SSOAR

Open Access Repository
www.ssoar.info

Twitter as a first draft of the present - and the challenges of preserving it for the future

Bruns, Axel; Weller, Katrin

Postprint / Postprint
Konferenzbeitrag / conference paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Bruns, A., & Weller, K. (2016). Twitter as a first draft of the present - and the challenges of preserving it for the future. In W. Nejdl, W. Hall, P. Parigi, & S. Staab (Eds.), *WebSci '16 : proceedings of the 8th ACM Conference on Web Science* (pp. 183-189). New York: Association for Computing Machinery (ACM). https://doi.org/10.1145/2908131.2908174

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46886-1

Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Bruns, Axel & Weller, Katrin
(2016)
Twitter as a first draft of the present – and the challenges of preserving it for the future. In
Nejdl, Wolfgang, Hall, Wendy, Parigi, Paolo, & Staab, Steffen (Eds.)
Proceedings of the 8th ACM Conference on Web Science, ACM, Hannover, Germany, pp. 183-189.

This file was downloaded from: http://eprints.qut.edu.au/95296/

© Copyright is held by the owner/author(s). Publication rights licensed to ACM.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

http://doi.org/10.1145/2908131.2908174

Twitter as a First Draft of the Present – and the Challenges of Preserving It for the Future

Axel Bruns
Digital Media Research Centre,
Queensland University Technology
Phone: +61 7 3138 5548
a.bruns@qut.edu.au

Katrin Weller
GESIS Leibniz Institute for the Social Sciences
Computational Social Science
Phone: +49 221 47694 472
katrin.weller@gesis.org

ABSTRACT
This paper provides a framework for understanding Twitter as a historical source. We address digital humanities scholars to enable the transfer of concepts from traditional source criticism to new media formats, and to encourage the preservation of Twitter as a cultural artifact. Twitter has established itself as a key social media platform which plays an important role in public, real-time conversation. Twitter is also unique as its content is being archived by a public institution (the Library of Congress). In this paper we will show that we still have to assume that much of the contextual information beyond the pure tweet texts is already lost, and propose additional objectives for preservation.

CCS Concepts
• Collaborative and social computing → • Collaborative and social computing theory, concepts and paradigms → • Social media

Keywords
Twitter; social media; user-generated content; cultural heritage; archiving; history; historical sources.

1. INTRODUCTION
Scholars in the field of history have so far rarely studied user-generated online content from social media. Although historians are well represented in digital humanities initiatives, their focus is mainly on transferring existing sources into digital formats (through digitization projects, e.g. for digital editions) rather than on discovering new sources of historically important material online. But recently, interest in understanding the value of social media or other Internet data as historical sources has emerged amongst the historian community [15; 18; 40]. Social media can be considered as a new entry point to our "collective memory" and thus broaden historical practices such as those suggested by [2]. This paper demonstrates why Twitter should be considered a particularly valuable source on history-as-it-happens – and what challenges we need to solve today in order to avoid losing the information it contains for future historians.

User-generated content is increasingly interwoven with traditional journalistic sources and mainstream media. Within the overall media ecology, Twitter has by now established itself as a key social media platform which – compared to its major competitor, Facebook – plays an especially important role in public, real-time conversation and information exchange [12; 20; 16]. Twitter should therefore be of interest not only for today's journalists, but also for future historians: Twitter users comment on acute events (e.g. elections, natural disasters, protests, televised events) as well as on everyday life in real time; in addition, important contemporary figures like the President of the United States or the Pope are using Twitter to communicate with the public (or conversely, the public use Twitter to contact or discuss @barackobama or @pontifex).

This paper reconceptualizes Twitter as a historical source. It addresses scholars from the digital humanities who should be enabled to then transfer concepts from traditional source criticism to new media formats and user-generated content, and experts at cultural heritage institutions who could start to implement some of the suggested steps for preserving Twitter. In contrast to previous work by Risse et al. [37], this paper takes a primarily conceptual and less technical perspective, and includes more recent information on the state of archiving Twitter at the Library of Congress as well as general thoughts on preserving Twitter as a cultural artifact in its own right.

2. THE CASE FOR ARCHIVING TWITTER

2.1 Why Twitter?
Twitter is one of a number of currently prominent social media platforms that emerged in the early to mid-2000s; amongst this group are also the market leader Facebook, the visually focused Instagram (now itself owned by Facebook), and the predominantly Chinese-language platform Weibo. Common to all these platforms is a very large number of registered accounts (for Twitter, that number is now above one billion), of whom a still very substantial subset are classified as "monthly active users" who post to the platform at least once per month; the number of users who use social media platforms simply to "listen" and follow other users' updates [14] rather than post their own is greater still. Twitter, Inc. itself currently estimates its number of monthly active users to be around 320 million [49].

Unique to Twitter is the specific communicative structure underlying the platform. While Facebook's network – known as its "social graph" [17] – is built fundamentally around reciprocal 'friend' relationships and therefore privileges smaller-scale, stronger-tie networks, on Twitter it is possible to follow any of the 95% of Twitter accounts that have chosen to make their posts globally public [10], without a need for these accounts to follow back. This creates a network structure that enables considerably larger, weaker-tie networks in which substantial numbers of users

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
WebSci '16, May 22 - 25, 2016, Hannover, Germany
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4208-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2908131.2908174

follow globally recognized individuals and organizations including political leaders, celebrities, news outlets, and commercial entities, as well as their own friends and acquaintances. Comprehensive studies of entire national Twitterspheres show the presence of large clusters centered around shared thematic interests from politics to sports, for instance [5].

This more widely connected, weaker-tie structure has proven to be especially responsive to breaking news events and similar rapid information cascades, facilitating the broad dissemination of emerging information within very short timeframes. Such rapid dissemination is often described using metaphors of information contagion, and 'viral' transmission processes have been observed for major political crises and natural disasters [21; 31; 32; 33], but also for Internet memes and other forms of popular cultural content. Their speed of dissemination is further enhanced by the use of hashtags to amplify their visibility: the inclusion of hashtags (short topical keywords prefixed with the hash symbol '#') in tweets enables users to track all public tweets containing the same hashtag independent of whether they already follow the accounts posting those tweets. Hashtags and keywords or phrases that are rapidly coming into widespread use amongst global and local Twitter populations are in turn tracked and highlighted through Twitter's own 'trending topics' functionality, which further amplifies their visibility and thus increases the likelihood of relevant tweets being distributed more widely. The dynamics of such rapid information dissemination have been highlighted in a number of studies [13; 8], and Twitter has been shown to be a more effective medium for such processes than its major competitor Facebook [16].

But the viral distribution of breaking news on Twitter does not constitute a purely linear, outward transmission of information to an ever-wider circle of receivers; rather, at every step of the process such information is also likely to trigger further individual responses and interpersonal discussion. Such responses may not always themselves be as widely disseminated as the original information, but they offer important first-hand insights into how individuals and groups process, interpret, and contextualize the information they are exposed to, as indeed the dynamics of dissemination themselves point to the extent to which different items of information are perceived, evaluated, and considered to be worthy of sharing on further. For each individual decision as well as in a global aggregate, these observations thus point to the relative importance accorded to specific issues and topics by individual users and societies as a whole.

Additionally, Twitter is of course not only, and perhaps not even predominantly, a medium for the rapid dissemination of breaking news; rather, the current state of Twitter research [cf. 50; 34] shows it to be a widely and diversely used platform that caters to practices ranging from interpersonal communication to political campaigning. Recent contributions to Twitter scholarship [11] have pointed out a need to move beyond a focus on breaking news, and on the hashtags that often signify such events, and towards studies of more everyday user practices, in order to document more fully the lived experience of using the platform. Such work has already served to "debanalize" Twitter, as Richard Rogers has pointed out [39], by combatting the early misperception that the platform was dominated by streams of "pointless babble" [25] or of users posting pictures of their lunch; it has argued convincingly that what was misunderstood and dismissed as "pointless babble" is indeed deeply meaningful (if highly phatic and ephemeral) communication that maintains social relationships and contributes considerably to the persistence of connections within the Twitter community [29].

2.2 Twitter as a First Draft of the Present

The 500 million tweets posted every day [24] thus serve one or both of at least two major functions: in general, they constitute a first-hand record of what aspects of their daily lives and their responses to the events around them any of its 320 million monthly active users found relevant and important enough to address and share with their followers and the wider world; and in particular, they document how this widespread international userbase collectively responded to and engaged with the major local, national, and global breaking news events that they were exposed to through this and other media channels.

Both of these functions – and there is no reason to believe that these are the *only* possible functions the platform could serve for any one of its users – make Twitter an important medium for historians to observe and preserve. Further, the 140-character limit for individual tweets especially encourages brief, quick responses in the moment rather than the longer, more fully considered reflections after the fact that blogs, email, or offline genres such as diaries or letters would favor. This very real-time, conversational nature of communication on Twitter means that what we may observe as we engage with communication data from this platform represents a dynamic first take on experiences in everyday life and events in the news – an *ad hoc* interpretation that is not yet settled, and whose gradual transformation in response to further experiences and information we are able to observe, at the individual as well as at a group, community, or societal level.

Journalism has long been described as "a first rough draft of history" [41], because its need to cover the news in (at the time) daily installments necessarily meant that it had to engage in the writing of very recent history while that history itself still continued to unfold. The cycles of journalism itself have sped up with the introduction of broadcast media and 24-hour news channels in the meantime, but real-time social media platforms such as Twitter have served to increase that speed yet again, arguably beyond the capabilities of even the best-resourced news organizations: breaking news now regularly surfaces on Twitter through the "random acts of journalism" [27] committed by users who happened to be on the scene of an event, well before journalists could be mobilized to cover it, and the first Twitter comments and reflections on such events follow soon after. In so doing, Twitter serves as a medium for "ambient news" [20; 12].

We argue, therefore, that Twitter and similar social media should be understood, in a play on the earlier phrase, as providing *a first draft of the present*: they offer a rich, diverse, collectively authored, and comprehensive stream of real-time updates on what is happening in the world at this very moment, and on how the people to whom it is happening think and feel about it. In time, through the interplay of personal reflection and further social and mainstream media coverage, such thoughts and feelings are likely to settle into a more firmly held interpretation of these events that agrees with or opposes the journalistic coverage of the same events; even later, such accounts become the material basis upon which history itself is written. But at the time that they are observed on, and can be captured from, the Twitter stream, they remain fresh, immediate, and as yet unsettled responses.

Given the considerable role that social media play as platforms for public communication in contemporary society, there is therefore now an acute need to comprehensively archive and preserve public expression through these platforms in order to afford future historians the best possible chance to write the history of our present not just from the perspective of those few whose stories are covered in the news media, but rather by taking into account

the much broader and more diverse responses made by the many hundreds of millions who participate in platforms such as Twitter each month. To have access to these users' immediate responses to the events around them means not to have to rely on the journalistic process and on the official historiography to reconstruct past events and their likely impact on the general population, but to have access to individuals' and communities' evolving views on the world around them first-hand, in their own words, on a day-by-day basis. And while all mainstream social media platforms make important contributions to that overall, fine-grained, immediate picture, we argue that because of its specific structure and affordances, and the comparative potential accessibility of its data on users' public communicative acts, Twitter serves as an obvious starting-point for this task.

3. CURRENT APPROACHES FOR PRESERVING TWITTER

3.1 Major Challenges

Twitter therefore constitutes an important source about our present that should be preserved. Technically, it is already used to preserve collections of tweets, e.g. collections based on hashtags, users or other search criteria. However, this is restricted by regulations around the Twitter APIs, e.g. the restriction to collecting at most 1% of the current Twitter volume from the public APIs. And once the data is collected, the Twitter Terms of Service [48] impose substantial constraints on efforts to share archived data. In particular, they only allow the sharing of collections of tweet IDs instead of tweet texts themselves (with the exception of smaller collections, e.g. as Excel sheets). While this may have been established as a means to protect user privacy, it significantly hinders the preservation of Twitter data and working with archived tweet collections (see section 4.1). Despite these challenges, there are more and more individual approaches to archiving single datasets collected from Twitter for research purposes (e.g. at the ICWSM conference series [22]), or publicly shared thematic tweet collections, e.g. offered through the Internet Archive (like a collection of tweets commenting on the shooting of Michael Brown in Ferguson, MO, in August 2014 [44]). Also, first initiatives exist to solve more technical challenges in archiving social media content, such as the ARCOMEM project [37]. In the long run, it is desirable that such individual efforts are more and more embedded in collaborative initiatives in order to avoid them being dependent on the funding available to individual institutions or projects – and in order to make them more easily retrievable. Both the Web Observatory Community [46] and the Web Archiving community in general and the International Internet Preservation Consortium [23] in particular could act as such frameworks in the future (and have already begun to address some of these challenges).

Archiving individual datasets (e.g. defined by shared keywords or hashtags) is already challenging, although it has led to some first useful approaches. But it is an entirely different challenge to preserve Twitter itself as a cultural artifact. For this purpose, one has to face the challenge that Twitter is not only a collection of text strings with additional metadata stored in CSV or JSON files. Twitter also consists of various non-textual elements, such as (audio-)visual information (e.g. user profile pictures and the images and videos shared in tweets), of connections to content outside of Twitter (via hyperlinks in tweets or in user profiles), and of various functionalities that are embedded in the user interface (which again may vary if used on different types of devices) and that through the look and feel of the platform may influence the way in which users can interact with each other. Twitter is an interactive environment that is evolving over time, with new features being introduced and new practices emerging with its users. As it happens, Twitter users themselves have also already influenced the look and feel of the platform: hashtags, @messages and retweets were all initially invented by ordinary users, before Twitter turned them into platform features [19]. Currently, no feasible concept exists to preserve the Twitter experience itself as a whole, and not just parts of its user-generated content.

For specific case studies it does make sense to preserve single collections of tweets, but there is always the challenge of how to decide on which data to collect and preserve. It may not be possible to predict what is considered of value for the future – we can assume that the Twitter activity around some of today's news events will be of interest for future historians, such as the Arab Spring movement or the protests at Gezi Park. But we cannot anticipate all of the possible research questions that future historians would want to ask of a Twitter database. Historians are used to working with the material they find to be available at archives or in other collections – which have never been complete. Archivists and librarians are also used to selecting the material they consider to be the most appropriate and important for preservation. One current problem is that no-one is taking care of this selection process for social media content yet, and that it might be too late to do so in the future. Preserving Twitter in its entirety would thus be the ideal scenario to ensure that it will still be of value in the future, as even collections of tweets might be of limited use if their context is lost.

By *entirety*, we mean the following dimensions: (1) the cultural artifact that is Twitter, with (1a) its look and feel and technical affordances over the course of time, and (1b) the broader societal context into which Twitter is embedded, including user numbers, demographics and usage practices, and (2) the Twitter data consisting of (2a) the complete collection of all user-generated content, including non-textual information and hyperlinks, and (2b) contextual information like collections of hashtags for important events or lists of usernames for important groups of users.

In light of this situation, many hope for the Library of Congress to resolve these problems [26]. However, we will show that this is currently not the case.

3.2 The Library of Congress

In 2010, Twitter Inc. announced that it would donate its entire collection of tweets to the Library of Congress (LC) for 'preservation and research' [43]. With this approach, Twitter became the first (and still is the only) major social media platform whose content is being archived in its entirety by a public institution. However, little is known to the public about the exact nature of this archive, and the LC has not yet opened the archive to any users. The public announcements concerning the Twitter Archive at the LC are rare. At about the same time as Twitter made its announcement about the collaboration in 2010, the LC also released a blog post and some FAQ [35; 36]. The next public announcement followed three years later with another blog post [1] and a white paper [28] on the state of the Twitter Archive – but without any details about its future availability. Zimmer [51] has discussed in some detail the challenges that the LC may be facing with this unprecedented endeavor, including privacy issues.

In the meantime, one of the authors of the present paper has conducted a research fellowship in Digital Studies at the LC's John W. Kluge Center. An initial announcement that the Twitter Archive for the years 2006-2010 could be used during the Fellowship was soon revoked – the archive at the Library of Congress is

still not ready to be used by researchers, and it remains as uncertain as ever whether and when it will become available [30]. Based on the experiences obtained while staying at the LC we must also assume that most of the contextual information beyond the pure tweet texts and formal metadata is already lost: visual features of the platform itself are not captured, shortened URLs from tweets are not resolved, and the Websites, images and videos referenced by such URLs are not stored, for example. To the best of our knowledge, the Library of Congress is obtaining the data via Twitter, Inc.'s subsidiary data reseller GNIP in a specific, text-based format, and is currently storing incoming tweets on various tapes (for tweets collected in a certain period of time) in a non-searchable way. The preservation of such core materials is certainly welcome in its own right, but an archive that remains inaccessible and omits key contextual elements from capture is only of limited value to future historians. We therefore have to consider additional approaches and initiatives for truly preserving Twitter.

3.3 Observing the Evolution of a Social Media Platform?

In order to preserve social media platforms such as Twitter as evolving cultural artifacts of their own, we need more critical thoughts in line with what Rogers [38] calls "site biography". Little practical guidance is available for approaches to this mission, and those who want to attempt it will have to find useful sources that enable them to understand the evolutionary processes in the first place. Such resources may be announcements from the official Twitter blog on the introduction of new features (e.g. [47]), or related news reports (e.g. recent reports on the Twitter, Inc.'s changing of the symbol for favorited tweets [3]). It may also be interesting to study introductory books on Twitter usage for different audiences (such as "Twitter for Dummies") as they may contain screenshots of what Twitter looked like in the past. Similarly, various YouTube videos exist that demonstrate how to use certain features of Twitter – which can even more vividly illustrate the evolving look and feel of the platform over the course of time. An additional source for such a project might be the Internet Archive's captures of the Twitter Website – although this is highly limited by the fact that the underlying archiving software Heretrix is not designed for capturing dynamic content. But in addition to the other sources mentioned so far, the Internet Archive's Wayback Machine includes a structured timeline which may enable tracing back single features or interface changes to a particular point in time. Finally, it might be useful to capture usage practices by interviewing different types of Twitter users.

4. WHAT DO WE LOSE?

4.1 The Dark Ages of Twitter: What is already lost?

Those historians who care for born-digital documents as historical artifacts already consider the early years of the WWW as the "Dark Ages of Internet History" [15; 18]: even with the current efforts of the Web Archiving community it will not be possible to restore some of the very early contents of the WWW. The same is likely for the history of Twitter: with the current setup, we can assume that we are already losing much of Twitter's unique look and feel. However, if considerable efforts are undertaken, it may still be possible to reconstruct the major steps of Twitter's evolution. Nevertheless, the following data are highly at risk of being permanently lost, with all of them contributing to the diffusion of context and thus complicating all present and future efforts to make sense of single tweets or tweet collections:

1. Deleted tweets: if a user decides to delete single tweets, or his/her entire Twitter account with all tweets, those tweets are supposed to also be deleted from all datasets collected from Twitter. By archiving the tweet ID only one will have to go back to the Twitter API and collect all again based on their ID (a process sometimes referred to as rehydrating of tweets) – which will only return tweets that have not been deleted in the meantime. This creates a major challenge, as in some cases, deleted tweets might be the most interesting to future historians – e.g. in cases of controversial statements by important public figures like politicians, or in case of eye-witnesses of critical events (while in other cases tweets may just be deleted to correct a typo). This current practice means that in principle, all tweet collections that respect the practice of archiving and sharing IDs only are at risk of becoming useless – either because at some point there might not be a Twitter API from which to collect the tweet content, or because crucial information is no longer available after tweets have been deleted. First experiments show that even within a few months, large numbers of IDs no longer retrieve any tweets from the API. Summers reports his attempts to 'rehydrate' a dataset about the attacks on the offices of Charlie Hebdo in early 2015: less than two months after the release of the initial dataset, more than 8% of the tweets could no longer be rehydrated [45].
2. URLs: Links to external Websites are an important element of tweets. Often, these links have been processed by URL shorteners in order to better fit the limit of 140 characters per tweet. If the URL shortener services are suspended, it may no longer be possible to look up a shortened URL from a tweet, and the link is broken. But even if the link is still working, the Website that was linked to might already be gone. Bruns & Highfield encountered this problem in processing a dataset on the 2012 US presidential election: shortly after the election, the losing Mitt Romney campaign decommissioned its custom URL shortener, which made it impossible to resolve the *mi.tt* short URLs contained in tweets posted by the campaign [7].
3. Audiovisual information: As of the time of writing we are not aware of any approaches to preserve the images or videos which are included in tweets. The free Twitter APIs and the Firehose access through GNIP provide text-based information only, so that this is the typical format that researchers work with today and that is being archived for future use. If no additional measures are considered to preserve such audiovisual information, it may be lost completely.

4.2 Next Steps for Preserving Twitter as a Historical Source

Based on the previous analysis, we suggest the following practical steps to support the preservation of Twitter and its data in a form that will be useful for future historians and critical source studies.

1. Document Twitter's evolution from today on and trace back its history as completely as possible (starting with the steps described). Ideally this should also comprise technical details, such as API functionalities.
2. Collect and document available information about Twitter's role in society and in the broader media ecology. A first step would be detailed information on user demographics (also in relation of different populations, e.g. by countries).
3. Important actors (persons as well as institutions) should be identified on Twitter, and their Twitter handles should be recorded. In some case, even official accounts change their Twitter user name (we have observed this for the Twitter handles of major football clubs, for example [9]) – and information

about which actor used which handle at a certain point in time will be both highly valuable and difficult to identify in the future. The profile pages of these actors should be accessed and documented on a regular basis. This includes visually capturing the look of the profile page, including profile images and texts used in self-descriptions. Lists of elite users could be provided and shared, including political and societal leaders.

4. One may even consider whether all verified accounts should be treated this way. A useful step would be if Twitter would exclude tweets from verified accounts from its policy that only tweet IDs may be shared. Verified accounts are less afflicted by privacy concerns (as their owners should be well aware of the publicness of their tweets), and are more likely to be of importance for studies on contemporary events.
5. Important events should be monitored. Apart from the possibility to collect tweets for single events, it will be useful to document important background information such as hashtags (e.g. for cases in which different hashtags are used over the course of time, such as the Arab Spring – which started with the hashtag #jan25 before turning to different hashtags such as #egypt and #arabspring) or other contextual information, such as that Twitter was shut down in Turkey in light of the protests at Gezi park.
6. Finally, guidelines for applying concepts of source criticism need to be discussed and validated with the help of historians or the digital humanities community. A basis for this effort could be the guidelines for verifying sources or trustworthy information used by journalists (e.g. [42]).

The measures listed here would mainly be useful to preserve information about Twitter as a cultural artifact and about specific Twitter use cases – which both constitute important contextual backgrounds needed for working with Twitter data in the future. In terms of archiving the actual Twitter data we note here that none of these efforts are able to do more than capture specific, highly limited subsets of the entire Firehose of all global tweets, however, and would therefore omit substantial volumes of material that may be considered to be of future historical significance. A convincing argument could be made that the considerable efforts that would have to go into determining the specific populations of user accounts or selections of topical keywords and hashtags to track would be better spent on developing the means to reliably archive the *full* Twitter Firehose on an ongoing basis, rather than on the technologies required to filter it down to smaller, incomplete subsets. Such comprehensive rather than selective capturing of research data, to be searched or filtered by researchers at a later date, is in line with the overall "computational turn" towards the digital humanities [4].

As the much-anticipated Twitter Archive project at the Library of Congress by now appears to have stalled indefinitely, there remains a significant need for renewed efforts from other parties, including both national libraries and major research institutions. One of the authors of the present paper is leading an effort to identify and track the public tweeting activities of all Australian Twitter accounts, in collaboration with the National Library of Australia; to do so preserves at least one entire national subset of global Twitter activity [6]. To date, this project has captured some one billion tweets from 2.8 million identified Australian accounts.

5. CONCLUSIONS

The saying that "history is written by the winners" has been variously attributed to Winston Churchill, George Orwell, and Walter Benjamin; ironically, that uncertainty demonstrates that a more accurate rephrasing might state that history is written by those whose written records survive intact. This is certainly true throughout most of human history: we know a great deal less about the illiterate peasants than about the literate ruling classes of earlier ages, and even more recent, more enlightened centuries have still seen a highly divergent level of preservation for the historical records of political and societal leaders as compared to everyday citizens, much to the frustration of social historians who are seeking to understand the full range of the lived experiences of people at any given time in recent history.

As we have demonstrated here, the wholesale archiving of social media content from platforms such as Twitter, which from a technical perspective is now well within our capabilities, would contribute substantially to capturing and preserving present-day history even before it becomes history, generating a rich and detailed record of the everyday communicative activities of users ranging from ordinary people to world leaders in their fields. Such live archiving of social media content could begin with Twitter not because it is the largest and most important social medium presently in existence, but because for practical, ethical, and privacy reasons it poses the least significant challenges: the full global Firehose of all tweets is available at least in principle, if not readily accessible to everyone, and the simple privacy controls embraced by the platform mean that it may be reasonably assumed that the vast majority of the content posted by users that is public is also meant to be public, and thus appropriate to archive and preserve. (Further extensions of social media archiving initiatives, however, should certainly strive to progress well beyond Twitter and tackle the challenges inherent in archiving other currently leading platforms such as Facebook, Instagram, and Weibo, as well as yet others that may emerge in the future.)

As we have shown, the continuous archiving of the Twitter Firehose will serve to capture an unfolding history-as-it-happens, with unprecedented immediacy: the very liveness of Twitter's real-time stream of posts that range from highly personal comments to updates on global news events means that it represents a historical record that has yet to settle and be accepted as history in the full sense of the term – as we have described it above, what Twitter provides is not even a first draft of history, as journalism has been said to do, but *a first draft of the present*. Far from devaluing the archiving task we have sketched out here, the unsettled nature of the live history that the Twitter stream depicts is one of its greatest assets: it is fundamentally impossible to determine at the time that events unfold what historical impact they may come to have, so the only option that remains is for us to preserve today whatever we can, in order to give future historians access to the richest and most comprehensive archives of life in the early twenty-first century that it is possible to create.

We suggest, therefore, that there is an acute need to tackle the project of comprehensively archiving Twitter – and its broader communicative context – as soon as possible; not just because of their past performance, but also as a matter of principle we cannot, and should not, wait for the Library of Congress and Twitter, Inc. to solve this issue for us. Recent estimates suggest that more than 500 million new tweets are posted each day [24]; this figure indicates the magnitude of historical information that is lost to future generations the longer we delay addressing this task. We would argue that the present-day community of Internet researchers and digital humanities scholars has a duty to tackle these issues without delay: we owe it to future historians to do so.

6. ACKNOWLEDGMENTS

This research was supported by the Australian Research Council through the Future Fellowship project "Understanding Intermedia Information Flows in the Australian Online Public Sphere."

Furthermore, part of this work was conducted as part of a Digital Studies Fellowship by the Library of Congress's John W. Kluge Center.

7. REFERENCES

[1] Allen, E. 2013. Update on the Twitter archive at the Library of Congress. Retrieved from http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/ (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6f6RfTg63)

[2] Assmann, J. 1995. Collective memory and cultural identity. *New German Critique* 65 (1995), 125–133.

[3] BBC. 2015. Twitter's 'favourite' stars change into 'like' hearts. Retrieved from http://www.bbc.co.uk/newsbeat/article/34713811/twitters-favourite-stars-change-into-like-hearts (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6fC8Aqjgw)

[4] Berry, D. 2011. The computational turn: thinking about the digital humanities. *Culture Machine* 12. http://www.culturemachine.net/index.php/cm/article/view/440/470.

[5] Bruns, A., Burgess, J., and Highfield, T. 2014. A "big data" approach to mapping the Australian Twittersphere. In *Advancing Digital Humanities: Research, Methods, Theories*, P.L. Arthur and K. Bode, Eds. Palgrave Macmillan, Houndmills, 113–129.

[6] Bruns, A., Burgess, J., Banks, J., Tjondronegoro, D., Dreiling, A., Hartley, J., Leaver, T., Aly, A., Highfield, T., Wilken, R., Rennie, E., Lusher, D., Allen, M., Marshall, D., Demetrious, K., and Sadkowsky, T. 2015. *TrISMA: Tracking Infrastructure for Social Media Analysis*. Retrieved from http://www.trisma.org/

[7] Bruns, A., and Highfield, T. 2016. May the best tweeter win: the Twitter strategies of key campaign accounts in the 2012 US election. In *Die US-Präsidentschaftswahl 2012: Analysen der Politik- und Kommunikationswissenschaft*, C. Bieber and K. Kamps, Eds. Springer Fachmedien, Wiesbaden, 425-442.

[8] Bruns, A., and Sauter, T. 2015. Anatomie eines Trending Topics. In *Digitale Methoden in der Kommunikationswissenschaft*, A. Maireder, J. Ausserhofer, C. Schumann, & M. Taddicken, Eds. Freie Universität Berlin, Berlin, 141-161. DOI= http://doi.org/10.17174/dcr.v2.7

[9] Bruns, A., Weller, K., and Harrington, S. 2014. Twitter and sports: football fandom in emerging and established markets. In *Twitter and Society*, K.Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, Eds. Peter Lang, New York, 263-280.

[10] Bruns, A., Woodford, D., and Sadkowsky, T. 2014. Exploring the global demographics of Twitter. Paper presented at the Association of Internet Researchers conference, Daegu, 22-25 Oct. 2014. Retrieved from http://snurb.info/node/1963

[11] Burgess, J., and Bruns, A. 2015. Easy data, hard data: the politics and pragmatics of Twitter research after the computational turn. In *Compromised Data: From Social Media to Big Data*, G. Langlois, J. Redden, and G. Elmer, Eds. Bloomsbury Academic, New York, 93-111.

[12] Burns, A. 2010. Oblique strategies for ambient journalism. *M/C Journal* 13, 2 (2010). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/view/230

[13] Chowdhury, A. 2011. Global pulse. *Twitter blog* (29 June 2011). Retrieved from http://blog.twitter.com/2011/06/global-pulse.html (accessed Feb 6, 2016).

[14] Crawford, K. 2009. Following you: disciplines of listening in social media. *Continuum* 23, 4 (2009), 525-535. DOI= http://doi.acm.org/10.1080/10304310903003270.

[15] Crueger, J. 2013. Die Dark Ages des Internet? In *Was Bleibt? Nachhaltigkeit in der digitalen Welt*, P. Klimpel and J. Kneiper, Eds. iRights:Media, 191–197. Retrieved from http://irights-media.de/publikationen/was-bleibt-nachhaltigkeit-der-kultur-in-der-digitalen-welt/

[16] Dewan, P., and Kumaraguru, P. 2014. It doesn't break just on Twitter: characterizing Facebook content during real world events. *arXiv* 1405, 4820 (2014). http://arxiv.org/abs/1405.4820

[17] Facebook, Inc. 2007, 24 May. Facebook unveils platform for developers of social applications. Retrieved from http://newsroom.fb.com/news/2007/05/facebook-unveils-platform-for-developers-of-social-applications/ (accessed Feb 6, 2016).

[18] Haber, P. 2011. *Digital Past: Geschichtswissenschaft im digitalen Zeitalter*. Oldenbourg.

[19] Halavais, A. 2014. Structure of Twitter: social and technical. In *Twitter and Society*, K.Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, Eds. Peter Lang, New York, 29–41.

[20] Hermida, A. 2010. From TV to Twitter: how ambient news became ambient journalism. *M/C Journal* 13, 2 (2014). http://journal.media-culture.org.au/index.php/mcjournal/article/view/220

[21] Hermida, A., Lewis, S.C., and Zamith, R. 2014. Sourcing the Arab Spring: a case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions. *Journal of Computer-Mediated Communication* 19, 3 (2014), 479–499. DOI= http://doi.org/10.1111/jcc4.12074

[22] ICWSM. 2012. ICWSM Dataset Sharing Service. Retrieved from: http://icwsm.cs.mcgill.ca (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6fC7JfFyR) .

[23] IIPC. International Internet Preservation Consortium. Retrieved from http://www.netpreserve.org/ (accessed Feb 6, 2016).

[24] Internet Live Stats. 2016. Twitter usage statstics. http://www.internetlivestats.com/twitter-statistics/

[25] Kelly, R. 2009. Twitter study. Pear Analytics, San Antonio. Retrieved from http://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf (accessed Feb 6, 2016).

[26] Landgraf, G. 2010. Historians await access to the Library of Congress's Twitter archive. American Libraries. Retrieved from http://americanlibrariesmagazine.org/2010/05/17/historians-await-access-to-the-library-of-congresss-twitter-archive/ (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6f6SjXzT2).

[27] Lasica, J.D. 2003. Blogs and journalism need each other. *Nieman Reports* (Fall 2003), 70-74. Retrieved from http://1e9svy22oh333mryr83l4s02.wpengine.netdna-cdn.com/wp-content/uploads/2014/04/03fall.pdf (accessed Feb 6, 2016).

[28] Library of Congress. 2013. Update on the Twitter archive at the Library of Congress. Retrieved from http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6f6TGRHnp).

[29] Marwick, A.E., and boyd, d. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13, 1 (2011), 114-133.

[30] McLemee, S. (2015) The archive is closed. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research (accessed Feb 6, 2016).

[31] Mendoza, M., Poblete, B., and Castillo, C. 2010. Twitter under crisis: can we trust what we RT? 1st Workshop on Social Media Analytics (SOMA '10), Washington, DC.

[32] Palen, L., Starbird, K., Vieweg, S., and Hughes, A. 2010. Twitter-based information distribution during the 2009 Red River Valley flood threat. *Bulletin of the American Society for Information Science and Technology* 36, 5 (2010), 13-17.

[33] Papacharissi, Z., and de Fatima Oliveira, M. 2012. Affective news and networked publics: the rhythms of news storytelling on #Egypt. *Journal of Communication* 62, 2 (2012), 266–282. DOI= http://doi.org/10.1111/j.1460-2466.2012.01630.x

[34] Rambukkana, N., Ed. 2015. *Hashtag Publics: The Power and Politics of Discursive Networks*. Peter Lang, New York.

[35] Raymond, M. 2010a. How tweet it is! Library acquires entire Twitter archive. *Library of Congress Blog* (14 April). Retrieved from http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/ (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6f6TUANAj).

[36] Raymond, M. 2010b. The Library and Twitter: An FAQ, *Library of Congress Blog* (28 April). Retrieved from http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/ (accessed Feb 6, 2016, archived by WebCite® at http://www.webcitation.org/6f6TbjMyG) .

[37] Risse, T., Peters, W., Senellart, P., and Maynard, D. 2014. Documenting contemporary society by preserving relevant information from Twitter. In *Twitter and Society*, K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, Eds. Peter Lang, New York, 207–219.

[38] Rogers, R. 2012. *Digital Methods*. MIT Press, Cambridge, MA.

[39] Rogers, R. 2014. Debanalising Twitter: the transformation of an object of study. In *Twitter and Society*, K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, Eds. Peter Lang, New York, xi-xxvi.

[40] Schreiber, C. 2012. Genuine Internetdaten als Historische Quellen: Entwurf einer korrealistischen Quellentheorie. *Zeitschrift für digitale Geschichtswissenschaften* 1. Retrieved from http://universaar.uni-saarland.de/journals/index.php/zdg/article/view/292/357

[41] Shafer, J. 2010. Who said it first? *Slate* (30 Aug. 2010). Retrieved from http://primary.slate.com/articles/news_and_politics/press_box/2010/08/who_said_it_first.html (accessed Feb 6, 2016).

[42] Silverman, C., Ed. 2014. *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage.* European Journalism Centre, Maastricht.

[43] Summers, E. 2014. ferguson-tweet-ids. Retrieved from https://archive.org/details/ferguson-tweet-ids (accessed Feb 6, 2016).

[44] Summers, E. 2015. Tweets and deletes: silences in the social media archive. Retrieved from https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed#.pay32r3eu (accessed Feb 6, 2016; archived by WebCite® at http://www.webcitation.org/6f6KxoikL)

[45] Stone, B. 2010. Tweet preservation. *Twitter Blog* (14 April 2010). Retrieved from https://blog.twitter.com/2010/tweet-preservation (accessed Feb 6, 2016).

[46] Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D. Contractor, N., and Hendler, J. 2013. The web science observatory. *IEEE Intelligent Systems* 28, 2 (2013), 100–104.

[47] Twitter, Inc. 2014. Share a tweet through direct messages. *Twitter Blog*. Retrieved from https://blog.twitter.com/2014/share-a-tweet-through-direct-messages (accessed Feb 6, 2016).

[48] Twitter, Inc. 2015. Developer agreement & policy. Retrieved from: https://dev.twitter.com/overview/terms/agreement-and-policy (accessed Feb 6, 2016).

[49] Twitter, Inc. 2016. Twitter usage: company facts. Retrieved from: https://about.twitter.com/company (accessed Feb 6, 2016).

[50] Weller, K., Bruns, A., Burgess, J., Mahrt, M., and Puschmann, C., Eds. 2014. *Twitter and Society*. Peter Lang, New York.

[51] Zimmer, M. 2015. The Twitter archive at the Library of Congress: challenges for information practice and information policy. *First Monday* 20, 7 (2015). DOI= http://doi.org/10.5210/fm.v20i7.5619