Heather West, Google
Position paper for IAB/W3C/ISOC/MIT
Internet Privacy Workshop:
How can Technology help to improve Privacy on the Internet?

More and more information about individuals is available online, leading some to worry that the Internet never forgets. It is easy to articulate a “privacy nightmare” that results from a Wild West online. The cultural process of adjusting to a world where information never disappears may prove difficult, as we surmise that no child engaged in social activity now could ever become president. Creating adequate, useful tools to allow users to signal their intentions with the data they share is an important step towards avoiding this dystopian vision of an Internet that never forgets, and never respects user wishes.

New, innovative services based on the information you provide continue to be introduced to a ravenous audience. In short, success in the information age has relied on how personal information can be accessed and shared quickly, easily and in a way that makes it portable and accessible to any community from anywhere on the globe.

Privacy as “the right to be let alone” or to “be forgotten” is still available to each and every citizen of the information age. However, taking that right is inherently opting out of the information society, and all the myriad benefits that it provides. After all, a society requires investment in order to reap the rewards – in this case, investment in the connected community. The information age demands new and better tools to describe how we choose to interact, and a new definition of data control that doesn’t define privacy as isolation, but as empowerment.

The information age is all about how you define yourself. The information that you choose to share and the communities that you choose to join tells the world who you are, or who you want to be. Online, identity is established as the information we present about ourselves, not what we hide. To make the next step forward, we need tools that will allow us to tell others how we intend information or data to be used. Some sites do implement these tools - privacy controls on social networks, ad network opt-outs, but they are limited to the individual site and are as implemented take a great deal of effort to find and understand. More useful would be a standard set of preferences that could be expressed to services as the user’s preferred default settings.

Since information online is a flow, rather than a one-to-one relationship, it is hard to codify ownership or value as users share and interact with information online. Rather than arguing about how to mandate “forgetting”, we should enable users to signal their intention in visible and computer-readable ways across operators. Providing users with a way to stake a claim to their information online, and say what they are comfortable with others doing with it, would easily improve the experience where it was implemented.

Tools that would allow users to signal to all service operators how they expect their data will be used or treated - in order to establish known defaults - would move the debate forward. Protocols for this kind of signalling, like robots.txt or P3P, are not new - but some have worked better than others. Looking at these protocols, we can come up with a few attributes that will make any future signalling protocol more likely to succeed.

These protocols seek to solve similar problems - how a user, unassociated with an operator, can signal

their intentions and preferences around data. Robots.txt is used to determine which parts of a site are excluded from a search index, thus allowing a user to signal to a service what content, if any, they want to make easier for others to find - and what they would like greater control of who sees. While these are not technically enforceable, they are widely respected by major crawlers. Sitemap.xml, an inclusion protocol, is less-known. It is intended to allow webmasters greater granularity in declaring which pages on a site might change often and ensuring that crawlers find each page they want indexed. Several major search engines use this protocol. P3P is intended to inform the user (through their browser) about the practices of a website that gathered their data and allow them to make a meaningful choice about the way that their information was to be used. Instead, however, the protocol did not involve a range of choices for users, and was largely directed at users rather than service providers. The protocol never found wide adoption or use.

Robots.txt benefitted from the assumption that it is a reasonable way to signal the intention that part of a website not be accessible via third parties, like search engines. This meant that rudimentary security through obscurity could be achieved, and provided some safety for providers from copyright claims - after all, as long as the user could easily tell all operators their preferences around the information being indexed and re-presented, they did not have a strong copyright claim. This provides a helpful signal for those indexing content that they do their best to respect user intentions, providing some help in lawsuits over indexing.

P3P never benefitted from this legal advantage, as users never changed their behavior based on the signals of P3P. At best, they would understand that the settings in their browser were not letting them access the services online. Despite P3P's general failure, the lessons learned from P3P can be helpful in synthesizing a way to implement a new protocol. There were never enough operators involved to get critical mass for the protocol, and it was not effective in communicating to users. While there were many choices that could be expressed through P3P, it was not easy to do so or to understand the implications of choices that were made. Operators were not given the incentives to implement it, the way that robots.txt provided incentive to respect the protocol via fair use law.

Sitemaps.xml serves almost an opposite purpose from robots.txt, signalling more granularly what a site owner would like indexed, and how often it typically changes content and should be reindexed. Through a combination of robots.txt and sitemaps.xml, a site owner can comprehensively signal to all operators what they expect to be crawled, and how often - creating a net benefit for both the site owner and operators.

A widely adopted inter-operator signaling protocol for users, in the model of robots.txt and sitemaps.xml, would go a long way towards the adoption and use of this signaling and the clear exercise of user privacy preferences. Ensuring that the protocol benefits users and operators, and ensuring that it is not seen as an industry runaround, is important in building user trust in a protocol. These tools will eventually be necessary to allow users to share their expectations around data. In a diverse and healthy ecosystem, we need to give users easy ways to signal their expectations of the services that they use.