Privacy and the W3C: Questions towards a direction

For the W3C Privacy Workshop, London, July 2010

David Singer

Multimedia and Software Standards, Apple Inc.

1 Overview

"What should the W3C do about privacy?"

This paper attempts to take this fairly open-ended question and break it down into a series of questions that are more specific, in the hopes that dialog about these – and other – questions will yield results.

We are searching for an answer to the overall question that both (a) falls within scope for the W3C, and (b) to which the W3C can reasonably be expected to develop a useful answer. The way ahead may not be clear, but workshops and dialog – to which this document hopes to contribute – may help show more of the road.

2 The Problem of Definition

The first immediate problem that arises is "What do we mean by privacy?" It is interesting to note that neither the W3C [1] nor the IETF [2] documents attempt a crisp definition. This may be because it is one of those terms notoriously in the category "I cannot give a definition, but I know it when I see it." Or rather, in this case, "I feel it when it has been violated".

What do we mean by 'privacy', specifically at the W3C?

3 The Problem of Layer

Recent round-table discussions on privacy have explored some of the (very real) problems with some existing systems, protocols, and specifications. However, the W3C is responsible for neither:

- the privacy implications of the protocols that W3C specifications rest upon (though it might be responsible for the choice of protocols and the way they are used);
- nor the way that the W3C specifications are used (though it probably is responsible for the privacy implications of its own specifications).

When the W3C sees privacy questions in the protocols its specifications rest upon, it should probably bring them to the attention of the owners of those protocols (usually the IETF).

The W3C is generally responsible for specifications at the presentation layer; it should be aware of where those specifications have a privacy implication (‘link visited’ styling in CSS is an example). Is it also appropriate to alert users to possible privacy issues in the use of the specifications (i.e. at the applications layer)? For example, personal data is often collected using form input on web pages, but it’s not clear that the HTML specification should carry warnings in the ‘forms’ section about the privacy implications of collecting personal data.

Clearly, there are a myriad ways in which W3C-specified technologies can be used, ranging from the innocuous to the obviously harmful. We cannot construct specifications with which it is impossible to do harm; and the definition of what is a potential privacy violation is clearly a question of policy, not technology.

Does the W3C have a clear idea of the ‘layer(s)’ for which it is responsible?

4 The Problem of Policy

This brings us to the questions of policy.

There is considerable cultural, and hence geographic, variation in what is expected or normal in this area. The W3C is a ‘technology provider’ and in that sense it needs to remain neutral; we cannot embody a single ‘policy’ into our specifications, without favouring some and implicitly dis-favouring others. Privacy policies are, ultimately, the responsibility of web sites, the organizations to which they belong and respond to (including companies, trade associations, and the like), and ultimately governments.

The W3C and the IETF have both looked at electronic expression of privacy policies, as already noted. Electronic, formal, privacy policy expressions may suffer from a number of problems, however. Do users understand what their own ‘formal’ privacy expressions actually say, or more importantly, what they imply? Given a variety of sites, all of which ‘respect your privacy’, how happy are users with the possibility that what that ‘respect’ means varies from site to site? (This is akin to rights expressions; users generally are not keen on having to read a formal expression of what rights they get when they buy something, every time they buy.) How likely is it that users’ formal expressions will match their desires? How likely is it that sites’ formal expressions will match their intentions?

This is a fluid area, and there is much ingenuity being applied to provide users with useful, innovative services – which often have novel, and sometimes challenging, privacy questions. The default mental model of users is probably that ‘unless I permit something, it is forbidden’, whereas the default mental model of services is probably closer to ‘unless something is forbidden, it can be considered permitted’. In a context of innovation – doing the unexpected – any policy language, and perhaps even specific policy, is liable to have grey areas. Given that, this tension of higher-level expectations is likely to result in tensions in practice as well.

Given the rapid evolution of services and understanding, can the W3C do anything useful, at the moment, with either the definition or the expression of policy?

5 The Problems of Degree and Context

It is tempting to think of privacy as a simple binary yes/no problem, and whether it is retained as an obvious 'fence' that is either respected or crossed. However, it is not clear that this simple mental model applies.

There is a question of degree. One example is when other people take photos that include me when I am in a public place: most people do not object to that, even though they never see the photos. However, their unease rises sharply when the number of images increases, or it is continuous video, and they start to talk of 'surveillance societies' (with all the negative overtones).

This is closely related to the problem of context. One converse of 'private' is 'public', but just as private is not a simple binary question, nor is public. Certain facts about me may indeed be 'public', in that people who know them are under no secrecy constraint – such as the person behind me in line at a store seeing what I purchased). However, if that person publicizes what I buy, for many to read about – perhaps not even identifying me – I may be less happy.

The problems of degree and context come up together when data is correlated and combined. Each individual fact may be of little consequence, but when presented together become disturbing.

Many specifications are written with the assumption of limited, un-correlated disclosure. For example, the HTTP 'cookie' facility was expressly designed to attempt to keep facts isolated, but there has been much (successful, it seems) work at re-correlating cookies. 'Links visited' seems fairly harmless when it exposes the state of a handful of visible links on a page, but less so when it exposes the state of hundreds of thousands of links invisibly run through a script.

There is also work in taking 'anonymized' or 'de-identified' data and 're-identifying' the individual from which it came, as well as work on deriving characteristics of that individual (such as their probable gender or sexual orientation) from their usage style.

Any analysis of a specification has to consider not only the privacy implications of a single transaction, but also how that transaction might be replicated, or the data combined with other data, or used in a different context.

Do we know how to manage the degree and context of privacy exposure?

6 The Problems of Choice and Awareness

A related field to privacy management is risk management. One of the older lessons in risk management is that it is not all about the statistics of risk (the objective likelihood of harm); people care greatly about whether the risk is (in)voluntary and (im)perceptible. The analogy is that people may be upset when something is done without their knowledge, when they would probably have given reasoned permission if asked. Similarly, they are more likely to be upset when something unforeseen happens to their privacy than if it was foreseen (and accepted) as a possibility.

In contrast to this is the tedium people experience reading policies, agreements, and 'click-throughs'. Even short license agreements are rarely read, it seems. There is an apparent paradox here: users react badly to both surprises and to detailed information.

Do we know how to keep privacy implications both informed and voluntary?

7 Conclusions

In the spirit of the rest of this document, in which the predominant punctuation is a question mark, this section concludes with some more general questions about the way ahead.

As noted above, some of the difficult privacy questions on the world-wide web concern not so much the specifications that make it cohere, but the way in which those specifications are used.

Does the W3C have enough membership from those deploying sites and services that encounter privacy questions?

When a specification is written at the W3C, is the responsible working group aware of the privacy implications of the requirements etc. in that specification?

When a specification is written at the W3C, does it alert implementers or users of the specification to possible privacy implications of using it?

Does the W3C have a sufficient 'body of knowledge' in the form of examples, and taxonomies, to be able to catalyze privacy awareness in both those making and using its specifications?

8 References

[1] The Platform for Privacy Preferences, W3C Working Group Note 13 November 2006, http://www.w3.org/TR/P3P11/

[2] Common Policy: A Document Format for Expressing Privacy Preferences, RFC 4745, H. Schulzrinne et al., Feb. 2007