

# Shortcuts and Dead Ends: Control Issues with Online User-Generated Content

**Abstract:** Increasingly, internet users are creating and sharing content through a variety of socially-based websites. This includes sharing images, videos, stories, diary-like text, and personal information with others. With all of this information being created, what happens to user content once it has been uploaded to a site and effectively removed from the user's hands? We set out to explore this question within some of the popular user content hosting sites on the Internet. In doing so, we discovered a flawed paradigm wherein sites offer little guarantees as to service, but limit the precautionary measures that users can take themselves.

**Résumé :** De plus en plus, les internautes créent et partagent du contenu sur des sites web sociaux, notamment des images, des vidéos, des histoires, des entrées de journaux « intime » et des renseignements personnels. Par contre, qu'arrive-t-il lorsque le contenu de l'utilisateur est placé sur un site, hors du contrôle du créateur? Nous avons décidé d'explorer cette question sur certains sites sociaux populaires. Les résultats démontrent un paradigme imparfait où les sites offrent peu de garanties de service, mais limitent les mesures de précaution que peuvent prendre les utilisateurs.

## 1. Problem

We set out to determine, *What control do users have once they put their content online?* This stemmed from a perceived discord between what responsibilities user generated content sites absolves themselves of and what options users are given to save their data. At stake are issues of intellectual property, copyright, privacy, presumed and real data security, and the ease with which the user can identify their rights and obligations both on paper and in practice. What sort of security and philosophy is instilled on users, and how does the experiential practice compare? Additionally, how are conflicting interests of users and websites resolved? Cyberlaw is continually undergoing change in order to adapt to new social spaces for user-generated content (see Síthigh 2008), and there appear to be inconsistencies in how websites handle it.

## 2. Method

To examine this problem, we identified a series of categories for websites that host user generated content. These categories include: Social Networking, Blogging, Image Sharing, and Video Sharing. After the research team developed a common definition of each category, we sampled the three most trafficked websites of each, based on the Alexa.com statistics for the date of March 5th, 2009 (Figure 1). Though based on Canadian statistics, these rankings were found to be generally transferable to American rankings for the same period. To accommodate niche sites that did not fit into the categories, there was also a "Miscellaneous" category of four purposefully chosen sites.

Figure 1: Website Categories and Sample

Social Networking	Blogging	Image Sharing	Video Sharing	Miscellaneous
Myspace.com	Blogger.com	Flickr.com	Youtube.com	Twitter.com
Facebook.com	Wordpress.com	Photobucket.com	Megavideo.com	Plentyoffish.com
Skyrock.com	Livejournal.com	Deviantart.com	Dailymotion.com	Wikipedia.org
				Docs.google.com

A codebook was created for evaluating the websites according to common criteria. It looked at deletion policies, deletion processes, import and export functionality, per-item malleability of content, data portability features, and instructional material. There was also a qualitative examination of the Terms of Use for all websites. All researchers had a strong background with computing, though not necessarily familiarity with specific websites. To increase validity, sites in the sample were coded multiple times, with the results reviewed by the lead coder. To increase inter-coder reliability, a test sample was coded first, and the codebook adapted to any discrepancies revealed in the test. Finally, the lead coder went through the full sample first, so as to verify the clarity of coding questions.

### 3. Outcomes

We discovered that, with little exception, content hosting websites do not offer users flexible control over content. Backup functionality is rarely included, terms of service were absolvent, and deletion policies inconsistent. While websites are quick to emphasize that users retain their own copyright, they simply do not consider users' content beyond their servers. The most positive sign, however, is that an increasing amount of sites offer application programming interfaces (APIs) and other methods for data portability. This means that, in principle, technologically apt users can subvert the priorities of the service and build their own tools to fill in the gaps.

Most providers examined do not offer any sort of comprehensive export services. Only two sites - Wordpress.com and Blogger - offer full exports. In both of these cases, the exports allow a user to export an entire blog, but not all blogs in an account. Overall, five of the sixteen sampled sites offered any form of data export. While much of the content exists in the form of text and images, which by their nature can be saved from a browser, this process can be tedious. None of the video sharing websites offer downloads of a user's videos once they were uploaded, though Megavideo - a site that appears to have reached popularity as a source of piracy - offers a pays service for downloading *all* videos.

In contrast to export abilities, half of the sites surveyed offered batch import functionality. What results is a virtual brick wall for user content; like a dead-end on a one-way street, the observed paradigm is one where content feeds in and never gets out. The exception appears to be blogging, where all three web services provided standardized means for exporting data. The worst of these was Livejournal, which allows export of up to one month. The best import/export functionality was provided by Wordpress.com, which in addition to comprehensive export, allows import from a wide variety of standards. Wordpress.com is one of two sites that operate on open-source platforms. The other, Wikipedia, supports comprehensive data movement functionality in its platform,

but has a number of these features turned off.

Deletion policies were unevenly provided by the sampled websites. Five services offered no account deletion whatsoever, while four offered a "soft-delete". The soft-delete allows users to change their minds at a later time and reactivate their accounts, while a permanent delete immediately destroys the data. Most soft-deletes eventually roll over into permanent deletions, but the timeframes we observed ranged from 48 hours to 6 months. Treatment of deleted material and links to it was similarly inconsistent. Despite these problems, the majority of sites in the sample offered privacy settings, either public/private or more complex.

Online service providers find themselves with no clear answers as to how to share users' public content. This issue surfaces most apparently with the handling of comments. Though providers emphasize retention of copyright, they also do not want discussions broken by disappearing comments. We found that most of the services in the sample do not delete comments when a user destroys their account. Some even suggest that a user delete their comments prior to their account, as they cannot afterward. In Blogger and Wordpress.com, users can only delete their blogs, rather than their full account, thus keeping comments linked to their accounts. This also factors into copyright issues, and some Terms of Service agreements include an indefinite license to use the content. Shortly before this study, Facebook was subject to a controversy regarding overreaching wording of such a clause, but we found that it is not uncommon.

The terms of service offers an insight into the service's view of such issues. As expected, most websites renounced any responsibility to service reliability and claimed the right to terminate accounts and content at their will. This clauses are off-putting in light of the space support for content backup. Some sites, such as Facebook and Megavideo, explicitly place responsibility for backups on the users.

#### **4. Future Research Directions**

There are a series of questions that still remain to be explored. One such direction is a look into now-defunct websites, exploring their policies and how these affected their users. When content hosting websites die, do they give thought to the data of their users? How did their principles affect usage of the site and, if data *was* lost, did former users regret its loss?

In order to determine the facets of content control from a more representative sample of users, we would conduct a qualitative study of users ranging in computing ability and familiarity to each website initially sampled. Further user study would ask:

- Do people know or even care about the level of control they have over their content?
- What kinds of choices about their content would users rather have?
- Is this affected by the user's general familiarity with computing, particularly social computing?

A final research direction is to run this study again, for a longitudinal comparison of how content control is changing.

## 5. Conclusion

Ultimately, we found that the current state of the web is lacking in reliability as a place for hosting user-generated content. Users should be aware of these limitations and keep multiple copies of content that they share. Due to bad export practices, any content created natively in the service should be periodically saved elsewhere.

There does appear to be an improving trend and, in the months since this study, the landscape has been changing quickly. Google, which owns three of the sites studied, has an internal "Data Liberation Team" that works to encourage free movement of data amongst the company's project teams. Google Dashboard, for example, shows a user what they have stored in all their services, and Google Docs now supports multiple file export. There are also movements from below to encourage good data practices and fill in the gaps lefts by other services. One such movement is the Archiveteam (<http://www.archiveteam.org>), who keep a deathwatch of the internet and try to scrape data from dying sites before they close.

While practices have been improving quickly, the damage of the current paradigms is becoming apparent. A weak economy has resulted in numerous services being closing or scaling back on features. Two weeks after this study's completion, Yahoo announced the closing of Geocities, once the largest web-hosting service in the world, leaving the burden of archiving to others. Until content hosting practices improve and standardize, the responsibility of tending to their content unfortunately falls on users. Users are advised to be proactive in understanding the limitations and working within them.

## Bibliography

- Hargittai, Eszter and Gina Walejko 2008. The participation divide: Content creation and sharing in the digital age. *Information, Communication & Society* 11(2): 239-256.
- Shah, Rajiv C. and Jay P. Kesan 2008. Setting online policy with software defaults. *Information, Communication & Society* 11(7): 989-1007.
- Síthigh, Daithí 2008. The mass age of internet law. *Information & Communications Technology Law* 17(2): 79—94.