Ethical Considerations in Web 2.0 Archives

Antoinette E. Baker
San Jose State University, tonibaker@mac.com

Follow this and additional works at: https://scholarworks.sjsu.edu/slissrj
Part of the Library and Information Science Commons

Recommended Citation

This article is brought to you by the open access Journals at SJSU ScholarWorks. It has been accepted for inclusion in School of Information Student Research Journal by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Ethical Considerations in Web 2.0 Archives

In April of 2010, digital media company Twitter announced that it had donated its entire archive of “Tweets,” 140-character messages distributed via the Internet or telephone text message, to the Library of Congress (Stone, 2010). The gift consists of all public Tweets from 2006 on, numbering more than 50 million Tweets per day (Library of Congress, 2010). When we think of archives it is often about public figures – artists or politicians – who voluntarily donate their papers to an institution long after any expectation of personal privacy has passed. Or we may imagine materials created by ordinary people, such as letters and diaries that have been discovered or collected, rather than voluntarily donated. In these cases, the privacy of the affected persons is not a concern because the affected persons are deceased. Nonetheless, scholars have typically addressed ethics and privacy for these types of archives.

However, the decision by the Library of Congress to create and maintain a Twitter archive was not made by the users who generated the content; rather, the archive was donated by a corporate sponsor. Most users understand that their Web 2.0 creations are disseminated publicly. The Twitter Privacy Policy says as much (Twitter, 2011a), but as the volume of Web 2.0 materials grows, so does the interest in preserving it. One can imagine archives of Facebook posts, YouTube videos, or Flickr photographs that were not sanctioned for donation by the users who created the content or the users who are described or featured in the content. As Web 2.0 grows, the need for definite ethical guidelines for information professionals becomes acute.

This paper explores three issues related to the ethics of archiving Web 2.0 materials, and interprets professional ethics codes and scholarship to suggest possible answers to these questions. First, I consider whether Web 2.0 materials such as Tweets merit archive preservation and conclude that there is an ethical imperative to preserve these materials. Second, I explore whose privacy concerns must be considered and protected when users who generate Web 2.0 materials do not consent to, or have knowledge about, the archive; I conclude that there is an ethical obligation to protect the privacy of these so-called blind donors. Finally, I explore historical approaches for protecting the privacy of blind donors, and suggest ways to adapt privacy protections in the Web 2.0 era.

Discussion

Ethical Codes

The Society of American Archivists (SAA) approved its most recent ethical code in 2005, at the start of the rise of Web 2.0 (SAA, 2005b). There are several provisions in the code that are applicable to digital archives. Under Professional
Relationships the code states, “Archivists select, preserve, and make available historical and documentary records of enduring value” (SAA, 2005a, para. 5).

Two separate sections address privacy. Under Access, the code states, “Archivists strive to promote open and equitable access. … Archivists may place restrictions on access for the protection of privacy or confidentiality of information in the records,” (SAA, 2005a, para. 9); and under Privacy the code states:

Archivists protect the privacy rights of donors and individuals or groups who are the subject of records [emphasis added]. They respect all users’ right to privacy by maintaining the confidentiality of their research and protecting any personal information collected about them in accordance with the institution’s security procedures. (SAA, 2005a, para. 9)

To summarize, the SAA (2005a) articulates three broad concerns relevant to archives of Web 2.0 materials: selecting historically valuable material; permitting open access to the material subject to reasonable privacy restrictions; and protecting privacy of donors and subjects of the records. The code does not spell out how to weigh these concerns when they are in conflict, and in fact the code states in its preamble, “It does not provide the solution to specific problems” (SAA, 2005a, para. 2).

The International Council on Archives (ICA) adopted its most recent code of ethics in 1996. The ICA (1996) addresses the same general concerns about selection, access, and privacy as the SAA code and includes additional detail; one helpful recommendation is,

Archivists negotiating with transferring officials or owners of records should seek fair decisions based on full consideration – when applicable – the following factors: authority to transfer, donate, or sell; financial arrangements and benefits; plans for processing; copyright and conditions of access. (para. 5)

This provision reminds archivists that they should not passively accept gifts, and should instead evaluate gifts in accordance with the recommended factors. This provision empowers archivists to negotiate with donors the conditions under which gifts are made. The ICA (1996) privacy provision also includes a special admonition on third-party privacy rights, stating, “they must respect the privacy of individuals who created or are the subjects of records, especially those who had no voice in the use or disposition of the materials” [emphasis added] (para. 7).

The emphasized clause seems particularly relevant in the Web 2.0 era, wherein

---

1 SAA Code of Ethics for Archivists is currently being revised with proposed text submitted for member comment in spring 2011 (SAA, 2011). The proposed new Code contains provisions about archive purposes and restrictions to protect privacy with guidance to archivists; however, it does not yet address particular challenges presented by Web 2.0 archives.
corporations will be establishing archives of material created by third-party users who do not explicitly consent to or are unaware of the archive gift. Curiously, the SAA code contained a similar clause about those with no voice in the use or disposition of the archive, but the provision was deleted when the code was revised in 2005 (SAA, 2005b). Despite the amendment, the SAA code values third-party privacy.

There can be advantages to the ICA approach of including a specific reference to third-party privacy in the ethical code. Dingwall (2004) discussed the role of ethics codes for archivists, particularly as new technologies impact the profession. He argued that ethics codes, when adhered to, can preserve the public trust in archives, and accordingly, drafters of ethics codes must acknowledge that the public is a primary audience for ethics codes. Put simply, if ethics codes emphasize third-party privacy, those same third parties may be more trusting of digital archives generally. As archivists try to navigate new ethical questions that arise as a result of Web 2.0 the profession should endeavor to revise and update its ethics codes, both to provide guidance to professionals and to persuade the public that archives can protect public interests.

**Ethical Imperative to Preserve Digital Archives**

According to the SAA Code of Ethics for Archivists (2005a) archivists must preserve “historical and documentary records of enduring value” (para. 5). Several authors have made persuasive cases for maintaining archives despite potential violations of personal privacy. Danielsen (2005) described the archives of the East German State Security Service (Stasi) after the fall of the Berlin Wall. These archives raised enormous concerns about personal privacy, as East German citizens could have been exposed as either subjects of police investigation or police informants. Members of the Communist Party shredded a part of the archive, and an ad hoc government agency advocated for its destruction (p. 94). Nonetheless, Danielsen (2005) argued that taking extensive steps to preserve the archive with appropriate protections for individual privacy, including restoring the shredded materials and providing individuals with copies of their Stasi files, made the Republic stronger (pp. 110-111).

Allen (2005) argued that privacy should not come at the expense of other societal values; privacy can be over-valued, and “people not subject to [censure] may lack incentives for avoiding antisocial conduct” (p. 402). Allen (2005), referencing the writings of John Stuart Mill, argued that in certain circumstances, exposing criminal actions justified invasion of privacy. Privacy then, in and of itself, is not a sufficient reason not to maintain a Web 2.0 archive. The Danielsen and Allen examples justify invading privacy in the name of politics, crime, and justice – issues that our society values more than the subject of the average Tweet.
Balkin (2005) addressed this distinction in his essay on freedom of speech in the digital age. Specifically he argued the “republican” model of freedom of speech, by which citizens require freedom of speech to debate political topics, has been usurped by a digital model, by which citizens use speech to comment on arts, culture, media, and corporate control (pp. 325-332). Applying Balkin’s (2005) analysis, a movie review is speech that is as essential to our democracy as a political endorsement.

Scholars have stressed the growing importance of archives of ordinary people in historical and sociological studies, including diaries, scrapbooks, letters, and popular culture (Berger, 2009). Schwarz (2005) argued that these non-traditional materials have been essential in preserving the history of marginalized communities, such as gays and lesbians. Hookway (2008) encouraged the use of blogs in social science research as preferable to diary research or interviews. As a source for research blogs are easy to access and provide insight into different cultures and communities without requiring the researcher to travel. Furthermore, blogs are not “contaminated” by the interests of the researcher; they are not constructed in response to an interview question (Hookway, 2008).

The Library of Congress makes an impressive case for the importance of the Twitter archive: “individually, [T]weets might seem insignificant, but viewed in the aggregate, they can be a resource for future generations to understand life in the 21st century” (Raymond, 2010, para. 6). Raymond (2010) referenced the past activities of the Library of Congress in archiving oral histories of veterans and citizens and argued that the Twitter archive will be a continuation of that, serving as a record of “citizen journalism” and “glimpses of the lives of ordinary people” (paras. 6-7). Digital archives of Web 2.0 technologies have intrinsic value as a record of our society’s exercise of its freedoms and as a snapshot of everyday life. Accordingly, these meet the ethical criteria for records that merit preservation.

Questions of Access

MacNeil (1992), one of the foremost scholars in the area of individual privacy and archives, focused her early research on private data in government archives. She explored the archives of mundane information we provide to the government as a matter of course, either to fulfill our legal obligations or to obtain government benefits. MacNeil (1992) described privacy protections, including limiting access to personal information entirely, anonymizing data, and screening access to the archives based on the qualifications of the researcher or the nature of her project. Perhaps with thanks in part to MacNeil’s work, statutes now protect much of the private information she was concerned about.
Anonymizing Data

Archives should be established with a purpose in mind, and should be administered in a way that is useful to the community the archive will serve. For example, MacNeil (1992) was concerned about archives of census data or records of welfare agencies. For those materials, anonymizing data and limiting access to personal information by masking out names, addresses, and other identifying information are sensible solutions. Visitors to those archives presumably will be more interested in data trends than in a particular individual’s census form. Anonymizing data might thwart visitors who wish to learn personal details about their neighbors, but archives do not have to serve every interest of the public at large.

A Web 2.0 archive may have to include more identifying information in order to make it useful to its intended audience. It may be interesting to know the number of Tweets corresponding to a past news event, but it will be of greater historical interest to read the Tweets of the individuals personally involved in the news event. On the other hand, a Web 2.0 archive can restrict personal account information, such as users’ addresses and credit card numbers. MacNeil’s (1992) suggestions of anonymizing data and limiting access to personal information will have to be considered in connection with the mission of the archive.

Scholars-only Access Restrictions

MacNeil’s (1992) suggestion of limiting access to an archive based on qualifications of the researcher and the nature of the project merits further discussion. When discussing electronic privacy, the public is typically concerned about employers, investigators, family members, or past and future romantic partners accessing digital information. No one seems too concerned about academic scholars, which makes this type of restriction attractive.

Geselbracht (1986) described the history of archival access restrictions at the Library of Congress Manuscript Division. The first archival manual, authored by the Division’s assistant chief and published in the United States in 1913, warned against access to archives by a ‘scandal-seeking’ press. Later publications at the Division encouraged archivists to read the notes of scholars who had accessed archives, and permitted archivists to withhold quotations of libelous or personal nature. These restrictions were not lifted in entirety until 1963, in favor of an open access policy (Geselbrecht, 1986). Scholars-only restrictions can also impose a research monopoly on particular topics. For example, a past curator of the Rodin Museum in Paris restricted access to materials while she wrote what she hoped would be a definitive work on Rodin (Case & Xu, 1994). Similarly, when the Huntington Library made a complete photocopy of the Dead Sea Scrolls available to the public in 1991, the scholars who had been granted exclusive access to the
materials for 40 years opposed the move. Nevertheless, democratic access can only benefit the public and the academic community (Case & Xu, 1994).

The United States Supreme Court has twice attempted to restrict its archives to scholars (Cox et al., 2009). After portions of oral arguments in a controversial case were broadcast on the radio, former Chief Justice Warren Burger learned that all recordings of Supreme Court oral arguments had been publicly available at the National Archives for decades; subsequently, he asked the National Archives to require anyone accessing the arguments to sign a statement agreeing that they would only use the records for private research and teaching. Eventually, the Court lifted restrictions on the tapes (Cox et al., 2009). Controversy arose again when former Supreme Court Justice Thurgood Marshall died and his papers were donated to the Library of Congress. The terms of the gift restricted access to “scholars or researchers engaged in serious research” (Cox et al., 2009, p. 44) but reporters immediately accessed the papers. Librarian of Congress James Billington defended its Open Access policy as consistent with Marshall’s wishes, acknowledging that reporters can be engaged in serious research (Cox et al., 2009).

A scholars-only restriction is antithetical to the mission of many archives, as well as the SAA (2005) and ICA (1996) codes of ethics, which advocate open and equitable access to archives. These restrictions place a burden on administrators to evaluate scholars’ credentials and the scope and breadth of their research and future projects. Ultimately, it is preferable to initially craft the mission of both the archive and the collection, and then permit all interested persons to view it, rather than restricting access to appropriate persons and projects.

Electronic Access

Electronic archives of digital material raise further concerns about security of information. Speck (2010) argues that all donors to archives in this era have an expectation that their materials may be digitized and placed on the Internet, and that archivists must address this issue in their agreements with the donor in order to reassure the donor that the archive is “forthright and sophisticated in its use of technology” (p. 49). Hodson (2006) is more circumspect, acknowledging, “it is one thing to make available possibly sensitive letters for research in a library reading room, but quite another to post them on the Internet, where they can be read potentially by millions of people” (para. 25). If we allow scholars to conveniently access a digital archive with a username and encrypted password, they might still provide access to others. This risk is not entirely a product of the digital era; certainly scholars have misused archival information in the past. But concerns about privacy are best served by imagining all the possible security
breaches that could occur. Archives in the Internet age should rely on privacy protections that are not dependent on the integrity of the user.

**Sealing Archives to Protect Privacy**

Hodson was one of the first scholars to discuss the issue of third-party privacy in manuscript collections, which are a step removed from government data collections described by MacNeil. Hodson (2005) wrote about archives of celebrity authors as a special case; unlike archives of political figures, writers’ archives are often valued for the private nature of the contents. Furthermore, writers may include correspondence with third parties (lovers, friends, colleagues) who did not give permission or do not even know that their personal information is archived. Hodson (2005) advocated a commonly used procedure to protect individual privacy - namely, “sealing archives for a ‘reasonable’ period of time, possibly past the lifetime of the author” (p. 142).

Gaudette (2003), after exhaustively analyzing different types of protection for third-party materials in archives-materials created by persons she terms “blind-donors” (p. 21), concluded that “the simplest and most sensible solution” (p. 30) to the dilemma is to apply time restrictions to the entire archive or the portion that contains blind donor material. Gaudette (2003) proposed a restriction based on United States Copyright Law of life of the author plus 70 years, unless the blind donor consents to earlier inclusion in the archive.

Web 2.0 users are similar in nature to blind donors. They may understand that they are posting material on the Internet for their friends and family to access, and that strangers may also see the information as a matter of course; this is not necessarily the same as understanding that their web postings will be preserved, collected, and studied, perhaps even past their death. Regardless, both the Hodson (2005) and Gaudette (2003) proposals for sealing material in archives beyond the life of the author, while laudable for their simplicity and heady respect for privacy, would not work in the Internet era. Web 2.0 archives include the material of too many blind donors – the archivist would not be able to track all the users, nor their dates of death.

In a few decades of widespread public use, the Internet has transformed several times over. Internet communities have risen, fallen, and been forgotten in the space of years, and archives of Internet material should reflect how rapidly the medium changes. A time restriction on an Internet archive that is as lengthy as Hodson (2005) and Gaudette (2003) suggest would effectively result in an archive of questionable value.
Donor Agreements

Hodson (2006) argued that monitoring archives for privacy violations requires particular attention by the archivist:

Archivists must examine the content of this material, must be aware of the donor’s or family’s knowledge of the material and their sensitivities about it, must keep in mind the currency of the material (i.e., is it contemporary?; how many people are still alive who might suffer invasion of privacy from the revelation of information in the material?), and must probe the motivations for considering restricting the material. (para. 20)

As previously described Hodson’s (2005) work involved manuscript collections of famous authors. These collections, by definition, are smaller and circumscribed. While it may be desirable for archivists to know their collections inside and out, it is unlikely that Web 2.0 archivists will be able to master the contents of their archives, let alone the sensitivities and motivations of millions of people who participate in Web 2.0 activities.

Mark Greene, Director of the American Heritage Center and a Fellow of the Society of American Archivists, has criticized Hodson’s approach as impractical for it is “implying archivists can afford to focus immense attention on making fine distinctions and decisions for small numbers of VIP collections” (Greene, 2005, p. 13). Archival collections of ordinary people merit the same attention and concern for third-party privacy as celebrity archives, but as a collection gets larger, knowing the collection intimately and making case-by-case decisions becomes more impractical (Greene, 2005).

In a separate article Greene (1993) considered whether questions of third-party privacy are best addressed by donor restrictions, suggesting that most archivists simply do not and should not impose institutional restrictions to protect third-party privacy. Greene acknowledged that archivists cannot know what material might be sensitive, damaging, or offensive to third parties, and that no two third-parties will necessarily view these issues similarly. Green (1993) writes,

I consider it a fond delusion to believe that we can realistically find “sensitive” material in large, modern manuscript collections, that we can make tenable judgments about whether material represents a potential invasion of privacy, or that (as an alternative), we must self-impose lengthy restrictions on large portions of our collections. (p. 34)

Greene (1993) suggested that donor-imposed restrictions are ideal because donors presumably understand the privacy concerns of affected third parties whose names appear in the archives, and donors can impose restrictions to address privacy, including unequal restrictions, without violating any archival ethics codes. To draft thoughtful donor agreements that address third-party privacy,
Greene (1993) proposed a “donor version of the Miranda card” (p. 37), which would assure the donor understands their obligations and alternatives.

Greene’s (1993) argument that archivists should not be taxed with protecting privacy on an item-by-item, person-by-person basis is very compelling. Requiring archivists to make judgment calls on materials before releasing them would be time-consuming and prone to error. Multiple archivists could reach different conclusions about the same material. Administration of this privacy scheme could bottleneck the operation of an archive, and excessive restrictions on sensitive material could hamper research efforts on sensitive topics. And yet, Greene’s (1993) total confidence in donor agreements might be overestimated. Historically, donor agreements were used to restrict access to archives based on religion or gender (Case & Xu, 1994). When excerpts from former Supreme Court Justice Thurgood Marshall’s papers appeared in the Washington Post, then Chief Justice Rehnquist explicitly threatened the Library of Congress, stating that “future donors of judicial papers will be inclined to look elsewhere” (Cox et al., 2009, pp. 44-45). When making agreements, donors may be motivated by factors that archivists rightly find unsavory or immaterial.

Particularly, I am troubled by Greene’s (1993) suggestion that archivists can end-run around ethical obligations by permitting donors to place restrictions that an archivist cannot. In the Internet era it seems that many user privacy protections come about as a result of legislation or as the aftermath to large-scale public relations disasters following a breach of privacy. Corporations may have little incentive to create protections for third-party privacy, and when they do create those protections, they are doing so to protect their corporate interests rather than blind donor interests. Although donor agreements are one of the most practical tools to protect privacy, the thoughtful archivist must be prepared to suggest and influence corporate donors, and provide options and alternatives, including novel technical solutions. Archivists must also be prepared to turn down donated archives if the donor is not willing to address fundamental privacy concerns.

**Twitter: A Case Study**

Matt Raymond, Director of Communications for the Library of Congress, wrote a thoughtful blog post approximately two weeks after receiving the gift of the Twitter archive. In the post he described some safeguards that the Library of Congress would employ to protect the privacy of Twitter users. For example, private account information and deleted Tweets will not be included in the archive, and there will be a minimum six-month delay before the publication of a Tweet and its availability in the archive (Raymond, 2010). These precautions reflect scholarly suggestions made about protecting privacy in archives. The rules will be applied consistently and universally; this is a sensible approach where
affected blind donors will presumably number in the millions, and may be anonymous or pseudonymous. In the age of Web 2.0 archives, tracking privacy to the date of death of the user creator simply is not practical. Implementing some delay is wise, because it allows users to delete their account or individual Tweets if they post something they regret. A six-month delay is short, but in the fast-paced Web 2.0 world, a period of decades or even years might diminish the relevance or value of the archive.

Raymond (2010) states that the archive will be available to researchers, but it is not yet clear whether this is an institutional policy or a condition of the gift from Twitter. Given the Library of Congress’s actions when the Thurgood Marshall papers were restricted to scholars, it is likely that this provision will be interpreted broadly. Since much of the Twitter archive will be available to search on Google (Stone, 2010), a scholars-only restriction has less practical import. Encouragingly, the Library of Congress (2010) plans to collaborate with researchers to assess issues of third-party privacy, using the Twitter archive as a case study. Use of the archive may expose some serious privacy violations as a result of the new technology, but the Library of Congress is cognizant of the potential for problems and committed to protecting privacy as an ongoing goal (Raymond, 2010).

Twitter’s documentation of the archive issue however, is not as reassuring. Its Privacy Policy informs users, “Your public information is broadly and widely disseminated. … You should be careful about all information that will be made public by Twitter, not just your Tweets. … Tip: What you say on Twitter may be viewed all around the world instantly” (Twitter, 2011a, paras. 6-7). In its Information Sharing and Disclosure section, Twitter (2011a) does not reference the archive or the Library of Congress, merely stating, “We may share or disclose your non-private . . . information, such as your public Tweets” (para. 19). Similarly, in its Terms of Service, Twitter (2011b) reminds users they “consent to the collection and use (as set forth in the Privacy Policy) of this information, including the transfer of this information to the United States and/or other countries for storage, processing and use by Twitter” (para. 7).

Although it is not providing users with incorrect or false information, Twitter is capable of disclosing the Library of Congress Twitter archive in a more straightforward way. Explicit references to the archive institution and the restrictions placed on the archive would educate users and enable them to make more informed decisions about what they post. Twitter does include specific information about the archive, clearly stating, “It should be noted that there are some specifics regarding this arrangement. Only after a six-month delay can the Tweets be used for internal library use, for non-commercial research, public display by the library itself, and preservation” (Stone, 2011, para. 3). The
reference to “specifics regarding this arrangement” (Stone, 2011, para. 3) does suggest that Twitter and Library of Congress negotiated the terms of the gift. By including this information in a blog post rather than in its official Privacy Policy or Terms of Service, Twitter could be suggesting to users that the restrictions on the use of the archives are malleable or unimportant, and this may damage public trust in the archive, or in digital archives generally.

Conclusion

Archives of Web 2.0 information are essential and merit preservation. In order to protect the privacy of third parties who create the information, archivists should consider the following protective measures:

• Anonymize sensitive personal information, such as account registration information, that will not be useful to users, keeping in mind the mission of the archive;
• Seal archives for an appropriate period of time;
• Implement technical solutions to permit access to the archive while protecting privacy;
• Establish policies for addressing complaints and monitoring potential privacy violations;
• Revise and update professional ethics codes to address Web 2.0 archive challenges; and
• Educate other archivists about Web 2.0 archive problems and solutions.

Archivists should use their expertise to educate and influence donors of archive materials. In the case of a corporation donating material created by third parties, archivists should work with corporate donors to create policies that allow third parties to make informed decisions, such as:

• The archival institution, the existence of the archive, and the restrictions on its use should be clearly disclosed on the website’s terms of service or privacy policy;
• Users should be given an opportunity to opt-out (or opt-in) to archives via a checkbox;
• Changes to the archive policy should be disclosed to users in the same way changes to the website’s terms of service or privacy policy are disclosed; and
• In the event that archivists cannot persuade corporate donors to adopt measures reasonably protective of third-party privacy, and cannot protect third-party privacy completely in the archive in practice, the archive gift should not be accepted.
References


