What Happens When Facebook is Gone?

Frank McCown
Harding University
Computer Science Dept
Searcy, Arkansas, USA 72149
fmccown@harding.edu

Michael L. Nelson
Old Dominion University
Computer Science Dept
Norfolk, Virginia, USA 23529
mln@cs.odu.edu

ABSTRACT

Web users are spending more of their time and creative energies within online social networking systems. While many of these networks allow users to export their personal data or expose themselves to third-party web archiving, some do not. Facebook, one of the most popular social networking websites, is one example of a "walled garden" where users' activities are trapped. We examine a variety of techniques for extracting users' activities from Facebook (and by extension, other social networking systems) for the personal archive and for the third-party archiver. Our framework could be applied to any walled garden where personal user data is being locked.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*

General Terms

Design, Experimentation, Management, Human Factors

Keywords

digital preservation, social networks, personal archiving

1. INTRODUCTION

A few months ago, a former graduate from Harding died after surgery complications stemming from a traffic accident. His girlfriend, who was "friends" with this student's Facebook persona, was able to view and notify his other Facebook friends of what happened. His account is still active on Facebook, and it is likely his family is saving and printing whatever items of interest they find publicly available. Unless they have access to his password, they will likely be unable to recover some truly important messages that are only accessible to the account owner. As a thirty-something-year-old when he died, it is safe to assume a good amount of this former student's life is well documented in Facebook. It is also likely he was not prepared to die at such a young age, and much of his personal life, which lies in the digital "cloud", may never be accessible to his loved ones [15].

Today's generation of college students, and increasingly a larger percentage of the US population, is investing significant amounts of time in Facebook[1], a social networking website which has become one of the more popular development platforms as well. Facebook claims to have over 175 million active users, more than half outside of college, who are spending 3 billion minutes a day on the site [4]. Facebook replicates a number of traditional applications like email, instant messaging, photo and video sharing, and blogging. Thus Facebook is becoming a world unto itself, capturing large quantities of valuable personal information and interaction.

National libraries, archives, and non-profit institutions like the Internet Archive have been working for years on archiving the Web for posterity. As the Web has begun transitioning into a Web 2.0 world, archivists have taken note and adapted [6], developing new techniques to archive websites like YouTube [9] and MySpace [10], for example. But a growing amount of personal (and what will be historically significant) information is locked behind the walled garden of Facebook.

From a third party perspective, archiving Facebook presents a number of obstacles. Access to the website is mostly closed to web crawlers (except for scaled-down personal profile pages) and is password-protected. Privacy issues surrounding Facebook profiles introduce ethical dilemmas for archiving [13, 22], and even Facebook's Terms of Use prohibits "data mining, robots, scraping or similar data gathering or extraction methods [2]," regardless of what the data is to be used for. Researchers unable to obtain data directly from Facebook have resorted to creating fictitious accounts in order to crawl Facebook for useful information, only to have their accounts disabled by Facebook [11].

Archiving one's personal Facebook data is also not currently possible. Facebook does not provide a mechanism to locally archive one's profile, activities, or messages or to export one's profile to other social networking sites. This is despite efforts like the *Bill of Rights for Users of the Social Web* [23], a manifesto espousing the opinion that all data from social networks should be transportable, and public statements made by Mark Zuckerberg (founder of Facebook) in 2007 supporting that opinion [21].

[1] `http://www.facebook.com/`

JCDL'09, June 15–19, 2009, Austin, Texas, USA.

Facebook makes all undeleted messages, wall posts, photos, and other interactions accessible, from today back to the time the user creates their account, but Facebook does not guarantee such items will always be accessible. A growing number of individuals have had their accounts deactivated without warning or indications as to how they broke Facebook's Terms of Use (e.g., [25]). These individuals have understandably expressed alarm over the loss of their personal data. A recent data-loss incident at Facebook has also produced speculation that other personal data could be in danger [5].

Facebook may not always be around. Although this statement might make today's college students sneer with disbelief, librarians, archivists and historians are more likely to nod in agreement. A number of Web 2.0 sites provided by "stable" companies have disappeared over the last few years (e.g., Yahoo Mash [16], Yahoo Photos [8], and Google Lively [3]), taking many users' data along with them. One may only speculate how many web services would be terminated in a Microsoft-Yahoo merger. Even websites designed for archiving the Web are not guaranteed to persist [17], especially in today's increasingly harsh economic climate which may force a number of Web 2.0 websites out of existence.

We have explored the types of personal data being stored in Facebook and have examined methods by which this data can be extracted and archived for personal use or for posterity. As far as we know, this work is the first effort to examine the problem of archiving Facebook accounts and to propose possible solutions.

2. FACEBOOK DATA

There are several types of personal data stored in Facebook as described below:

1. **List of Friends** – This list constitutes Facebook users who have been accepted as "friends". These friends may usually view all Wall posts, photos, and Notes produced by the Facebook user, depending on security settings.

2. **Personal Information** – This includes whatever information the user wants to make publicly known such as: name, birthdate, political and religious affiliations, personal likes (movies, books, quotes, etc.), work and education, and group memberships.

3. **Wall posts** – These are public messages received from other users and applications. The Wall includes the user's *status updates* which are short, personal updates entered by the user that often express what the user is currently doing or thinking (e.g., "Becky is excited the Cowboys won!"). Comments that are made about other's Wall posts, status updates, photo thumbnails, etc. are placed here, and applications may post information here (e.g., the Texas Football Fans application shows a taunting message from another Facebook user).

4. **Messages** – These are private messages received from other Facebook users, similar to email messages.

5. **Photos** – Photos may be posted by the user in albums and tagged to indicate who is pictured in the photo. Others may also leave comments about the photo.

6. **Notes** – These are blog-like entries that users can create containing text and photos. Others may post comments about the note. Notes may also be imported from external web feeds (e.g., a user's external blog can be imported as Facebook Notes automatically).

Facebook maintains all this information in their databases and only purges data by user request. Old Wall posts can be viewed (somewhat painstakingly) by repeatedly clicking the "Show More Posts..." link on the bottom of the Wall screen. Old messages that have not been deleted can also be viewed by clicking on links at the bottom of the Inbox screen. An unlimited number of photos can be posted on Facebook, and old Notes are always accessible.

While some of these items may be judged to be mere ephemeral remarks, some of it will likely have great personal value to the owner after a significant amount of time has gone by. Much of the personal artifacts stored in Facebook accounts will likely prove valuable to users' surviving family, children or grandchildren, and certainly to historians and sociologists.

3. WHAT AND HOW TO ARCHIVE

When archiving a Facebook account or other web content, there are several strategies which may be employed [7]:

1. **Archiving of bits** – Retain the bit patterns of the web pages as they are rendered.

2. **Archiving of content** – Retain only the text, images, etc. that appear in the web pages.

3. **Archiving of experience** – Retain the look-and-feel and interactivity of the web pages.

Archiving the bits is the most straightforward approach and is the approach most often taken by web archivists when archiving general websites. This requires downloading all HTML, style sheets, JavaScripts, Flash, and any other content required to generate each view of the personal digital artifacts making up one's Facebook account. The files would be viewable in a web browser and would allow the user to see their account as it appeared when archived.

Archiving the content only requires us to grab the textual and image content from the account. This captures the essence of the account, but the original formatting is lost.

Archiving the experience may be the most daunting archival task since the user should be able to not only navigate between pages but also perform search functions. This requires having all the HTML, style sheets, etc. making up the website as well as replicating the database and information retrieval components of Facebook. A less demanding form of archiving experience would involve link re-writing to just maintain navigational functionality of the site. The look-and-feel and interface may also be altered in subtle ways as web browsers change in the future, so an emulation environment may also be needed to archive the full experience.

There are several methods we have devised that could be used to archive one's personal Facebook account:

1. Enable email notifications from Facebook to archive textual data in an email account.

2. Manually copy and paste screenshots or the text and images from all Facebook screens into a word processor.

3. Use the web browser or browser extension to save complete snapshots of each Facebook screen to disk.

Table 1: Facebook API support for data extracting

Data Type	API access	API functions
List of friends	yes	friends.get
Personal info	yes	users.getInfo
Wall posts	no	
Messages	no	
Photos	yes	photos.get, photos.getAlbums, photos.getTags
Notes	no	

4. Use the Facebook API to mine user data.

5. Automate saving snapshots of Facebook data using a web crawler or browser extension.

Facebook can be configured to send emails to users containing some Wall posts and messages from others as they are posted (Option 1). Having this option enabled will allow a user to archive these textual messages to their email, but this option must be enabled *before* the messages are sent. Although this solution is ideal for archiving only a limited amount of the account content (status updates and images are not included), it is the most promising for an individual if they are shut-out of their account.

Option 2 is probably the most labor-intensive, but it could be used to archive content and visual appearance. The user would have to manually navigate through all messages, wall posts, etc. and then transfer the screen-shots and/or text to a word processor or other software that could accept this information. For users who have accumulated several years worth of activity, this could be a very long and tedious operation.

Option 3 would be helpful in archiving the bits and partially the experience. Most popular web browser like Internet Explorer or Firefox have the capability to save the HTML, images, style sheets, JavaScripts and other related files to disk and re-write internal links and references to point to the archived material. The ScrapBook Firefox extension [12] enhances this capability by providing the ability to crawl a website to multiple depths. It also provides quick access and search on archived pages. Like previous options, this would not violate the Facebook Terms of Use if the crawling option were not used, but it would be equally slow to archive a Facebook account. And although ScrapBook has the ability to crawl a website to various depths, it cannot handle Facebook's AJAX functionality and produce a complete archive of a Facebook account. The advantage over the previous options would be having the material in its original format.

Facebook provides a freely available API for writing Facebook applications[2]. The API could be used to archive some content (option 4), but since it is very security-conscious, it lacks access to some pivotal information that would be necessary to produce a complete content archive of a Facebook account. A listing of Facebook data types and API support for accessing the data is provided in Table 1. Because the API lacks access to Wall posts, messages, and Notes, a complete archiving of content is not possible.

The final option would allow the bits, content, and most of the experience to be archived in the least labor-intensive manor. However, we are unaware of any existing web crawler or browser extension (or robots) which is currently able to archive a complete Facebook account. Facebook uses many AJAX requests to navigate the website which is problematic for any automated robot. For example, the "Show More Posts..." link which allows access to older Wall posts issues an asynchronous POST request which returns HTML composing the next listing of friends. A robot cannot simply mine the HTML and JavaScript of the Wall page for a link to produce this content.

[2]`http://developers.facebook.com/`

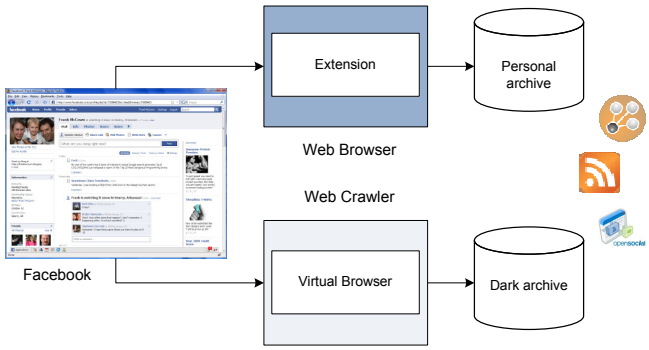

Figure 1: Framework for capturing Facebook accounts.

Automating requests against Facebook violates the Facebook Terms as mentioned earlier. The question of whether to screen-scrape a third party's website or not is a topic of debate [18, 19]. We believe the limited use of automated access to one's Facebook account would put little strain on Facebook's resources and maximize the utility of their service. An archiving service would possibly induce users to invest more effort and time in the Facebook environment since they would be less worried about their data being trapped or lost. But there is also the risk that an archiving service might make it easier for people to migrate their content out of Facebook and into other social networks, something Facebook leadership would likely frown upon. However, the existence of such an application, even if it is viewed negatively by Facebook leadership, would hopefully cause them to reconsider the long-term ramifications of locking away so much personal data.

4. ARCHIVING FRAMEWORK

We have developed a framework for capturing Facebook accounts for a personal archive via a browser extension and by third party web archivers (Figure 1). A browser extension or web crawler with an embedded virtual web browser can be built specifically for extracting Facebook data by leveraging a web browser's JavaScript interpreter and DOM to take complete snapshots of fully rendered pages after making AJAX requests. After producing an archive, the archived data could be mined and bundled with an OAI-ORE Resource Map [14]. Or it could be made accessible as an Atom/RSS feed which would make it easy to read in a feed reader. The data could be made accessible to the OpenSocial API, a standard social-networking API developed by Google that is supported by a number of social networking sites [1]. A dark archive with limited access could be made available for historians and other researchers.

The Browser Monkeys project [24] for the Heritrix web crawler [20] is one effort that could be leveraged to crawl multiple Facebook accounts. It uses instances of the Firefox browser to extract links and perform other actions, and it could be modified to archive rendered pages from a Facebook account. The crawler would need access to an account directly via username/password or via a friend profile (although some data like Messages would be hidden).

The ScrapBook Firefox extension [12], mentioned earlier, could be used in the given framework to produce an archived Facebook account. We are currently in the process of producing a prototype called Facebook Archiver which uses a modified version of ScrapBook to perform specific AJAX requests in order to capture each screenshot of a Facebook account. We are modifying the internal linkage of the captured pages to make the archived collection browseable. Facebook Archiver could be used by an individual to archive their account or by a third party to archive a friend's account (although access to private portions would not be possible, as in the opening example in this paper).

Facebook Archiver must delay a number of seconds between each request to be polite (a regular practice for web crawlers) and in order to avoid tripping any automated detection of robots that Facebook has implemented. There are no published statistics for how much information is stored on the typical Facebook user's Wall, but given the author's account which is probably typical of an occasional Facebook user, it would take twelve AJAX requests to cycle through the author's entire Wall contents (2.25 years of Wall posts with approximately 600 postings). This equates to four minutes of time to archive the entire Wall contents.

5. ACKNOWLEDGEMENTS

We would like to thank Jacob Cantrell for his investigation and insight regarding the Facebook API.

6. CONCLUSIONS

Reliance on any large organization for the continued access of valuable personal information is unwise. In the event that Facebook should disappear, we speculate optimistically that the rich quantities of information they have accumulated will be made exportable to Facebook users or perhaps turned over to an organization like the Internet Archive or Library of Congress. In the meantime, we maintain that users should be given the opportunity to obtain without much effort a viable copy of their life on Facebook. Our work is the first step in making this possible.

7. REFERENCES

[1] OpenSocial - Google Code. http://code.google.com/apis/opensocial/.

[2] Facebook terms of use, Sept. 2008. http://www.facebook.com/terms.php.

[3] Lively no more, Nov. 2008. http://googleblog.blogspot.com/2008/11/lively-no-more.html.

[4] Facebook statistics, 2009. http://www.facebook.com/press/info.php?statistics.

[5] R. Adhikari. Facebook gets a spanking over data loss, Dec. 2008. http://www.internetnews.com/webcontent/article.php/3788386.

[6] P. Anderson. What is web 2.0? Ideas, technologies and implications. *JISC Technology and Standards Watch*, February 2007.

[7] W. Y. Arms, R. Adkins, C. Ammen, and A. Hayes. Collecting and preserving the Web: The Minerva prototype. *RLG DigiNews*, 5(2), Apr. 2001.

[8] M. Arrington. Breaking: Yahoo to shut down Yahoo Photos in favor of flickr, May 2007. http://www.techcrunch.com/2007/05/03/breaking-yahoo-to-announce-closure-of-yahoo-photos-tomorrow/.

[9] R. G. Capra, C. A. Lee, G. Marchionini, T. Russell, C. Shah, and F. Stutzman. Selection and context scoping for digital video collections: an investigation of YouTube and blogs. In *Proceedings of JCDL '08*, pages 211–220, 2008.

[10] E. Cook. Web archiving in a Web 2.0 world. In *Proceedings of Dreaming 08*, Sept. 2008.

[11] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking Facebook: characterization of OSN applications. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 31–36, 2008.

[12] Gomita. Scrapbook Firefox extension. http://amb.vis.ne.jp/mozilla/scrapbook/?lang=en.

[13] R. Gross and A. Acquisti. Information revelation and privacy in online social networks (the Facebook case). In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80.

[14] C. Lagoze, H. Van de Sompel, M. L. Nelson, S. Warner, R. Sanderson, and P. Johnston. Object re-use & exchange: A resource-centric approach, Apr. 2008. arXiv:0804.2273v1.

[15] C. Marshall, S. Bly, and F. Brun-Cottan. The long term fate of our personal digital belongings: Toward a service model for personal archives. In *Proceedings of IS&T Archiving 2006*, pages 25–30, May 2006. arXiv:0704.3653v1.

[16] C. McCarthy. Yahoo Mash gets smashed, bashed, quashed, Aug. 2008. http://news.cnet.com/8301-13577_3-10028716-36.html.

[17] F. McCown. Archive spam, June 2008. http://frankmccown.blogspot.com/2008/06/archive-spam.html.

[18] F. McCown and M. L. Nelson. Agreeing to disagree: Search engines and their public interfaces. In *Proceedings of JCDL '07*, pages 309–318, June 2007.

[19] J. McHugh. Should web giants let startups use the information they have about you? http://www.wired.com/techbiz/media/magazine/16-01/ff_scraping?currentPage=all, Dec. 2007.

[20] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. An introduction to Heritrix, an archival quality web crawler. In *IWAW '04: Proceedings of the International Web Archiving Workshop*, Sept. 2004.

[21] J. C. Perez. Web 2.0: Facebook wants to make members' data portable. http://www.macworld.co.uk/digitallifestyle/news/index.cfm?newsid=19410, Oct. 2007.

[22] A. Rauber, M. Kaiser, and B. Wachter. Ethical issues in web archive creation and usage – towards a research agenda. In *Proceedings of IWAW 2008*, Sept. 2008.

[23] J. Smarr, M. Canter, R. Scoble, and M. Arrington. A bill of rights for users of the social web, Sept. 2007. http://opensocialweb.org/2007/09/05/bill-of-rights/.

[24] B. Tofel and E. Vahlis. Browser monkeys: Leverage browsers for link-extraction, 2006. http://webteam.archive.org/confluence/display/SOC06/Leverage+browsers+for+link-extraction.

[25] M. Warren. Facebook deleted my account - find out why they could do the same to you, Aug. 2008. http://www.mattwarren.name/2008/08/17/facebook-deleted-my-account-they-could-do-the-same-to-you/.