which were often initially developed to capture traditional websites. The majority of institutions archiving the national web domain at scale, including the UK Web Archive (UKWA) and the National Library of France (BnF), use the Internet Archive's Heritrix crawler (Aubry 2010; Bingham and Byrne 2021). Despite a few technical issues with more dynamic sites using JavaScript, web crawlers like Heritrix have become a widely accepted method to successfully preserve a comprehensive snapshot of national TLDs (Brügger 2018). However, archiving social media platforms using these tools still poses many challenges. Most web crawlers struggle to correctly interact with the complex layout of social platforms and thus capture the highly dynamic content shared on social media, with institutions reporting gaps in the materials collected. As remarked by curators at the BnF[6], the inability of web crawlers to interact with elements on the page, such as buttons to expand hidden sections or scroll down pages to prompt the loading of more posts in the feed, can lead to the loss of relevant information from public profiles selected for their enduring cultural value that tend to share numerous posts daily. When harvesting social media content for the special collection dedicated to the COVID-19 pandemic, the BnF registered, overall, a higher success rate on Twitter compared to other platforms (Gebeil and Schafer 2020). According to the BnF, Facebook crawling scored instead a particularly low success rate that required additional efforts to obtain adequate captures, so much so that the BnF has decided to temporarily pause collection activities on this site until new, more sustainable archiving solutions are found (Gebeil and Schafer 2020). Indeed, the reduced quality and persistently unsuccessful outcomes obtained when archiving social platforms can ultimately result in institutions deciding to focus their time and resources on other, easier-to-archive sites. This means, however, that important evidence about contemporary events and culture will be lost, potentially increasing already existing biases and gaps.

In terms of technical challenges surrounding the capture of popular social media platforms, the Internet Archive's Archive-it Help Center page offers an interesting summary of known archiving issues using Heritrix. Among these, it is worth noting how Facebook and Instagram, both owned by Meta, are identified as platforms that hinder the capturing of many organizational profile pages and some Facebook Groups pages. As a result, this leads to the exclusion from the collection of countless relevant accounts, including small groups that only exist on these platforms. Besides, after Elon Musk

[6] Vladimir Tybin (National Library of France), interviewed by Beatrice Cannelli, Paris, 19–20 April 2022.

acquired Twitter in 2022 and implemented various changes from both technical and rebranding perspectives, the Archive-it Team updated the Twitter archiving status. They initially reported issues with harvesting some Twitter seeds[7] in March 2023, and then advised Heritrix users to pause archiving activities on the site as "recent changes to visibility of content on Twitter present multiple archiving challenges."[8] The combined legal, curatorial, and technical challenges generated by the complexity of social media platforms require archiving institutions to constantly adjust and find bespoke solutions to the latest variations in the field. In addition, institutions have to gauge the scale of preservation activities often based on the limited funds.

2.3 Resources and sustainability

Resources available play an important role in the development of representative collections of social media material and its long-term sustainability. Social media platforms and technologies are always shifting, requiring a substantial amount of resources to support the improvement and implementation of strategies and technologies that can successfully capture these platforms (Bingham and Byrne 2021). It is important to consider that many institutions preserving social media content are operating under legal deposit mandates that require them to archive, preserve, and provide access to this material. However, public archiving institutions are often developing social media collections with financial resources that do not always commeasure with the scale of the endeavor.

Apart from some exceptions, national web and social media archives are the result of small teams' efforts, which sometimes comprise only a few curators tasked with sifting through the sheer amount of information published on social sites, and carefully selecting profiles or hashtags that fit the scope of the collections. Moreover, a considerable share of resources necessarily flows into the technical side of social media archiving: developing ad hoc tools or implementing existing ones requires engineers or highly specialized technicians that institutions with limited budgets fail to attract as they struggle to offer competitive salaries. The alternative is either to outsource collecting activities to third parties or subscribe to external web archiving services (e.g. Archive-it) that offer a set of tools, training, and technical support for preserving and providing access to the archived data,

[7] Archive-it Help Center "Social media and other platforms status", 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-

[8] Archive-it Help Center "Social media and other platforms status", archived on 02/01/2024 https://web.archive.org/web/20240102174300/https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

both of which can still be expensive. Nevertheless, because of the many challenges and constraints that social media archiving entails, selecting and collecting social sites is still largely a manual process, which requires time, curators, and specially dedicated resources. Moreover, curators occupied with handpicking content from social platforms are often also simultaneously working on selecting websites for ongoing web archive collections, further stretching the capacity of curators to singlehandedly ensure a well-balanced, broad-spectrum representation of the various strata of society.

The long-term sustainability of (representative) social media archiving collections is an open issue. While recent studies, such as the one conducted by the BESOCIAL project[9] at the Royal Library of Belgium (KBR), have shed light on opportunities for the development of sustainable social media archiving strategies (Messens et al. 2021), questions persist on how to sustainably tackle inclusivity and diversity concerns in the context of social media collections.

3. Mitigating representativeness concerns through participation practices

To mitigate inherent institutional biases and concerns about representativeness of social media collections, many archiving institutions have experimented and consolidated specific participatory archiving strategies designed to make the most of the limited budgets, staff, and time available (Pendergrass et al. 2019). Web and social media archiving initiatives have been successfully using participatory collection practices to record specific events or topics, seeking the contribution of the public or researchers that could bring their own unique perspective to the archive.

Web archives have increasingly adopted participation to expand the catchment area of the web-based material to be archived as part of national collections and help address known problems of representativeness (Cui et al. 2023; Schafer and Winters 2021). When it comes to participatory approaches, web and social media archives tend to turn to forms of collaborative curation and crowdsourcing. In this context, it is important to reiterate that, as social media are frequently included in existing web archive collections, distinctions between social sites and traditional websites are often minimal in the development of participatory practices or campaigns.

Popular types of crowdsourcing practices related to the appraisal and selection of valuable web materials include open calls for suggesting content to be incorporated in specific collections. Some web archives have

[9] Further information about the BESOCIAL project can be found here: https://web.archive.org/web/20240214181009/https://www.kbr.be/en/projects/besocial/

dedicated pages on their portals where individuals can fill in a form providing information and the URL of websites they would like to nominate for preservation, such as the "Save a UK website" feature available on the UK Web Archive portal. However, most of these forms appear to be structured and formulated in favor of submitting website URLs rather than social media content, likely due to legal and ethical concerns surrounding the latter. Nevertheless, recent campaigns promoted between 2020 and 2021 in light of the COVID-19 outbreak have encouraged members of the public to nominate meaningful hashtags and social media content alongside traditional websites. Moreover, curators at the Luxembourg Web Archive[10] have noted that suggestions received through the campaign they launched at the beginning of the first lockdown helped them uncover small religious groups that were particularly active in disseminating official information among their members. These groups had not been included in their national web collection before (Schafer and Els 2020; Schafer and Winters 2021). Besides, institutions undertaking pilot or short-term projects to test the feasibility and sustainability of social media archiving have found in these practices a means to discover new themes and areas of interest to integrate the initial 'top down' approach. The BESOCIAL project at KBR[11], for example, decided to test different approaches including a crowdsourcing campaign to ask the public to nominate social media material (including text-based material, hashtags, and public accounts) that should be preserved as part of the online national heritage collection they were developing. The BESOCIAL team not only received hundreds of responses helping them fill in the gaps and mitigate representativeness concerns that emerged from the initial archiving approach, but also observed how the campaign supported the promotion of social media archiving activities in Belgium. Certainly, the effectiveness of such campaigns in terms of increasing representativeness of social media collections is linked to how they are promoted and among which communities. A meticulous dissemination strategy among specific target groups is indeed essential for obtaining contributions that can truly enrich the content already being preserved. While public involvement in these campaigns may vary, each individual URL can still be crucial for uncovering underrepresented themes or marginalized communities.

At some institutions, the selection of born-digital content to be added to the archive is the result of the combined effort of web and social media curators as well as a network of contributors identified both within and outside the cultural heritage institution. This is exemplified by the system established at the National Library of France (BnF) 12, where curators of

[10] Ben Els, interview.
[11] Fien Messens and Friedel Geeraert (Royal Library of Belgium), interviewed by Beatrice Cannelli, 13 July 2022.
[12] BnF, *Cooperer autour de l'archivage du Web*:

the digital legal deposit team, contributors from other BnF departments, and associated regional centers (e.g. regional archives, libraries, and research institutes) have been collaborating to support the capture of a diverse representation of the French territory and society. To facilitate the process and management of web and social media materials to be collected, the BnF has also developed an application called "BnF Collecte du web[13]"(BCweb), that allows contributors to independently perform actions such as entering, modifying, or deactivating URLs in the seed list.

Similarly, other institutions have invited collaboration or established partnerships with researchers who are both qualified information professionals and representatives of certain minority groups. Researchers-curators are often sought for their participation in the development of thematic and special collections focusing on capturing the many-sided reality of small communities on the national territory. For example, in the UK Web Archive[14] several thematic collections have been created through participatory curation practices, including the "Black and Asian Britain", "French in London" and the "LGTBQ+ lives online". These participatory practices aim to bring a diverse range of material into the web and social media archive, helping preserve communities' own viewpoints, experiences, and stories.

However, even in the context of co-curation practices, these collections might still face criticism due to certain curation choices. While curatorial decisions are made in collaboration with the archiving institution, documenting episodes of hate, discrimination, and violence can raise concerns among community members, despite these unfortunate occurrences often being integral parts of minority groups' lives. Nevertheless, constructive discussions between the parties involved can still lead to positive outcomes, such as the production of extensive collection descriptions featuring any potentially controversial aspects and content warnings, which are of great value to both the institution and users.

Conclusion

Social media has radically changed the way individuals interact and communicate online, offering unique insights into contemporary events and providing environments for minority groups to self-represent. While the inclusion of social media content of enduring value in national cultural

---

https://web.archive.org/web/20240107200657/https://www.bnf.fr/fr/cooperer-autour-de-larchivage-du-web#bnf-un-r-seau-de-partenaires-pour-encourager-les-recherches-sur-les-archives-du-web

[13] BnF, Collecte du web homepage: https://web.archive.org/web/20231208191553/https://collecteweb.bnf.fr/login

[14] Nicola Bingham (British Library), interviewed via Zoom by Beatrice Cannelli, 23 June 2022.

heritage preservation strategies is on the rise, numerous unsolved archiving challenges persist, prompting questions about representation and inclusivity.

In this chapter, I have provided an overview of the manifold limitations influencing the representativeness of social media collections. I began by considering the social media archiving landscape and how this is shaped by dynamics ascribable to geopolitical and social media divides. Following that, I have described how legal frameworks, technical matters, social media policies and their ephemerality can deeply affect the degree of representativeness of social media collections at a national level, including the type and rate with which different platforms are collected.

I have illustrated how, in order to mitigate representativeness concerns, many social media archiving institutions have adopted specific curatorial strategies that seek the participation of researchers, networks of contributors, and the wider public. Engaging with the public and external contributors has proved to be a valuable approach to uncover stories from underrepresented communities that might escape the large links of the net used by archiving institutions to sift through content in scope. The impact of these participatory practices on the overall enhancement of the representation of national social media collections, and especially their sustainability in the long term, still needs to be fully assessed. In the meantime, documenting how these participatory collections have been developed and making collection scoping documents publicly available or upon request would be a significant step towards helping researchers fully understand the potential implications of such curatorial processes.

Nevertheless, the participatory practices described in this chapter present a good opportunity to raise awareness of the significance of social media archives among the wider public, also contributing to continual engagement with these collections. Encouraging members from different groups of society to actively participate in developing national social media archives, can truly support the preservation of the multifaceted impact these communities have on the national cultural landscape, letting them tell their stories—through content they suggested—using their own voices.

References

Acker, Amelia, and Adam Kreisberg. 2020. “Social media data archives in an API-driven world.” *Archival Science* 20 (2): 106–123. https://doi.org/10.1007/s10502-019-09325-9.

Aubry, Sara. 2010. “Introducing web archives as a new library service: The experience of the national library of France.” *Liber Quarterly* 20 (2). https://liberquarterly.eu/article/view/10584/11316.

Barrowcliffe, Rose. 2021. “Closing the narrative gap: Social media as a tool to reconcile institutional archival narratives with Indigenous counter-narratives.” *Archives and Manuscripts* 49 (3), Article 3. https://doi.org/10.1080/01576895.2021.1883074.

Ben-David, Anat. 2021. “Critical web archive research.” In *The Past Web: Exploring Web Archives*, 181–188. Springer. https://doi.org/10.1007/978-3-030-63291-5.

Bergis, J., Summers, E., and Mitchell, V. J. 2018. “Documenting the Now White Paper: Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations. Documenting the Now, Documenting the Now.”

Bingham, Nicola, and Helena Byrne. 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). https://doi.org/10.1177/2053951721990409.

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge: MIT Press.

Bruns, Axel, and Katrin, Weller. 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

Cannelli, Beatrice. 2022. “Mapping social media archiving initiatives: State of the art, trends, and future perspectives.” IIPC Net Preserve Blog. https://netpreserveblog.wordpress.com/2022/11/30/mapping-social-media-archiving-initiatives-state-of-the-art-trends-and-future-perspectives/

Caswell, M., Migoni, A. A., Geraci, N., and Cifor, M. 2017. “‘To be able to imagine otherwise’: Community archives and the importance of representation.” *Archives and Records* 38 (1). https://doi.org/10.1080/23257962.2016.1260445.

Chambers, S., Birkholz, J., Geeraert, F., Pranger, J., Messens, F., Lieber, S., Mechant, P., Michel, A., and Vlassenroot, E. 2021. “BESOCIAL: final report WorkPackage1 an international review of social media archiving initiatives.” 91. https://www.kbr.be/wp-content/uploads/2020/07/202012_BESOCIAL_Report_WP1_Review_of_existing_social_media_archiving_projects.pdf

Clayton, J., & Hoskins, P. 2022. “Elon Musk takes control of Twitter in $44bn deal.” BBC News. October 28. https://www.bbc.com/news/technology-63402338

Colin-Arce, A., Fernández-Quintanilla, S., Benítez-Pérez, V., & García-Monroy, A. 2023. “Web Archiving en español: Barriers to Accessing and Using Web Archives in Latin America.” https://www.youtube.com/watch?v=plQURfARGBc

Cui, C., Pinfield, S., Cox, A., & Hopfgartner, F. 2023. “Participatory Web Archiving: Multifaceted Challenges.” In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity,* edited by I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, and R. D. Frank, 79–87. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28035-1_7

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). Available at: https://ijoc.org/index.php/ijoc/article/view/7464

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). https://doi.org/10.1515/pdtc-2018-0011.

Gebeil, Sophie and Valérie Schafer. 2020. “Exploring special web archives collections related to COVID-19: The case of the French National Library (BnF).” *WARCnet Papers*.

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., and Hargreaves, S. 2022. “Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: A systematic review.” *BMJ Open* 12 (11). https://doi.org/10.1136/bmjopen-2022-061896.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. 2014. “Assessing the bias in samples of large online networks.” *Social Networks* 38: 16–27. https://doi.org/10.1016/j.socnet.2014.01.004.

Good, K. D. 2012. “From scrapbook to Facebook: A history of personal media assemblage and archives.” *New Media & Society*: 15 (4). https://doi.org/10.1177/1461444812458432.

Graham, M., Hale, S. A., and 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.

Harris, Verne. 2002. “The Archival Sliver: Power, Memory, and Archives in South Africa.” *Archival Science* 2, no. 1–2: 63–86. https://doi.org/10.1007/BF02435631.

Hegarty, Kieran. 2022. “The Invention of the Archived Web: Tracing the Influence of Library Frameworks on Web Archiving Infrastructure.” *Internet Histories* 6 (4): 432–51. https://doi.org/10.1080/24701475.2022.2103988.

Henninger, Maureen, and Paul Scifleet. 2016. “How Are the New Documents of Social Networks Shaping Our Cultural Memory.” *Journal of Documentation* 72 (2): 277–98. https://doi.org/10.1108/JD-06-2015-0069.

Hino, Airo, and Robert A. Fahey. 2019. “Representing the Twittersphere: Archiving a Representative Sample of Twitter Data under Resource Constraints.” *International Journal of Information Management* 48 (October): 175–84. https://doi.org/10.1016/j.ijinfomgt.2019.01.019.

Jackson, T. 2020. “‘I’ve never told anybody that before’” In *Communities, Archives and New Collaborative Practices*, edited by S. Popple, A. Prescott, and D. H. Mutibwa, (1st ed., 93–106). Bristol University Press. https://doi.org/10.2307/j.ctvx1hvvd.13.

Jimerson, Randall. 2006. “Embracing the Power of Archives.” *The American Archivist* 69

(1): 19–32. https://doi.org/10.17723/aarc.69.1.r0p75n2084055418.

Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 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.

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.

Maemura, Emily. 2023. "Sorting URLs out: Seeing the Web through Infrastructural Inversion of Archival Crawling." *Internet Histories* 7 (4): 386–401. https://doi.org/10.1080/24701475.2023.2258697.

Masanès, Julien, Daniela Major, and Daniel Gomes. 2021. "The Past Web: A Look into the Future." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 285–91. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_22.

Messens, F., Birkholz, J. M., Chambers, S., Geeraert, F., Michel, A., Mechant, P., Vlassenroot, E., Lieber, S., Dimou, A., and Watrin, P. 2021. "BESOCIAL–Towards a sustainable strategy for archiving and preserving social media in Belgium." Digital Humanities Benelux 2021 Conference.

Pehlivan, Z., Thièvre, J., & Drugeon, T. 2021. "Archiving Social Media: The Case of Twitter." In *The Past Web: Exploring Web Archives,* edited by D. Gomes, E. Demidova, J. Winters, and T. Risse, 43–56. Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5.

Pendergrass, Keith L., Walker Sampson, Tim Walsh, and Laura Alagna. 2019. "Toward Environmentally Sustainable Digital Preservation." *The American Archivist* 82 (1): 165–206. https://doi.org/10.17723/0360-9081-82.1.165.

Pietrobruno, S. 2013. "YouTube and the social archiving of intangible heritage." *New Media & Society* 15 (8), Article 8. https://doi.org/10.1177/1461444812469598.

Richardson, Allissa V. 2020. "The Coming Archival Crisis: How Ephemeral Video Disappears Protest Journalism and Threatens Newsreels of Tomorrow." *Digital Journalism* 8 (10): 1338–46. https://doi.org/10.1080/21670811.2020.1841568.

Ringel, Sharon, and Roei Davidson. 2022. "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* 24 (5): 1216–33. https://doi.org/10.1177/1461444820972389.

Schafer, V., and Els, B. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL." *WARCnet Papers*.

Schafer, Valérie, Gérôme Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. "Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives." *Media, War & Conflict* 12 (2): 153–70. https://doi.org/10.1177/1750635219839382.

Schafer, Valérie, and Jane Winters. 2021. "The Values of Web Archives." *International Journal of Digital Humanities* 2 (1–3): 129–44. https://doi.org/10.1007/s42803-021-00037-0.

Schwartz, Joan M., and Terry Cook. 2002. "Archives, Records, and Power: The Making of Modern Memory." *Archival Science* 2 (1–2): 1–19. https://doi.org/10.1007/BF02435628.

Simon, R. I. 2012. "Remembering together." In *Heritage and Social Media: Understanding Heritage in a Participatory Culture*, 89–106. Routledge.

Sinn, Donghee, and Sue Yeon Syn. 2014. "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.

Statista.com. 2023. "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/

Storrar, T. 2014. "Archiving social media." May, 8. *The National Archives Blog*. https://blog.nationalarchives.gov.uk/archiving-social-media/

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

Tromble, Rebekah, Andreas Storz, and Daniela Stockmann. 2017. "We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3079927.

Van Dijck, José. 2011. "Flickr and the Culture of Connectivity: Sharing Views, Experiences, Memories." *Memory Studies* 4 (4): 401–15. https://doi.org/10.1177/1750698010385215.

Wu, Siqi, Marian-Andrei Rizoiu, and Lexing Xie. 2020. "Variation across Scales: Measurement Fidelity under Twitter Data Sampling." *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May): 715–25. https://doi.org/10.1609/icwsm.v14i1.7337.

Yakel, Elizabeth. 2003. "Archival Representation." *Archival Science* 3 (1): 1–25. https://doi.org/10.1007/BF02438926.