

Is Web 2.0 Privacy Stuck in 1999, and Can They Do Better?

Jon Hyman and Kevin Bombino

{ jhyman , bombino } @ eecs.harvard.edu

The “Web 2.0” revolution is in full swing and entirely new classes of interactive web services are now being used by millions of people. Instead of simply browsing web pages, people are interacting with these websites by posting their own personal data, such as photos, videos, appointments, and more.

People are taking their personal interactions online: social networking sites such as Facebook, MySpace, and LinkedIn claim millions of users each. Personal productivity applications are also moving online: applications for messaging, chat, accounting, and contact management that used to live on individual desktops are now being offered as hosted services. Even wholesale data is being put on the Internet: services such as XDrive and Apple’s .Mac are offering online file storage and backup.

As a result, more people are entrusting third parties with their data than ever before.

“Web 2.0 is about controlling data.”
– Tim O’Reilly (who coined the phrase Web 2.0)

Because of activism by the privacy community in the early days of the Internet, it has become a given that all commercial websites will likely have some sort of privacy policy. However, most of these privacy policies are built around the (now antiquated) notion that the World Wide Web is a collection of static content and the primary threat to privacy is for someone to track what you have been looking at. Now that users are providing content and data to these Web 2.0 companies, privacy policies should be revisited to ensure that they explain the privacy practices with regard to this data. How is this it being protected? Are these companies mining it to look for trends? Can it be sold without our knowledge?

We scoured the web and took a look at the privacy policies of many Web 2.0 companies in the areas of personal applications, communications and networking, and online backup. We looked at the following personal applications: Google’s Apps and Calendar, 37signals’ Basecamp (project management), Stikipad (online wiki), and QuickBooks (accounting software). The communications tools we looked at were GMail, Facebook, MySpace, Meebo (web-based instant message gateway), and LinkedIn. In the backup space, we looked at XDrive, Apple (.Mac), and Amazon Web Services (S3). All of these services have hundreds of thousands of users each, although most reach into the millions. All privacy policies examined were the current versions posted on the websites for these services as of May 4, 2007.

Analyzing a Web 2.0 Privacy Policy

A privacy policy is very useful for understanding how a company will use the data it collects. Plenty of work has been done on evaluating privacy practices of companies in many respects, such as user tracking, and we will not be undertaking a full analysis of the privacy practices of each of these websites. Instead, we would like to focus our study on how these privacy policies treat the “Web 2.0” aspect of their respective websites – that is, how they are treating the new types of data, such as private messages, photos, video, tags, and social network interactions that consumers are trusting them with.

Additionally, we want to look at how these policies are treating the difference between (what we’re going to call) sensitive and non-sensitive data. The idea is that there is a fundamental difference between data that is posted to a Web 2.0 application for the purpose of publishing (non-sensitive) and data that is for personal use only (sensitive). For example, the contents of a personal profile on MySpace are intended for publishing, so we would consider that to be non-sensitive data. However, users of TurboTax online store their financial data on TurboTax's servers so you they use TurboTax's online processing software, so this should be considered sensitive data.

It is important to realize that most Web 2.0 websites actually process and store both sensitive and insensitive data for its users. Take Facebook as an example: a user's picture is insensitive data (the user wanted it to be published), while a private message that the user sent to his girlfriend is sensitive data (it was processed by the system for the exclusive use of those two users). The privacy policy should reflect a distinction between these different types of data being stored.

Our findings

We found that although some websites do seem to be on the mark with regard to their privacy policy, many of the privacy policies we looked at do not address the differences between sensitive and non-sensitive data, and some don't address much of anything at all.

One thing that we did find abundant in most of these privacy policies was information about cookies and banner image tracking – the things that were originally concerning to people when privacy policies were first created back in the late nineties. We suspect that most current privacy policies are written using an existing privacy policy as a template, and since these issues have been covered exhaustively since the dawn of the privacy policy, lawyers continue to include these provisions. It's hard to blame just the lawyers -- most online “privacy policy generators¹” ask questions such as “does your website have links?” instead of “what are the different types of data that the user might provide and how do you handle it differently?”

¹ <http://www.the-dma.org/privacy/creating.shtml>
<http://www.enbs.com/privacy-policy-generator.php>

But we want to focus on user-submitted data. MySpace lists in its privacy policy that it collects the following user submitted information: name, email address, age, and profile data. Is that all? What about personal messages, photos, photo comments, and videos? You know, all the Web 2.0 stuff – the stuff O'Reilly talks about when he says that Web 2.0 is about data. Presumably they lump all of that stuff under the term "profile data," even though the uses of these different types of data should vary greatly. Private messages should be protected more than public "comments" (the MySpace equivalent of the Facebook "wall").

A few of these companies do make an explicit distinction between sensitive and non-sensitive data. For example, XDrive specifically states in their privacy policy that they do not look at the files users upload, but files that are shared through the service are subject to different expectations of privacy than files that are simply backed up:

"[XDrive] do[es] not use the data, information or files that you submit, upload, post, or download on or through Xdrive ("Xdrive Files"). If, however, you disclose your Xdrive Files on public areas of the service, or make such Xdrive Files available to others, such as by sharing, you understand and agree that other online users will have access to any information you provide."

Google takes a similar stance with its Calendar and Docs and Spreadsheets services, saying that by default, information is private and will only be made visible to the user, but if the user chooses to disclose it, it can "be read, copied, used and redistributed by people [he] know[s] or, again if [he] choose[s], by people [he] do[es] not know." As a result, Google includes a disclaimer in their privacy policy supplement, saying to "use care when including personal data". However, Google does admit that they use a user's account content "internally to deliver the best possible service to [their users], such as improving the...user interface and maintaining a consistent and reliable user experience." Therefore, after reading the these two privacy policies, it is easy to determine that it is safe to backup my financial statements to XDrive, but it might not be entirely safe to make a Google Document with all of my family's social security numbers since it may be looked at by Google employees.

Opposite to acknowledging what they use and do not use, some sites such as Facebook make no distinction between what they will do with insensitive public data versus what they will do with sensitive private data. They do not explicitly promise to treat private communications any differently from public communications with regard to privacy, but rather only focus on the privacy a user can expect with respect to other users on the site. As for disclosing which data (sensitive or insensitive) Facebook uses for its business practices, Facebook is as broad as MySpace by saying that they use "information in [user] profile[s]" for advertising and marketing purposes. Could Facebook use a user's private messages for targeting advertisements? Of course – "profile information" is not explicitly defined in the privacy policy, so it appears that Facebook has free reign over using any user data they wish.

Many smaller Web 2.0 companies indicate that the users are in control of all their data. For example, 37signals says:

"Although 37signals owns the code, databases, and all rights to the Basecamp application, [users] retain all rights to [their] data. [37signals] claim[s] no intellectual property rights over the material [users] provide to the Service."

[User] profile[s] and materials uploaded remain [property of the user]. However, by setting...pages to be shared publicly, [users] agree to allow others to view and share [user] content."

This is reassuring because it lets the user know that 37signals will not sell or use the data they collect from users because they explicitly state that they do not own that data, so it is not theirs to use or sell.

In fact, we found that in general, the best privacy policies (and best privacy practices) tended to come from smaller startups in the space. There are several possible reasons for this. A cynical way to look at it would be that the smaller start-ups simply do not have as many opportunities to aggregate or sell your data, so they don't worry about it. A more plausible explanation would be that these companies recognize that while many users are basically forced to use large services like Facebook regardless of their privacy policy (because they need to connect to their peers), users aren't cornered into these smaller services. As such, many of these smaller websites are marketing their privacy protections as a feature. 37signals's Basecamp advertises on its homepage that it is "SAFE, PRIVATE, SECURE" and explains that they will make no intellectual property claim at all against your data. Stikipad also goes above and beyond the call of duty, presumably as a marketing attempt, by offering its users access to all information about them, including proprietary information that they maintain about that user and his accounts, just by making a simple email request.

Additionally, the privacy policies from these large corporations tended to be thick in legalese, which is certainly tough for the average consumer to understand. On the other hand, many of the small startup companies we studied seem to be making a conscious effort to make their privacy policies as short and digestible as possible. For example, Stikipad's privacy policy consists of seven bullet points in addition to an introductory and concluding paragraph. Other small Web 2.0 sites as well are simplifying their privacy policy: the company 37signals offers bullet points beneath large, easy-to-understand headers. Meebo even went so far as to create a human-readable privacy policy on their blog².

One troubling trend that we found among the smaller companies was a tendency to build clauses into their privacy policies saying that upon such transfer of ownership, all data will become subject to the privacy policy of the purchasing entity. This is dangerous because, as we have shown, the policies of the big guys tend to be a lot less protective.

However, this isn't universally the case. Plaxo, a strong supporter of user-owned data, claims that even if there is a business transition or policy change, users will still own their data. In the event that Plaxo has a change of ownership, users will be notified of the new privacy policy, but users who signed up under Plaxo's existing privacy policy will still be treated according to that policy. If some action is going to be taken on a user's data that was not in the original privacy policy, Plaxo will notify the user and the user will have the option to delete his data or not.

² http://blog.meebo.com/?page_id=174

Can they do better?

We think that these vendors can improve on two fronts: by being very specific in the privacy policy about each type of data shared by users and also by ensuring that these privacy policies are easy to read. One way to ensure that all data collection features are accounted for while simultaneously improving user understanding of privacy practices would be for companies to provide summaries of their privacy practices in line with the actual websites themselves.

An example of this might be on the MySpace send message screen, a little privacy icon would appear where you could hover your mouse and see a message saying "MySpace will never use the contents of this message for any purpose other than to deliver it to the recipient. We will delete all copies of it when the recipient clicks the Delete button." On the edit profile screen, that icon would display something like "MySpace will show this information to everyone who can view your profile. We may also analyze it for demographic and ad targeting purposes."

We think that by implementing this as a hover over a "privacy" icon, it remains unobtrusive to users that are uninterested in the details while being easily accessible to those who are interested. Furthermore, if this icon was standardized across websites (like the RSS icon, for example), users would automatically know what it means and how to use it.

Additionally, we would like to see more Web 2.0 companies stand firmly behind the promises that they make in their privacy policy, even in the event of a change of ownership. To sell your company to a new owner with the intention of renege on your promises of privacy is a terrible breach of the trust instilled in Web 2.0 by the community of users.

Conclusion

We set out to determine if the privacy policies of Web 2.0 websites accurately reflected the various types of sensitive and non-sensitive data collected by those sites. We found that while some of the startups came close, in general most websites do seem to have their privacy policies stuck in 1999.

On one hand, we are disappointed by the lack of specific clauses in most privacy policies to deal with the notion of how users are uploading all sorts of data and content to Web 2.0 sites. On the other hand, we are happy to see that the startups in the space are taking these privacy issues seriously, and we hope that they will continue this attitude as they expand (instead of reverting when they get bought out).

We remain excited about the interactive web and look forward to many new innovative applications and mash-ups in the coming months and years. Our plea to these websites is to make sure that the privacy policy they publish is reflective of all of the novel types of data collection that the application enables.