doi:10.2218/ijdc.v7i1.224

The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data

Heather Small,
Graduate School of Education and Information Studies,
University of California, Los Angeles

Kristine Kasianovitz,
Government Information Librarian,
International, State and Local Documents
Stanford University

Ronald Blanford and Ina Celaya,
Graduate School of Education and Information Studies,
University of California, Los Angeles

Abstract

Social networking sites and other social media have enabled new forms of collaborative communication and participation for users, and created additional value as rich data sets for research. Research based on accessing, mining, and analyzing social media data has risen steadily over the last several years and is increasingly multidisciplinary; researchers from the social sciences, humanities, computer science and other domains have used social media data as the basis of their studies. The broad use of this form of data has implications for how curators address preservation, access and reuse for an audience with divergent disciplinary norms related to privacy, ownership, authenticity and reliability.

In this paper, we explore how the characteristics of the Twitter platform, coupled with an ambiguous and evolving understanding of privacy in networked communication, and divergent disciplinary understandings of the resulting data, combine to create complex issues for curators trying to ensure broad-based and ethical reuse of Twitter data. We provide a case study of a specific data set to illustrate how data curators can engage with the topics and questions raised in the paper. While some initial suggestions are offered to librarians and other information professionals who are beginning to receive social media data from researchers, our larger goal is to stimulate discussion and prompt additional research on the curation and preservation of social media data.

International Journal of Digital Curation (2012), *7*(1), 174–197. http://dx.doi.org/10.2218/ijdc.v7i1.224

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The use of social networking sites and other forms of social media by a global population is a relatively recent phenomenon that has enabled new forms of narrative; researchers have a growing interest in accessing, mining, and analyzing this participatory documentation of real time events and social trends. Yet conversations between academics, librarians, archivists, online service providers and other interested parties have been relatively limited in terms of how this content should be collected, preserved and accessed, or whether it warrants preservation at all.

These issues came to our attention when the UCLA Library agreed to take on one faculty researcher's collection of Twitter-based content, which documented the early days of the 2011 protests and revolutions in Egypt and Libya. The Library is interested in using this data as a test case for its new Islandora repository, and our team (originally made up of graduate students and professionals in UCLA's *Information Studies, Data Practices* seminar) was tasked with outlining the challenges involved in curating such data. Starting from that original analysis, this paper probes more deeply into the implications that multidisciplinary use has for how curators address the preservation, access and reuse of this data for an audience with divergent disciplinary norms related to the issue of privacy. We examine privacy as it relates to and is affected by notions of ownership, authenticity and reliability, in the context of Twitter and in social media more generally.

Twitter[1] is a popular social media platform.

> "Social media is defined as Internet-based platforms and technologies that permit user interaction and/or facilitate the creation and exchange of user-generated content…. Social media data refers to the information (photos, comments, etc.) that users generate or share while engaged in or with social media." (ESOMAR, 2011).

Characteristics common to most social networking sites include the construction of user profiles, the specification of relationships with other users and access to information as a result of that relationship (Pike, Bateman, & Butler, 2009).

Social media data are not homogenous; each platform has its own socio-technical structure, which facilitates distinct types of content. These differences result in the creation of platform-specific capabilities, limitations, practices and norms. Each platform also has its own distinct audiences and users with their own unique expectations and forms of interaction (Ess & AoIR Ethics Working Group, 2002). Acknowledging these differences, we focus predominately on Twitter. We explore how the characteristics of the platform, combined with an ambiguous and evolving understanding of privacy in networked communication, and divergent disciplinary understandings of the nature of the data under analysis, create complex issues for curators in trying to ensure broad-based and ethical reuse of Twitter data, and other social media data more generally.

[1] Twitter: http://twitter.com

Our initial discussion of these issues is followed by a case study of the Twitter data set collected by Dr. Todd Presner, the UCLA faculty researcher mentioned above. In demonstrating how the issues raised in this paper apply to a specific data set, we hope to illustrate the social and technical challenges that this seemingly simple type of data can invite. Our research has left us convinced that providing for the collection, preservation and reuse of social media data necessitates an inclusive conversation among libraries, archives, institutional review boards, scholarly societies, and other national and international organizations concerned with the production and preservation of scholarship.

While we offer some initial recommendations for institutional repository staff who are beginning to receive social media data, our larger goal is to initiate a broader conversation and research agenda around this topic in order to devise appropriate standards, guidelines and practices to promote an ethical and sustainable approach to archiving social media data.

Getting the Lay of the Land: Twitter Users, Twitter Miners and Twitter Keepers

What is Twitter?

Launched in 2006, Twitter is a micro-blogging platform that allows authors to post short, 140-character messages, or “Tweets” to their network of followers.

> “Twitter prompts users to answer the question ‘What are you doing?’ creating a constantly updated timeline, or stream of short messages that range from humor... to links and breaking news.” (Marwick & boyd [sic], 2010).

The functionality and structure of Twitter revolves around illustrating connections. Twitter enables users to view content by, and connect to, people and organizations of interest that they “follow”. Users are able to view who follows whom in the Twitterverse, or Twitter user community, and read public discussions around specific topics designated by a “hashtag” (e.g. #idcc11). Hashtags function as a folksonomic keyword system for organizing topic-based posts. The networked structure of Twitter allows users to view the content in different ways by clicking on the hashtag, author, time or follower.

While individual Tweets may only consist of 140 characters, the data that can be extracted from any particular Tweet is far greater than just the Tweet text. Public information includes: (a) data pertaining to the specific Tweet, including a time stamp and unique Tweet ID; (b) the profile information of the Tweet author, including a unique user ID, the author’s username or handle, the author’s actual name, the location of the author as entered in the profile, the URL link to the author’s profile image, the design settings of the author’s Twitter page, and information about who the author is following, as well as who is following the author; and (c) geo-location data specific to the Tweet, such as the latitude and longitude coordinates of the mobile location, if the author has enabled this feature. The data miner can also use Twitter’s "entities" attribute to parse the text of Tweets and extract structured data about images, urls, @mentions, and/or hashtags.[2]

The fuzzy boundaries of capturing networked data, such as web pages or social media, quickly become clear with the example of Twitter. How many, if any of the posting, following, viewing and tagging behaviors should be preserved when collecting Twitter data? Any imposition of curatorial boundaries is somewhat artificial and necessitates decisions as to what is important. Moreover, bounding networked data at all rips the material out of its natural context, affecting the way future audiences will understand the data. However, the definition of boundaries is crucial to the preservation of this information.

Who is Using Twitter Data?

Once Tweets have been downloaded, they can be used for any number of purposes. There is a growing body of scholarly (and commercial) research that makes use of Twitter and other social media data. The number of articles in *Scopus* in which Twitter was a keyword increased from nine articles in 2007, to 48 in 2009 (De Longueville, Smith & Luraschi, 2009). We recently searched *Web of Science* for articles in which Twitter was a topic and found roughly 150 articles, almost half of which used Twitter as a primary source of data.

Recent studies based on Twitter data demonstrate the multidisciplinary value of almost every facet of the content. Researchers have used Twitter's time and location parameters to track the flows of information around events (Cross, 2010; De Longueville, et al., 2009; Lundquist, 2011; Yardi & boyd, 2010); they have collected samples of Tweets to assess how conversations and demographics shift over time (Golbeck, Grimes & Rogers, 2010); they have conducted hashtag and conversational analysis to get a sense of how conversations and impressions are developing around certain topics (Ifukor, 2010; Marwick & boyd, 2010; Ross, Terras, Warwick & Welsh, 2011; Yardi & boyd, 2010); and researchers have collected the entire stream and network of a single user to conduct deep studies of individual usage of social media (Gruzd, Wellman & Takhteyev, 2011; Starbird, Palen, Hughes, & Vieweg, 2010). In several cases, researchers combined data from multiple social media platforms, illustrating the additional value that will result from making Twitter data searchable and interoperable (Cross, 2010; Starbird, et al., 2010).

Accessing and Mining Twitter Data

Any information the user provides to Twitter, with the exception of their email address, password, and direct messages to other users, is considered public, unless you choose to "protect" your Tweets. The Tweets and personal information contained in a protected account are not accessible and will not appear in search results (Twitter, 2011e). Twitter offers two different Application Programming Interfaces (APIs) by which applications can search and download information. The API allows users to restrict results to specific hashtags, @userids, languages, or geographic areas. The REST (or Search) API is used for queries that return a one-time response of limited

[2] See the Twitter glossary for a complete list of Twitter terms and explanations at: http://support.twitter.com/articles/166337-the-twitter-glossary. See also the Twitter Developers page on Tweet entities for more information on this functionality at: https://dev.twitter.com/docs/tweet-entities

information from Twitter. The Streaming API keeps the connection open and continues sending new information as it becomes available on the Twitter servers (Twitter, 2011d, 2011f).[3] There are also numerous tools to aid individuals who may not have the programming skills to capture and archive twitter data using the APIs.

Using the public APIs has its limitations. Twitter restricts the number of requests that can be made to its servers within a given time period. Additionally, boyd and Crawford (2011) note that Twitter makes only a fraction of its data available through the public APIs. Rather than having access to the entire Twitter archive, boyd and Crawford contend that researchers more often only have access to between 1% and 10% of public Tweets. Researchers can license access to 10% or 50% of Tweets through one of Twitter's partners, but such access may be prohibitively costly.[4] Finally, Twitter's API returns only the 1500 most recent Tweets; notable gaps in data collection may occur when Twitter is being used heavily (Lotan, Graeff, Anany, Gaffney, Pearce, & boyd, 2011).

Data collectors for academic use have primarily been individuals or small teams of researchers, who have collected Twitter data for use in specific projects. These partial and specific methods of collection inherently impose selection parameters on the data and inhibit some forms of reuse. Data curators versed in the capture and reuse of different types of social media content can provide valuable guidance in this area, ensuring that detailed documentation of capture methods (including their limitations) is included with collections in order to make the data sets as accessible to future users as possible.

Curatorial Issues of Social Media Data

> "For the first time in human history, the day-to-day interactions between people are being permanently recorded and formatted in easily organizable segments of information" (Parr, 2008).

Value

Creating, and capturing social media data is such a new practice that assessing its long-term value is particularly difficult. Researchers are only beginning to mine and interrogate the data, and curators are not yet certain if the data sets these scholars are collecting will be useful for future study. Twitter data are ephemeral, making appraisal decisions even more challenging. Currently it is only possible to access Tweets less than a week old using Twitter's Search API. Researchers who think they perceive the start of an important trend are often under pressure to capture the data before Twitter removes the content from publicly available servers (Starbird, Palen, Hughes, & Vieweg, 2010). While curators may wish to exercise great care and judgment about what will be collected, as well as the best methods of collection, the timeframe for decision making is short.

[3] For an example of the type of information captured by the API, see Raffi Krikorian's *Map of a Twitter Status Object* at http://www.scribd.com/doc/30146338/map-of-a-tweet which shows a Tweet rendered in JSON. Also see http://www.slideshare.net/KrisKasianovitz/tweet-marked-up-in-xml for an example of a Tweet marked up in XML.

[4] As of 2010, access to 50% of Tweets from Gnip cost about $30,000, and access to 10% cost about $5,000 per month, while Google reportedly paid $15 million for access to the full stream of Tweets (Gannes, 2010a, 2010b).

Institutions may also doubt the wisdom of preserving the partial social media data sets collected by researchers when the Library of Congress (LC) has agreed to preserve the Twitter archive as a whole (Raymond, 2010). However, the LC has not yet provided a timetable for when the Twitter archive will be released, but they have confirmed that access to the archive will be restricted to "known researchers" who will need to go through an approval process before getting access to the data (O'Keeffe, 2011; Watters, 2011). Thus, there may always be a need for some local collections. Once the LC Twitter archive is up and running, institutions can reassess the value of keeping separate collections. For now, the ephemerality of Twitter data demands action.

Twitter Identity: Can this Data be Trusted?

> "Online identities cannot be understood as linear or static.... In an online social community, the look, feel, perceived appearance, even location of an online identity can be changed, edited, augmented, or deleted at a moment's notice.... Online identities need to be understood as continually changing representations, never fixed in one position, and perpetually in a state of assembly." (Tyma & Leonard, 2011).

Authenticity is a highly subjective concept, made all the more contentious by the anonymity issues presented in social media. While some researchers are more interested in the linguistic, or networked properties of Twitter data than in the identity or location of the author, for other researchers, problems in authenticating such information may render the data unusable (Zimmer, 2010b). In several of the studies we reviewed, researchers accessed user profiles, and followed links to confirm identities, or add context to the data (Starbird, et al., 2010; Yardi & boyd, 2010).

The authenticity of a Tweet may be based on an imagined or anonymous author; while the Twitter @username may be considered the "author" of record, the "real" individual who created the Tweet may remain anonymous, making it difficult for some users to trust the data. For example, if we want to archive Tweets from Hosni Mubarak, former President of Egypt, can we trust that the Hosni Mubarak @EgyptState is the real Mubarak? The avatar image seems to be Mubarak, the background image of the account homepage looks official and the link provided leads us to his Wikipedia page. However another user, @HosniMubarrack is described as the "Former President, Leader, and currently unemployed!" located in Sharm El Sheikh. The misspelling of the name, and the lack of a profile image provide hints that the account may be a parody, but if the textual information is all that is retained, such content might be misinterpreted. Indeed it would require close, Tweet-by-Tweet analysis (and even this may not always be enough) to determine whether the author is who he or she purports to be. Our sense is that researchers may not be evaluating and analyzing their data at a granular enough level to catch such "fake users", particularly when they have amassed collections of hundreds of thousands of Tweets. In the future, when accessing the user's profile is no longer feasible, confirming the authenticity of Tweets may become impossible.

Such issues challenge information professionals to define best practices for social media digital archiving, with an emphasis on the ethics of disseminating personal information, such as photographs, text and geographic location, even if completely imagined or fabricated by the author. Archiving Twitter content also raises important questions about the extent to which researchers, institutions, and the general public can and should trust social media records lacking any verifiable author or creator. At a minimum, archives should note these issues with provenance and authenticity in all Twitter collections.

Significant properties

The multidisciplinary demand for social media data, in addition to disciplinary debates about the "publicness" of Twitter data, can make the task of determining which properties of the data to preserve even more challenging. Knight (2008), and Wilson (2007) have argued that "significant properties" are those characteristics of digital objects that must be preserved over time to ensure its continued accessibility, usability, meaning and its capacity to be accepted as evidence of what it purports to record. Since the Twitter API returns only text fields, and does not necessarily preserve the "relationship" network within Twitter, most of the contextual information about any given Tweet and its author may be lost when it is downloaded. While the API returns a link to the user's profile, it does not preserve a snapshot of the web page, which could be used to verify the user's identity, and may retain important historical information related to the context of the Tweet, or the topic/event. For example, after the protests surrounding the 2009 Iranian election, many Twitter users tinted their profile images green, and switched their location to Tehran to show solidarity with the protestors (Lotan et al., 2011). In such cases, it may be important to archive additional elements, such as screenshots, web archives of Twitter, and other contextual information that cannot be captured through the API.

To further muddy the waters, certain properties deemed by some researchers as essential to verify the provenance of the data (such as photographic content, profile content or location data), may be viewed as privacy violations by others. Researchers and other data collectors should be encouraged to think broadly about potential future uses that they, or others, might conceive of for the data, while also taking potential privacy concerns into account when considering how the data will be captured and accessed. Curators should collaborate with researchers in determining which properties are important to preserve to ensure that the data are meaningful and trustworthy, and expect that those properties will vary widely according to the researchers' field of study and specific research goals.

Defining roles: Collaboration with researchers

At the 2010 *Archiving Social Media Conference*, held at George Mason University, there was some debate as to whether the researchers themselves, or archivists and librarians, should initiate social media data collection (Center for History and New Media, 2010). The debate revolved around the argument that while information professionals would likely be more focused on best practices and collections that facilitated broad-based reuse, such collections would also be less useful to individual scholars than collections they captured themselves according their specific research interests. Others argued that collections had always been, and likely would continue to be, a mixture of both archive and researcher-created content (Theimer, 2010).

Libraries and archives have historically valued professional neutrality in serving their communities (Koehler, 2003). Libraries are charged with providing access to a broad spectrum of opinions to a diverse community of users with varying attitudes, beliefs and practices. It has not been our place to tell researchers how to do their job, but how can we adhere to our professional neutrality when our users have diverging views about the ethics of collecting and accessing the same set of data? Data curators have realized the necessity of being involved early in the research lifecycle, collaborating with researchers from the moment of data collection, or even research design, in order to ensure that research data are appropriately preserved and documented (Abbott, 2008) These collaborative moments (if we take advantage of them) provide insight into the ethical traditions and perspectives of the researchers we support, and allow us an opportunity to discuss the privacy and other ethical issues related to preservation and access in order to reach an informed conclusion about how best to serve both the original researcher(s) and a broad segment of future users.

"But the Data are Already Public": Privacy and Twitter Data[5]

Analyzing Twitter Data: Person, or Text?

Determining whether or not public Twitter data has privacy issues largely comes down to a matter of who is defining the object under analysis. Does analyzing a person's Tweets constitute researching a human subject? The question can be broadly understood as a debate between humanistic and social science understandings of the data. Social scientists, having a long history of working with sensitive data about human subjects, see potential privacy issues with social media data, and seek to protect the "subjects" from harm. Humanists, who have not historically worked with living subjects, tend to see Tweets and other forms of online content as textual in nature, and see Tweet authors as content creators deserving of recognition (Bruckman, 2002; Thelwall, 2010). Bruckman insists that privileging one disciplinary view of the data and its authors over another legitimizes some forms of research, while excluding others.

O'Riordan and Basset (2002), posit that it is the hybrid nature of the medium that lends itself to confusion, suggesting that where researchers see the Internet and social media platforms as a *space* in which human beings interact, they will tend to see the content as being a virtual representation of those beings. For others, they argue, the Internet is simply a new *medium* for cultural production, similar to television, radio, and the newspaper. In the end, O'Riordan and Basset reason, it is not a binary choice: Internet content, including social media data can often be viewed both as text and as a virtual representation of the author. How is the curating institution to deal with these differing interpretations of the data when planning for access and reuse?

One of the key issues these differences raise is a tension over notions of ownership, citation and attribution. The Twitter Terms of Service[6] state that the author retains the rights to all content they produce. Additionally, Twitter's guidelines for the media

[5] The title of this section is a nod to Michael Zimmer's influential article: "But the Data is Already Public": On the Ethics of Research in Facebook.

[6] Twitter Terms of Service: http://twitter.com/tos

explicitly request that users receive attribution for their content by requiring that the text and user ID are not deleted, obscured, or altered when displayed or published (Twitter, 2011b). Yet Twitter also acknowledges the hybrid (public/private) nature of Twitter data by suggesting that the media contact Twitter directly where privacy or security risks are perceived as a result of making the user ID or other identifying information available. Those who see Twitter data as data that contains potentially identifying information about human subjects may want to anonymize the data for the authors' protection, and may see displaying user names as unethical.

What is the balance between proper attribution and protection of user privacy? Many Twitter users (such as major media outlets) see Twitter as a platform for publicity and may want, or demand, to be cited. The copyright status of Tweets has yet to be tested in court, but Reinberg (2009) argues that Tweets would most likely be unprotected, as the 140-character limit probably prevents Tweets from meeting the criteria for originality. Issues of ownership are made more complex when the researcher or data collector is added to the mix. Disciplinary interpretations about the nature of the data alter how researchers believe their own contribution should be acknowledged, and how willing they are to share the data they have collected.

At the time of writing, UCLA's, Dr. Presner sees his collection of Tweets from Egypt and other countries as public, primary source material that should be made broadly available, and while he desired some sort of credit for collecting and adding value to the data, he did not see himself as the "owner of the data" (T. Presner, personal communication, May 25, 2011). However, social scientists have a long history of sharing quantitative data, while closely guarding qualitative data (Parry & Mauthner, 2005). It is not entirely clear where Twitter and other social media data would fall on that spectrum; however, Parry and Mauthner observe that even when qualitative data are made available, it typically has low rates of reuse (primarily by the original researcher) due to the fact that the "recovery of context can only ever be partial." Thom-Santelli and Millen (2010) argue that it is anonymization that remains a stumbling block to sharing social media data, as it raises both ethical issues and issues of data quality for some researchers. As we will continue to point out, the multidisciplinary uses of Twitter and other social media data will likely expose data curators to passionate, yet divergent beliefs about ownership and privacy.

Legal and Ethical Perspectives

> "Technology frequently runs ahead of existing laws and ethical guidelines, and at least some of the solutions to the problems this can cause are likely to lie outside traditional approaches to handling legal and ethical issues." (Charlesworth, 2009)

In addition to differing interpretations of the nature of Twitter data, much of the debate about the capture, reuse and display of Twitter and other social media data can be framed as an argument between the legality of collecting this content, and the ethics of doing so. Twitter's Privacy Policy[7] and Terms of Service seem to provide a clear indication that researchers, and curators, are free to collect, aggregate and use this data. For example, the Twitter Privacy Policy cautions: "Most of the information you provide to us is information you are asking us to make public. This includes not

[7] Twitter Privacy Policy: http://twitter.com/privacy

only the messages you Tweet … but also many other bits of information. Your public information is broadly and instantly disseminated… You should be careful about all information that will be made public by Twitter, not just your Tweets.” Many researchers have argued that as it is possible for users to protect their Twitter accounts, those users who have opted to make their accounts public have no grounds for complaint about the collection and reuse of their content, even if they did not anticipate reuse by researchers or commercial firms (Thelwall, 2010; Vieweg, 2010).

Those who argue about data collection from an ethical rather than a legal rights standpoint argue that just because we are legally able to capture the data, this does not necessarily mean that we should (boyd, 2007). Zimmer (2010a) argues that even within ethics-based approaches to privacy, many researchers take a harm-based view of privacy, in which the goal is to protect users’ information from negative actors. Zimmer argues instead for a dignity-based view of privacy that sees having one’s personal information stripped from the intended sphere of the social networking profile, and amassed into a database for external review as an affront to the users’/subjects’ human dignity and their ability to control the flow of their personal information. While the legalistic arguments for capturing public Twitter data are relatively easy to understand, the ethical arguments against making “public” data public are more nuanced and will be considered in greater detail below.

Defining and Contextualizing Privacy in Social Media Environments

Privacy is an ambiguous term, the meaning of which varies broadly over space and time. As boyd (2007) observes: “What it means to be public or private is quickly changing before our eyes and we lack the language, social norms, and structures to handle it.” The researchers that data curators serve have different interpretations of privacy, and even within cultural heritage institutions there are often conflicting notions about the balance between privacy and openness. What scholars increasingly agree on is the fact that “private” vs. “public” cannot be seen as a dichotomy, but rather must be seen as a continuum influenced by numerous factors (Barth, Datta, Mitchell & Nissenbaum, 2006). ESOMAR (a market research organization) categorizes Twitter and other social media data as “semi-public” (2011). Although the content is technically available for anyone to read, its authors would not necessarily expect that it would be viewed by an audience that was not involved in the particular topic or discussion for which it was created.

International flows of information add to the confusion about privacy norms. Cultural definitions and expectations of privacy are diverse and it cannot be expected that all Twitter users share the same perspective as the researcher or curator. Markham and Buchanan (2011) claim that one of the most common violations of privacy committed by researchers is a lack of attention to local, regional, or cultural laws and perspectives. Mauthner and Parry (2005) go further, arguing that extracting and abstracting data flows from international sources may, in worst cases, constitute a form of neo-colonialism.

Definitions and expectations of privacy also shift over time; the persistence of digital objects may present problems as the authors’ status and vulnerability shift over time, and unknown future audiences access their data (Markham & Buchanan, 2011; Pike, Bateman & Butler, 2009). Boyd (2011) observes that definitions of privacy may also vary on an individual level. Accepting the legal argument that public Twitter data are indeed public and up for grabs at face value obscures how individual users understand their rights and their notions of what constitutes acceptable reuse of their data.

Contextual integrity is a framework that seeks to explain these divergent attitudes by asserting that individuals operate in distinct social contexts, each with their own relative norms and expectations of privacy (Barth, Datta, Mitchell & Nissenbaum, 2006; Nissenbaum, 2004). Different social media platforms may constitute differing social contexts; users may use specific platforms for specific purposes, or to reach different audiences (O'Riordan & Basset, 2002; Zimmer, 2010d). While acknowledging the legal status of Terms of Service agreements, it is important to understand how communities view privacy in the contexts of specific platforms (Buchanan & Johnson, 2011). Author views and imagined audiences are crucial to understanding contextual norms within social media platforms. Boyd and Crawford (2011) argue that: "There is a difference between being in public and being ʻpublic' that is rarely acknowledged by big data researchers." In social media, the imagined audience (friends and like-minded thinkers) can differ markedly from a user's actual audience (researchers, commercial firms, media, and others) (boyd, 2011).

One way of getting at user expectations is examining the types of privacy controls enabled by the platform. Schmidt, Trepte and Reinecke (2011) observe that there are shared routines and expectations about how to self-disclose, and whom to address, noting that privacy management is performed for a specific audience. Facebook, for example, enables users to select privacy settings on a post-by-post basis, choosing who is able to read, comment and interact with specific content, and allowing the user fairly granular control over the flow of their information. Twitter allows only binary control; users can designate their account as "protected" (i.e. Tweets are only visible to approved followers), or "public" (enabled by default), which makes a user's profile and timeline accessible to anyone, even those without a Twitter account. The consequence of having only two levels of privacy control (all private, or all public) is the collapse of temporal, social and spatial boundaries, making it difficult to distinguish public and private interactions (boyd, 2011). Users who want to make any of their content public must make all content public, creating an environment where "public" content contains a full spectrum of communications from personal and private, to mass personal, to traditional mass media; yet Twitter does not have a mechanism that allows these different uses to be segregated based on intent and audience (Lotan et al., 2011; Wu, Hofman, Mason, & Watts, n.d.).

Researchers who have examined Twitter data have noted their own discomfort when coming across "public" data that seemed obviously private (Vieweg, 2010). One researcher who used to harvest social media data claims he wouldn't do so today, noting that: "...people that are using Twitter nowadays may actually want to go back and delete their accounts or take those things out of the public at a later date, and they no longer can" (Parry, 2011). While acknowledging the legality of harvesting public data, it is important for researchers and curators to understand and reflect on these issues before planning for capture and future access to social media data.

Risks: Security Issues, Loopholes, and Re-Identification

Boyd (2007) suggests that social networking sites, like Twitter, constitute new forms of "mediated publics." According to boyd, mediated publics have four unique properties: persistence, searchability, replicability and invisible audiences. Users' content, produced within a certain context, may persist long after the context that gave rise to it has past. Full text search engines and open social media collections make it relatively easy to trace the content back to the user, even if they have deleted the content from the Twitter server. Meanwhile, the mutability of content makes it difficult to authenticate content, but easy to doctor it. Finally, content viewers are often invisible to the user, and persistence and searchability ensure a future audience that may be far removed from the time, space,and other circumstances surrounding the content's creation. These properties introduce risks to users that researchers and curators may want to mitigate, even if the users do not initially protect themselves.

Users may be unaware of the very real security concerns that may be present. Content and profiles posted on social media sites have been used by governments to crack down on dissident groups and to track individuals (Cross, 2010; De Longueville, et al., 2009; Lundquist, 2011). Capturing and archiving this data may complicate and compound these risks.

In addition to security concerns, numerous loopholes may allow private data to be extracted with public Twitter content (Zimmer, 2010c). A user may tweet something that only their followers can see, yet if one of those followers retweets it, the content could then become public.[8] Additionally, while users can choose to protect their accounts at any time, their previous Tweets remain public. Tweets that users have deleted will not be deleted from researchers' collections, or the archives that have already captured them. Finally, it is impossible to identify vulnerable populations, such as children, through data collected in the search API (Markham, 2005). While Twitter states that users must be over 13, there is no way of verifying the age of a user based on the tweet content.

While researchers familiar with human subject research may believe that anonymizing the data can mitigate many of these concerns, numerous studies have shown that anonymizing social network data are nearly impossible. Acquisti and Gross (2009) note that public records can be seen as "breeder" documents of more sensitive data, and Backstrom, Dwork and Kleinberg (2007) argue that: "...anonymous social network data almost never exists in the absence of outside context, and an adversary can potentially combine this knowledge with the observed structure to begin compromising privacy." Zimmer (2010d) goes the furthest, insisting that anonymization is not achievable. In a networked world, "anything can potentially become the missing link to re-identify an entire data set."

A recent study based on data from Facebook (another social media platform) illustrates the ease with which networked data can be re-identified. Researchers anonymized data from a study that collected the Facebook profiles and posts of one

[8] It should be noted that users would have to bypass Twitter's ReTweet function, which does not allow private content to be retweeted, but this would be relatively easy to do.

college class from their freshman to senior years (Parry, 2011). Within days of the data being made available to other researchers, Internet Privacy scholar Michael Zimmer was able to identify Harvard University as the source of the data based on the students' majors, many of which were unique to Harvard (Zimmer, 2010a). Identifying individuals would have been a relatively easy next step. One researcher noted that the Facebook data collected from Harvard was now a bit like "kryptonite" and claims that dealing with data that has been de-anonymized means putting researchers' own ethical stances at risk (Parry, 2011).

Harvard's Institutional Review Board (IRB) had approved the study, illustrating the fact that even organizations charged with protecting data privacy and security face new challenges in dealing with networked data (Parry, 2011). According to a survey conducted by Buchanan (2010), over 70% of IRBs in the United States do not have guidelines for Internet research. The Harvard case also shows that what researchers and IRBs consider an appropriate level of privacy protection is wide-ranging. For some, anonymizing the data may be enough, others may believe that the data should be restricted or closed for the lifetimes of the users (Center for History and New Media, 2010), and some would not even think of consulting an IRB, believing that the data should be open and public.

Data repositories are caught in the middle of these divergent view points when trying to determine the best methods of providing access to the data. The norms of individual research disciplines often provide guidance for curators, but when researchers with divergent norms seek access to the same data, it can be difficult to determine how best to serve the broadest number of users. While we cannot point to clear solutions, it should be clear that researchers, libraries and IRBs need guidance in dealing with networked data that contains identifying information, and a broad conversation among researchers from multiple disciplines, archivists, campus legal counsel and IRBs is necessary to set clear guidelines for Internet research.

The Concerns and Risk-Avoiding Behaviors of Tweet Authors

One way users can mitigate their privacy risks is by deleting Tweets or deleting location information from Tweets, which the Twitter platform allows. However, if these Tweets are archived, deleting them from the Twitter servers does not delete them from the archives that have already downloaded them. Tweet authors will have no idea who has archived their Tweets or where they may be stored. Some users are quite concerned about having their Tweets archived. In response to the announcement that the Library of Congress (LC) would archive all public Tweets, many users expressed privacy concerns on the LC blog (O'Keeffe, 2011; Raymond, 2010). A service that automatically deletes all Tweets before they can potentially be placed in the LC's archive almost immediately sprung up. NoLoc.org[9] claims that users rely on the relative ephemerality of Twitter data to keep potentially embarrassing information from resurfacing. Repositories should be sure to clearly post guidelines and contact information for users who wish to make a takedown request.

[9] NoLOC.org: http://noloc.org/

Boyd (2007) and Zimmer (2010b) argue that many users rely on the fact that their posts will simply get lost in the huge volume of Tweets. Public listings are enabled by default and less than 1% of users opt out (Bonneau, Anderson & Danezis, 2009). Nissenbaum (2011) believes that part of this behavior is due to the burden of reading and keeping up with privacy policies, which are frequently changed, and which are written in legalese that is difficult for the average user to fully comprehend. Others have pointed out that there may be social pressure to use a specific platform; in such cases there is a social cost to *not* sharing that may override users' privacy concerns, undermining the notion that users are necessarily freely choosing to make all of their content public (Carey, Burkell, Kerr, Steeves & Lucock, 2009; Ellison, Vitak, Steinfield, Gray & Lampe, 2011; Nissenbaum, 2011; Raynes-Goldie, 2010).

Scholars have observed that many users exhibit a "functional illiteracy" about Internet privacy (Carey, Burkell, Kerr, Steeves & Lucock, 2009). Users need to think critically about privacy issues, make informed choices about what personal information they choose to share, and understand that almost all information put on the Internet must be considered public due to its persistence and mutability (Debatin, 2011). Mackey and Jacobson (n.d.) call for metaliteracy: a framework that considers the acquisition, production and sharing of knowledge in collaborative, online communities to be a key component of modern information literacy. To what extent should libraries and researchers recognize that users might not be agreeing to what we think they are, and to what extent is their content fair game because it is legally public?

Privacy and Information Professionals

Finally, libraries and archives have their own complex views on privacy. Patron privacy and confidentiality remain core values of libraries (American Library Association [ALA], 2008), while the Society of American Archivists (SAA) *Code of Ethics* (2005) states: "Archivists protect the privacy rights of donors and individuals or groups who are the *subject* of records." At the same time, libraries, particularly in recent times, have increasingly committed to and promoted openness in terms of access to publicly funded research data. The SAA *Code of Ethics* also states: "Archivists strive to promote open and equitable access to their services and the records in their care… in accordance with legal requirements, cultural sensitivities, and institutional policies… Archivists may place restrictions on access for the protection of privacy or confidentiality of information in the records." Deciding how to balance privacy and openness has long been an issue for archival institutions.

Case Study: HyperCities Egypt

The following case study illustrates the issues discussed above, as they apply to a specific data set. We explored potential legal and ethical issues with the content, how it was captured, and how the outcomes impact plans for preservation, access and reuse. We conducted this risk assessment by examining the dataset and its surrounding context in light of Twitter's legal guidelines and policies, and the Association of Internet Researchers *Guidelines on Ethical Research*.[10]

[10] These Guidelines are currently accessible as a DRAFT at http://aoirethics.ijire.net. They have been placed online to encourage comments and discussion from researchers, before they are finalized.

Background

Dr. Todd Presner, Professor of Germanic Studies and Co-Director of the new Digital Cultural Mapping Program, David Shepard, technical team lead, and Yoh Kawano, UCLA GIS Coordinator, harvested a "special collection of distributed data that documents the Arab spring from the first days of protest on January 25th to the ousting of [former President Hosni] Mubarak to the on-going struggle to build a democratic future for the country" (T. Presner, personal communication, April 11, 2011). The collection, called HyperCities Egypt is part of HyperCities Now, a collection of Twitter data and other content that has continued to grow as events progress in Egypt, Libya and elsewhere around the Middle East.[11]

This data set is a specific subset of the overall Twitter data available, delineated by time, place and subject. The HyperCities team used the Twitter Search API to pull data based on the location parameter (within 200 km of the center of Cairo), time period (January 30, 2011 through February 24, 2011), and one of three hashtags (#jan25 or #egypt or #tahrir). Presner, Shepard and Kawano had downloaded approximately 420,000 public Tweets during the initial phase of this analysis. Based on the search parameters, the data set captures eight out of approximately forty possible Twitter data fields, revealing how the method of capture and search parameters profoundly shape the resultant data. Additionally, the team began collecting data several days into the protests, and was unable to access and capture the earliest Tweets. These facts must be documented in order for future users to have a clear understanding of the data set.

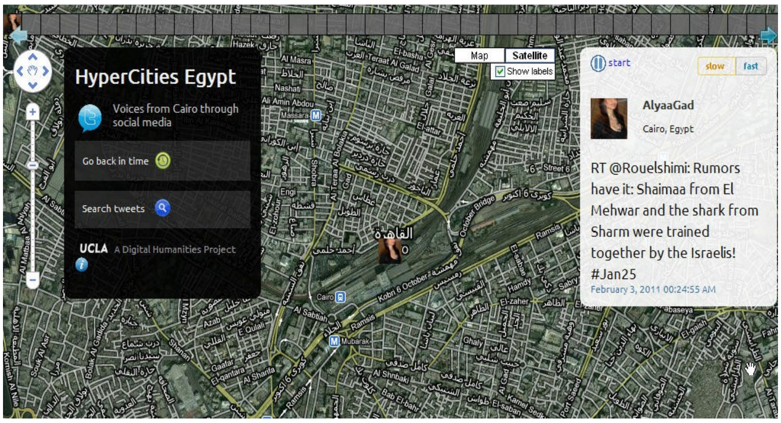
Compliance with Twitter Guidelines

The HyperCities team downloaded the data in accordance with the allowances and rate limits of the Search API, though their access was blocked for one day for making too many calls on the API. The access restriction shows that Twitter is indeed monitoring these downloads and will take measures to ensure compliance with download rates. The access restriction also resulted in the potential loss of data for that time period. Additionally, at one point the team had their personal server crash due to the volume of content that was being downloaded, resulting in further potential data loss. While these losses were acceptable to the team, they also reveal the types of gaps that may occur in data collection.

The HyperCities Egypt interface displays the content according to Twitter's display guidelines (Twitter, 2011a). The Tweets are displayed with the Twitter icon, the user's username, and the unaltered Tweet text (illustrating the team's connection to the view that Twitter content is a publication whose author's deserve attribution, rather than seeing the users as human subjects).[12] As shown in Figure 1, all Tweets with the location "Cairo" are displayed in the same location on the map. Practically speaking, the HyperCities team has not violated any Twitter policies, and therefore the team should not be exposed to any legal challenges.

[11] HyperCities is a collaborative research and educational platform for travelling back in time to explore the historical layers of city spaces in an interactive, hypermedia environment. See http://hypercities.com/ for more information.

[12] View the collection at: http://egypt.hypercities.com/



AlyaaGad
Cairo, Egypt

RT @Rouelshimi: Rumors have it: Shaimaa from El Mehwar and the shark from Sharm were trained together by the Israelis! #Jan25
February 3, 2011 00:24:55 AM

Figure 1. Tweet from the HyperCities Egypt collection displayed in the HyperCities platform.

Comparison with the AoIR Guidelines on Internet Research Ethics

The Draft AoIR Guidelines on Internet Research Ethics (Markham & Buchanan, 2011) suggest that researchers ask themselves several questions when undertaking Internet Research in order to evaluate risks to the subjects/authors. We attempted to answer them, using our knowledge of the HyperCities Egypt data set. We did not go through these questions with the research team, but sought to determine whether the questions were helpful in assessing the curatorial risk of ingesting the data. These guidelines take a social science approach to Internet Research ethics, viewing content creators as human subjects, rather than as content authors. Guidelines and suggestions from other disciplines should also be reviewed and discussed.

The Guidelines encourage researchers to closely examine the cultural and social context of the research environment, including legal privacy expectations, cultural privacy expectations and the researcher's understanding of user privacy. While legally the captured Tweets are "public" data, as we discussed in detail above, it is less clear what the authors of the data set expect. Security risks are very real for these users. A recent survey found that Middle Eastern activists faced high levels of attacks stemming from their online activities, and often use false identities and other means of obscuring their identity (Faris, Roberts, Heacock, Zuckerman and Gasser, 2011). The current political situation in Egypt is in flux, and users' attitudes and concerns about privacy may alter as the governance of the country evolves.

Presner views the data as a new type of special collection, or primary source material, that documents a moment of historic importance. While the HyperCities team did not find it necessary (or practical) to obtain informed consent from the authors, they attempted to address privacy and security concerns by removing geo-coordinates, and aggregating location data to the city level (showing that they did fear potential privacy and security risks in this "public" data), while continuing to display usernames, and profile images (showing their desire to give credit to the content authors). The research team also discussed potential privacy and security risks with various university representatives in an attempt to mitigate any potential risks involved in displaying the data.

When viewed against the AoIR guidelines, publicly displaying these data in their current form may indeed result in security and privacy risks for the Tweet authors. While the danger is probably not enough to merit completely restricting the data, the repository should appraise this data set accordingly with an eye to risk assessment and management. It will be crucial for libraries and other entities that are archiving Twitter, and other user-generated social media data, to engage with the issues surrounding the data and the ways in which research ethics are evolving.

Recommendations

Andrew Charlesworth provides several recommendations geared toward the preservation and curation of personal digital archives (PDArcs) in the Digital Lives Legal and Ethical Issues report. Although intended for PDArcs that potentially contain social media, the assessment provides a framework that is generalizable to social media data sets. Charlesworth suggests that curators create policies and infrastructure that are "pragmatic, flexible, and standards-based" (2009). The following recommendations are based on Charlesworth's recommendations for the Digital Lives Project. They strive to create a balance between respecting the privacy of content creators and fostering openness in the access and reuse of social media data.

1. **Libraries or other data repositories will need to decide if archiving social media data fits with their overall institutional mission and goals.**

   An internal dialogue with relevant library departments and key campus personnel and offices should be initiated, followed by an external dialogue with the community of scholars and institutions also grappling with social media data. Regular dialogue and communication with these players should continue.

2. **Libraries should determine the overall risks associated with collecting and archiving social media data and design strategies to mitigate those risks.**

   Libraries must understand whether there are any security or privacy issues that pertain to the social media data sets and, based on those risks, determine how they will provide access to the data set. Risk management strategies may include providing tiered access to the data (e.g. the repository could allow public access to Tweet texts, while restricting access to profile and location information), and should also include a strategy for users to request that their data to be removed from or suppressed within the dataset. Repositories should undertake a risk assessment that includes dialogue with appropriate institutional players, such as the IRB, general counsel, academic senate, social science data archive, digital humanities centers and campus IT.

3. **Libraries choosing to archive social media data should develop clear and easy to use collection and deposit policies, forms and tools.**

   Collection policies will need to clearly articulate which content and formats will be accepted, and should include criteria for determining which content will receive long-term preservation. Deposit forms should be created to ensure that the researcher's data capture methods are well documented. A list of captured data fields, the context surrounding the data, and its potential issues, problems and limitations must also be recorded to ensure future users can fully understand and trust the data set.[13] Deposit forms should serve not only as the means to collect the metadata about the data set, but also as a mechanism to instruct the depositors and raise awareness about privacy, provenance, authenticity and re-use issues. Access tools for social media data sets should include a detailed "codebook" generated from the deposit form, re-use parameters, and standardized statements about issues related to provenance, authenticity and privacy of such datasets.

4. **Libraries should engage researchers as early as possible in the research process.**

   While this may not be feasible in all cases, curatorial intervention at the outset can add value to the collections, which will have lasting benefits for archiving, preservation and interoperability with other datasets. Raising the researchers' awareness of emerging practices and guidelines, like the AoIR *Guidelines for Internet Research Ethics*, may also serve to encourage researchers to fully think through and engage with the privacy issues related to collecting social media data. Libraries and curators should seek to foster communication between and collaboration with other social media archiving institutions and researchers, facilitating meetings and workshops that bring these parties together.

Conclusions

Do Your Tweets Tell Us Too Much About You?

Notions of privacy in social media data are complex. There is just enough personally identifying information available within Twitter to cause concern. The expectations of content authors will always be in flux; privacy norms are temporal, cultural, contextual and personal. There is no silver bullet in determining how to balance openness and the research needs of the academic community with the ethical treatment of social media data and the potential concerns of content creators.

Research ethics are tied to disciplinary culture. The multidisciplinary interest in social media data will require curators to navigate researchers' conflicting views regarding the treatment of social media data, particularly as it pertains to access and permission for reuse. Because of this, curators are presented with a golden opportunity to collaborate with researchers near the beginning of the research lifecycle. By partnering with scholars, providing and receiving guidance, we can facilitate the creation of collections that balance openness with privacy concerns, and encourage broad reuse. We encourage library, archive, or repository staff to initiate a cross-disciplinary conversation about these issues in order to attempt to create a collection policy that is amenable to all.

So, will your tweets be around in 100 years? Time, technology, and evolving academic and curatorial practices will tell.

[13] See http://www.slideshare.net/KrisKasianovitz/sample-twitter-data-deposit-form for an example of such a deposit form.

Acknowledgements

Peter Broadwell, Mimi Calter, Stephen Davison, Elijah Meeks, Todd Presner, David Shepard, and Elizabeth Stephenson were generous with their time and patience in helping us understand the issues related to social media data. Ronald Blanford, James Jacobs, Lisa Snyder, Lynn Yarmey, and Shinjoung Yeo offered valuable suggestions and feedback on this paper. We'd also like to thank Dr. Christine Borgman for teaching the *Data, Data Practices, and Data Curation* course which allowed us to grapple with such interesting issues. Special thanks to Todd Presner, David Shepard and Yoh Kawano for collecting Twitter data, and allowing us to road test the many issues of social media data curation with their collection.

References

Abbott, D. (2008). *What is digital curation?* Briefing paper by the Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation

Acquisti, A. & Gross, R. (2009). Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences of the United States of America, 106*(27). Retrieved from http://www.pnas.org/content/106/27/10975.full

American Library Association. (2008). *Code of ethics of the American Library Association.* Retrieved from http://www.ala.org/ala/issuesadvocacy/proethics/codeofethics/codeethics.cfm

Backstrom, L., Dwork, C. & Kleinberg, J. (2007). Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. Paper presented at the 16th International Conference on World Wide Web, Banff, AB, Canada. Retrieved from http://dl.acm.org/citation.cfm?id=1242598

Barth, A., Datta, A., Mitchell, J. C. & Nissenbaum, H. (2006). Privacy and contextual integrity: Framework and its applications. Paper presented at the IEEE Symposium on Security and Privacy, Oakland, CA.

Bonneau, J., Anderson, J. & Danezis, G. (2009). Prying data out of a social network. Paper presented at the International Conference on Advances in Social Network Analysis and Mining, Kaohsiung City, Taiwan. doi:10.1109/ASONAM.2009.45

boyd, d. (2007). Social network sites: Public, private, or what? *The Knowledge Tree, 13*. Retrieved from http://kt.flexiblelearning.net.au/tkt2007/edition-13/social-network-sites-public-private-or-what/

boyd, d. (2011). Dear voyeur, meet flaneur... Sincerely, social media. *Surveillance & Society, 8*(4). Retrieved from http://library.queensu.ca/ojs/index.php/surveillance-and-society/article/view/4187 boyd, d., & Crawford, K. (2011). Six provocations for big data. Paper presented at A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, Oxford Internet Institute. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Internet Research Ethics.* Retrieved from http://www.nyu.edu/projects/nissenbaum/ethics_bru_full.html

Buchanan, E. (2010). Internet research ethics and IRBs. Paper presented at the OHRP Research Forum, Chicago, IL. Retrieved from http://www.slideshare.net/InResEth/internet-research-ethics-and-irbs-4159809

Buchanan, E. & Johnson, M. (2011). Internet research ethics. Paper presented at the 2011 Institutional Review Board Annual Educational Conference, Columbia University. Retrieved from http://www.cumc.columbia.edu/dept/irb/education/2011Agenda.html

Carey, R., Burkell, J., Kerr, I., Steeves, V. & Lucock, C. (2009). A heuristics approach to understanding privacy-protecting behaviors in digital social environments. *Lessons from the Identity Trail: Anonymity, privacy and identity in a networked society.* Oxford: Oxford University Press.

Center for History and New Media. (2010). *Archiving social media: Workshop notes*. Retrieved from http://archivingsocialmedia.org/themes/index.html

Charlesworth, A. (2009). *Digital Lives >> Legal & Ethical Issues*. Report for the British Library. Retrieved from http://britishlibrary.typepad.co.uk/files/digital-lives-legal-ethical.pdf

Cross, K. (2010). Why Iran's green movement faltered: The limits of information technology in a rentier state. *The SAIS Review of International Affairs, 30*(2). Retrieved from http://muse.jhu.edu/journals/sais_review/v030/30.2.cross.html

De Longueville, B., Smith, R. S. & Luraschi, G. (2009). *OMG, from here, I can see the flames!* Proceedings of the 2009 International Workshop on Location Based Social Networks. ACM Press. doi:10.1145/1629890.1629907

Debatin, B. (2011). Ethics, privacy, and self-restraint in social networking. In S. Trepte & L. Reinecke (Eds.) *Privacy Online: Perspectives on Privacy and Self-disclosure in the Social Web*. Heidelberg: Springer.

Ellison, N. B., Vitak, J., Steinfield, C., Gray, R. & Lampe, C. (2011). Negotiating privacy concerns and social capital needs in a social media environment. In S. Trepte & L. Reinecke (Eds.) *Privacy Online: Perspectives on Privacy and Self-disclosure in the Social Web*. Heidelberg: Springer.

ESOMAR. (2011). *Guidelines on social media research. ESOMAR World Research.* Retrieved from http://www.esomar.org/index.php/professional-standards-codes-and-guidelines-guideline-on-social-media-research.html

Ess, C. & AoIR Ethics Working Group. (2002). *Ethical decision-making and Internet research: Recommendations from the AoIR Ethics Working Committee.* Association of Internet Researchers. Retrieved from https://aoir.org/documents/ethics-guide/

Faris, R., Roberts, H., Heacock, R., Zuckerman, E. & Gasser, U. (2011). *Online security in the Middle East and North Africa: A survey of perceptions, knowledge and practice.* Research Publication No. 2011-04. Harvard University: The Berkman Center for Internet & Society. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1942857

Gannes, L. (2010a). Twitter firehose too intense? Take a sip from the gardenhose or sample the spritzer. *All Things D*. Retrieved from http://allthingsd.com/20101110/twitter-firehose-too-intense-take-a-sip-from-the-garden-hose-or-sample-the-spritzer/

Gannes, L. (2010b). Gnip becomes Twitter's first authorized data reseller. *All Things D.* Retrieved from http://allthingsd.com/20101117/gnip-becomes-twitters-first-authorized-data-reseller/

Golbeck, J., Grimes, J.M. & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology, 61*(8). doi:10.1002/asi.21344

Gruzd, A., Wellman, B. & Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist, 55*(10). doi:10.1177/0002764211409378

Ifukor, P. (2010). "Elections" or "Selections"? Blogging and Twittering the Nigerian 2007 general elections. *Bulletin of Science, Technology & Society, 30*(6). doi:10.1177/0270467610380008

Knight, G. (2008). *Deciding factors: Issues that influence decision-making on significant properties.* Report for CeRch: Center for e-Research. Retrieved from http://www.significantproperties.org.uk/deciding-factors.html

Koehler, W. (2003). Professional values and ethics as defined by "The LIS Discipline." *Journal of Education for Library and Information Science, 44*(2). Retrieved from http://www.jstor.org/stable/40323926

Lotan, G., Graeff, E., Anany, M., Gaffney, D., Pearce, I. & boyd, d. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *The International Journal of Communication, 5*. Retrieved from http://ijoc.org/ojs/index.php/ijoc/article/view/1246

Lundquist, E. (2011). Crisis tool: Social media can provide situational awareness during disasters... in 140 characters or less. *Sea Power 54*(2). Navy League of the United States, Arlington, VA.

Mackey, T.P., & Jacobsen, T.E. (n.d.). Reframing information literacy as a metaliteracy. *College & Research Libraries, 72*(1).

Markham, A. (2005). The methods, politics, and ethics of representation in online ethnography. In N.K. Denzin & Y.S. Lincoln (Eds.) *The SAGE Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications Inc.

Markham, A. & Buchanan, E. (2011). *Ethical decision-making and Internet research (version 2.0): Recommendations from the AoIR Ethics Working Committee.* AOIR Ethics Guidelines Review Draft 2011. Retrieved from http://aoirethics.ijire.net/

Marwick, A.E., & boyd, d. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1). doi:10.1177/1461444810365313

Mauthner, N. & Parry, O. (2010). Ethical issues in data archiving and sharing. *eResearch Ethics.* Retrieved from http://eresearch-ethics.org/position/ethical-issues-in-digital-data-archiving-and-sharing/

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=534622

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4). Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/DAED_a_00113

O'Keeffe, H. (2011). *Legal issues in building social media collections.* Presentation to the Association of Research Libraries. Retrieved from http://www.arl.org/bm~doc/mm11sp-okeeffe.pdf

O'Riordan, K. & Basset, E.H. (2002). Ethics of Internet research: Contesting the human subjects research model. *Internet Research Ethics*. Retrieved from http://www.nyu.edu/projects/nissenbaum/ethics_bas_full.html

Parr, B. (2008). Five ways social media will change recorded history. *Mashable.* Retrieved from http://mashable.com/2008/11/18/consequences-of-social-media/

Parry, M. (2011). Harvard's privacy meltdown: Technology. *Chronicle of Higher Education.* Retrieved from http://chronicle.com/article/Harvards-Privacy-Meltdown/128166/

Parry, O. & Mauthner, N. (2005). Back to basics: Who re-uses qualitative data and why? *Sociology, 39*. doi:10.1177/0038038505050543

Pike, J.C., Bateman, P.J. & Butler, B.S. (2009). I didn't know you could see that: The effect of social networking environment characteristics on publicness and self-disclosure. Paper presented at the 15th Americas Conference on Information Systems, San Francisco, California.

Raymond, M. (2010). *The library and Twitter: An FAQ*. Library of Congress blog. Retrieved from http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/

Raynes-Goldie, K. (2010). Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook. *First Monday, 15*(1-4). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2775/24

Reinberg, C. (2009). Are tweets copyright-protected? *WIPO Magazine*. Retrieved from http://www.wipo.int/wipo_magazine/en/2009/04/article_0005.html

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation, 67*(2). doi:10.1108/00220411111109449

Schmidt, J.-H., Trepte, S. & Reinecke, L. (2011). (Micro)Blogs: practices of privacy management. *Privacy Online: Perspectives on Privacy and Self-disclosure in the Social Web*. Heidelberg: Springer.

Society of American Archivists. (2005). *Code of ethics for archivists.* Retrieved from http://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics

Starbird, K., Palen, L., Hughes, A. L. & Vieweg, S. (2010). Chatter on the red. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work. ACM Press. doi:10.1145/1718918.1718965

Theimer, K. (2010). Reflections on "archiving social media." *ArchivesNext*. Retrieved from http://www.archivesnext.com/?p=1655

Thelwall, M. (2010). Researching the public web. *eResearch Ethics.* Retrieved from http://eresearch-ethics.org/position/researching-the-public-web/

Thom-Santelli, J. & Millen, D. (2010). Characterizing social data sets: Why so hard to share? Paper presented at Revisting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA. Retrieved from http://www.cc.gatech.edu/~yardi/ethics-cscw2010_files/AcceptedPapers.htm

Twitter. (2011a). *Display guidelines*. Twitter Developers. Retrieved from https://dev.twitter.com/terms/display-guidelines

Twitter. (2011b). *Guidelines for use of Tweets in broadcast or other offline media.* Twitter Help Center. Retrieved from http://support.twitter.com/entries/114233

Twitter. (2011c). *Partner providers of Twitter data.* Twitter Developers. Retrieved from https://dev.twitter.com/docs/twitter-data-providers

Twitter. (2011d). *REST API resources.* Twitter Developers. Retrieved from https://dev.twitter.com/docs/api

Twitter. (2011e). *Streaming API.* Twitter Developers. Retrieved from https://dev.twitter.com/docs/streaming-api

Twitter. (2011f). *Streaming API concepts.* Twitter Developers. Retrieved from https://dev.twitter.com/docs/streaming-api/concepts

Tyma, A.W. & Leonard, L.G. (2011). It's not all zeroes and ones: Constructing online identity assembly theory. Paper presented at Internet Research 2.0: Performance and Participation, Seattle, WA.

Vieweg, S. (2010). The ethics of Twitter research. Paper presented at Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA. Retrieved from http://www.cc.gatech.edu/~yardi/ethics-cscw2010_files/AcceptedPapers.htm

Watters, A. (2011). *How the Library of Congress is building the Twitter archive.* O'Reilly Radar. Retrieved from http://radar.oreilly.com/print/2011/06/library-of-congress-twitter-archive.html

Wilson, A. (2007). *Significant properties report.* Arts and Humanities Data Service. Retrieved from http://www.significantproperties.org.uk/wp22_significant_properties.pdf

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D., J. (n.d.). *Who says what to whom on Twitter.* Yahoo Research. Retrieved from http://research.yahoo.com/pub/3386

Yardi, S., & boyd, d. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society, 30*(5). doi:10.1177/0270467610380011

Zimmer, M. (2010a). "But the data are already public": On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4). doi:10.1007/s10676-010-9227-5

Zimmer, M. (2010b). *Is it ethical to harvest public Twitter accounts without consent?* Michael Zimmer.org. Retrieved from http://michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/

Zimmer, M. (2010c). *How your private tweets might be included in the Library of Congress public archive.* Michael Zimmer.org. Retrieved from http://michaelzimmer.org/2010/04/14/how-your-private-tweets-might-be-included-in-the-library-of-congress-public-archive/

Zimmer, M. (2010d). SACHRP Presentation: Research ethics in the 2.0 era: Conceptual gaps for ethicists, researchers, IRBs. Presented at the Secretary's Advisory Committee on Human Research Protections. Retrieved from http://michaelzimmer.org/2010/07/20/presentation-research-ethics-in-the-2-0-era/