An Overview of Social Media Archiving Tools

Zefi Kavvadia

International Institute of Social History
Netwerk Digitaal Erfgoed/Dutch Digital Heritage Network

December 2020
Version 1.0

Table of Contents

Context and summary

This report is the outcome of a project funded by the Dutch Digital Heritage Network/Netwerk Digitaal Erfgoed (DDHN/NDE) between 2019 and 2020, which called for cultural heritage institutions in the Netherlands to conduct research in order to build knowledge and capacity in social media archiving. The projects that were selected ranged from looking into criteria for appraising and selecting social media collections, to legal issues in social media archiving and the problematics of offering access to such collections.

As part of developing our born-digital archiving expertise, the International Institute of Social History in Amsterdam proposed to NDE and undertook a project on social media archiving tools. The project looked specifically at tools for the capturing of social media content, and its aim was to assist professionals working mainly but not exclusively at cultural heritage organizations to select appropriate tools for social media archiving. The result is an overview of a range of tools for social media capturing falling into two broad categories: those preserving the “look and feel” of social media, and those preserving structured social media data. The tool survey is accompanied by a contextualization of social media archiving tool selection as an archival issue, and by the description of use cases that display tool usage in different contexts.

1. Introduction

Archiving social media is a new endeavour for most cultural heritage institutions. While web archiving practice has progressed and gained a more or less established position, at least in organizations involved with digital preservation, social media archiving is still very much an open question that archival organizations often seem reluctant to tackle. Of course, researchers, activists, and other interested communities have been collecting social media content for various purposes, as the already rich corpus of research using such data shows. But the systematic and principled archiving of social media content by organizations whose purpose is to preserve it for the long term is still nascent.

The complexity of acquiring, preserving, and giving access to social media collections has deterred organizations from engaging in it, except in one-off projects or as part of their wider web archiving programs, thus not giving attention to its specificities. The issues that block progress are several, with the most prominent being the legal and ethical restrictions that these collections create, but also the technical difficulties of acquiring and giving access to them. The present report focuses exactly on this process of acquisition, and specifically on selecting appropriate software for the capturing of social media content.

The goal of the research project that we undertook at IISH was first to determine a list of social media archiving tools that would then be used to experiment on capturing various social media material. This report is the result of this project, and consists of a general framing of social media archiving issues from an archival perspective that relates features of social media with features of tools and their implications, a detailed review of the examined tools, and a series of use cases to exemplify their potential usage. The report is aimed at new practitioners in social media archiving, especially those in the cultural heritage field, to assist them in choosing social media capturing tools appropriate to their needs and level of available expertise.

2. Methodology

The methodology for selecting and testing social media archiving tools was based on the requirements and resources of a middle-sized cultural heritage institution like the IISH. This means that some of the requirements we put forward might be less or more pressing for organizations of different sizes, and also that the usage of the tools we describe would need to be tailored to their respective capacity and use cases. A local government archive, for example, might want to invest more time in ensuring that their collections only include content by governmental agents, and filtering out other types of content – this would mean that they might be interested in tools that have more sophisticated filtering functionality. Alternatively, a non-profit with less in-house IT expertise might assign more importance to selecting tools that can be used reliably without constant need for instruction or maintenance. The functional requirements described below cover a reasonable amount of such possible uses, and the non-functional requirements represent the more general quality criteria that we think the tools should fulfil.

2.1. In Scope: Free and Open-Source Capturing Tools

Early on, we decided that the research would focus on tools for capturing social media content for long-term preservation. We made this decision because it made practical sense: looking into the multitude of systems and applications available for all stages of the archival workflow would require a longer timeframe for research and experimentation than the one we had currently available. We decided that a more immediate response to the increasing significance of social media archiving was preferable, seeing as most organizations in the Netherlands are still at a very nascent stage of their social media archiving efforts. By focusing on capturing, we believe we can offer organizations the opportunity to make their first steps into preserving social media, and we are opening up the lines of communication and co-operation for more projects and collaborations that will perhaps focus on different aspects of the social media archiving process.

We also opted for open-source tools exclusively, even though this was not a formal requirement from the NDE for our research. Recent policy planning by IISH (IISH 2018, pp. 25) indicates that whenever possible, we will attempt to make use of open-source tools, so that we will be able to have better control of our data, but also to even contribute to further development of the tools themselves. Thus, it made sense for us to continue the research for the present project under this premise. This necessarily means that there could be valuable and interesting paid tools that could possibly resolve some of the issues we encountered, but we believe it is vital for cultural heritage organizations to support open-source software when they can afford to do so, as it can offer more freedom of choice and tailor-made options.

Consequently, out of the scope of this project were:

- Paid tools
- Free and open-source tools for archival processing, description, long-term storage, and publication and access

2.2. Functional requirements

The list of functional requirements below is based on an earlier project carried out within IISH which focused on workflows for acquiring and preserving born-digital materials in the broad sense, and another project that looked into web archiving tools and workflows specifically. We made our proposal to NDE based on the experience we had gained from these two projects and we also used our Strategic Plan (IISH 2019a), Collection Policy (IISH 2018), and Digital Preservation Policy (IISH 2019b) documentation to further specify requirements.

We were particularly interested in testing tools and their outcomes with a focus on how they can be used in cultural heritage settings, and for this reason we took a view as broad as possible of what the tool outputs could be (webpage snapshots, structured data, even screenshots). While there is increasingly more and more interest in finding out what humanities and social science researchers need from web and social media collections (Hockx-Yu 2014a; Jackson et al. 2016; Winters 2017), the truth is we are still at a stage where a lot of the decisions made in cultural heritage institutions are necessarily based on various degrees of informed guesswork. The type of research (quantitative, qualitative, mixed methods, etc.) they perform, as well as their digital skills and background, affects the kinds of access that researchers desire. Experimenting with different tools allows us to gain experience of the possibilities for access that the output of each tool creates before we finalize our strategies and methods of collecting.

At the same time, it is also worth noting that the final number of social media archiving tools examined in this report is rather small, compared to the deluge of applications and addons one can find online. From open-source repositories to browser plugins, professional and amateur archivists have created and/or appropriated many tools to archive social media activity. After looking at various options, we decided that we would be looking at tools that could satisfy the requirements laid out below, either individually or in combination with other tools:

1. Preservation of "artefactual" value of content (the "original look and feel" of the page as it browsed by end users of the platforms)
2. Preservation of "informational" value of content (the "informational content" of the page, i.e., the textual content of posts and comments, the links, the usernames, and the metadata associated with those)
3. Capture of password-protected content (what sits behind a log-in screen as opposed to publicly available social media content)
4. Media capture and/or extraction (e.g., capturing images and videos embedded in the page but also downloading them as separate files)
5. Rich media (i.e., interactive graphics and other dynamic content) capture
6. Snapshot captures (one-time capture of the page) vs. scheduling of periodic captures
7. Output in accepted archival formats and/or widely adopted and/or open-source formats
8. Internal logging and documentation capabilities (e.g., logs, change tracking, capture session metadata, including ability to extract this information in a usable format)
9. Integration with existing workflows (e.g., the degree to which the tool can be more or less easily integrated within existing systems and workflows in use)

2.3. Non-functional requirements (quality attributes)

Functional requirements refer to what the software can do, its actual affordances, while non-functional requirements refer to how the software achieves these goals by using specific mechanisms, which themselves affect the quality of the user experience it offers and the tool’s sustainability (Chung et al. 2000). The non-functional requirements, also known as quality attributes, listed below were taken from the practice of software testing and software selection. While most of the academic and professional literature on software testing and software selection seems to be geared towards software used in business (and thus targets developers, analysts, and information architects), there is still merit in making use of tested, used, and widely documented principles and adjusting them to our purposes.

There are a few different criteria for software testing listed in different resources, and one could make a case for or against using some or all of them, but the approach taken here is based on the relevance of the criteria to the needs of a heritage and research organization like IISH involved in digital preservation generally, and specifically in web and social media archiving. The following non-functional requirements were distinguished for choosing tools for social media capturing:

1. **Usability**

   While one should ideally not let the difficulty of using a tool necessarily become an obstacle, it is understandable that a less usable tool might result in slower or decreased adoption and desired outputs. It is thus more beneficial, if and when possible, to select tools which can be used efficiently with less effort in order to make the best of the often-restricted time, human, and financial resources that cultural heritage organizations involved with digital preservation have.

   Usability requirements in social media capturing software can refer to simplicity in the GUI[1] or even the presence of a GUI, the responsiveness of the software e.g., how quickly it loads and completes tasks, and the extent to which users with average IT proficiency[2] can (semi)-intuitively learn to use the software without considerable learning curves. It might also refer to the presence and quality of the tool's documentation.

[1] Graphical User Interface. Some tools have one, other only have a CLI (Command Line Interface), and some could have both.

[2] The definition of average IT proficiency used is the one proposed by the NetLab project of the DIGHUMLAB (Digital Humanities Lab Denmark): "people in this skill level are experienced with programs they use for daily tasks, often including some advanced functions, but they will usually require some time and help for learning new routines. They usually are aware of the importance of security, and have a basic understanding of some data types and data handling, but when learning new routines and programs they are usually best served with also establishing a number of fixed routines and precautions in how the data should be stored, handled, and protected. Advanced programs and tasks can be learned, but the learning process may prove difficult and time- demanding" (NetLab, accessed 9 November 2020).

2. **Scalability**

   When doing social media archiving it is especially important to be able to configure a workflow and its tools in order to deal with possibly fluctuating amounts of data to be captured, often on short notice (e.g., on the occasion of important socio-political events). Optimal use of institutional resources might mean strategically increasing or decreasing the time spent, machines devoted, or storage space assigned for a particular workflow.

   Scalable social media archiving software is able to withstand capturing large amounts of data without failure, and to be integrated with other existing systems.

3. **Reliability**

   Reliability is a relevant attribute in social media archiving software as it indicates that the tool can be trusted to become part of established workflows in an organization.

   Reliability in a social media archiving tool could translate to the tool being consistent in how much time it needs to capture similar kinds of data under similar conditions (machines, bandwidth, configurations). Additionally, and in a broader sense, reliability could also refer to how well-supported the tool is in terms of continued development and maintenance.

4. **Security**

   Security is a relevant requirement because of the sensitive nature of social media data. Personally Identifiable Information (PII) is at the core of this material and calls for attention to privacy, during capturing[3] and when publishing. Privacy and copyright issues could arise if an organization unwittingly shares captured social media content without the platform's and/or users' consent. Additionally, the security of the organization itself could be at risk.

   Secure social media archiving tools could enable the protection of those using them to perform captures (e.g., by not storing the archivist's usernames and passwords in the output files when harvesting password-protected pages), and also the protection

[3] Under the General Data Protection Regulation (GDPR), capturing social media content and preparing it for preservation can be considered personal data processing, defined as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (Article 4, Chapter I, Regulation (EU) 2016/679).Thus, legal considerations must be thought of preferably as soon as possible in the social media archiving workflow.

of users through filtering and scoping[4] to weed out sensitive information from a capture. If the tool is open-source, its code could also be examined for vulnerabilities and improvement.

These quality attributes can best be thought of as features on a continuum, rather than something that the tools either have or do not have. After all, different tools could be combined to fit one's purposes. Finally, it must also be noted that we did not measure the adherence of the tools to these non-functional requirements in a quantitative way or with formalized tests. Rather, we used them as considerations to keep in mind: we were interested in usable tools, which would scale in different circumstances, would perform systematically well when maintained reasonably, and that would enable the protection of the security and safety of the data subjects included in the social media collections and of the organization and its staff.

[4] Scoping is the process through which we determine rules for a capture session that will enable us to capture exactly the materials we want and to avoid capturing materials that are not useful or relevant. Scoping might involve using keywords and regular expressions to exclude or include URLs, blocking entire domains, capturing only specific file types, etc.

3. Defining social media and social media archiving

What is social media? The term is used widely, but a little research quickly indicates it is used differently by different people. In most cases, social media refers to websites like Facebook, Twitter, and Tumblr, in which users create accounts and connect with one another by producing, sharing, and interacting with content. There are precedents to this, such as online forums and blogs; for example, Kaplan and Haenlein classified as social media not only blogs but also collaborative projects like Wikipedia, and even virtual worlds like Second Life and online role-playing games like World of Warcraft (2012, pp. 101). This is a very expansive definition, but it drives home the point that social media is an all-encompassing term for various kinds of communicative and technical infrastructures that exist on the Internet.

Social media are identified with the rise of Web 2.0 and the era of increasing online interactivity, personalization, the use of mobile devices and cloud computing. A broad definition like the one proposed by Treem et al. (2016, pp. 770), who see social media as technologies that “create a way for individuals to maintain current relationships, to create new connections, to create and share their own content, and, in some degree, to make their own social networks observable to others” could be helpful for archivists and other information professionals that are looking for ways to define what to include in their social media archiving policies. Recently, the use of WhatsApp to organize protests in Egypt and Sudan (Shendi 2020) or Telegram’s role in spreading right-wing ideologies through multi-user chatrooms (Urman and Katz 2020), or again, the use of the Discord messaging service for “Corona-time” virtual parties, or the Twitch streaming platform to stream yoga sessions as well as PhD defences during the coronavirus pandemic, all indicate that social media usage is always evolving beyond initial intended purposes. One of the challenges for tomorrow’s archival institutions will be to remain open to the shifts happening in the information and digital communications landscape, in order to keep collecting what is valuable to be preserved beyond limitations of medium or intended use.

Understanding the breadth of what can possibly constitute eligible content for social media collections is important for tool selection and assessment, but also for acknowledging that at times the solution might not lie with tools per se. The many differences between social media platforms in terms of technical affordances, styles of interaction and demographics influence the choice of tools for capturing content – it is very likely that to capture Instagram content without capturing its media will result in a collection that is less rich than it could be, considering how integral images and videos are to this platform; thus, a tool should be selected that can capture Instagram media, if possible. Conversely, WhatsApp and its use as a communication tool in protest movements indicates that archiving social media is not only about choosing appropriate tools, but also about building relationships of trust and accountability with those who create the data we are interested in archiving (Jules, Summers, and Mitchell 2018) – sometimes, this will be the only way to even get access to data sitting behind a log-in screen or private chat room, like in the case of WhatsApp.

Finally, another important aspect of what constitutes social media archiving practice is the realization that social media platforms are not necessarily, if ever, eager to co-operate with cultural heritage organizations for the preservation of their content. In terms of choosing and using tools, this translates to sudden changes in platform design that render tools unusable, and in the need for the archivist to adapt to these conditions. This situation makes social media archiving challenging and poses a significant threat to the safekeeping of the individual, social, and political memory of today. As the contemporary historical record, now to a great extent comprising of born-digital content published on proprietary online platforms, is potentially restricted according to corporate will (Bruns 2019), the future of digital heritage and research could be endangered.

3.1. Social media archiving tools and archival principles

Social media archiving, as argued above, can be seen as an extension of web archiving, and both of them fall under the umbrella term of born-digital archiving. While archiving born-digital content has its own peculiarities (Littman et al. 2018; Hockx-Yu 2014b), archival principles that apply to preserving digital content in general can help us identify a basis from which to derive some criteria for assessing the quality of social media archival collections. By doing this, we can also apply the criteria to our tool selection, since the choices of tools when archiving social media have great bearing on the nature of the collections that we will be able to create with them.[5]

[5] A note at this point must be made about the application of the term “electronic records” to talk about social media archives in this section. On the one hand, the use of the word “electronic” in place of the now more common “digital” sounds perhaps slightly anachronistic; on the other hand, speaking of social media records definitely requires us to define precisely what we mean by this. This is indeed a fascinating subject that calls

A lot of archival organizations that maintain archival repositories make use of ISO standards like the *Open Archival Information System (OAIS) Reference Model* (ISO 14721:2012) or the *Records Management-Part 1: Concepts and Principles* standard (ISO 15489-1:2016) for the design of their archival and digital preservation systems. For the purposes of the present overview, four essential features that must characterize electronic archival records according to the International Council on Archives (ICA) will be used. ICA uses the ISO Records Management standard, as well as previous work by groups such as InterPARES (International Research on Permanent Authentic Records in Electronic Systems) to define authenticity, reliability, integrity, and usability as essential requirements for digital records (ICA 2005, pp.12). It is useful to revisit these definitions in thinking of social media archiving tools, because ultimately, our selection of software should hopefully result in producing sound archival collections.

Authenticity is a fundamental requirement for archival collections. Authentic archival records are what they claim they are and they have been created by those who claim to have created them. While archivists have long used tested and well-established processes to determine and safeguard authenticity in paper records, digital records are of course trickier, given the many possible reproductions a digital file may have, and the ease with which these can be exchanged. For social media collections, authenticity is a difficult subject; strictly speaking, as with most digital records, we can make no claims to authentic social media records in the traditional sense of the term. Consider the archiving of a Facebook group: a Facebook group is a dedicated forum in which Facebook users can interact with each other on the basis of a shared interest or specific topic. Groups exist within the Facebook platform, and they are identified by a distinct URL. When we use scraping and harvesting tools to archive a Facebook group and store the material in a digital file, we create a wholly new entity that did not exist beforehand. Thus, claims to authenticity have to be revised when applied to social media collections, and geared towards making sure that we document any discrepancies between the archival social media record and the social media page at the time of capture.

for more enquiry, but the commonly accepted archival definition of records as information produced or received by organizations or individuals in the course of completing a particular activity or function (Reitz 2004, pp. 722), can definitely apply to social media archival collections. The particular unit which will be defined as record e.g., a tweet, a collection of Facebook pages, a dataset with structured textual information and metadata about posts with a particular hashtag, will depend on the context and purpose of collecting.

Reliability, defined as the degree to which a record reflects the range of activities or facts that it claims to describe, is related to authenticity, in the sense that if we can establish the identity and origin of a record reliably, we can also establish whether it represents as fully as possible the activities or events it is seen as describing. Knowing the context of the online content we are archiving, why it was created, when, and how, can help us make some judgement about whether the content is reliable as a source for the topic it relates to. Once again, reliability in the digital environment has to be established by tracing as accurately as possible the conditions in which the record came to be – the tools, systems, stakeholders, and circumstances that contributed to its creation. This means that ensuring reliability in a social media collection is tied to the capturing process.[6] When a social media post or hashtag feed is captured via an API, it can be difficult to determine whether it is complete or reflective of what it describes,[7] other than by documenting the tools, processes, and context in which the capture was made. Additionally, especially with API collecting, it is important to document the terms and conditions of the platform at the time of capture, and any restrictions imposed by the API on the harvest. This also has to do with the nature of the web and social media archiving process itself, even if we are using methods other than APIs to capture content: because in many cases, without interaction from the user or the browser, it is impossible to even load the content we want on our screen. This means that many, if not all, the tools used to mimic user behaviour and capture social media content via the browser, will necessarily introduce minor alterations in the captured files in order to make harvesting possible. Only by being explicit about what is missing or what cannot be guaranteed can we claim reliability of social media collections.

Integrity also relates to authenticity and to reliability, as the requirement for records to be untampered – by proving authenticity and reliability, we can also prove integrity, because we have the necessary evidence that the record has not been altered after it entered the collection, and the necessary procedures to make sure that it stays unaltered in the future. Integrity in the strict sense of making sure the records are free from any intervention is once again difficult to claim in social media collections, as for various reasons the collections might

[6] And additionally to the capturing process, it also depends on wider processing activities, including quality assurance, normalization, validation, ingestion, etc. and how well they are documented.

[7] This is because, for example, Twitter does not publish information that will allow archivists and researchers to make sure that what they capture from the API is indeed representative of the data that is available, and also does not give free access to historical data beyond 7 days in the past.

indeed be edited, e.g., to protect the identity of a group or individual, or because the terms of the platform require that content deleted on the platform must also be deleted on any captured data sets (Littman 2019). Similarly, integrity can be established by performing fixity checks on the harvested files, but in terms of securing the collections completely against being altered, social media collections are perhaps more at risk than others. An example can be taken again from the practice of API harvesting of Twitter content, in which instead of sharing the content of tweets, the platform only allows the publication of their unique identifiers (Twitter Developer Policy, "Let's get started," accessed 15 November 2020). Using these identifiers, researchers could query the API again and retrieve the relevant content for their use – minus any content that was deleted or blocked after the harvest was made. These restrictions effectively limit the ability of archivists to guarantee the integrity of API harvested data.

Finally, usability refers to the quality of a record to be retrieved, viewed and interpreted; in a paper environment, usability would translate to appropriate finding aids and an efficient system of storage and retrieval. In the digital realm, these two requirements still hold, but they can be complicated to satisfy. How could we make sure that a file remains accessible, when the software needed to render it often has a short lifecycle, and technical developments vastly outstrip the capacity of most archival institutions to keep up? Usability in social media collections is intimately tied to the different formats in which the materials can be captured. Apart from making sure that we have the necessary hardware and software to render and use the materials, the choice of format significantly impacts the kinds of uses a social media collection can be put into. Usability thus becomes a matter of degree: usable to who and for what kind of purposes?

It then becomes apparent that there are many technical and non-technical factors that can play into archiving social media successfully, and this success will also be defined according to context and purpose. These factors also play into the selection of social media archiving tools, as different tools result in different outputs, and different outputs could serve different sorts of archival purposes.

3.2. Approaches to archiving social media

3.2.1. The "look and feel" approach

Considering its close ties with web archiving, it is reasonable to assume that we can perform social media archiving using similar methods. The most common method of web archiving, namely web crawling or web harvesting, attempts to preserve the so-called "look and feel" of online content, meaning the layout, structure, and style of a website, as well as its navigational features, like buttons and menus. However, this has proven to be relatively limited in what it can accomplish with social media. In their study of the diachronic evolution of website archivability, Kelly et al. examined how the increasing reliance of web designers on JavaScript to create websites has seriously impacted the web's potential to be archived successfully (2013). JavaScript is used to create more interaction between the user and the page, and nowadays is one of the many mechanisms that contribute to the dynamic nature of social media. So-called traditional crawlers, like Heritrix, capture online content by performing requests to host servers based on the URL we want to archive. Like a browser requesting whatever is at www.example.com with an HTTP request, and then displaying what it receives from the servers, crawlers request the pages in the URLs we provide to them, the so-called seeds, and then they store the content they receive. Unfortunately, these static websites of old are becoming more and more uncommon. Where it was previously possible to capture an entire website by gradually visiting all its URLs one by one, e.g. www.example.com, www.example.com/faq, www.example.com/help, etc., nowadays the dynamic features of most websites, including social media, very often require user interaction in order for pages to load. Basically, the user clicking and interacting with the page triggers the construction of URLs for the elements of the website – without the interaction, these URLs are not constructed at all. This puts most contemporary social media content out of traditional crawlers' reach.

In response to this, alternative methods of capturing social media content have been in development and are gaining ground. To capture the "look and feel" of social media content, i.e., the pages that make up the social media website that an end user can peruse, scrapers and harvesters based on browser emulation are used. As the front-page content of a social media website is the most recognizable form of social media experience for the majority of people, being able to preserve this component of social media is important. Browser-based crawlers, e.g., Brozzler, Webrecorder Desktop, etc., mimic the interaction of a human user on a social media website, playing media content, expanding nested comments, and generally triggering all the interactions required by the page for it to be fully rendered. The goal is to recreate and store the visual outlook and functionality of the page as it was at the time of capture.

3.2.2. The “structured data” approach

There is however another way of capturing social media content, that focuses less on the “artefactual” quality of the material, its visual form and its multimedia affordances, and more on its “informational” quality, on the raw data deriving from it. Arguably, it could be claimed that a fully rendered Instagram feed can be very informational on its own; a case in point is Amalia Ulman who, in the process of creating and sharing her digital performance art work “Excellences and Imperfections” on Instagram, made use of the colouring scheme and shape of the platform’s 2014 interface. Her personal account and posts are archived by the Rhizome team[8] and preserved in the form they were first published in, including the now outdated Instagram interface – but without access to the look and feel of her profile, a significant aspect of the artwork would be lost (Thomson 2017, pp. 7-8). Even so, differentiating between the artefactual and the informational quality allows us to characterize an approach to social media archiving that speaks to the ubiquitous and rapidly developing fields of data analytics, visualization, etc., that are now triggering a change towards viewing collections as data (Milligan 2016; Padilla 2018; Vlassenroot et al. 2019).

The output of an approach focused on informational qualities is structured textual data, usually in tabular form. Structured data are easier to analyse and process with computational tools, making them highly valuable for research. Social media platforms make structured data derived from their websites available via API services, i.e., specific interfaces created for applications and tools to connect and interact with the back-end of the platform. By connecting to an API, an interested party is able to access information not normally available to the end user of the social media website, such as aggregated numbers of likes and reposts, metadata about location, unique identifiers for each post, etc. While primarily targeted towards commercial users such as developers, web designers, market analysts, etc., social media platform APIs have been heralded as a valuable source of social media data by social scientists, policy makers, journalists, and others.

[8] https://rhizome.org/

Even so, the reality is that social media platforms still restrict research use and archiving of their data to a great extent: Twitter, as one of the most popular data sources for social media research, has taken into consideration the non-commercial users interested in its API, and allows for researchers to create accounts for academic purposes; the restrictions for sharing and publication though still hold and limit the possibilities of what can be done with the data. Unless one is willing to pay for premium access, Twitter only allows access to data from up to 7 days in the past from its public API – which means that content that was published even a month before the time of capture is inaccessible. The Twitter API terms and conditions also impose limits on the quantity of content that can be requested, by rate-limiting, i.e., controlling and restricting the number of requests that can be made in specific time intervals from the same account (Twitter Developer Documentation, "Rate limits," accessed 1 November 2020). Other APIs, such as that of Facebook and Instagram, are not even accessible to researchers without a painstaking application process and they do not offer any accommodations for researchers and by extension archivists (Ben-David 2020). These restrictions may seriously impact the ability of an organization to create social media collections with authenticity, reliability, integrity and usability.

Regardless the limitations and obstacles, API-based harvesting of social media content is still a very popular and useful way of preserving social media records, and it could be performed side by side "look and feel" harvesting, if and when it is possible, for more complete and well-rounded collections.

3.3. File Formats and Designated Communities

The two general approaches outlined above also have implications for file format selection, which by extension has implications for preservation and collection quality. The choices made when capturing and preserving will affect the possible uses the collection can be put into, which means these choices are also directly connected with the nature of the designated communities that the collections are intended for.

The "look and feel" approach to social media archiving will almost always result in WARC files. One might also opt to download media content, or to take screenshots,[9] but

[9] Taking screenshots is a valid method of archiving online content, especially if all else fails. It is however also useful for performing quality reviews, allowing the archivist to compare the screenshots taken by the capturing tools with the captured content and with the live website and discover discrepancies.

WARC is widely accepted enough to be considered one of the default formats for storing captured content from the web. The WARC followed its predecessor, the ARC, as the main file format in use by the Internet Archive, and is maintained by the International Internet Preservation Consortium (IIPC). The rationale behind the WARC format is that one file format for web archiving should preferably be able to hold not only the archived resources themselves, but also metadata about the resources and the capture. The WARC is thus an aggregator format that combines all the segments of a crawled website plus the HTTP requests and responses performed during the crawl together (The WARC Format 1.1, accessed 9 November 2020).

Even though the WARC is popular, there can be discrepancies between WARCs created by different software and systems, e.g., if there are inconsistencies between the application of the format standard across tools. This is a reason why WARC file format validation could be important for a social media preservation workflow (Veenendaal 2020). Nevertheless, one of the most important things to note about WARCs is that the kind of access that they enable is meant to reproduce the experience of browsing the original website. The assumption is that the collection user will navigate from page to page and website to website, consuming the content similarly to how the users of the page's live version would. This assumption is reflected in the interfaces of WARC replay tools, most characteristic of which is the Wayback Machine of the Internet Archive: a search bar for URLs, and then various versions of the desired URL arranged chronologically. While useful for those interested to go through a relatively small number of resources, or to perform a close reading on the content, such access interfaces do not make it easy to discover and manage the vast volumes of data that is often contained in web and social media archival collections. This is why the WARC itself, and access methods based on browsing single websites, have received some criticism in later years.

There are now calls for more attention to data-intensive access methods to web and social media archives, and the structured data approach to capturing social media produces output appropriate for this. Usually in formats like JSON, XML, CSV, XLSX, and others, collections made up of structured data are more amenable to computational methods such as network analysis, topic modelling, and many other visualization and analysis methods. While critics of WARC-based collections claim that a lot of extra work is needed in order to make WARC data machine-actionable (Wang and Xie 2020), collections in JSON or CSV could potentially lower the threshold of that effort. However, there is a catch that involves the integrity and reliability of social media archives: such structured data, if they derive from social media APIs as they usually do, are not easy to control in terms of provenance, because as it was mentioned earlier, the platforms themselves are not transparent in their policies of data exchange and publishing. If we do not know how the platforms choose the data that they give to us, we cannot make claims as to its completeness and integrity. Being required as we are by Twitter to delete data we have captured if a Twitter user deletes it from their live profile, puts the reliability of collections at risk, and also creates an extra task for archivists to monitor social media websites for changes.

The differences between WARC on the one hand and formats like JSON, CSV, etc. on the other hand, enable different kinds of users to access social media collections for different purposes. The concept of the designated community, referring in the OAIS model to the group or groups of users that the cultural heritage organization intends to make the collections accessible for, seems simple enough but can be complicated, especially when the object is as multi-faceted as social media. If the aim of establishing a designated community for one's collections is to make these collections understandable to that community (ISO 14721:2012 OAIS), how do we define understandability? In the case of IISH, for example, whose designated community is mainly academic researchers, one could argue that the method that is more useful for efficient and valid research should be chosen. But what kind of research? An economic history researcher that uses historical online data to research inflation rates will probably have different needs than a researcher of urban sociology who wants to read blogs and comment sections by city residents to understand gentrification. Even though both are in the humanities and social sciences, they might have different levels of technical skill that will allow them to use collections in different ways.

If we made social media collections available exclusively as structured data, would we contribute to the usability and comprehension of the content or would we make it more inaccessible to those without the digital skills to manipulate it? Additionally, where are the data subjects in this discussion, the users of the social media platforms that, usually without their knowledge, might end up in these collections? One of the thorniest issues in social media archiving these days, and arguably one that is seriously blocking progress, is the various legal and ethical issues surrounding the re-use of personal data. The requirement to delete captured Twitter data that is deleted on the live website for example, is on the one hand problematic for the integrity and reliability of the collections, but on the other hand, it could be fundamental to protecting the rights of vulnerable people. Institutions that collect social media, with their selection of tools, formats, workflows, and designated communities, should consider how they might be harming or benefitting the individual users and/or communities in the collections they create, at the same time as they are trying to create social media collections that are valuable and usable for research and memory.

4. Tools

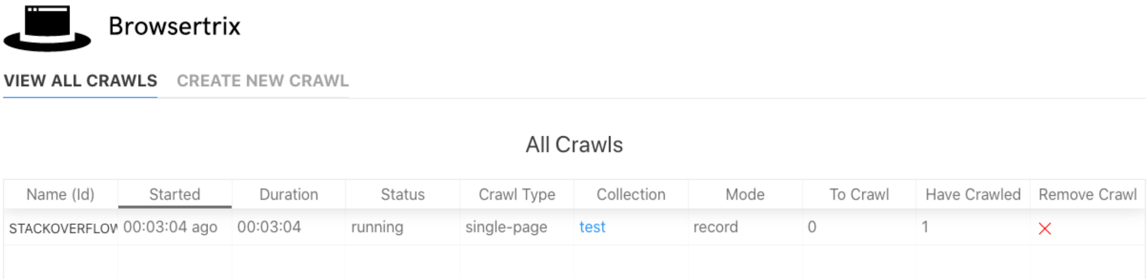
In the beginning of this research project, an attempt was made to choose a manageable number of tools out of those available. The tools to be tested were selected based on how they fit the criteria we set, but also on whether it was possible for us to install and use them on our available machines, and on whether they were in fact still functional. The tools presented below are the final selection. They have been divided into two categories ("look and feel" and "structured data") and they are presented in alphabetical order.

4.1. "Look and Feel" Output

4.1.1. Browsertrix[10]

Browsertrix is a project by the team behind Conifer/Webrecorder Desktop (see Webrecorder Desktop/Conifer), and it is basically a set of browsers, capturing behaviours, and a system for WARC replay, bundled together in Docker containers. Docker,[11] the containerization technology that allows for packages of software to be used and transferred between computer environments without requiring users to install all the necessary dependencies themselves, gives Browsertrix its flexibility as a capturing tool that is meant for more large-scale crawling. Compared to Webrecorder Desktop, Browsertrix requires in principle much less human intervention in order to perform crawls.

Browsertrix is a CLI-based tool, even though a GUI exists and can be used to control some of the basic operations; however, the GUI was seen to be a bit unreliable sometimes, freezing during large crawls, or incorrectly displaying crawls that had already finished or had been removed as active. Nevertheless, it is useful especially if the crawl to be done does not require a lot of configuration, e.g., if it is a matter of crawling three Twitter profiles, without special scoping rules, the GUI can be used to enter those quickly and start the harvest.



All Crawls

Name (Id)	Started	Duration	Status	Crawl Type	Collection	Mode	To Crawl	Have Crawled	Remove Crawl
STACKOVERFLOV	00:03:04 ago	00:03:04	running	single-page	test	record	0	1	✕

Figure 1: Browsertrix GUI

[10] https://github.com/webrecorder/browsertrix

[11] https://www.docker.com/

The differentiating feature of Browsertrix, however, is the flexibility it offers, compared to Webrecorder Desktop, in deploying different crawling options. Using seed lists, one can specify URLs to be crawled, and then choose if they want only the specified seeds to be captured, or all the URLs belonging to the same domain or sub-domain as the seeds (useful if, say, a company website has different language versions), or set any kind of custom scoping rules they like. Defining the scope of a website or a social media page might be tricky, because in order to preserve the context and provenance of the material, the archivist might feel they need to include everything that appears on the page: images, videos, links to other websites, embedded content, etc. Putting aside the technical issues this might cause, it also brings about the conceptual question of how much of this content "belongs" to the page in question, and how far we should go when scoping our crawls. The answer is not easy to determine and will probably depend on a lot of situational factors, but the tool giving us the ability to choose what to include and what to exclude from a social media crawl is very important. It must be noted, however, that a learning curve does exist when learning to apply scoping rules, especially for a person only now familiarizing themselves with these technologies (see Brozzler for more information on this issue).

Browsertrix is able to capture password-protected websites, thus it is suitable to crawl logged-in versions of social media pages. This can be done by setting specific browser profiles, that call up a new instance of the remote browser that Browsertrix uses for crawling. This browser can be used to log in and capture content as needed, and it was successfully tested to capture logged-in versions of Facebook pages and Instagram profiles. When reviewing the captured logged-in pages, the login details do not be to re-entered. However, there is a possibility that the WARC might retain the credentials used to log into the platform, thus care should be taken when offering access to the collections.

4.1.2. Brozzler[12]

For those looking for large-scale harvesting solutions, Brozzler, like Browsertrix, is an interesting choice. Brozzler was developed and is still being maintained by the Internet Archive, and it is already used by organizations such as the Portuguese Web Archive.[13] It is

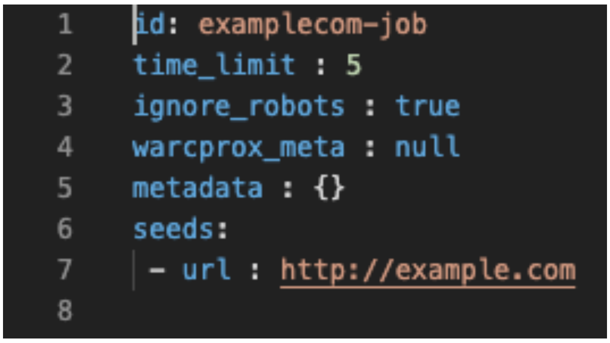
[12] https://github.com/internetarchive/brozzler

[13] https://github.com/arquivo/arquivo-brozzler

a browser-based crawler which uses Chrome or Chromium to access web content and harvest it in a WARC file. Brozzler is one of the newer-generation capturing tools which leverages browser technologies to interact with pages and overcome the difficulties that dynamic content poses for traditional crawlers. It is enhanced with youtube-dl, a video download tool that is able to extract media from crawled content.[14] To display captured content, Brozzler makes use of a custom version of the pywb web archive replay tool.[15]

For Brozzler to work, a database must be deployed which is used to store and manage the crawl data that make harvest configuration and replay possible. Currently, the tool uses RethinkDB for this purpose. The option to use another database exists, but would probably require some tinkering in order to find out exactly how it interacts with RethinkDB and replicate that with an alternative one.[16]

Brozzler also comes with a simple GUI that offers an overview of running crawl jobs, and can be useful to monitor concurrent captures. The actual crawl configuration though is done via YAML, a data serialization language that is often used to write configuration files. There is a number of examples available on the Brozzler GitHub repository, but through testing on various social media platforms, it is clear that the YAML files need to include more specific scoping rules than just a seed list to successfully harvest the content. For example, the following could be used for a simple website capture.



```
id: examplecom-job
time_limit : 5
ignore_robots : true
warcprox_meta : null
metadata : {}
seeds:
  - url : http://example.com

```

Figure 2: Example of a simple YAML crawl specification

The above website is simple enough to warrant only a time-limit configuration and a request to ignore the robots.txt policy if it attempts to block the crawling. However, a lot of social media content is unfortunately not that simple to capture. One particularly difficult case was Facebook, which kept resulting in WARCs that only contained the rudimentary user interface

[14] https://github.com/ytdl-org/youtube-dl
[15] https://github.com/webrecorder/pywb
[16] https://github.com/internetarchive/brozzler/issues/159

of the page, but no actual content. The solution was to employ some scoping rules, e.g., to exclude the domain “facebook.com” from being crawled, and to block URLs which contained specific strings. These settings were based on the instructions provided by the Archive-It team on their Help Center,[17] targeted at users of the paid Archive-It service they provide. Archive-It actually leverages Brozzler to archive dynamic websites such as social media, thus the tips mentioned on the Help Center can come in handy when using the standalone version of Brozzler. Please note, however, that the instructions given are meant for the Archive-It service users who are given access to a GUI to configure the crawler – in order to configure the free and open-source Brozzler, the instructions must be written within a YAML file and implemented via the CLI.

One of the most notable missing features in Brozzler, as in Browsertrix, is the native capability to schedule crawls for the future, either one-time or recurring. While such a practice could create a significant amount of harvested data requiring (temporary) storage and possibly appraisal, and brings with it the risk of redundant content being captured, it could greatly benefit organizations that would like to automate their social media archiving workflows. Though not available natively, it could be achieved by remotely controlling the browser e.g., with something like Puppeteer[18] and scheduling jobs to run through it.

4.1.3. Crocoite[19]

As browser-based crawling seems to become central in the practice of archiving the dynamic web, there is a concurrent increase in interest in using headless browsers to crawl and capture online content. Headless browsers are in essence browsers stripped of their GUI - they are able to perform all other functions of a regular browser, but they do so in the background without displaying them to the user. Headless browsers are often used when testing a page to make sure all the interactions run smoothly without using up a lot of system resources (which GUIs often do). Their flexibility and speed have made them attractive to the web and social media archiving community, and more and more tools are experimenting with them e.g., Brozzler.

[17] https://support.archive-it.org/hc/en-us/articles/208333113-Archiving-Facebook
[18] https://github.com/puppeteer/puppeteer
[19] https://github.com/PromyLOPh/crocoite/tree/master/crocoite

Crocoite is one such tool that, unlike others in this list, uses exclusively a headless browser for all its operations. Using Chrome in headless mode, crocoite is able to fetch JavaScript-heavy online content, such as social media, and store it in WARCs. Operating the tool happens exclusively via the CLI, and specifically on a Linux machine (testing on Mac was not successful). Crawl configuration is not extremely granular, but it is still useful, and allows a quick capture of a single seed, or more detailed instructions to follow and capture links from that seed.

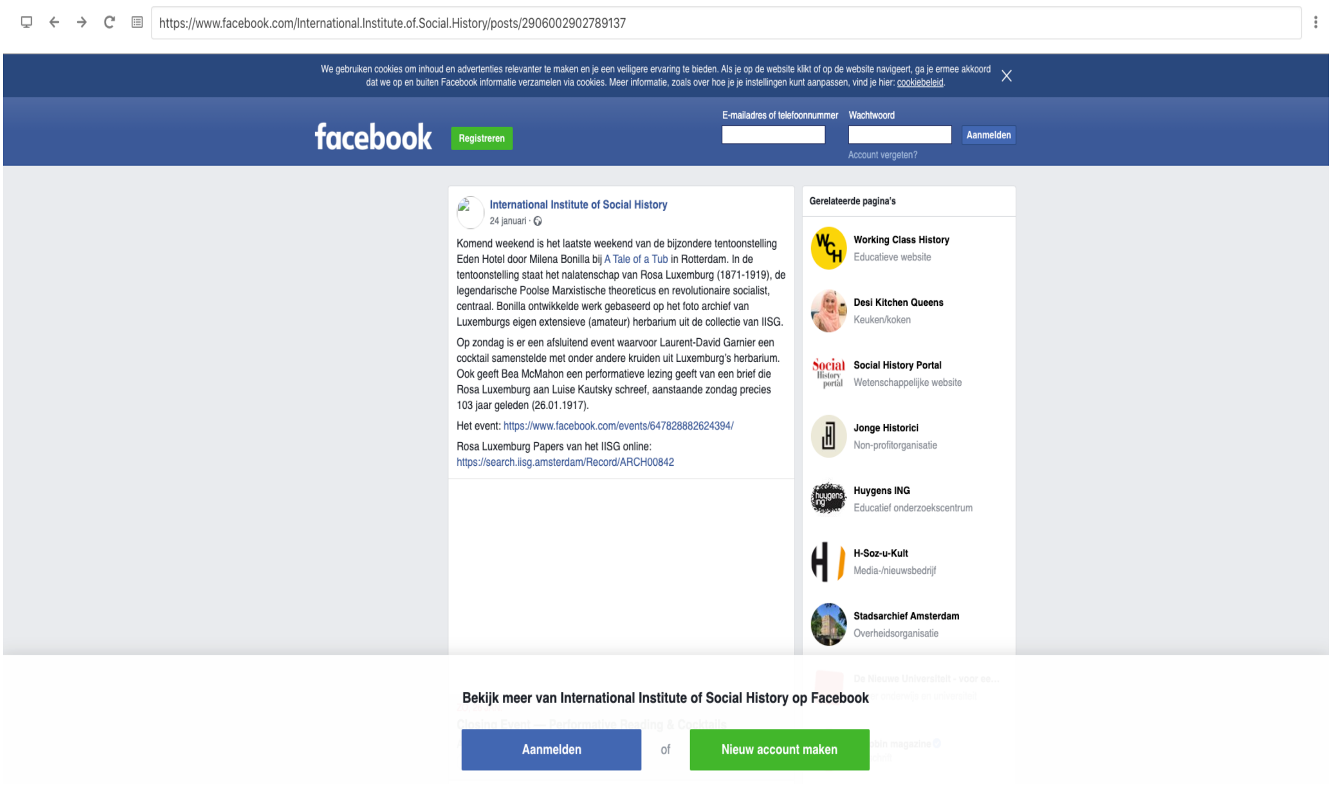
According to the developer,[20] crocoite is able to archive the dynamic web so successfully because it bases its function on picking up the network traffic between the headless browser and the page and using it to reconstruct the URLs to be captured. This means that, in essence, what is captured by crocoite is not necessarily what the website server sent to the client/browser: it is a reconstruction based on the data that crocoite picks up by listening to network events. The reconstruction might more often than not be accurate, at least on the level of the end user browsing an archived page, and for most collection users and archivists it will probably be undetectable as well, unless perhaps they forensically examine the harvested files and compare them against the live website traffic. Nevertheless, it does underline the fact that what we archive when we archive the web is almost never an "original" – it is rather a reconstruction of elements from the content as it was at the time of capture combined with materials necessarily introduced during the archiving process to make it possible (Brügger 2011, pp. 32). This is the case not only with crocoite but with practically any tool we use to capture and reproduce online digital content: strictly speaking, even the automated "behaviours" we must use to programmatically trigger content that requires interaction to be loaded, are in a way an intervention, slightly altering the captured content to make it replayable. It is useful to keep this mind as we begin capturing with any tool.

Finally, crocoite is a good example of a tool arising from the open-source community that could prove problematic to use in a professional setting because of lack of ongoing support.

[20] https://6xq.net/crocoite/rationale/

4.1.4. Munin-Indexer (Munin)[21]

Munin (Munin-Indexer) uses Docker to wrap different scraping and archiving tools together and offer a scraping solution for Facebook, Instagram, and VKontakte. It indexes and scrapes posts, then crawls and captures them, and finally uses pywb to display them. The important thing to note about Munin is that it is only able to archive public posts, i.e., only posts that do not sit behind a log-in. Consequently, this means that it is useful for archiving public Facebook pages, public Facebook groups, or the public posts on a personal Facebook account, but cannot archive private Facebook group content or private posts. Likewise, for Instagram, if the posts belong to an account that is restricted, Munin cannot get to them and archive them.



https://www.facebook.com/International.Institute.of.Social.History/posts/2906002902789137

We gebruiken cookies om inhoud en advertenties relevanter te maken en je een veiligere ervaring te bieden. Als je op de website klikt of op de website navigeert, ga je ermee akkoord dat we op en buiten Facebook informatie verzamelen via cookies. Meer informatie, zoals over hoe je je instellingen kunt aanpassen, vind je hier: cookiebeleid.

International Institute of Social History
24 januari ·

Komend weekend is het laatste weekend van de bijzondere tentoonstelling Eden Hotel door Milena Bonilla bij A Tale of a Tub in Rotterdam. In de tentoonstelling staat het nalatenschap van Rosa Luxemburg (1871-1919), de legendarische Poolse Marxistische theoreticus en revolutionaire socialist, centraal. Bonilla ontwikkelde werk gebaseerd op het foto archief van Luxemburgs eigen extensieve (amateur) herbarium uit de collectie van IISG.

Op zondag is er een afsluitend event waarvoor Laurent-David Garnier een cocktail samenstelde met onder andere kruiden uit Luxemburg's herbarium. Ook geeft Bea McMahon een performatieve lezing geeft van een brief die Rosa Luxemburg aan Luise Kautsky schreef, aanstaande zondag precies 103 jaar geleden (26.01.1917).

Het event: https://www.facebook.com/events/647828882624394/

Rosa Luxemburg Papers van het IISG online:
https://search.iisg.amsterdam/Record/ARCH00842

Gerelateerde pagina's

Working Class History	Educatieve website
Desi Kitchen Queens	Keuken/koken
Social History Portal	Wetenschappelijke website
Jonge Historici	Non-profitorganisatie
Huygens ING	Educatief onderzoekscentrum
H-Soz-u-Kult	Media-/nieuwsbedrijf
Stadsarchief Amsterdam	Overheidsorganisatie

Bekijk meer van International Institute of Social History op Facebook

Figure 3: Preview of WARC containing a single post harvested from the IISH Facebook account

Additionally, Munin is not a scraping tool meant for archiving historical content per se, i.e., it cannot capture posts from days, weeks, or months before the crawl is initiated. Munin is in fact a monitoring tool that, after you have entered a URL, will detect each new post starting from the moment you begin the crawl – posts made before this moment will not be captured.[22] It will then store each post in a separate WARC file. While most of the tools in

[21] https://github.com/peterk/munin-indexer
[22] However, while using it to capture posts from Facebook pages and groups, it was noticed that some content from a few days before was actually captured. The reason for this is unclear, and personal communication

this category create one WARC file for the entire page they crawl, Munin is different in that it creates individual WARCs for each individual post. This is reasonable as, unlike the rest of the look and feel tools we looked at, Munin indexes and scrapes individual public posts, and not entire pages.

Having each post archived in its own WARC might be an issue if you intend to archive a great number of materials, as it can make ingest and storage potentially more complicated. Additionally, if the ultimate goal is to preserve an account or page's look and feel, the approach Munin takes could be said to affect the provenance of the material, and ultimately its authenticity, as it extracts posts from their originating context and presents them as individual pieces of content.

Nevertheless, especially for the notoriously difficult to archive Facebook, Munin could be a viable choice.

4.1.5. Webrecorder Desktop/Conifer[23]

Note: This entry provides information about Webrecorder Desktop, but the basic functionality and characteristics of the app do not differ significantly between the desktop version and the online version (Conifer, formerly known as Webrecorder). Using one or the other will depend on preference, convenience, and sometimes specific needs, as the desktop version does not include as of yet all the simulated browsers that the online version does.

Webrecorder Desktop is the desktop version of the popular Conifer service for user-friendly web page recording. Webrecorder Desktop depends on a series of browsers that access a page and store the content that is loaded. In order to fully capture it though, human intervention is needed: by browsing the page, clicking on links, expanding comments, and playing videos, each of these elements is recorded and included in the archived version of the page. This way, one is able to choose the extent to which they would like to archive a given website e.g., perhaps a particular social media account contains lots of interesting external links that are all deemed worthy of preservation, and another account is only interesting for the main posts it publishes and nothing else, so not much time needs to be spent to record it in its entirety. This tool stores the captured data locally and not online as

with the developer indicated that he is also not certain about the reason for this– one more token of the uncertain nature of social media archiving tools.

[23] https://github.com/webrecorder/webrecorder-desktop, https://conifer.rhizome.org/

Conifer does – this could be important for reasons of security and compliance in an archival organization. At the same time, it allows more flexibility in adding storage space and managing the data.[24]

One of the highlights of Webrecorder Desktop is the presence of a mobile device emulation mode, which allows the archivist to capture the mobile version of a website. This could be valuable for anyone interested to preserve different facets of social media platforms, as opposed to just the desktop website view, since after all more and more social media users prefer mobile apps over using the social media website. Additionally, the presence of lists that can be created to organize captured materials for public viewing could also come in handy if an organization chooses to use Webrecorder Desktop as an access tool e.g., in a reading room. On the other hand, one downside to the way data is organized within the tool is the fact that, while one can create a collection and keep adding new captured pages to it, they cannot delete a URL from the collection if it was captured by mistake – in this case, either the resulting WARC will need to be modified with external tools, or the collection will need to be re-captured.

Having been created with digital art preservation in mind, the manual approach that it takes to harvesting makes sense for Webrecorder Desktop. It is an appropriate approach for the archiving of digital artworks, which are usually relatively finite and/or reasonably sized. But when it comes to social media, it can be problematic to use. How can we capture the many nested comments below an Instagram post, or the threads and sub-threads of a popular tweet? Having to manually click all those is extremely time-consuming, thus the developer team of the Webrecorder Project created the handy feature of Autopilot. Autopilot enables the user to perform an automatic capture of a page, and it includes pre-made “behaviours” that allow it to interact with the website in a similar way as a human user would; however, it is not always reliable and might need to be complemented by manual capturing. There are some caveats when using Autopilot in the online version of Webrecorder,[25] and as of the publishing time of this report, some of the behaviours have been rendered almost wholly unusable due to changes in the social media platforms’ UIs and back-ends.

[24] As with practically all of the tools in this report, Webrecorder Desktop too can be even further tailor-made by building local instances not only to manage storage, but also to modify functionalities.

[25] See the Autopilot guide. https://guide.conifer.rhizome.org/docs/autopilot/

This is one of the most pervasive issues in the practice of social media archiving, that repeatedly came up during the course of the research: platforms will change their design and make tools temporarily or even permanently unusable. On the bright side, there is a lively community behind Webrecorder Desktop and Conifer that is actively supporting the tools and performing the necessary upkeep. Even without automation wholly available, both Conifer and Webrecorder Desktop still allow for reasonably reliable capturing of the original look and feel of social media pages. Of course, it is still true that issues can arise even with a user-friendly tool like this: during the testing phase, and due to the aforementioned issues with changing social media platforms, we encountered various problems e.g., with loading Twitter pages, capturing Instagram accounts without being logged into Instagram, etc. Solving them was a matter of trying out some solutions, or in many cases, simply waiting for the developers to attempt to address the issue. This is not reliable of course, and it must be taken into account when using community-supported, open-source social media archiving tools in an institutional setting.

4.2. Structured Data Output

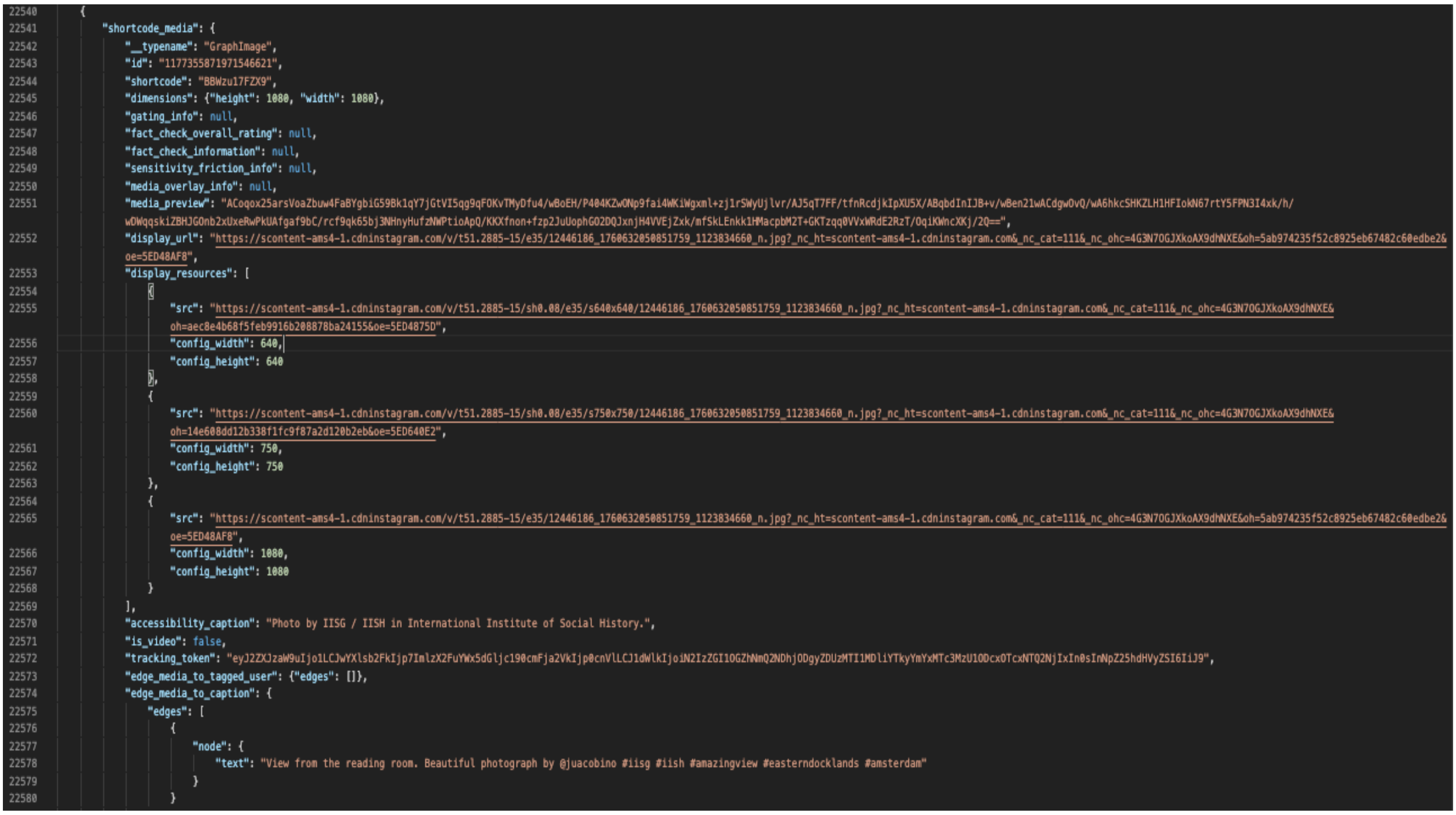
4.2.1. Instamancer[26]

Instamancer, as its name indicates, is a tool specialized in Instagram capturing. It does not, however, connect directly to the Instagram API, as most other similar tools do. It uses the Chromium browser and, instead of attempting to mimic user behaviour on that browser, like the tools examined above, Instamancer's Chromium intercepts the traffic between the Instagram URL that the user has provided as a seed, and the API that provides the data that are presented on that URL. For example, one would like to harvest all posts by the International Institute of Social History (https://www.instagram.com/iisg_amsterdam/), Instamancer would not need to create an app on the Instagram API and connect to it.[27] It just listens for requests that its browser makes to the API to grab the data it needs in order to load the page to the users, and captures the data that is sent from the API in response to these requests. This makes it easier to harvest API data because, in essence, it does not require the crawler to even connect directly to that API.

[26] https://github.com/ScriptSmith/instamancer

[27] That is, after all, quite difficult at the moment as the Instagram Graph API is built on the Facebook API and is governed by the same restrictive policies that the Facebook API is.

The output of Instamancer consists of two kinds: structured data files and extracted media content. The data files are in JSON and CSV format and include information like post and image URL, time of posting, etc. The other kind of output Instamancer may produce is the actual posted content on Instagram, e.g., images and videos, including whole albums of those. This is a useful feature that might come in handy if capturing the actual look and feel of the Instagram content is not possible or desirable.



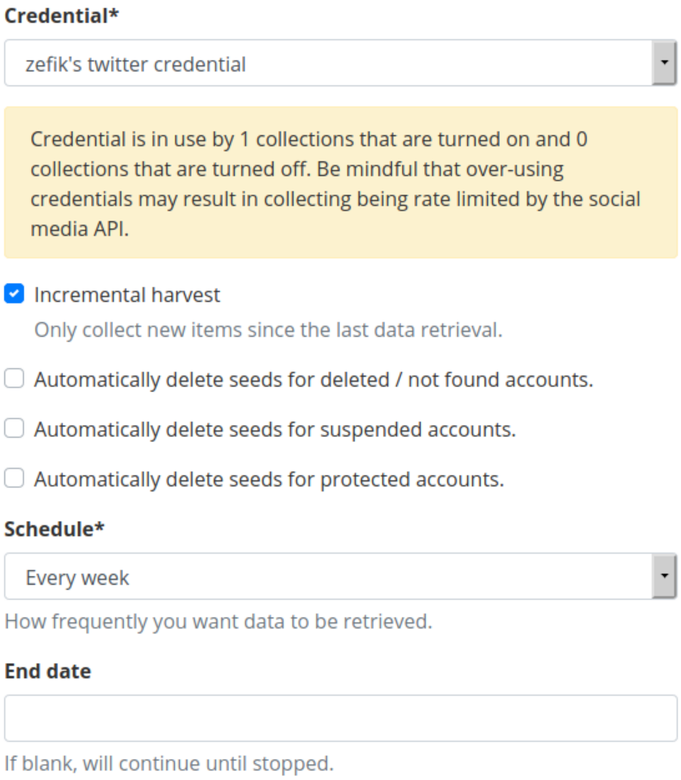
```
{
    "shortcode_media": {
        "__typename": "GraphImage",
        "id": "1177355871971546621",
        "shortcode": "BBWzu17FZX9",
        "dimensions": {"height": 1080, "width": 1080},
        "gating_info": null,
        "fact_check_overall_rating": null,
        "fact_check_information": null,
        "sensitivity_friction_info": null,
        "media_overlay_info": null,
        "media_preview": "ACoqox25arsVoaZbuw4FaBYgbiG59Bk1qY7jGtVI5qg9qFOKvTMyDfu4/wBoEH/P404KZwONp9fai4WKiWgxml+zj1rSWyUjlvr/AJ5qT7FF/tfnRcdjkIpXU5X/ABqbdInIJB+v/wBen21wACdgwOvQ/wA6hkcSHKZLH1HFIokN67rtY5FPN3I4xk/h/
        wDWqqskiZBHJGOnb2xUxeRwPkUAfgaf9bC/rcf9qk65bj3NHnyHufzNWPtioApQ/KKXfnon+fzp2JuophGO2DQJxnjH4VVEjZxk/mfSkLEnkk1HMacZ1fIzA1wPOBoSqXwdwVv9U/Sjs7sjsSUyMUY+Gk0AY5XZTrV9xpkjDSuCYTkwsIoPDnqmhqn8qFfGTFcH1bAafSqOtfyk3bH6yWnk//2Q==",
        "display_url": "https://scontent-ams4-1.cdninstagram.com/v/t51.2885-15/e35/12446186_1760632050851759_1123834660_n.jpg?_nc_ht=scontent-ams4-1.cdninstagram.com&_nc_cat=111&_nc_ohc=4G3N7OGJXkoAX9dhNXE&oh=5ab974235f52c8925eb67482c60edbe2&
        oe=5ED48AF8",
        "display_resources": [
            {
                "src": "https://scontent-ams4-1.cdninstagram.com/v/t51.2885-15/sh0.08/e35/s640x640/12446186_1760632050851759_1123834660_n.jpg?_nc_ht=scontent-ams4-1.cdninstagram.com&_nc_cat=111&_nc_ohc=4G3N7OGJXkoAX9dhNXE&
                oh=aec8e4b68f5feb9916b208878ba24155&oe=5ED4875D",
                "config_width": 640,
                "config_height": 640
            },
            {
                "src": "https://scontent-ams4-1.cdninstagram.com/v/t51.2885-15/sh0.08/e35/s750x750/12446186_1760632050851759_1123834660_n.jpg?_nc_ht=scontent-ams4-1.cdninstagram.com&_nc_cat=111&_nc_ohc=4G3N7OGJXkoAX9dhNXE&
                oh=14e608dd12b338f1fc9f87a2d120b2eb&oe=5ED640E2",
                "config_width": 750,
                "config_height": 750
            },
            {
                "src": "https://scontent-ams4-1.cdninstagram.com/v/t51.2885-15/e35/12446186_1760632050851759_1123834660_n.jpg?_nc_ht=scontent-ams4-1.cdninstagram.com&_nc_cat=111&_nc_ohc=4G3N7OGJXkoAX9dhNXE&oh=5ab974235f52c8925eb67482c60edbe2&
                oe=5ED48AF8",
                "config_width": 1080,
                "config_height": 1080
            }
        ],
        "accessibility_caption": "Photo by IISG / IISH in International Institute of Social History.",
        "is_video": false,
        "tracking_token": "eyJ2ZXJzaW9uIjo1LCJwYXlsb2FkIjp7ImlzX2FuYWx5dGljc190cmFja2VkIjp0cnVlLCJ1dWlkIjoiN2IzZGI1OGZhNmQ2NDhjODgyZDUzMTI1MDliYTkyYmYxMTc3MzU1ODcxOTcxNTQ2NjIxIn0sInNpZ25hdHVyZSI6IiJ9",
        "edge_media_to_tagged_user": {"edges": []},
        "edge_media_to_caption": {
            "edges": [
                {
                    "node": {
                        "text": "View from the reading room. Beautiful photograph by @juacobino #iisg #iish #amazingview #easterndocklands #amsterdam"
                    }
                }
```

Figure 4: Entry for a post by the IISH Instagram account captured in JSON with Instamancer

Instamancer does not require API credentials to be used, which is convenient for ease of use. At the same time, recent (at the time of writing) developments in the Instagram design have made Instamancer's scraping capabilities slightly less reliable: while there is no certain explanation, the reason seems to be that Instagram detects requests made by cloud-hosted browsers, like Instamancer's, and blocks them to deter scraping of its content. This is a common obstacle in social media archiving and it requires a degree of alertness and even improvisation to deal with the sudden loss of a valuable tool. Nevertheless, as of November 2020, Instamancer was able to harvest posts by public Instagram accounts, with the caveat being that different crawl sessions returned different numbers of posts, even when using the same query.

4.2.2. Social Feed Manager (SFM)[28]

Social Feed Manager is an application created by George Washington University Libraries to assist in archiving social media content via APIs. The difference with the rest of the API-based tools examined here is that it does so by providing a GUI, which allows the archivist to control the parameters of each capturing session quite easier. The ultimate usability of its interface is one of its strengths, as well as its ability to schedule capture sessions to happen automatically. Other features include the ability to harvest Twitter incrementally e.g., only the new tweets since the last capture, and to automatically detect if there are seeds whose tweets have been deleted so as to delete them as well, in order to comply with Twitter's policies.



Credential*

zefik's twitter credential

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

☑ Incremental harvest

Only collect new items since the last data retrieval.

□ Automatically delete seeds for deleted / not found accounts.

□ Automatically delete seeds for suspended accounts.

□ Automatically delete seeds for protected accounts.

Schedule*

Every week

How frequently you want data to be retrieved.

End date

If blank, will continue until stopped.

Figure 5: Tweet collection configuration on SFM

SFM can capture data from the Tumblr, Twitter, Flickr, and Sina Weibo APIs. To do this, it combines various modules, e.g., twarc and others, and orchestrates them using Docker. In essence it manages, as its name suggests, the access and retrieval process of structured social media data by storing and using the required API credentials for each social media platform. These API credentials can be obtained relatively easily in many cases, with the important exception being Facebook and Instagram, which require users to authenticate

[28] https://gwu-libraries.github.io/sfm-ui/

themselves and provide all kinds of documentation and even a business case for why they want to access the API. As such, these two platforms are relatively inaccessible for archivists and researchers. Twitter and other platforms make it slightly easier. Twitter, for example, requires one to apply for a Twitter Developer account, but applications that request access for research and educational purposes get granted almost automatically. In order to use SFM, one needs to first apply for and obtain the required API credentials.

SFM allows control not only of the captures, but also of their documentation. SFM automatically tracks activities such as creation of new collections, adding of seeds, requests to APIs, so that a record can be kept to trace back the creation of the collection – this is a powerful feature that can be very useful in an organizational setting. Some descriptive information about the harvests can also be added manually, and all of it can be exported later to be used as part of the collection's provenance.

The collected datasets can be exported in a variety of file formats, including JSON and CSV. Additionally, and to aid transferability and sustainability, the datasets are all stored in WARC files, out of which the JSONs and others are exported. These WARC files are not like the ones mentioned earlier, containing the look and feel view of the seeds; in this case, the WARCs are used literally as containers, to store the collected structured data and allow it to be easily transferred between different instances of SFM or between different systems, as well as to remain readable.

SFM can be really useful, but it does take a considerable amount of technical skill to install, set up and implement in production, which is why it might be a more attractive choice for organizations that can afford spending time and resources on this. Having experimented with earlier versions of the tool as well, we observed that its usability and stability has improved, but it remains a more demanding choice than others. Additionally, improper configuration could result in a lot of unwanted data that will crowd your storage, as it happened to us. Thus, it could be useful to spend time to familiarize oneself with the specifics of this tool if it going to be used.

4.2.3. TAGS[29]

It is safe to say that most of the tools that output structured data are not the easiest or most intuitive to use. One notable exception then is TAGS (Twitter Archiving Google Sheet). TAGS

[29] https://tags.hawksey.info/

is in essence an app built on Google Sheets, that uses the Twitter API to fetch structured data based on queries the user inputs in the spreadsheet. TAGS makes use of an already authenticated Twitter API app for its operation, but you are able to use your own Twitter API app if you prefer. The strongest point of the tool is definitely its user-friendliness, which only requires one to log into their Google Drive, open up the TAGS spreadsheet, fill in their search query, and wait for the captured data to be downloaded.

The usual restrictions of Twitter API usage apply, e.g., rate limiting and a 7-days-in-the-past window for capturing older tweets, but all in all, the tool works very smoothly. It was tested to capture an individual user's tweets from their timeline, as well as tweets based on keyword searches e.g., "Amsterdam," "#coronavirus" and others. The tool also allows you to harvest all of a user's favourited tweets.

TAGS can be configured to capture tweets for extended periods of time, and does not require monitoring or even your machine to be on, like most tools mentioned above do. Plus, a neat extra are the Summary and Dashboard tabs, that allow you to inspect the content you harvested in graphs and numbers e.g., how many unique tweets vs. tweets in total, number of links, the popularity of a particular term over the harvesting period, etc. These features could come in handy for performing an initial appraisal of the harvested content and determining whether it is suitable for preservation, or if additions and/or filtering are needed.

However, ease of use comes at the expense of flexibility, as TAGS is not as granular and configurable in its search and crawling capabilities as SFM or twarc. Nevertheless, it is definitely recommended for starters, and possibly for use in educational projects involving web and social media archiving trainees, donors, and collaborators that are not (yet) comfortable with using more technically demanding tools. Another important note to be made about this tool is that it does not seem to be actively developed and/or maintained –as such, there is no telling how long its viability will be. Nevertheless, similar tools could be custom-made using apps in Google Cloud or other providers, or on an organization's own infrastructure.

4.2.4. Twarc[30]

Twarc is another instance of a community-developed tool that has enjoyed wide recognition from professional circles as well. Developed by the Documenting the Now (DocNow) group,

[30] https://github.com/DocNow/twarc

which is active in promoting the ethical collection and use of social media data, twarc is a command-line tool for the capturing of Twitter API data. Within the framework of limitations posed by Twitter, it is, in fact, quite versatile. One can use twarc as a standalone tool on any OS, or even as a library when building other tools. Like SFM, twarc requires you to create a Twitter Developer account and register an application on the Twitter API. After you authorize twarc to use this application, it can connect to the API and pull the requested data in JSON format.

The tool is able to collect data based on search queries e.g., a hashtag or a keyword, but it can also collect entire user timelines, follower lists, retweets, and others, while it can also filter the results based on location (useful for collections focusing on a specific city, region, or other locale). Learning how to form appropriate and more complex queries takes some practice,[31] but it allows for a lot of control and fine-tuning of captures.

All the functionality mentioned above refers primarily to using the Twitter Standard API, which can be accessed for free; if you are looking for historical data older than 7 days before, you could purchase access to the Premium Search API. Nevertheless, whichever API service you use to collect your Twitter data, the important limitation is that they cannot be published in their current state. Twitter API terms of use do not allow for the tweet full text to be shared without the platform's permission, but rather only the Tweet, Direct Message, and User IDs of each element in the collection. Because of this, twarc allows you to de-hydrate collected datasets, i.e., transform them into an ID set, as well as re-hydrate them, i.e., use the ID set to retrieve the equivalent tweets back from the API. This can be done via the command-line, but DocNow have also made a handy tool with an interface specifically for tweet hydration called Hydrator.[32]

While twarc is quite barebones in terms of usability features, or functionalities like scheduling and separate profiles for different users, its reliability makes it an attractive choice for those interested in Twitter API data. During this research project, we used it to collect various datasets based on keywords and hashtags related to the Coronavirus pandemic, as well as personal timelines and follower lists, without any hiccup. The lively DocNow

[31] Refer to twarc's GitHub page for some instructions, as well as to this tutorial (please be aware some information could be outdated): https://github.com/alblaine/twarc-tutorial

[32] https://github.com/DocNow/hydrator

community behind it is an additional reason why this tool should be seriously considered, as it is a possible indicator of its sustainability.

5. Tool Use Cases

A list of tools and their characteristics is handy, but where does one start? In order to accommodate both those who are already on the path of creating their social media archiving workflows, but also those who are still looking for ideas and ways to implement these tools in their situations, we are providing a small number of use cases in which the tools described above could be used. We initially considered the creation of a few user guides, but the rapid pace in which new tools are created and existing ones are going out of use have led us to decide not to create tool-specific guidelines, as these would surely become near-obsolete very soon.

The use cases are illustrative and they are based on the author's experience of testing the tools and working in the environment of a cultural heritage and research organization like IISH. As mentioned earlier, social media archiving tool selection is highly context-dependent. Care was taken to consider different aspects of using these tools in different contexts, but it is unavoidable that some aspects were excluded and others magnified because of IISH's specific circumstances.

5.1. Archiving a current socio-political event

In early 2019, the Constitutional Court in Thailand decided that the Thai Raksa Chart Party, a political party which had recently announced that its candidate for Prime Minister would be Princess Ubolratana, had to be dissolved ("Constitutional Court Disbands Thai Raksa Chart," 2019). The reason given was that members of the royal family, even those that had officially relinquished their title like the Princess, could not be allowed to enter into governmental positions. The decision indeed resulted in the party's dissolution, which was also accompanied by a requirement to immediately deactivate its official communication channels, e.g., website and social media.

As part of our collecting activities on social and political issues in Southeast Asia, we decided to respond as quickly as possible and capture the official online presence of the Raksa Chart party. This included a website, a Twitter account, and a Facebook account. The website was captured using the Web Curator Tool (WCT), a web archiving tool meant for small and large-scale web harvesting.[33] As WCT does not fare especially well with capturing

[33] https://webcuratortool.org/

social media content, other tools were selected to capture the Twitter and Facebook pages, according to the following criteria:

- Ease of use to allow capturing as fast as possible because of the time-sensitive nature of the acquisition – considering we had limited staff capacity to devote to this, it was deemed important to make the best use possible of their time
- Completeness of the captured content, both in terms of including all elements i.e., images, videos, comments, likes, retweets, shares, as well as in terms of including the entirety of the content posted by each account without gaps or jumps[34]
- Integrity of the collections created because of the sensitive and possibly controversial nature of the captured materials, i.e., if called for, we wanted to be able to guarantee that the collected materials were reliable as evidence by making sure they remained unchanged and secure

To fulfil the above criteria, the online Webrecorder tool (now Conifer) was selected. Being the most easy-to-use out of the tools we had available, it allowed us to begin collecting immediately. Webrecorder allowed us to acquire as much of the surrounding environment (layout, navigation tabs, etc.) as well as the embedded social media content as possible. For Twitter, it also ensured that we would be able, in theory at least, to capture the entire account feed from the present up to its creation (something not possible with API-based harvesting of Twitter accounts, which only captures a limited number of tweets and not those older than 7 days); for Facebook, which has severely limited its API accessibility, Webrecorder was virtually the only readily available and straightforward solution. And for reasons for integrity and reliability, Webrecorder also seemed to be a good choice, because it did not involve the Twitter API and its limitations of sharing and re-using data via Tweet IDs, that could force us to have our entire collection practically disappear when the Raksa Chart Twitter account was deactivated.

At that point in time, the handy Autopilot feature in Webrecorder had not yet been released, thus the process of recording the pages would necessarily have to be entirely

[34] In a previous experiment with Webrecorder, we encountered a case where a Facebook page captured on Autopilot appeared complete in the WARC, but when replayed there was some content that was completely missing when scrolling down the page, content that was actually observed to be harvested successfully during the capture. This might have been a file rendering issue, but in any case, we wanted to be sure that the Raksa Chart content would be captured in its entirety because if the pages would be taken offline, our copy could be one of the few remaining ones.

manual. The process of manually recording a Twitter and a Facebook feed was definitely time-consuming and would definitely not scale if the target collections were larger; for this particular use case though, we estimated that the time we would invest in capturing the content was worthwhile, considering its potential uniqueness and fit with our collection policy.

In more than one cases, the process had to be restarted due to the tab running Webrecorder crashing. This happened when we were running Webrecorder in a minimized window, or simultaneously with other processes on our browser, or when the content being recorded was simply a lot. The endless scrolling functionality that most social media platforms utilize can be especially taxing on computer memory, and as Webrecorder needs to take advantage of this functionality to access older content on a social media feed, it might cause malfunctions on the browser. This is again one more reason that Webrecorder does not easily scale to more large-scale archiving efforts, for which one would have better luck using tools such as Browsertrix, Brozzler, crocoite, and others.

Additionally, Webrecorder without Autopilot required the user to manually click on comments and replies to expand them and include them in the capture; this is something that was deemed too time-consuming even for the modest amount of data gathered for this use case, thus our manual capture of these accounts does not include comments and replies. We decided instead to include videos, which entailed playing them in order to include them in the capture. Images were captured as part of the recording process without effort on our part, yet a few were missing when we replayed the WARC files afterwards.

Regarding documentation, we kept a record of the seed URLs, the capturing times, and the Webrecorder version used, as well as the names of each archivist involved in the process. We used a simple form to do this, to which we added details as needed: duration of recordings, possible errors and crashes, QA review results e.g., missing media, possible access restrictions, etc. The ultimate goal was to document all of these details to later transfer them to an archival information management system e.g., ArchivesSpace, that would allow us to track the provenance of the collections effectively.

5.2. Archiving an individual's or an organization's personal social media account

The preservation of personal accounts, either belonging to individuals (citizens, public figures) or to organizations (museums, governmental agencies, non-profit organizations, businesses), is a relevant use case for the majority of cultural heritage institutions interested in social media archiving. Different purposes tied to different organizational mandates will determine tool selection and usage, but in this use case, the steps that IISH took as a research and collecting institution are presented.

Some considerations we made before selecting which tool to use to archive personal social media accounts were:

- Are we more interested to preserve the “look and feel” of the account, e.g., layout, friend suggestions, trending topics, ads, or would API-based collecting of structured data satisfy our collection requirements?
- Is API harvesting reasonably doable for the social media account we want to preserve, i.e., is there an API and can we use it?
- Do we have permission to capture the account, and to give access to the collections later?

It has been mentioned frequently in this report, but one of our most significant findings is that the purpose of the collecting activity should determine tool selection. If the purpose is compliance and accountability, the appropriate tool to capture social media might be different than if the purpose is preserving heritage and memory. For our use case, we focused on our capture of Leo Lucassen’s Twitter account. Leo Lucassen is a well-known Dutch public intellectual and political commentator with a lively social media presence. Preserving his Twitter account is part of the collecting policy of IISH on social and political developments in the Netherlands, as well as part of the institution’s own corporate memory, since Leo Lucassen, currently the IISH director, was then the IISH research director. Thus, both accountability and memory would be the aims of this collecting activity.

Therefore, we decided that we wanted to preserve both the look and feel of the account and a structured dataset. We used Webrecorder Desktop, the desktop counterpart to Conifer. This allowed us more freedom in terms of storage, compared to the more limited capacity of the free online account. We were able to employ Autopilot, which reduced the amount of manual labour significantly. We faced another issue however, and this issue is quite illustrative of the problems that might arise when doing social media archiving. Before we began recording Leo Lucassen’s account, we performed a number of test captures of the latest tweets posted on the account. No problem arose during these tests, so a few days later we moved on with the full capture. When the recording of the account was completed, we viewed the WARC files using Webrecorder Player and discovered that instead of the Twitter account, all we could see was a white page with the Twitter logo on it and the message “Something went wrong – try again.” This issue went on for a couple of days and after consulting with colleagues in other organizations, it seems that it was not exclusively a problem on our side. Communication with representatives of the Webrecorder project indicated that Twitter itself could be limiting the amount of data we were able to capture through rate-limiting. The fact that the issue was resolved a few days later, and we were then able to capture an almost full copy of the account, makes it more difficult to know exactly what has gone wrong but indicates the kind of contingencies that web and social media archivists are continuously dealing with.

Additionally, this incident highlights another significant aspect of social media archiving work, which is the need to constantly keep track of the progress of the workflow in order to document it. In our case, when we first discovered the non-functional captures of Leo Lucassen’s account, we decided to at least hold onto the test captures we had made earlier as samples; they were not what we were going for, but they were better than nothing. Had we not been able to capture a full version of the account, we could have kept these files, document that they were the result of a test capture, as well as that a complete capture was not possible and reasons why. Especially for content that might become unavailable by the time we are able to capture it properly (like the Raksa Chart accounts above), such documentation of harvesting processes could be very useful.

For our capture of structured data, we used twarc. The Twitter API is fairly accessible, if one creates an account for educational or research purposes, which is what we did. Capturing a user’s Twitter timeline with twarc can be achieved with a command like `$ twarc timeline Leolucassen > leolucassen_twitter.json` which captures approximately 3000 of the most recent tweets by Leo Lucassen and stores them in a JSON file. We also used the filter command that connects to the Statuses/Filter Twitter API to capture tweets by a Twitter user in real time, as opposed to tweets made until the moment of capture. We used this option to monitor Leo Lucassen’s account for new tweets for about two weeks – this was an experimental implementation, but we foresee that we could make use of this functionality to monitor Twitter accounts for new tweets in a longer time frame.

We then dehydrated the tweet collection to be able to share it with third parties outside the IISH if needed, and next to technical details and the rationale of the capture, we also documented the version and terms of use of the Twitter API at the time of capture to keep more information about the collection's context. This was deemed necessary especially since we predicted we would not be able to offer access to the collection for the foreseeable future, because of the API limitations, but also because of the presence of lots of other people's data in the collection, who had commented on Lucassen's tweets but had not given their consent to be included in the capture.

5.3. Archiving community social media

One of the most important use cases for IISH is preserving the social media presence of grassroots organizations, ad-hoc groups, and generally of any kind of community formation that is relevant for our social history collections. For a cultural heritage organization that aims to preserve the activity and presence of contemporary social movements online, archiving their social media is paramount.

In this use case, we opted to experiment with two different kinds of capturing activities: the first one was to attempt to harvest copies of the Facebook and Twitter account of a grassroots organization that had recently decided to deposit some of its materials with us. The second was to create a curated collection of various social media accounts, mainly Facebook, Twitter, and YouTube, related to a popular social and political movement.[35] In the latter case, the accounts belonged to various groups, collectives, and sometimes also individuals, involved with the movement, and they also contained a number of Facebook groups (some public, some closed), as well as the Twitter feeds of several hashtags. All the Twitter and Facebook accounts, Facebook groups, YouTube channels and Twitter feeds contained a large number of images, videos, and links, that were important to be preserved if the purpose and context of this content was to be safeguarded. Additionally, in both cases, the languages that the social media users in these accounts and pages used were written in Arabic script.

As we started out planning the captures, the considerations were to a large extent same as those for capturing individual accounts, i.e., figuring out our specific purpose for these collections, deciding whether specific aspects of the look and feel and/or specific API-

[35] Because of the sensitive nature of the materials collected, and the potential risks associated with exposing information about groups, parties, and organizations operating in national and international contexts, we decided to retain the anonymity of our creators/collaborators. The grassroots organization will remain unnamed throughout the report and will hereby be referred to as "political group", while the creators of the materials of the social and political movement collection will be collectively referred to as "social movement."

based data were the most appropriate to fulfil our purposes, and determining whether all the content we wanted was available via API at the time we wanted to start the capture. We also had to think about:

- How long in the past should we capture the hashtag feeds (when using a tool that allows us to capture older content, like for example most of the "look and feel" tools)?
- Do we need to include nested Facebook comments, Twitter sub-threads, and YouTube channel comments in our captures?
- How do we gain access to closed Facebook groups and should we even capture content in these closed Facebook groups without permission?
- How do we make sure that most of the media content (images, videos, links) is captured, and how do we scope the captures, especially with regard to including external links?
- Especially for YouTube, is extracting the videos as stand-alone files enough, and if not, with which criteria should we determine to also preserve a copy of each video's page?
- How are the capturing tools going to handle non-Latin scripts?
- What is our relationship with the account owner(s)? Would it be possible to request access to copies of their account data, if other capture options are not preferable or possible?

Responding to these considerations was an interesting thought experiment, as we tried to address most of them before we started the harvesting process. Especially the fact that explicitly restricted content like closed Facebook groups was included in our collection plan led us to consider carefully what the best steps would be to create collections, but also to protect the social media users and IISH itself.

Firstly, we decided we would need to clarify our intentions and the intentions of the political group and the social movement about the archiving of their materials. We wanted to create collections that would be useful for researchers, but also to respect the rights and wishes of the creators themselves. This was important if we were to create inclusive and balanced collections. In the first case, we had explicit communication with the political group, which only expressed the desire for the collection to be kept restricted unless they give access permission. The social movement did not express any particular wishes directly, as we did not have communication with representatives of the groups the social media pages belonged to – our communication was with a curator who appraised and selected content to be included in the collection as a specialist. Through the curator's input, we were able to determine which pages, channels, etc. were more sensitive and had to be restricted fully, and which could potentially be accessible – even though, as in the use cases described above, we knew we would not be giving access to the collections in the near future anyway.

For the political group's Facebook, we had to choose the look and feel approach, since accessing the Facebook API was not possible and the capture was small enough. However, an added complication was that the Facebook page we wanted to preserve had at that moment already been un-published, i.e., it had been made invisible to all Facebook users but to its owners. We made arrangements with the admins to securely receive their credentials and access the page to capture it, but this was definitely something that had to be carefully negotiated, planned, and validated with a donor agreement: we had to let them know immediately when we would have completed the capture, so that they would revoke our access and change their credentials. This goes to show that relationships of trust and mutual respect can go a long way to making social media archiving possible.

We used Webrecorder Desktop to capture the Facebook page, and since it was only one, we captured it manually to get it in the highest quality possible. A compromise we had to make was in the capture of the Photos tab of the page. Because we did not click to enlarge every image, but only browsed through each album, what was captured was the photo previews but not the full images themselves. During the quality assurance process, when we clicked on an image of the harvested page to view it, it would just load indefinitely. This is how we discovered that it is not enough to browse the albums, but that each photo needs to be opened separately if it is going to be included. In the end, we decided that having access to the image previews, considering that the rest of the page was fully captured, was acceptable. We also decided that all external links of the home page were in scope, which meant they had to be clicked one by one in order to be captured – this was definitely a painstaking process that cannot happen regularly, but it was deemed acceptable because of the uniqueness of the material.

Webrecorder Desktop also presented us with an issue we had not anticipated, when we attempted to capture the Twitter account: because its URL contained Arabic characters, the tool was not able to display the page correctly. The URL and the Arabic text in the Twitter page itself appeared as garbled text. The solution was to enter the URL first in the Twitter search bar, find the page, navigate to it, and then start the capture. Communication with the Webrecorder team confirmed that the problem was indeed caused by a bug in the software that they were working to address, and it was indeed fixed in future versions.

For the social movement, because the content was quite a lot, we chose to preserve the look and feel of the Facebook and Twitter accounts, mainly due to the fact that, on the one hand, we could not access the Facebook API, and on the other hand the Twitter content went far enough in the past that the 7-day access window offered by the Twitter Search API would have not been enough. We used Brozzler to capture Twitter accounts and Facebook pages, but had to spend considerable time tweaking our configuration YAML files to manage quality captures of the Facebook content that could be scrolled down, included videos, etc. – although we still had to make do with several files that were simply less functional compared to the live counterpart at the moment of capture. We decided to limit the capture only to the same domain, thus no external links outside the social media platforms would be followed; this was deemed necessary because we could end up with massive amounts of harvested content – this was a trade-off between potentially preserving less of the context of the social media accounts, and preserving content that could potentially be trivial, duplicate, or potentially harmful (copyrighted materials, multiple reposts of the same links, etc.). A potential solution to this problem could be to use social media monitoring tools to scan relevant accounts and hashtags before harvesting in order to get an idea of the content circulating in them.

We used youtube-dl to download all the videos from the YouTube channels that we were collecting, and we included the downloaded metadata in JSON with our captured videos. After consulting with our curator, we decided that extracting only the videos was enough for most of the channels, because in most cases the comments were simply expressions of endorsement or agreement with the video, usually in one word or an emoji. We could document these without much difficulty, and instead we opted to do full captures of only one YouTube channel which contained lots of substantial discussions in the comment sections of its videos.

Finally, after a series of consultations with the curator and some legal advice, we chose not to go forward with capturing the private Facebook groups that were included in the acquisition plan for the social movement collection. We performed a couple of test captures of them with crocoite, Brozzler, and Webrecorder Desktop to compare results, but ultimately, we deleted these test harvests; after all, we were not able to capture the full content of these groups because would need to be explicitly invited as members to most of them to access them anyway. For those groups that we did not need to be invited to, we also decided that capturing without permission or notification was deemed a risk for the IISH and the people whose data we would be collecting. For now, we have only documented some information about these Facebook groups and we will be looking at further options in the future.

6. Conclusion and recommendations

This report has attempted to give professionals in cultural heritage organizations that are now starting to archive social media some pointers regarding the tools that are available to capture this type of content. At the same time, it has attempted to combine this tool survey with a more holistic overview of the issues surrounding the acquisition of social media as archival collections, e.g., when it comes to ensuring their authenticity, reliability, integrity, and usability, determining designated communities, and laying out features of social media platforms, but also of the available tools themselves, that could prove challenging for a social media archiving workflow.

One important observation that came out of this research is that a document like this, and any resource attempting to focus on specific tools or techniques, is doomed to become outdated very fast. It is no exaggeration to say that significant parts of this report could become obsolete in a matter of months, and even during the course of the project we had to modify the testing schedule or our use cases because of new tool features, or because changes made some tools non-functional. Still, this report is hopefully a living document that we can modify and extend in the future to address changes in the landscape of social media archiving tools.

Nevertheless, and in place of a conclusion, we want to close with a brief series of recommendations that pertain to choosing social media archiving tools specifically, but also to implementing a workflow for capturing social media more generally. After all, one of the outcomes of our experience in this research project is the realization that even though social media archiving practice is varied and context-specific, it should ideally be as closely integrated with the rest of an organization's processes as possible, in order to benefit from the systematic and coherent workflows already in place that will balance out its fluid nature. For this reason, our recommendations are not strictly tool-specific, but they definitely have bearing on the selection of tools and capturing approaches.

1. Determine purpose of social media archiving early and plan according to your goals, in order to choose the appropriate tools and approaches
2. Design or modify documentation processes to record the steps taken during capturing, because otherwise important information might be lost

3. Consider the legal and ethical dimensions of social media archiving and, if applicable, invest in relationships with social media users that will give you access to content, because sometimes capturing with tools might not be possible
4. Be flexible and open to challenges and learning new skills or modifying existing and well-established processes, because the existing methods available of archiving social media do not always fit easily with accepted archival practice and principles

7. Tool Matrix

Tools are listed in order of appearance in the report.

TOOL	Output format(s)	Media extraction	Capturing password-protected content	Capture logs and metadata*	GUI	Capture scheduling**
Browsertrix	WARC	✗	✓	Logs	✓	✗
Brozzler	WARC	✗	✓	Metadata	✓	✗
crocoite	WARC	✗	✗	Logs	✗	✗
Munin	WARC	✗	✗	✗	✓	✗
Webrecorder Desktop	WARC	✗	✓	✗	✓	✗
Instamancer	JSON, CSV	✓	✗	Logs and post metadata	✗	✗
SFM	JSON, CSV, TSV, XLSX, Dehydrated IDs list, WARC***	✗	✗	Logs and post metadata	✓	✓
TAGS	XLSX, CSV, TSV, PDF	✗	✗	Post metadata	✓	✓
twarc	JSON, JSONL	✗	✗	Logs and post metadata	✗	✗

* By capture logs and metadata, we refer here to logs and metadata that document the capturing process itself to allow the archivist to track it. By “post metadata,’ we refer to metadata relating to the actual social media content itself. By definition, all API-based collections have this sort of metadata.

** Capture scheduling refers to the ability to schedule captures to begin automatically in the future, and to recur automatically. Tools that can be left to run indefinitely are not considered to enable scheduling.
*** As mentioned in the SFM entry, this is not a WARC which contains the crawled social media page. It rather contains the API data all aggregated in one container for ease of storage and transfer.

References

Ben-David, Anat. “Counter-archiving Facebook.” *European Journal of Communication* 35, no. 3, pp. 249-264. SAGE, 2020. https://doi.org/10.1177/0267323120922069

Brügger, Niels. “Web Archiving: Between Past, Present, and Future.” *The Handbook of Internet Studies*, ed. Mia Consalvo and Charles Ess, pp. 24–42. Blackwell, 2011.

Bruns, Axel. “After the ‘APIcalypse’: Social media platforms and their fight against critical scholarly research.” *Information, Communication & Society* 22, no. 11, pp. 1544-1566. Taylor & Francis, 2019. https://doi.org/10.1080/1369118X.2019.1637447

Chung, Lawrence, Brian A. Nixon, Eric Yu, and John Mylopoulos*. Non-Functional Requirements in Software Engineering*. Vol. 5. Springer Science & Business Media, 2012.

“Constitutional Court Disbands Thai Raksa Chart.” *Bangkok Post*. 7 March 2019. https://www.bangkokpost.com/thailand/politics/1640796/constitutional-court-disbands-thai-raksa-chart Accessed 12 November 2020.

Hockx-Yu, Helen. “Access and Scholarly Use of Web Archives.” In *Alexandria* 25, no. 1-2, pp. 113-127. SAGE, 2014a. https://doi.org/10.7227/ALX.0023

Hockx-Yu, Helen. “Archiving Social Media in the Context of Non-Print Legal Deposit.” Paper presented at *IFLA WLIC 2014*, 16-22 August 2014, Lyon, France. 2014b. http://library.ifla.org/999/1/107-hockxyu-en.pdf

International Council on Archives. *Electronic Records: A Workbook for Archivists*. Study 16. International Council on Archives, April 2005. https://www.ica.org/sites/default/files/ICA_Study-16-Electronic-records_EN.pdf Accessed 9 November 2020.

International Institute of Social History (IISH). “Collection Policy Plan.” Version 2.0. 2018. https://confluence.socialhistoryservices.org/display/CTS/List+of+documentation+mentioned+in+CTS+form Accessed 30 November 2020.

International Institute of Social History (IISH). “Digital Preservation Policy.” 2019a. https://confluence.socialhistoryservices.org/display/CTS/Digital+Preservation+Policy+2019-2022#DigitalPreservationPolicy2019-2022-5.3Preservationintent Accessed 30 November 2020.

International Institute of Social History (IISH). “Strategic Plan.” Version 6.0. February 2019b. https://confluence.socialhistoryservices.org/display/CTS/List+of+documentation+mentioned+in+CTS+form Accessed 30 November 2020.

International Internet Preservation Consortium. “The WARC Format 1.1.” https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/ Accessed 9 November 2020.

ISO 14721:2012 Space data and information transfer systems - Open archival information system (OAIS) - Reference model. 2012. https://www.iso.org/standard/57284.html Accessed 9 November 2020.

ISO 15489-1:2016 Information and documentation - Records management - Part 1: Concepts and principles. 2016. https://www.iso.org/standard/62542.html Accessed 9 November 2020.

Jackson, Andrew, Jimmy Lin, Ian Milligan, and Nick Ruest. "Desiderata for exploratory search interfaces to web archives in support of scholarly activities." In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pp. 103-106. IEEE, 2016. https://doi.org/10.1145/2910896.2910912

Jules, Bergis, Ed Summers, and Vernon Mitchell, Jr., *White Paper: Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations. DocNow*. Documenting The Now, 2018. https://www.docnow.io/docs/docnow-whitepaper-2018.pdf Accessed 9 November 2020.

Kaplan, Andreas M., and Michael Haenlein. "Social Media: Back to the Roots and Back to the Future." *Journal of Systems and Information Technology* 14, no. 2, pp. 101-104. Emerald, 2012. http://dx.doi.org/10.1108/13287261211232126

Kelly, Mat, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson. "On the change in archivability of websites over time." In *International Conference on Theory and Practice of Digital Libraries*, pp. 35-47. Springer, 2013. https://doi.org/10.1007/978-3-642-40501-3_5

Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. "API-based social media collecting as a form of web archiving." *International Journal on Digital Libraries* 19, no. 1, pp. 21-38. Springer, 2018. https://doi.org/10.1007/s00799-016-0201-7

Littman, Justin. "Twitter's Developer Policies for Researchers, Archivists, and Librarians." *Medium*. 2019. https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2 Accessed 9 November 2020.

Milligan, Ian. "Lost in the infinite archive: The promise and pitfalls of web archives." *International Journal of Humanities and Arts Computing* 10, no. 1, pp. 78-94. Edinburgh University Press, 2016. https://doi.org/10.3366/ijhac.2016.016

NetLab. "NetLab IT Proficiency Test." *NetLab*. https://www.netlab.dk/services/it-proficiency-test/ Accessed 9 November 2020.

Padilla, Thomas G. "Collections as data: Implications for enclosure." *College and Research Libraries News* 79, no. 6, pp. 296-300. University of Nevada, Las Vegas, 2018. http://dx.doi.org/10.5860/crln.79.6.296

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e40-1-1 Accessed 9 November 2020.

Reitz, J. M. *Dictionary for Library and Information Science*. Libraries Unlimited, 2004.

Shendi, Nadia. "Revolutionary Sudanese Social Media as an Indicator of the Changing Face of Online Activism in Sudan and the Remainder of the Middle East and North Africa."

Master's Thesis. *Leiden University Student Repository*. Leiden University, 2020. https://hdl.handle.net/1887/133583 Accessed 9 November 2020.

Thomson, Sara Day. "Preserving social media: Applying principles of digital preservation to social media archiving." In *RESAW 2017*, pp. 1-13. 2017. https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-Thomson-applying_principles_of_digital_preservation_to_social_media_archiving.pdf Accessed 9 November 2020.

Treem, Jeffrey W., Stephanie L. Dailey, Casey S. Pierce, and Diana Biffl. "What we are talking about when we talk about social media: A framework for study." *Sociology Compass* 10, no. 9, pp. 768-784. Wiley, 2016. https://doi.org/10.1111/soc4.12404

Twitter Developer Documentation. "Rate limits." https://developer.twitter.com/en/docs/twitter-api/rate-limits Accessed 1 November 2020.

Twitter Developer Policy. "Let's Get Started." https://developer.twitter.com/en/developer-terms/policy Accessed 15 November 2020.

Urman, Aleksandra, and Stefan Katz. "What they do in the shadows: Examining the far-right networks on Telegram." *Information, Communication & Society*, pp. 1-20. Taylor & Francis, 2020. https://doi.org/10.1080/1369118X.2020.1803946

Veenendaal, Remco van. "Ervaringen met WARC-validatie." Kennisnetwerk Informatie en Archief (KIA) Blog, 24 July 2020. https://kia.pleio.nl/groups/view/48637242/kennisplatform-webarchivering/blog/view/55815115/ervaringen-met-warc-validatie Accessed 9 November 2020.

Vlassenroot, Eveline, Sally Chambers, Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel, and Peter Mechant. "Web archives as a data resource for digital scholars." *International Journal of Digital Humanities* 1, no. 1, pp. 85-111. Springer, 2019. https://doi.org/10.1007/s42803-019-00007-7

Wang, Xinyue and Zhiwu Xie. "The Case For Alternative Web Archival Formats To Expedite The Data-To-Insight Cycle." In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (*JCDL '20*), pp. 177–186. ACM, 2020. https://doi.org/10.1145/3383583.3398542

Winters, Jane. "Coda: Web Archives for Humanities Research - Some Reflections." In *The Web as History: Using Web Archives to Understand the Past and the Present*, ed. Niels Brügger and Ralph Schroeder, pp. 238-248. UCL Press, 2017.