Archiving the Timeline:
Digital Archival Practices for Social Media Data
Willem Helf

Abstract

In the advent of the age of social media, social media content is considered “born-digital” heritage worth preserving for future scholarly research and historical documentation. Archives of users’ photos, videos, text, and more can provide insights into how people communicate with one another, react to major events, and experience cultural phenomena, and are being looked at with increasing importance as a form of digital record. However, putting together and managing such an archive does not come without its difficulties: enormous amounts of data, disparate and often lacking technologies, difficult ethical decisions, and social media platforms’ restrictive data sharing policies may all hinder building an archive that functions well for end users in the long-term. Knowledge of how to do so is crucial at this point in history, as the amount of social media content in the world only grows and runs the risk of being lost or deleted permanently. This paper is divided into three parts, essentially breaking the question of social media archives into the questions of why they are important, what their current pitfalls are, and what can be done to improve work on them in the future, with a focus on technology, collaboration, and questions of ethics. While every archive has its own unique issues and needs, this paper posits several ways that archivists and researchers can put together social media archives that are built and function smoothly for both users and makers.

Keywords:
Social media archiving, born-digital media, web archiving, big data

—

Introduction

The era of Web 2.0 is well underway, and with it the era of social media. Used by 5.22 billion people worldwide (DataReportal, n.d.), social media has become not just an immensely popular way for people around the world to communicate and share information with one another but also a rapidly-growing record of these communications, which can give researchers insight into human behavior and document the past from millions of different points of view (Vlassenroot et al 2021).

This record is more than just posts on a screen. From an archival point of view, social media content is rich with information that can be used for a variety of both quantitative and qualitative scientific research purposes in many different fields of study, give a granular look into the 21st century’s history and culture, and lend meaning to users’ memories and personal lives. By saving and analyzing social media data and metadata, researchers can gain fresh insights into human behaviors, conversations, and communities (Borji et al, 2022); cultural trends and major events can newly be witnessed as narrated by a massive online Greek chorus.

Yet social media data is not static: what does not eventually vanish due to a user's or platform's actions can still be difficult to access, whether due to constraints around sharing, capturing, or saving. (Thomson & Kilbride, 2015) Application programming interfaces, or APIs, enable social media posts to be acquired for archival purposes, but often at a very restricted rate in a way that may not be conducive to the goal of a project. Large varieties of data formats may be difficult to store in one place; platforms may drastically change how they function, store data, or allow data access. (Wong & Chiu, 2024) With few-to-no technological standards around acquiring or saving social media data, archivists and researchers are often left to their own devices to sort out which acquisition and archiving methods are available to them, and oftentime there are few tools that suffice.

This paper takes a positive tack towards the future of the social media archiving space, beginning with an overview of reasons why preserving social media data is imperative, delving into the various technical, legal, and ethical issues around doing so, and, finally, proposing several ideas as to how archivists can approach future social media archiving endeavors successfully. Per UNESCO's Charter on the Preservation of the Digital Heritage, "continuity of the digital heritage is fundamental. To preserve digital heritage, measures will need to be taken throughout the digital information life cycle, from creation to access." (UNESCO, 2009) It is crucial not only that our born-digital heritage be preserved, but also that the knowledge and prowess needed to do so is continually nurtured and encouraged.

Why Archive Social Media?

Social media is woven into the fabric of modern life. Now the primary source of human communication in the world (Jeffrey, 2012), as of October 2024 94.5% of internet users now use social media at least monthly (DataReportal, n.d.). Per minute, 4,146,600 YouTube videos are watched, 456,000 posts are made on X (formerly Twitter), 46,740 images are posted to Instagram, and 527,760 photos are sent on Snapchat, an enormous amount of content that contributes to the 2.5 quintillion bytes of overall internet data generated by humans every day (Marr, n.d.).

Yet this data exists in a state of ephemerality. After one year, roughly 10% of social media content has been lost or deleted; this number jumps up to 27% after two years (SalahEldeen, 2012). Certainly some forms of social media posts are intentionally transient – Snapchat photos and Instagram stories, for example, are only visible to users for a set amount of time. However, the ease with which platforms enable users to edit or delete content renders "permanent" posts just as fleeting, and some users may also opt to programmatically remove content in bulk in what Ringel and Davidson refer to as "proactive ephemerality." (Ringel & Davidson, 2022)

Not all content removal is user-instigated, however; social media platforms may experience accidental data loss or be taken down altogether. In 2019, MySpace, once a leading social network, permanently lost millions of users' photo, video, and audio files after a faulty data migration deleted all uploaded content that was over three years old (Binder, 2019). Vine, a popular short-format video platform launched publicly in 2013, was discontinued in 2017 by its parent company Twitter after it failed to monetize successfully; although an official online archive of Vine videos was initially created, it was taken down permanently in 2019 (EM360Tech, n.d.).

These losses can be personally devastating for users. The advent of Web 2.0 and social media has enabled its users to utilize various platforms as forms of storage for digital artifacts – many users now view their personal digital ephemera as their literal possessions (Odom et al, 2010), and a blog may hold just as much sentimental importance to someone as a traditional paper diary might (Steinhauer, 2015). Yet as social media platforms change, merge, and vanish over time, the risk of these ephemera being permanently lost continues to loom.

As it stands, these losses have as many consequences for history as they do for individuals. In the *UNESCO Charter on the Preservation of the Digital Heritage,* it is made clear that "born-digital" resources – that is, materials that exist only in digital form – "have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for future generations," and that they are "...frequently ephemeral, and require purposeful production, maintenance, and management to be retained." (UNESCO, 2009) Notably, web pages are listed in the charter as a born-digital format considered worth preserving; this line of thinking can logically be extended to social media platforms – though they are not static websites, they exist solely online, in both the digital realm and in a transient state.

As social media is preserved, so does it preserve: social media content can, over time, provide a record of events and communications that has potential cultural significance to researchers and historians. The clear and direct methods of conversation that social media facilitates leave us with logs, metadata, and posts that can give insights into everyday life and methods of communication previously unavailable to us, enabling us to document the past much more vividly (Vlassenroot et al, 2021). Major events and cultural moments are now richly documented online, from the point of view of hundreds of millions of everyday users – essential material for anybody with an interest in studying and analyzing the past (Jeffrey, 2012).

This data is also rich for harvesting for research purposes. Archived social media data provides researchers with reusable data sets that can generate new analyses and insights into human behavior once unavailable (Borji et al, 2022). In particular, metadata generated by a post or interaction – location, application used, exact time, and so forth – can provide insights into conversations in a coherent, easily-analyzed way (Lomborg, 2012), opening the door to studying phenomena such as the spread of hate speech across online communities and relationships between users and politicians (Thomson & Kilbride, 2015).

A notable example of digital archiving and data collection from social media is shown in The Syrian Archive, founded in 2014 in the aftermath of what is known as the Arab Spring.

Beginning in Tunisia in 2010, the Arab Spring was a series of protests and uprisings across the Arab world in which social media communication, documentation and surveillance played a significant role – people were able to organize and mobilize protests and broadcast events to the world through the internet, but social media accounts were frequently surveilled or suspended, and people could be arrested, tortured, and possibly even killed as a result of something posted online (Arnold & Sampson, 2014). In 2011, Hadi al Khatib noted that there was no storage or summaries of videos being shared, so founded the Syrian Archive in 2014 with the goal of saving as much online media as possible for both historical posterity and for evidence in legal proceedings around human rights violations and fact-finding (Kayyali). Over the course of several years, Hadi and colleagues worked to put together a functional archive, which expanded to cover several other countries by 2017. (Kayyali, 2022)

What challenges do social media archivists face?

In 2010, Twitter and the Library of Congress signed an agreement stipulating that Twitter would gift all public tweets spanning from 2006 to 2010 to the Library of Congress, acknowledging that saving born-digital media is crucial to preserving cultural memory (Fondren & McCune 2018). By the time the entire archive of tweets was delivered to the Library in early 2012, it contained 170 billion tweets, each of which came with 150 pieces of metadata; in total, this added up to over 133 terabytes of data, nearly doubling the amount of data the Library contained in its digital infrastructure as a whole. (Zimmer, 2015)

By 2017 the archive was still not available for use, the Library explaining in a blog post on its website that it would only be archiving selected tweets moving forward and that "...the Twitter collection will remain embargoed until access issues can be resolved in a cost-effective and sustainable manner." (LoC, 2017) As of 2024, the collection of tweets has still not been made available.

While the Library of Congress is not the only institution to collect a large amount of social media posts, its attempts to manage its archive of tweets are indicative of the infrastructural challenges behind archiving large collections of social media content. The sheer amount of data, along with the velocity at which it accumulates, renders traditional data collection methods inadequate – a single text search query to the Library of Congress's tweet archive took 24 hours to generate a response. (Zinaman 2024, Fondren & McCune 2018) Many institutions and researchers must, instead, selectively crawl social media websites and archive only specific, pre-chosen content if they are to build up an archive that is stable and usable.

Yet more granular methods of collecting social media data also have their own flaws. APIs, or application programming interfaces, allow two software entities to communicate with one another via a preset of rules; many social media platforms offer their own APIs with which users may harvest data directly from the platform. These APIs provide conveniently-formatted data in the form of JSON and XML and allow efficient collection methods, enabling precise quantitative research and computational processes. (Littman et al, 2018) However, every API is different:

while some allow for a broad amount of data harvesting, many others have strict limits on the volume, access, timeframe, and shareability of the data that can be collected. Social media platforms may also drastically change or altogether remove access to an API at their own discretion, greatly restricting what content may and may not be harvested. (Wong & Chiu 2024)

A common alternative to using an API is web scraping, a method of harvesting data that involves directly accessing a website's code and extracting its data. While scraping is low-cost and easily automated, it is often an unstable and user-unfriendly process, requiring coding skills and time; there are also legal risks around what data may be collected, as scrapers generally have fewer restrictions on what they can collect than an API does. (Chen et al, 2024) There is also the consistent risk of social media platforms changing and subsequently enforcing restrictions on data collection and sharing, which drastically affects the amount of data that a scraper can harvest.

Both APIs and web scrapers face the issue of how to crawl and harvest different media and website formats. An API or scraper may not have the technical ability to handle data in various formats such as HTML5, different image and movie file types, and JavaScript web pages, making it difficult to harvest and store content in a consistent manner. (Borji et al) Social media platforms are built and run by private companies, who are the sole deciders of how to shape their content; this can lead to a glut of differently-formatted data types as the number of proprietary data standards only increases. (Acker & Kreisberg, 2020) There is, overall, a lack of technical standards around storing social media data and managing its ephemerality, and researchers and archivists are often left to cobble together unique data management techniques on a per-project basis. (Vlassenroot et al, 2021)

Beyond strictly technical restrictions, there are also the complexities of ethics and legality surrounding archiving social media content. Throughout the stages of collection and analysis, archivists and researchers must contend with various copyright and privacy laws that can be particularly opaque in relation to social media data. (Zinaman, 2024) Section 108 of the U.S.' Copyright Act allows libraries and archives to reproduce and distribute copyrighted works for the purpose of preservation or research, but was enacted in 1976, asking the question of whether the law extends to online media; while it was recommended in 2016 that the Copyright Office review laws around archiving digital content, thus far there has been no official decision made. (Sharma, n.d.) This leaves the copyright question up in the air: while social media platforms frequently encourage resharing and remixing of content, a 2020 court ruling stated that media embedded in a tweet is subject to copyright protection, while another case ruled that a tweet itself could be copyrighted if it was considered "independently creative" and "creative" enough. (Zinaman, 2024, Sharma, n.d.)

There is also the issue of user privacy around archiving social media content. The Statement on the Right to be Forgotten by the International Federation of Library Associations and Institutions argues that information on the internet should "in general, not be intentionally hidden, removed or destroyed" while also acknowledging that some information may be damaging to a person's reputation or peace of mind (IFLA, 2016); however, this does not mean that people will automatically consent to their social media data being seen, saved, or used for research. Although users may have signed a platform's Terms of Service agreement allowing the platform to sell their data, they may not be aware of this caveat as many ToS agreements are intentionally long and opaque (Zinaman, 2024), and can still oppose the usage of their content. Laws around privacy may provide a loose guideline for how to approach these issues, but researchers and archivists must still contend with the fact that the ethics around collecting and storing users' social media data have no singular solution.

What can social media archivists do now?

Like social media itself, the social media archiving space exists in a still-nascent form – a lack of technical standardization, constantly-shifting ethical and legal concerns, and little awareness around the increasing importance of digital heritage slows down innovation and complicates what might otherwise be simple. As there is no singular "quick fix" for any of these issues, archivists and researchers looking to put together a social media archive in any sort of format must take into consideration the goals, limitations, and pitfalls of any project far before they embark on it.

Crucially, selection criteria for what exactly will be archived must be established – as the Library of Congress's attempts at managing a Twitter archive illustrates, it is not technologically feasible to archive the entirety of one social media platform's data in a useable way, so researchers must hone in on a clearly-defined data type or subject if the project is to be a success. (Vlassenroot et al, 2021) Questions around impact, ethics, and legality should also be considered early on. Amongst the users and subjects involved, who will be impacted by the project, and how? What legal and ethical questions do digitization, access, and reuse raise? Is selection bias imminent, and if so, how can it be mitigated? (Carbajal & Caswell, 2021)

Technologically speaking, there is also much to consider. While there is, as of yet, no perfect method of social media data collection, one must still be chosen and questions asked of its potential technological limitations, such as poor documentation, technical capabilities, or limits on amounts or types of collectible data. (Chen et al, 2024) A file format for storage and a long-term storage solution should also be carefully thought through. The constraints around the platform from which data is being acquired should be examined, seeing as the types and amount of data that can be harvested from a platform may change at whim. (Acker & Brubaker) These questions should all be considered in the context of both who is building the project and who will be accessing it in the future. (Borji & Naeini, 2022)

As it stands, new social media archiving technologies are currently being developed, many of which are large, flexible software services that enable comprehensive real-time data capture and storage. These services work by scraping and preserving data through an API, with the ability to archive a wide variety of data formats such as movies, text, events, and timelines, keeping the data in storage even after it is potentially removed from social media. (Borji & Naeni, 2022) While there is a wide range of these services to choose from, three that are notable for their popularity and functionality are ArchiveSocial, PageFreezer, and Smarsh, all of which are used by major companies on large amounts of data. (Jatheon, 2024) By connecting directly to a chosen social network, each of these services are able to record all of a user's public posts and interactions with the option to search for and export data easily while remaining compliant with social media retention policies and public record laws. (Borji & Naeni, 2022) While these platforms are not one-size-fits-all, their flexibility and ability to harvest and archive large amounts of data make them potentially useful tools for archivists and researchers both.

As useful as these tools may be, however, technology is never in stasis, and it is crucial to stay aware of changes and innovations in the born-digital archival realm. Beyond just outsourcing technical knowledge and skills to others, archivists and researchers should develop and hone their own skills and technical literacy in order to engage critically and thoughtfully with new and current archival methods. (Burgess & Bruns, 2012) This cannot be done in a vacuum; archivists and researchers must foster relationships with stakeholders, web content producers, users, and potential users to give and receive various forms of knowledge and raise awareness of the importance of preserving born-digital media. (Wong & Chiu, 2024)

Change in the social media archival space is inevitable. Technology, expectations, scholarly practices, and society as a whole shift and ebb over time, and archivists must continue building up the knowledge and skills needed to stay aware of these changes. (Carbajal & Caswell, 2021) For social media archiving to stay abreast of rapidly-changing social media platforms, societal needs and wants, and technological shifts, archivists must put in the time, collaboration, and thought to preserve our rapidly-growing digital heritage.

Conclusion

Data from social media content can give insight into questions asked by social scientists and uncover information about historical moments, cultural movements, and forms of conversation; however, archiving this content can be a thorny endeavor, with questions around technology, legality, ethics, and rapid infrastructural change consistently at the forefront. (Thomson & Kilbride, 2015) As social media usage continues to increase, so does the need for tools and techniques to collect its archival data accurately and at scale: with the rate of content disappearance, the amount of historical record captured and recorded by users, and continual need for research data, archivists now more than ever must have the knowledge and resources to build social media archives that function well in the long-term.

There is no perfect method with which to acquire and archive social media data, neither in the sense of a one-size-fits-all solution nor of a solution free of technical, ethical, and legal concerns. As the amount of born-digital content continues to rapidly grow, archival institutions are continually faced with the issue of how to juggle new technologies, anticipate the needs of future users, and contend with murky ethical and legal issues – yet things change at such a rapid rate that it is impossible to take this on independently. Rather, it is crucial to foster relationships with potential users of a given archive early on to anticipate future needs and exchange necessary knowledge and skills needed for the archive to succeed in the long term. (Thomson & Kilbride, 2015)

On a more macro level, it is imperative that professionals of many different backgrounds come together if social media archiving is to gain the tools and shared knowledge needed for the space as a whole to continue making headway. Archivists, librarians, technologists, researchers and publishers will need to work together to continue to develop new frameworks and tools for creating social media archives that last. (Hockx-Yu, n.d.) While it remains likely that individual projects and teams will have their own technological and logistical needs, crowdsourcing knowledge and long-term collaboration can build community in which methods, resources, and tools can be shared – a boon to all involved, and to the future of the social media archival space.

—

Bibliography

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. Archival Science, 20(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

Binder, M. (2019, March 18). MySpace lost 12 years of music and photos, leaving a sizable gap in social network history. Mashable. https://mashable.com/article/myspace-data-loss

Borji, S., Asnafi, A. R., & Naeini, M. P. (2022). A Comparative Study of Social Media Data Archiving Software. Preservation, Digital Technology & Culture, 51(3), 111–119. https://doi.org/10.1515/pdtc-2022-0013

Bruns, A., & Weller, K. (2016). Twitter as a first draft of the present: And the challenges of preserving it for the future. Proceedings of the 8th ACM Conference on Web Science, 183–189. https://doi.org/10.1145/2908131.2908174

Burgess, J., & Bruns, A. (2012). Twitter Archives and the Challenges of “Big Social Data” for Media and Communication Research. M/C Journal, 15(5), Article 5. https://doi.org/10.5204/mcj.561

Cannelli, B., & Musso, M. (2022). Social media as part of personal digital archives: Exploring users’ practices and service providers’ policies regarding the preservation of digital memories. Archival Science, 22(2), 259–283. https://doi.org/10.1007/s10502-021-09379-8

Carbajal, I. A., & Caswell, M. (2021). Critical Digital Archives: A Review from Archival Studies. The American Historical Review, 126(3), 1102–1120. https://doi.org/10.1093/ahr/rhab359

Chen, Y., Sherren, K., Lee, K. Y., McCay-Peet, L., Xue, S., & Smit, M. (2024). From theory to practice: Insights and hurdles in collecting social media data for social science research. Frontiers in Big Data, 7. https://doi.org/10.3389/fdata.2024.1379921

Fondren, E., & McCune, M. M. (2018). Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive. Preservation, Digital Technology & Culture, 47(2), 33–44. https://doi.org/10.1515/pdtc-2018-0011

Global Social Media Statistics. (n.d.). DataReportal – Global Digital Insights. Retrieved December 6, 2024, from https://datareportal.com/social-media-users

Hockx-Yu, H. (n.d.). Archiving Social Media in the Context of Non-print Legal Deposit.

IFLA Statement on the Right to be Forgotten (2016) – IFLA. (n.d.). Retrieved December 14, 2024, from https://www.ifla.org/publications/ifla-statement-on-the-right-to-be-forgotten-2016/

Jeffrey, S. (2012). A new Digital Dark Age? Collaborative web tools, social media and long-term preservation. World Archaeology, 44(4), 553–570. https://doi.org/10.1080/00438243.2012.737579

Jeffrey, S. (2012). A new Digital Dark Age? Collaborative web tools, social media and long-term preservation. World Archaeology, 44(4), 553–570. https://doi.org/10.1080/00438243.2012.737579

Kayyali, D. (2022). Digital Memory, Evidence, and Social Media. Lessons Learned from Syria. Sociologica, 16(2), Article 2. https://doi.org/10.6092/issn.1971-8853/15383

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., Vij, R., & Wrubel, L. (2018). API-based social media collecting as a form of web archiving. International Journal on Digital Libraries, 19(1), 21–38. https://doi.org/10.1007/s00799-016-0201-7

Matsuura, Koïchiro. “Charter on the Preservation of the Digital Heritage.” 2009. UNESDOC Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000179529. Accessed 9 December 2024

Osterberg, G. (2017, December 26). Update on the Twitter Archive at the Library of Congress | Timeless [Webpage]. The Library of Congress. https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2

Ringel, S., & Davidson, R. (2022). Proactive ephemerality: How journalists use automated and manual tweet deletion to minimize risk and its consequences for social media as a public archive. New Media & Society, 24(5), 1216–1233. https://doi.org/10.1177/1461444820972389

SalahEldeen HM, Nelson ML (2012) Losing my revolution: how many resources shared on social media have been lost? In: Zaphiris P, Buchanan G, Rasmussen E et al (eds) Theory and practice of digital libraries lecture notes in computer science. Springer, Berlin, Heidelberg, pp 125–137. https://doi.org/10.1007/978-3-642-33290-6_14

Sharma, S. (n.d.). “HOW TWEET IT IS!”: HAVE TWITTER ARCHIVES BEEN LEFT IN THE DARK?

Shiozaki, R. (2024). People’s perceptions on social media archiving by the National Library of Japan. Journal of Information Science, 50(4), 861–873. https://doi.org/10.1177/01655515221108692

Steinhauer, J. (2015, July 24). Preserving Social Media for Future Historians | Insights [Webpage]. The Library of Congress. https://blogs.loc.gov/kluge/2015/07/preserving-social-media-for-future-historians

Thomson, S. D., & Kilbride, W. (2015). Preserving Social Media: The Problem of Access. New Review of Information Networking, 20(1/2), 261–275. https://doi.org/10.1080/13614576.2015.1114842

Top 5 Social Media Compliance Archiving Tools Compared. (2024, February 5). https://jatheon.com/blog/pagefreezer-archivesocial-intradyn-smarsh-jatheon/

Tošić, J. (2024). Digital Mini-Archives: Social Media Users as Curators of an Architectural Utopia. Art + Media: Journal of Art & Media Studies, 34, 27–38. https://doi.org/10.25038/am.v0i28.562

Velte, A. (2018). Ethical Challenges and Current Practices in Activist Social Media Archives. The American Archivist, 81(1), 112–134.

View of Death, Memorialization, and Social Media: A Platform Perspective for Personal Archives. (n.d.). Retrieved December 15, 2024, from https://archivaria.ca/index.php/archivaria/article/view/13469/14791

Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J., & Mechant, P. (2021). Web-archiving and social media: An exploratory analysis. International Journal of Digital Humanities, 2(1), 107–128. https://doi.org/10.1007/s42803-021-00036-1

Wong, A. K., & Chiu, D. K. W. (2024). Digital curation practices on web and social media archiving in libraries and archives. Journal of Librarianship and Information Science, 09610006241252661. https://doi.org/10.1177/09610006241252661

Zimmer, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. First Monday, 20(7). https://doi.org/10.5210/fm.v20i7.5619

Zinaman, M. (2024). Social Media Archiving in Practice: A Troubled Landscape in Review.