Archival Science (2020) 20:105–123
https://doi.org/10.1007/s10502-019-09325-9

ORIGINAL PAPER

Social media data archives in an API-driven world

Amelia Acker[1] · **Adam Kreisberg**[2]

Published online: 24 September 2019

Abstract

In this article, we explore the long-term preservation implications of application programming interfaces (APIs) which govern access to data extracted from social media platforms. We begin by introducing the preservation problems that arise when APIs are the primary way to extract data from platforms, and how tensions fit with existing models of archives and digital repository development. We then define a range of possible types of API users motivated to access social media data from platforms and consider how these users relate to principles of digital preservation. We discuss how platforms' policies and terms of service govern the set of possibilities for access using these APIs and how the current access regime permits persistent problems for archivists who seek to provide access to collections of social media data. We conclude by surveying emerging models for access to social media data archives found in the USA, including community driven not-for-profit community archives, university research repositories, and early industry–academic partnerships with platforms. Given the important role these platforms occupy in capturing and reflecting our digital culture, we argue that archivists and memory workers should apply a platform perspective when confronting the rich problem space that social platforms and their APIs present for the possibilities of social media data archives, asserting their role as "developer stewards" in preserving culturally significant data from social media platforms.

Keywords APIs · Developer stewards · Platform perspective · Social media data archives

✉ Amelia Acker
aacker@ischool.utexas.edu

Adam Kreisberg
adam.kriesberg@simmons.edu

[1] School of Information, University of Texas at Austin, 1616 Guadalupe St, Suite 5.202, Austin, TX 78701, USA

[2] School of Library of Information Science, Simmons University, 300 The Fenway, Boston, MA 02115, USA

Introduction: social activity streams and APIs

In 2017, Clifford Lynch published a broad treatise on the challenges that algorithmic-intensive systems pose for archivists and the near futures of digital preservation in networked environments. In this vibrant call to action, Lynch makes a point of connecting the preservation mandates of traditional professional archival duties with the haziness of what is to come for stewardship in a society underwritten by collections of data and driven by an impulse to collect:

> "If archivists will not create, capture, curate the "Age of Algorithms," then we must quickly figure out who will undertake this task, and how to get the fruits of their work into the custody and safety of our memory organizations for long-term preservation. Traditional archivists seem most comfortable dealing with the outcomes of the work of various types of documenters, rather than creating the testimony: this is a professional constraint that needs to be explicitly recognized, considered, and if appropriate clarified and affirmed if it is to be the case going forward" (Lynch 2017).

Application programming interfaces (hereafter, APIs) are both gateways to access data and artifacts in their own right of our algorithmically driven information age. They are technologies of custody that enable the extraction and access to data. Our connected world is increasingly saturated with networked mobile computing devices that allow platform users to create data at tremendous rates that drive new markets, technologies, and social structures. Information policy, privacy, and law scholars have examined the functional sovereignty that these platforms now exert over society, democracy, and economies of scale by collecting personal data, providing access for third parties, and repurposing it for algorithms, personalization, and advertising technology (Pasquale 2016). However, most of this work is concerned with access to data collections presently and near-term implications; it does not concern long-term preservation contexts or future archives of data extracted from platforms with APIs. As archival scholars concerned with the future of preservation and digital cultural memory, we are concerned that platform APIs—as access points and technologies of custody—will have a significant impact on the abilities to preserve documentation from the algorithmic age and to create testimony about those impacts as Lynch has described because of their prohibitive control and access constraints to human activity data in the long term.

Since the early days of social media platforms, extracting data for secondary reuse, research, and auditing has been difficult for those interested in working with social media data. Social activity streams, as they are called now, are difficult to capture because of their content, context, and form ("Activity streams: specifications" 2019; Snell and Prodromou 2017a). As their name suggests, these streams of social media are not static documents. Social activity streams are frequently updated by platforms with features and algorithms, accumulating new elements of data containing many layers of context, multimedia objects, and engagement metadata. Often these social activity streams, in whole or in part,

remain locked into dynamic platforms with particular kinds of technical and legal gateways of access by way of extraction possibilities. Social platforms themselves identified this extraction "problem" as an opportunity to create a new marketplace in third-party data access to social media data by primarily providing access to user data (and metadata) through a range of APIs in the early 2010s (John and Nissenbaum 2019).

APIs allow third parties (known as "developers") the ability to query and gain access to portions of social activity streams from user data created in using and experiencing social platforms. APIs specify the rules by which software talks to each other, articulating which elements can be queried, how frequently, and how the results appear. Different APIs allow for purpose-driven access and extraction. For example, a social network API would allow a dating app to use information about user profiles to match people for dates. A content publishing API would allow a newspaper to promote breaking news articles on a social network's newsfeed, promoting content directly to specific users. Or an advertising API could allow a small business to promote their new menu on a restaurant rating and review platform. While APIs are often hidden or unknown to social media users themselves, they are part of the software development ecology of the social network infrastructures that drive our experiences of numerous platforms and apps.

Presently, access to much of the social media data from platforms is governed through this API-driven gatekeeping model set up by platform owners to establish the rules by which all third parties accessing data must abide (Bruns 2019). These models, while unique to each platform, are rigid in their permissions and restrictions on API users, failing to make distinctions between different kinds of data brokers such as developers who are researchers, journalists, or digital archivists. Not all API developer users have the same reasons or motivations for using them and accessing social media data, and not all API users are concerned with the long-term authentication, fixity, or reproducibility of results. The one-size-fits-all approach to developer access has increasingly become problematic for social platforms when providing access, measuring impact, and enforcing governance over developers' collections from APIs (Acker and Donovan 2019; Driscoll and Walker 2014). A series of gaps exists between the range of users that APIs are intended to serve, the account holders or creators that generate these data accessed in APIs, and third-party data brokers who use API extracted data in agreement with the platform's aim to keep users engaged. Indeed, the problem of extracted social media data collections is not just a concern for researchers and scholarly institutional repositories. As intermediaries between many kinds of users with different motivations, social platforms themselves grapple with issues of control over the records that users create, managing user data archives and the types of users who leverage them for data access (Glassman 2019).

Social platform APIs from Twitter, Facebook, Instagram, YouTube have not only changed the experience of the web for users who use and create social media, but also research methods in computational social science that allow researchers to create new models of instrumentation to gather social media data (Bruns 2013; Bruns and Weller 2014; Hargittai and Sandvig 2015). Just as platforms have changed instrumentation and research methods for scholars studying behavior online, they

have impacted our ability to collect and access research data (Freelon 2018). Thus, APIs also create new preservation and access issues for digital archivists, research data managers, scholarly communication repositories, and digital curation initiatives. The ascendency of the API access over other data extraction techniques such as web scraping has led to new models of digital collection, accessioning, and preservation in research archives and web archiving (Littman et al. 2016).

Digital preservation techniques, policy, social, and technical constraints have always been shaped by our ability to access information and bring it into custody, for example with proprietary formats or unique playback hardware (e.g., Faniel and Yakel 2011; Hedstrom and Lampe 2001; Rimkus et al. 2014). The current API access regime of platforms that depend on large collections of aggregated user data for access and reuse of these collections is not typically managed or oriented toward archival principles (Glassman 2019). Few preservation perspectives or non-commercial use cases have been subsumed into the development of platforms as part of their market strategies as for-profit corporations (Conway 2010; John and Nissenbaum 2019). Despite these preservation and access barriers, as collections of social media data grow (internally or through extraction), platforms are rapidly becoming sites of post-custodial archives. In post-custodial theory of archives, the power of the archivist over records is ceded to the contexts where records originate and circulate, thereby reorienting the role of traditional archival institutions to that of a "participant" in a (re)allocation of power very much [sic] at odds with traditional praxis" (Kelleher 2017, p. 23). Archival scholars have argued that post-custodial approaches democratize custody, power, and agency over archives by prioritizing context, allowing creators self-determination through the archival autonomy of asserting provenance, ownership, context, and function of records creation (Cook 1994; Evans et al. 2015). But platforms as data enablers and intermediaries that provide access through APIs can severely limit the archival autonomy of creators, researchers, digital stewards, and archivists by stripping context from activity streams as they provide access to platform data.

Social media data collections from platform APIs are a new post-custodial challenge for archivists because they result in a proliferation of decontextualized social activity streams as records and evidence of user data (Walker 2017). Digital archivists have commented on the drawbacks—there is a loss of context from the platform itself, data can be disconnected from the communities that have created it, there is a focus on machine-readable structured text over dynamic or visual content, and an over-reliance on platforms to shape the scope and acquisition through developers' APIs (Jules 2018; Littman 2019; Summers 2019). Another challenge is that while APIs present a means of access, there may be many different types of API users with different motivations for access who are all governed by the same terms of service.

Moreover, the design of social platform APIs follows changes in the platforms themselves for servicing users who create content and have knock-on effects that spread out into different areas of data access and reuse. While researchers, journalists, and stewards can gain access to platform data through APIs, the conditions of access strip context in important technical, social, and moral ways. And so, accurate reproducibility is not just a problem for researchers using developers' APIs; it poses

a threat to cultural memory stewards, such as archives and libraries, aiming to capture accurate slices of social media experiences from individuals, communities and filial groups, and society more broadly. Archivists, stewards, and repositories that collect social media data archives using APIs (as many now do) then regain some control as custodians of extracted social media data, but lose valuable context that the platform contains, arguably the rich context typically called for by post-custodial orientations. And further, the mechanism for which these data are harvested and collected, the API, is itself constantly changing—little is known about the long-term impact and unintended consequences of API rollbacks, updates, or their ever-changing terms of service. If the extraction and decontextualization of social media data are one kind of technical challenge, the developers' API that allows "one-size-fits-all" access remains another, larger ontological challenge for archivists who make use of APIs to assemble these collections.

In this paper, we examine features of social media data from platforms and discuss their long-term preservation consequences by focusing on the current landscape of data access through APIs. We begin the next section by introducing the preservation problems that occur with user data extracted from developers' APIs, and how these fit with existing models of archives and digital repository development. Then, we define and analyze the range of possible users concerned with extracting social media data from platforms. We make a distinction between platform users (such as account holders and creators of content) and API users, who may be platform users, but have API keys to make use of extracted data as "developers." This is because users who have social media platform accounts and developers who use platform APIs have separate terms of service, rights, and responsibilities when using social media platforms and developer APIs. Then, we discuss how platforms govern possibilities for access, and how the current access regime promotes persistent problems over stewarding personally identifying information, guaranteeing the reproducibility or fixity of content, and incredible amounts of energy use and resource consumption because of bottleneck redundancies. We finish by surveying early models for access to social media data archives, including community driven not-for-profit community archives, university research repositories, and early industry–academic partnerships primarily in the USA. We argue for applying a platform perspective in exploring the rich problem space that social platforms and their APIs present for efforts to collect social media data archives and manage them over the long term as digital cultural memory artifacts.

Persistent preservation problems

For decades, the digital preservation community has worked to develop standards, workflow, and infrastructure to manage a range of digital objects over time. These efforts, while vital for the future success of managing social media data, embody certain assumptions around digital preservation that do not neatly translate to an environment replete with complex digital objects, proprietary platform-specific standards, and dynamic content that is not discrete. These challenges can

be understood at three levels: the repository or infrastructure level, the format level, and the content workflow level.

Digital repositories are a vital element of infrastructure on the web and power research, learning, journalism, policy analysis, and the preservation of digital cultural heritage. These organizations are fundamentally concerned with building trust with users and demonstrating that the digital files on their servers are the same files that were created are the same ones deposited in the repository. One widely adopted standard for organizing the work of digital repositories is the Open Archival Information System (OAIS) model (CCSDS 2012). The model grew out of the work of the Consultative Committee for Space Data Systems (CCSDS) and initially focused on the requirements for stewarding data from large-scale, multi-site scientific research initiatives (Lee 2010). It has proved to be extremely influential and continues to shape much of current praxis around digital preservation for a wide range of objects, not only scientific research data.

A key element of the OAIS model, and of any digital repository that adheres to it, is the concept of fixity. This is the idea that a given file or digital object in a repository has not been altered since its deposit, and that the file can be computationally verified to be what it purports to be. The importance of file fixity is also underscored in other digital preservation frameworks including the National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation (Phillips et al. 2014). Digital preservation professionals use checksums to establish a cryptographic hash for every file that enters a repository; these checksums can then be rerun periodically to verify if a file has been changed in any way. Through these checks and the ability to claim that the files in their care have not been altered since deposit, digital preservation professionals build trust with their users, convincing them that they can be confident using the materials from a repository in whatever way they intend.

Platform data from social media sites that generate activity streams and allow for user engagement with multimedia and multitemporal content pose a challenge to the OAIS-based understanding of digital preservation. Activity streams are semantic descriptions of actions taken on a given website, social media platform, or piece of software (Snell and Prodromou 2017a), that is, an activity stream is a way to standardize the description of things people do in online spaces. For example, it offers a controlled vocabulary structured way to express an action on a platform such as "User X posted an update at time *T*." An Activity Stream is a type of metadata associated with digital environments that allows platforms to more consistently manage user data, but they can also be useful beyond short-term business needs. These types of digital objects do not neatly wrap themselves up into discrete files like PDFs, JPEG 2000s, or CSV data tables. The World Wide Web Consortium maintains an "Activity Streams" standard defining and describing a JSON-based method for describing a range of user actions which can be associated with related digital objects (Snell and Prodromou 2017b). While discrete surrogates such as these can stand alone and may include meaningful descriptive information, they fail to capture essential parts of platforms, as well as the platform APIs (particular versions or specific documentation), and are not as valuable over the long term once removed from their original context.

The files that different social media sites generate when a platform content creator requests to download a copy of their data contain a lot of metadata in addition to the text of the posts (such as tweets, or status updates) themselves. For example, a creator can see how many retweets a given tweet has, or how many likes a photograph on Facebook has, but these are only snapshots in time. If they revisit content on the platform after downloading an archive copy, the tweet or post may have a new number of engagements, reflecting ongoing activity on the platform (Acker 2018; Acker and Kriesberg 2017).

At the format level, preservation challenges for social media flow from the design of individual platforms and the business needs of the companies which operate them. While many platforms involve the creation and consumption of activity stream data, the standards around structuring data within a given network are driven by the business and organizational needs of the companies running these sites, not by their suitability for stewardship and long-term preservation beyond the life of the parent company (Helmond et al. 2017). Platforms such as Facebook, Twitter, and Pinterest do not make their data interoperable because their competitive advantages over each other lie in their differences, not their similarities. Actions such as posts, likes, replies, reposts, and comments each have their own definition on each platform and are reflected differently in the offline representations of content creators' data (Bucher and Helmond 2018).

Because social media platforms are built from software developed by private companies, they are able to shape their content in whatever way their engineers decide (Brügger 2017; Rosenzweig 2001). This can result in extremely different representations of social media data and significant challenges for digital preservation professionals seeking to document and manage an ever-increasing number of proprietary data standards. For example, Facebook's user data, when downloaded by an individual, are represented as a file directory with folders for photographs, posts, and other activities on the site. When accessing each of these types of downloaded data, users can view some interactions such as comments on photographs, but not others such as number of likes on a given post or photograph. See Figs. 1 and 2 for a visual representation of Facebook's profile download.

The content level presents additional challenges to digital preservation of social media data. By content level, we mean the text, engagement data, and multimedia digital objects that comprise the heart of social media platforms. This content is highly personal and cannot simply be made available in a digital repository for secondary use. Recently, institutional review boards (IRBs) have begun paying attention to these data and treating it seriously, as a form of Personally Identifiable Information (PII) (Considerations and recommendations concerning Internet research and human subjects research regulations, with revisions 2013). Furthermore, the General Data Protection Regulation (GDPR) in Europe suggests a more robust understanding of data protection and privacy of people's social media data, both on platforms and off. While the GDPR speaks to many types of privacy and user rights online, one aspect of this is enshrined in Article 20: Data Portability. This article states "The data subject shall have the right to receive the personal data concerning him or her…in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from

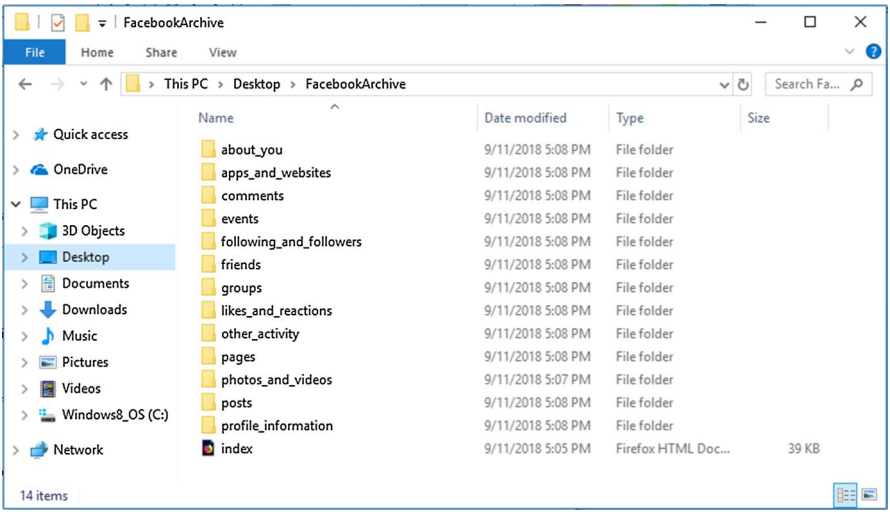

Fig. 1 Screenshot from the profile archive of one of the authors' personal Facebook profiles

Fig. 2 Screenshot of image captured from one of the authors' personal Facebook profiles showing comment from a friend but not number of likes

the controller to which the personal data have been provided." (European Parliament 2016). Here, the assumption is that platforms can and should provide account holders with their personal data upon request in a format which is meaningful outside of the context of the original platform. Account holders exercising their rights to data portability under Article 20 may want to engage on a new social media site using their data or networks from an existing account; they may also want to transfer their data from a platform to a digital repository for long-term preservation. This article implies the need for a common standard for social media metadata, which could be utilized by the preservation community to build long-term stewardship environments for data from a range of social media platforms, mapped to the standard.

Finally, the End User License Agreements (EULAs) agreed to by users of social media sites may not allow for preservation and publication of certain kinds of platform data outside the platforms themselves. In their Developer agreement and policy, in the section headed "Be a good partner to Twitter," Twitter recently stipulated that Tweet IDs, rather than tweets themselves, should be all that is published by researchers working with their data (Twitter 2018). This policy has affected the authentication, replicability, and verifiability of research data collected from Twitter and has, in some cases, pushed researchers out of compliance with the Developer Agreement. Recent work from Arkaitz Zubiaga demonstrates that retention of only Tweet IDs results in decreased availability of tweet content over time in research datasets, but that this does not affect overall interpretability for larger collections of tweets (Zubiaga 2018). However, for smaller and more targeted research collections created by either researchers, archivists, or repository managers, any loss of tweet content or associated media (i.e., images, video) represents a significant impact on the ability to authenticate or verify the contents of the collection because of missing data and/or metadata that cannot be verified or compared to historical platform data (Kerchner et al. 2016). While file formats, fixity, and portability of social media data continue to challenge archivists, leveraging APIs as a means of extraction confronts the potential that archivists have in asserting their role in creating documentation and exercising rescue functions when facing the vulnerability of culturally significant digital information (Garrett and Waters 1996, pp. 22–23).

The impact of APIs on access

Access to social media data with APIs is governed by platforms and their terms of service agreements with their users. As with most information systems, there is a range of motivations for access and it is worth clarifying the types of users that participate in social media platform ecologies, including APIs. Table 1 discusses these different users, as well as their conditions of access and motivations for using social activity stream data from platforms. We have characterized some of these various users and their conditions of access to API data based on observations from platforms terms of service, social media collections research, and the landscape of API utilization. While there are many different types of users, this article is primarily concerned with the status of researchers, stewards, and institutional repositories as users who make use of developer privileges when extracting data from platform

Table 1 Types of API users detailing the varying conditions of access and motivations to social media data

Types of API users	Conditions of access and motivations
Account holders are content creators that use the social media platform	• Create user data for APIs • Give developers permission to access and use personal information through the API • Access the API to retrieve account data for personal digital archives
Developers are third-party data brokers who access user data to create new technologies or apps, to advertise content on the platform, or to build data profiles to develop technology outside of the platform	• Create tools, clients, and plugins on the platform • Create advertising and promoted content to direct to users, contributing to filter bubbles • Use the API to achieve some kind of business goal through the platform itself, platform development/data enabler, and source • Motivated by profit generation
Developer researchers are third parties who collect data to investigate, collect evidence, and document social media phenomena, ranging from academics, journalists, civil society groups	• Access API to get more insights into how users communicate, read, consume content, and create social ties • Collect and analyze API data to make claims, to inform decisions, to create new knowledge, to inform policy and regulation • Access the API data to further understand the platform, underlying algorithms, code, data structures, in order to reverse engineer the system itself
Developer stewards are web archives, community archives, or institutional repositories who leverage APIs to collect social media data for preservation purposes and future use	• Produce stable, understandable facsimiles of platform content (get content off the platform for secondary use, research, preservation) • Want to know about the platform, data structure, format, content, context, form • Use the API for reliable, authentic data and to document actions and enforce accountability • Intend to provide long-term access

APIs. However, it is worth noting that API services are framed within a for-profit ecology of account holders and primarily for data brokers who innovate within or outside of the platform. The majority of enterprise developers are selling advertisements, innovating platform experiences, or gathering data for business analytics, while researchers, journalists, civil society organizations, government units are "secondary" developers in that they are not profit driven in access motivations. Here, it is worth drawing attention to the types of users that platforms serve, and the motivating conditions that platforms have in providing access to user data because of their impacts on data management, and governance over future use and long-term access. We define four different types of users in Table 1.

This access environment, with different user groups pursuing sometimes divergent goals on social media platforms, can be seen as a manifestation of post-custodialism. The platforms continue to manage information generated by their account holders indefinitely, thereby distancing archival and practitioners from materials they seek to obtain for preservation purposes. Post-custodial theory articulates a recordkeeping and information management vision in which archival institutions do

not possess records but rather provide management and oversight to records creators who retain control over materials (Bastian 2002; Ham 1981; Henry 1998). The conversation around the value of post-custodialism has long focused on electronic records, considering whether the proliferation of content types in digital environments necessitated a post-custodial approach or a recommitment to archival institutions gaining custody over records within their purview (e.g., Cook 1994; Upward and McKemmish 1994; Duranti 1996; Ngoepe 2017). While post-custodial models have proved transformative for community archives (Henningham et al. 2017) or a necessity for institutions without financial means to manage digital records (Ngoepe 2017), in social media contexts, the platforms themselves have not engaged with preservationists in order to ensure consistency and sustainability of social media data as digital information, the fastest growing form of digital cultural heritage.

Beginning in the early 2010s, methods for accessing social media data from developers' APIs such as Facebook's Open Graph API or Twitter's Streaming API were heralded as a promising new model of data access and management for social platforms (e.g., Iskold 2010). By opening the social activity streams to developers, developers' platform APIs would become a source of value and data capital for social media companies. For example, Facebook's Open Graph platform became a forum for technologists, publishers, and developers innovating on social networks to build more tools and leverage the power of Facebook's millions of users by creating new experiences within the platform.

This developers' API model also allowed Facebook to become a new kind of service provider for big data applications. As both enabler and source of social media data, platform APIs became the industry standard for forecasting new products, building tools, and interpreting the impact of social networks and the rise of mobile phones that are underwritten by the promise of constant data creation by users, increasing data collection by platforms, and more data access for second- and third-party data brokers. However, even during Facebook's Open Graph rollout in 2010, account holders immediately began to push back on the new terms of service and the platform's definition of personally identifiable information (Kang 2010). Most complaints involved the confusing ways that Facebook had tried to explain what data could be accessed when and by whom with a slew of different permissions, apps, and friend connections (Zuckerberg 2010). The backlash from individual account holders as users was initially more problematic for Facebook than the platform policies for developers collecting batches of user data. However, following the Cambridge Analytica scandal, focus returned to the higher level policies and repercussions of granting developers broad access to Facebook data and social graphs via the API (Cadwalladr and Graham-Harrison 2018; O'Sullivan 2018; Romm and Timberg 2018; Weaver 2018).

There continue to be unintended consequences to this gatekeeping model of data management after user data have been accessed and collected because of the collapsing of different types of developers into one group (typified by making money or targeting new users as measures of impact). A side effect of platforms' efforts to guard against data breaches, disinformation, and malicious actors has been to roll back API access and as a result researchers' ability to extract data for analysis suffers. These limitations show the social and technical hurdles that researchers and

institutional stewards face when governed by the terms of service built for developers concerned with innovation, engineering, and ultimately profits. Social media platforms restrict broad access to their data to discourage bad actors or protect potential revenue streams, but in the process cut off other types of API-driven activities including data collection by archives and libraries as well as research. As such, the impacts of the developers' API for social media data access are deep and wide for researchers, policy makers, and cultural institutions concerned with accurate documentation and evidence of social platforms as they become social infrastructures of public discourse, news, and information. There is tension here between access for research/innovation and access for user data protection. For example, the proposed solution to personal data breaches is to roll back access and increasingly remove PII (Personally Identifiable Information) from datasets. However, removing PII is increasingly a technical hurdle: We see that metadata are more identifying than the content itself (Perez et al. 2018). Capturing accurate social media collections for research and future access remains even harder when platforms are reactive and fixing an ongoing problem such as the spread of disinformation or computational propaganda; it creates problems for researcher communities in a variety of different ways concerned with documenting the experience or memory of users. The moral, technical and social challenges of documenting digital culture from social platforms continue to change. However, the reliance on APIs for access to social media data raises the stakes of gaining and maintaining access to social media data as social platforms themselves become post-custodial actors.

Existing initiatives to manage social media data

Challenges as profound as the ones we have articulated around digital preservation for social media data archives do not have one clear path forward toward a singular solution. But a series of projects and repositories exist which seek to help archivists and researchers manage social media data and ensure its ongoing preservation and access. Here we summarize four emerging models of access that exist leveraging APIs.

Social Feed Manager

Social Feed Manager is an open-source software tool developed by researchers and library staff at George Washington University to harvest social media data from platforms including Twitter, Tumblr, and Sina Weibo (Social Feed Manager 2018). This tool enables digital preservationists to set up queries to run and collect data from platforms with APIs that allow for such access. Given its development by information professionals working in a university library setting, the tool has a series of archival concepts built into it, including a mapping of archival lifecycle events such as records creation and appraisal onto social media digital objects (Littman et al. 2016). Social Feed Manager is currently available openly but developer support is limited to George Washington University community members but has been a

model for similar initiatives. It complies with the Twitter Terms of Service and only publishes Tweet IDs for publicly available datasets, but has more flexibility on other platforms around the collection and retention of actual content.

Documenting the now project

The Documenting the Now Project ("DocNow") emerged as a way to collect and preserve tweets related to Michael Brown's death in Ferguson, Missouri, in 2014 and has expanded to encompass a number of projects related to web archiving, social media, and norm setting around ethical practices for collection and preservation of social media data for community archives projects. DocNow is a collaborative project funded by the Mellon Foundation to support collaborators at the University of California—Riverside, the University of Maryland, and Washington University in St Louis. Their catalog of Tweet ID datasets adheres to "Twitter's terms of service [which] don't allow tweet datasets to be published on the web, but [do] allow tweet identifier datasets to be shared. This speaks to users' rights as content creators, while also allowing researchers to share their data with others" (Documenting the Now 2019). DocNow continues to lead the archival community in theorizing and researching the ethical consideration of documenting social movements, communities, and users online (Jules, Summers and Mitchell 2018).

The inter-university consortium for political and social research

The Inter-university Consortium for Political and Social Research (ICPSR), a data repository with significant experience curating and preserving a range of social science data, is also in the process of developing a social media data archive (Hemphill et al. 2018). This initiative will similarly reflect the terms of service (TOS) of existing platforms and, in the case of Twitter, only accept deposits in the form of Tweet IDs. This federated model will also include metadata enhancements to collections, providing additional descriptive context for the collections. While this system is still in development, its organizational home within one of the oldest research data repositories in the world signals that the social science research community is reckoning with the use of social media data as a valuable source for analysis and interpretation of society. To place collections of tweets and Facebook posts alongside social research datasets such as the National Crime Victimization Survey demonstrates the importance of social media data to the current practice of science.

Social Science One

Another model for providing access to social media data is that of Social Science One, a nonprofit enterprise bringing together academics and industry partners to provide access to private sector social media data for research purposes while allowing platform companies to maintain control of their data and ensure that personal information about users can be protected (Social Science One 2018). This initiative was born out of a recognition that academia and industry have fundamentally different

perspectives on the value and use of social media data in research. The organization has partnered with Facebook and will use a peer review-style process to screen proposals from researchers and provide access to privileged platform data (King and Persily 2018). Social Science One will use platform-developed APIs to grant researcher access, thereby protecting corporate interests through the propagation of new data streams. In this example, APIs beget APIs for access to social media data. The resulting situation leaves long-term preservation outside of platforms off the table, privileging access to industry-sanctioned partnerships over archival custodianship of these digital records. This model solely relies on platforms' invested interest, timelines, staff resources, and funding constraints. For those users, researchers, and archivists concerned with accountability strategies for platforms and long-term access, we find this industry-driven model to be opaque and the least sustainable of the models that we have surveyed.

Additional examples of social media preservation demonstrate the limitations of some existing frameworks for managing digital objects and web archives. The Wayback Machine, perhaps the largest and most well-known web archive, contains only sporadic coverage of social media sites in its large collection (Bohannon 2017). The Library of Congress Twitter Archive project attempted to collect and preserve every tweet posted, maintaining a version of this vast resource in a public institution with the goal of making its contents as available as possible (Zimmer 2015). However, scaling up the initiative as Twitter grew exponentially following the 2010 agreement proved both technically and organizationally difficult. In 2017, the Library announced its plans to scale back Twitter acquisition and collect tweets "on a selective basis" (Osterberg 2017). This significant scaling back of Twitter preservation efforts indicates the degree of difficulty in collecting, describing, preserving, and providing access to this information. The Library was not able to take the project to its completion and has instead re-framed its social media preservation as another of its web archiving projects with a clearly defined scope and mission (Fondren and Menard 2018).

Conclusion: confronting the platformization of digital cultural memory

Here we have shown how social media platform APIs, in particular, are used to provide access to different kinds of users resulting from a new, post-custodial problem space of content that is locked into platforms. Platform APIs and their terms of service with developers include a big tent of users—ranging from data brokers, to social scientists, to software developers and technologists, to digital archivists (among many more). As a result, many of our emerging preservation models and stewardship challenges are tied to API-driven access regimes with one-size-fits-all approaches to user motivations for data collections. Not only are there new kinds access motivations as a result of API collection contexts, it is becoming obvious that platforms are increasingly becoming post-custodial data managers with different motivations, agency, and autonomy exerted over social media data collections than

the users who create and access platform content, or the stewards who may want to document, or preserve this digital culture.

The API as an artifact and an access point to the so-called age of algorithms remains understudied by archivists and cultural memory institutions. As we articulated through our typology of users and motivations above, not all API users have the same reasons or motivations for accessing social media data, nor are they all concerned with the long-term authentication, preservation and archival access. Whether they are open-source tools for collecting data from APIs, federated collections from researchers, or partnerships with platforms, each of these models of social media data archives relies on APIs as technologies of custody in gaining access to data and creating collections. The impact of developers' APIs on the role of the archivists as "developer stewards" continues to unfold as they assert a rescue function, attempting to produce stable, authentic facsimiles of platform content for digital preservation with tools that are known to strip valuable context. What remains to be worked out, and Lynch (2017) identifies this prescient problem precisely, is the work archivists (among all kinds of various concerned documenters and memory workers) must do in order to clarify commitments to preservation mandates when confronting the "platformization" of digital cultural memory (Nieborg and Poell 2018). In their early work on the memorialization practices of Facebook users, Acker and Brubaker (2014) argued that traditional archivists and users should take a "platform perspective" when confronting the access barriers to social media data archives. Drawing on Gillespie's (2010) work on platforms as intermediaries, Acker and Brubaker argue that platforms represent "a tension of accountability and access between service providers and content creators, futures users, and […] archivists" (2014, p. 8). Identifying these accountability tensions between profit-driven platforms and the ecology of users platforms serve illustrates how platform lock-in and API access impact social media data archives, and clarifying problems of archival persistence is an ongoing, professional, and ultimately moral challenge for cultural heritage institutions and professional archivists.

In this paper, we have introduced a number of emerging US-based models for preserving and providing access to social media data extracted from platforms with APIs. Despite these efforts, social media data archives will continue to be vulnerable to shifting developers' Terms of Service for as long as all API users are subject to the same access and reuse policies. Early post-custodial theories of archives asserted visions where creators themselves would retain control over their own collections. However, with the rise of networked platforms and cloud storage, few individual creators assert meaningful control or even have access to the digital records, evidence, and information they create and stored in such platforms. Instead, platforms govern control and access to this culturally significant digital information without ensuring long-term preservation or taking archival responsibilities. We argue that in API-driven access regimes, it is developers who can assert control and custody, building social media data collections through extraction using APIs. Given the realities and challenges of the platformization of digital cultural memory, there is now a distributed responsibility for preservation between platform intermediaries, creators, and developers who use APIs to extract and access data. With a platform perspective toward accountability, the archival community should envision new models of

meaningful control and custody over social media data. We must reorient theories of digital preservation and professional vision to account for these data-driven, algorithmic-intensive platforms by asserting custody as developer stewards using APIs, for it is only through their active use that we can interpret, know, and critique their impact on the preservation of digital cultural memory.

References

Acker A (2018) A death in the timeline: memory and metadata in social platforms. J Crit Libr Inf Stud 2(1):27

Acker A, Brubaker JR (2014) Death, memorialization, and social media: a platform perspective for personal archives. Archivaria 77:1–23

Acker A, Donovan J (2019) Data craft: a theory/methods package for critical internet studies. Inf Commun Soc. https://doi.org/10.1080/1369118X.2019.1645194

Acker A, Kriesberg A (2017) Tweets may be archived: civic engagement, digital preservation and Obama White House social media data. Proc Assoc Inf Sci Technol 54(1):1–9. https://doi.org/10.1002/pra2.2017.14505401001

Activity streams: specifications (2019) http://activitystrea.ms/. Accessed 6 May 2019

Bastian J (2002) Taking custody, giving access: a postcustodial role for a new century. Archivaria 53:76–93

Bohannon L (2017) Wayback Machine archives websites for over 20 years. Spartan Newsroom 7 Dec 2017. https://news.jrn.msu.edu/2017/12/wayback-machine-archives-websites-for-over-20-years/ Accessed 27 Aug 2019

Brügger N (2017) Chapter 23: webraries and web archives—the web between public and private. In: Baker D, Evans W (eds) The end of wisdom? The future of libraries in a digital age. Elsevier, Amsterdam, pp 185–190. https://doi.org/10.1016/B978-0-08-100142-4.00023-3

Bruns A (2013) Faster than the speed of print: reconciling 'big data' social media analysis and academic scholarship. First Monday. https://doi.org/10.5210/fm.v18i10.4879

Bruns A (2019) After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Inf Commun Soc. https://doi.org/10.1080/1369118X.2019.1637447

Bruns A, Weller K (2014) Twitter data analytics—or: the pleasures and perils of studying Twitter. Aslib J Inf Manag. https://doi.org/10.1108/AJIM-02-2014-0027

Bucher T, Helmond A (2018) The SAGE handbook of social media. In: The affordances of social media platforms, pp 233–253. https://dare.uva.nl/search?identifier=149a9089-49a4-454c-b935-a6ea7f2d8986. Accessed 27 Aug 2019

Cadwalladr C, Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17 Mar 2018. http://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed 27 Aug 2019

Considerations and recommendations concerning Internet research and human subjects research regulations, with revisions (2013) Secretary's Advisory Committee on Human Research Protections. https://www.hhs.gov/ohrp/sites/default/files/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf. Accessed 27 Aug 2019

Consultative Committee for Space Data Systems (2012) Reference model for an open archival information system (OAIS): recommended practice. CCSDS 650.0-M-2. https://public.ccsds.org/pubs/650x0m2.pdf Accessed 5 Sept 2019

Conway P (2010) Preservation in the age of Google: digitization, digital preservation, and dilemmas. Libr Q Inf Commun Policy 80(1):61–79. https://doi.org/10.1086/648463

Cook T (1994) Electronic records, paper minds: the revolution in information management and archives in the post, custodial and post, modernist era. [Based on a presentation delivered by the author during his November 1993 Australian tour]. Arch Manuscr 22(2):300

Documenting the now (2019) Tweet ID datasets catalog. https://www.docnow.io/catalog/ Accessed 27 Aug 2019

Driscoll K, Walker S (2014) Big data, big questions| working within a black box: transparency in the collection and production of big Twitter data. Int J Commun 8:20

Duranti L (1996) Archives as a place [Paper presented at a half day seminar in Sydney on 19 October 1995]. Arch Manuscr 24(2):242

European Parliament (2016) 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. 2016/679 §

Evans J, McKemmish S, Daniels E, McCarthy G (2015) Self-determination and archival autonomy: advocating activism. Arch Sci 15(4):337–368. https://doi.org/10.1007/s10502-015-9244-6

Faniel IM, Yakel E (2011) Significant properties as contextual metadata. J Libr Metadata 11(3–4):155–165. https://doi.org/10.1080/19386389.2011.629959

Fondren E, Menard MM (2018) Archiving and preserving social media at the Library of Congress: institutional and cultural challenges to build a Twitter archive. Preserv Digit Technol Cult 47(2):33–44. https://doi.org/10.1515/pdtc-2018-0011

Freelon D (2018) Computational research in the post-API age. Political Commun 35(4):665–668. https://doi.org/10.1080/10584609.2018.1477506

Garrett D, Waters J (1996) Preserving digital information, report of the task force on archiving of digital information, pp. 1–71. https://www.clir.org/pubs/reports/pub63/. Accessed 27 Aug 2019

Gillespie T (2010) The politics of 'platforms'. New Media Soc 12(3):347–364. https://doi.org/10.1177/1461444809342738

Glassman D (2019) Facebook is creating records—but who is managing them? Arch Manuscr. https://doi.org/10.1080/01576895.2019.1614077

Ham F (1981) Archival strategies for the post-custodial era. Am Arch 44(3):207–216. https://doi.org/10.17723/aarc.44.3.6228121p01m8k376

Hargittai E, Sandvig C (eds) (2015) Digital research confidential: the secrets of studying behavior online. https://mitpress.mit.edu/books/digital-research-confidential Accessed 27 Aug 2019

Hedstrom M, Lampe C (2001) Emulation vs. migration: do users care? RLG DigiNews 5(6):5–11

Helmond A, Nieborg DB, van der Vlist FN (2017) The political economy of social data: a historical analysis of platform-industry partnerships. Proc 8th Int Conf Soc Media Soc 38:1–5. https://doi.org/10.1145/3097286.3097324

Hemphill L, Leonard SH, Hedstrom M (2018) Developing a social media archive at ICPSR. In: Proceedings of web archiving and digital libraries (WADL'18). Presented at the Web Archiving and Digital Libraries 2018, New York, NY. http://hdl.handle.net/2027.42/143185

Henningham N, Evans J, Morgan H (2017) The Australian Women's Archives Project: creating and co-curating community feminist archives in a post-custodial age. Aust Fem Stud 32(91–92):91–107. https://doi.org/10.1080/08164649.2017.1357015

Henry L (1998) Schellenberg in cyberspace. Am Arch 61(2):309–327. https://doi.org/10.17723/aarc.61.2.f493110467x38701

Iskold A (2010) Facebook Open Graph: the definitive guide for publishers, users and competitors. read*w*rite, 23 April 2010 https://readwrite.com/2010/04/23/facebook_open_graph_the_definitive_guide_for_publishers_users_and_competitors/. Accessed 1 Aug 2018

John NA, Nissenbaum A (2019) An agnotological analysis of APIs: or, disconnectivity and the ideological limits of our knowledge of social media. Inf Soc 35(1):1–12. https://doi.org/10.1080/01972243.2018.1542647

Jules B (2018) We're all bona fide. Medium, 5 January, 2018. https://medium.com/on-archivy/were-all-bona-fide-f502bdaea029. Accessed 3 Aug 2019

Jules B, Summers E, Mitchell V (2018) Ethical considerations for archiving social media content generated by contemporary social movements: challenges, opportunities, and recommendations. https://www.docnow.io/docs/docnow-whitepaper-2018.pdf. Accessed 5 Sept 2019

Kang C (2010) Facebook moves to fix privacy loophole after WSJ review. Post Tech, The Washington Post, 21 May 2010. http://voices.washingtonpost.com/posttech/2010/05/facebook_moves_to_fix_privacy.html. Accessed 1 Aug 2018

Kelleher C (2017) Archives without archives: (re)locating and (re)defining the archive through post-custodial praxis. J Crit Libr Inf Stud. https://doi.org/10.24242/jclis.v1i2.29
Kerchner D, Littman J, Peterson C, Smallen V, Trent R, Wrubel L (2016) The provenance of a tweet. https://gwu-libraries.github.io/sfm-ui/resources/provenance-of-tweet.pdf. Accessed 5 Sept 2019
King G, Persily N (2018) A new model for industry-academic partnerships. https://gking.harvard.edu/files/gking/files/partnerships.pdf. Accessed 5 Sept 2019
Lee CA (2010) Open archival information system (OAIS) reference model. In: Drake M (ed) Encyclopedia of library and information sciences. CRC Press, Boca Raton, pp 4020–4030
Littman J (2019) Twitter's developer policies for researchers, archivists, and librarians. Medium, 8 January, 2019. https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2. Accessed 3 Aug 2019
Littman J, Chudnov D, Kerchner D, Peterson C, Tan Y, Trent R, Vij R, Wrubel L (2016) API-based social media collecting as a form of web archiving. Int J Digit Libr. https://doi.org/10.1007/s00799-016-0201-7
Lynch C (2017) Stewardship in the "age of algorithms." First Monday, 22(12). http://firstmonday.org/ojs/index.php/fm/article/view/8097 Accessed 5 Sept 2019
Ngoepe M (2017) Archival orthodoxy of post-custodial realities for digital records in South Africa. Arch Manuscr 45(1):31–44. https://doi.org/10.1080/01576895.2016.1277361
Nieborg DB, Poell T (2018) The platformization of cultural production: theorizing the contingent cultural commodity. New Media Soc 20(11):4275–4292. https://doi.org/10.1177/1461444818769694
O'Sullivan D (2018) Scientist at center of data controversy says Facebook is making him a scapegoat. CNNMoney, 20 March, 2018. https://money.cnn.com/2018/03/20/technology/aleksandr-kogan-interview/index.html. Accessed 1 Aug 2018
Osterberg G (2017) Update on the Twitter archive at the library of congress. Library of Congress Blog, 26 December, 2017. https://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2/. Accessed 5 Sept 2019
Pasquale F (2016) The black box society: the secret algorithms that control money and information (Reprint edition). Harvard University Press, Cambridge
Perez B, Musolesi M, Stringhini G (2018) You are your metadata: identification and obfuscation of social media users using metadata information. http://arxiv.org/abs/1803.10133. Accessed 5 Sept 2019
Phillips M, Bailey J, Goethals A, Owens T (2014) The NDSA levels of digital preservation: an explanation and uses. http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf. Accessed 5 Sept 2019
Rimkus K, Padilla T, Popp T, Martin G (2014) Digital preservation file format policies of ARL member libraries: an analysis. D Lib Mag. https://doi.org/10.1045/march2014-rimkus
Romm T, Timberg C (2018) FTC opens investigation into Facebook after Cambridge Analytica scrapes millions of users' personal information. Washington Post, 20 March, 2018. https://www.washingtonpost.com/news/the-switch/wp/2018/03/20/ftc-opens-investigation-into-facebook-after-cambridge-analytica-scrapes-millions-of-users-personal-information/. Accessed 1 Aug 2018
Rosenzweig R (2001) The road to Xanadu: public and private pathways on the history web. J Am Hist 88(2):548–579. https://doi.org/10.2307/2675105
Snell JM, Prodromou E (2017a) Activity Streams 2.0. W3C Recommendation, 23 May, 2017. https://www.w3.org/TR/activitystreams-core/. Accessed 10 Sept 2018
Snell JM, Prodromou E (2017b) Activity Streams 2.0 core. W3C, 23 May, 2017. https://www.w3.org/TR/activitystreams-core/. Accessed 10 Sept 2018
Social Feed Manager (2018) Social Feed Manager: helping researchers and archivists build social media collections. https://gwu-libraries.github.io/sfm-ui/. Accessed 5 Sept 2019
Social Science One (2018) Overview. Social science one. https://socialscience.one/overview. Accessed 5 Sept 2019
Summers E (2019). Streams. Medium, 12 February, 2019. https://news.docnow.io/streams-e62c17b3fd0d. Accessed 3 Aug 2019
Twitter (2018) Developer Terms. Developer agreement and policy: be a good partner to Twitter. (25 May, 2018). https://developer.twitter.com/en/developer-terms/agreement-and-policy.html. Accessed 10 Sept 2018
Upward F, McKemmish S (1994) Somewhere beyond custody: literature review. Arch Manuscr 22(1):136
Walker S (2017) The complexity of collecting digital and social media data in ephemeral contexts (Thesis). https://digital.lib.washington.edu:443/researchworks/handle/1773/40612. Accessed 5 Sept 2019

Weaver M (2018) March 21). Facebook scandal: I am being used as scapegoat—academic who mined data. The Guardian, 21 March, 2018. http://www.theguardian.com/uk-news/2018/mar/21/facebook-row-i-am-being-used-as-scapegoat-says-academic-aleksandr-kogan-cambridge-analytica. Accessed 5 Sept 2019

Zimmer M (2015) The Twitter archive at the library of congress: challenges for information practice and information policy. First Monday, 20(7). http://firstmonday.org/ojs/index.php/fm/article/view/5619. Accessed 5 Sept 2019

Zubiaga A (2018) A longitudinal assessment of the persistence of twitter datasets. J Assoc Inf Sci Technol 69(8):974–984. https://doi.org/10.1002/asi.24026

Zuckerberg M (2010) From Facebook, answering privacy concerns with new settings. The Washington Post, 24 May, 2010. http://www.washingtonpost.com/wp-dyn/content/article/2010/05/23/AR2010052303828.html. Accessed 1 Aug 2018

Amelia Acker is an assistant professor of information at the University of Texas at Austin, School of Information. She currently studies data literacy, social media metadata, and information infrastructures that support long-term cultural memory with digital preservation.

Adam Kriesberg is an assistant professor of library and information science at Simmons University, School of Library and Information Science. His research focuses on digital preservation, digital curation, data management, and public sector information, and he has experience in teaching a range of courses in the areas of archives and digital curation.