List of Collected PING Privacy Review Questions

Abstract

A list of questions has been created based on earlier privacy reviews. This short writeup aims to compare the raised questions to the questions and guidance that is provided in the privacy guideline document. This is an attempt to determine the completeness of the available privacy guideline document and, if gaps are noticed, to update the description.

The privacy guideline document can be found here: http://tschofenig.priv.at/w3c-privacy-guidelines.html

Can the information be used (alone or in combination with other APIs / sources of information) to fingerprint a device or user?

[tlr, erin, npdoty, others]

This is described as part of the data minimisation guideline section and a reference to the more complete fingerprinting guideline document is provided:

> *Basic fingerprinting guidance can be found here. The following questions are relevant for specification authors: Is the designed protocol vulnerable to fingerprinting? If so, how? Can it be re-designed, configured, or deployed to reduce or eliminate this fingerprinting vulnerability? If not, why not?*

May I access to the information I created?

[karl]

Further clarification is needed here. Typically, privacy concerns are raised when information is shared or when unauthorized access to information is gained.

May I record it myself (locally)?

[karl]

Further clarification is needed here. Why is this is a privacy concern? Normally, everything can be "stored" that is available locally unless there is some DRM protection.

Am I able to have actions on this personal record?

[karl]

Further clarification is needed here. What type of actions should be applied to the personal data?

May I block partly or totally the record of the information?

[karl, tonyr]

I think this question concerns the ability to control sharing of data and the ability to share in a fine grained fashion. The guidelines on user participation relate to this aspect, namely:

> *Many Web protocols allow data to be made available through APIs and other protocol mechanisms. It is important for users to be able to control access to their data. What controls or consent mechanisms does the protocol define or require before personal data or identifiers are shared or exposed via the protocol/API?*
>
> *Does the user have control over their data after it has been shared with other parties, for example regarding secondary use? Are users able to determine what information was shared with other parties?*
>
> *What recommendations can be given to implementers and the deployment community regarding privacy-friendly sharing of information? This in particular refers to the ability to provide additional information about why the sharing is taking place (the purpose), what information is shared and with whom. Further relevant is to control the degree (or the granularity) of information sharing. Will the data subject be given enough context to make an informed decision? Will the decision about sharing with others be persistent? If so, for how long is the decision cached and how can it be revoked?*

May I fake it? (think about fuzzy geolocation or voluntary fake location)

[karl]

Further clarification is needed here. In general, information from end devices can be faked in a variety of ways. For information that is provided by a third party this might be more difficult. Which case are you referring to?

Is the data personally-derived, i.e. derived from the interaction of a single person, or their device or address? (If so, even if anonymous, it might be re-correlated)

[dsinger]

Further clarification is needed here. Is this question about whether a certain data item is actually personal data (which would then relate to the definition of personal data) or whether any data minimization techniques have been applied?

Regarding the former question about the definition of personal data the privacy guidelines document aims to provide an answer:

Personal Data: Any information relating to an individual who can be identified, directly or indirectly.

For the second question about data minimization some additional guidance is available:

Minimisation is a strategy that involves exposing as little information to other communication partners as is required for a given operation to complete. More specifically, it requires not providing access to more information than was apparent in the user-mediated access or allowing the user some control over which information exactly is provided.

For example, if the user has provided access to a given file, the object representing that should not make it possible to obtain information about that file's parent directory and its contents as that is clearly not what is expected.

Basic fingerprinting guidance can be found here.

In context of data minimisation it is natural to ask what data is passed around between the different parties, how persistent the data items and identifiers are, and whether there are correlation possibilities between different protocol runs.

For example, the W3C Device APIs Working Group has defined the following requirements in their Device API Privacy Requirements document.

APIs must make it easy to request as little information as required for the intended usage. For instance, an API call should require specific parameters to be set to obtain more information, and should default to little or no information.

APIs should make it possible for user agents to convey the breadth of information that the requester is asking for. For instance, if a developer only needs to access a specific field of a user address book, it should be possible to explicitly mark that field in the API call so that the user agent can inform the user that this single field of data will be shared.

APIs should make it possible for user agents to let the user select, filter, and transform information before it is shared with the requester. The user agent can then act as a broker for trusted data, and will only transmit data to the requester that the user has explicitly allowed.

Data minimisation is applicable to specification authors, implementers as well as to those deploying the final service.

As an example, consider mouse events. When a page is loaded, the application has no way of knowing whether a mouse is attached, what type of mouse it is (e.g., make and model), what kind of capabilities it exposes, how many are attached, and so on. Only when the user decides to use the mouse — presumably because it is required for interaction — does some of this information become available. And even then, only a minimum of information is exposed: you could not know whether it is a trackpad for

instance, and the fact that it may have a right button is only exposed if it is used. For instance, the Gamepad API makes use of this data minisation capability. It is impossible for a Web game to know if the user agent has access to gamepads, how many there are, what their capabilities are, etc. It is simply assumed that if the user wishes to interact with the game through the gamepad then she will know when to action it — and actioning it will provide the application with all the information that it needs to operate (but no more than that).

The way in which the functionality is supported for the mouse is simply by only providing information on the mouse's behaviour when certain events take place. The approach is therefore to expose event handling (e.g., triggering on click, move, button press) as the sole interface to the device.

The following questions arise concerning data minimalization:

a. Identifiers. What identifiers does the protocol use for distinguishing initiators of communications? Does the protocol use identifiers that allow different protocol interactions to be correlated? What identifiers could be omitted or be made less identifying while still fulfilling the protocol's goals?

b. Data. What information does the protocol expose about individuals, their devices, and/or their device usage (other than the identifiers discussed in (a))? To what extent is this information linked to the identities of the individuals? How does the protocol combine personal data with the identifiers discussed in (a)?

c. Observers. Which information discussed in (a) and (b) is exposed to each other protocol entity (i.e., recipients, intermediaries, and enablers)? Are there ways for protocol implementers to choose to limit the information shared with each entity? Are there operational controls available to limit the information shared with each entity?

d. Fingerprinting. In many cases the specific ordering and/or occurrences of information elements in a protocol allow users, devices, or software using the protocol to be fingerprinted. Is this protocol vulnerable to fingerprinting? If so, how? Can it be designed to reduce or eliminate the vulnerability? If not, why not?

e. Persistence of identifiers. What assumptions are made in the protocol design about the lifetime of the identifiers discussed in (a)? Does the protocol allow implementers or users to delete or replace identifiers? How often does the specification recommend to delete or replace identifiers by default? Can the identifiers, along with other state information, be set to automatically expire?

f. Correlation. Does the protocol allow for correlation of identifiers? Are there expected ways that information exposed by the protocol will be combined or correlated with information obtained outside the protocol? How will such combination or correlation facilitate fingerprinting of a user, device, or application? Are there expected combinations or correlations with outside data that will make users of the protocol more identifiable?

g. Retention. Does the protocol or its anticipated uses require that the information discussed in (a) or (b) be retained by recipients, intermediaries, or enablers? If so, why? Is the retention expected to be persistent or temporary?

Does the data record contain elements that would enable such re-correlation? (examples include an IP address, and so on)

What other data could this record be correlated with? (e.g. the ISP)

[dsinger]

The question about correlation is raised as part of the data minimization section:

f. Correlation. Does the protocol allow for correlation of identifiers? Are there expected ways that information exposed by the protocol will be combined or correlated with information obtained outside the protocol? How will such combination or correlation facilitate fingerprinting of a user, device, or application? Are there expected combinations or correlations with outside data that will make users of the protocol more identifiable?

If you had large amounts of this data about one person, what conclusions would it enable you to draw? (e.g. maybe you could estimate location from many ambient light events by estimating latitude and longitude from the times of sunrise and sunset)

[dsinger, tonyr]

This is currently not part of the provide guidance. Any idea how to phrase this aspect and where to put it?

Am I likely to know if information is being collected?

[wseltzer]

This aspect is part of the user participation section:

Many Web protocols allow data to be made available through APIs and other protocol mechanisms. It is important for users to be able to control access to their data. What controls or consent mechanisms does the protocol define or require before personal data or identifiers are shared or exposed via the protocol/API?

Does the user have control over their data after it has been shared with other parties, for example regarding secondary use? Are users able to determine what information was shared with other parties?

What recommendations can be given to implementers and the deployment community regarding privacy-friendly sharing of information? This in particular refers to the ability to provide additional information about why the sharing is taking place (the purpose), what information is shared and with whom. Further relevant is to control the degree (or the granularity) of information sharing. Will the data subject be given enough context to make an informed decision? Will the decision about sharing with others be persistent? If so, for how long is the decision cached and how can it be revoked?

How visible is its collection and or use?

[wseltzer, tonyr]

Further clarification is needed here. This question relates to the earlier question about the user participation when sharing information. Other than the user is asked to give consent what type of "visibility" do you ask for? Since this quickly moves to the user interface question about how the consent/sharing dialog could or should look like it is hard to make specifications authors to state anything explicit here.

Do I get feedback on the patterns that the information could reveal (at any instant, over time) so I can adjust behaviors?

[wseltzer]

Interesting idea. This is not currently in the provide guidelines document and for most services I would guess the answer is "yes, you get the feedback – from the media." I think that this is very hard to turn into an actionable item for someone who writes a protocol specification.

If a background event about the device is fired in all browsing contexts, does it allow correlation of a user across contexts?

[npdoty]

Also an interesting question but I do not know how to put this into the guidelines document. Nick, any suggestions?

Can code on a page send signals that can be received by device sensors on nearby devices?

[npdoty, tonyr]

Good question. Also something I would need more context to produce guideline specific text.

WHO can and does have access to the data?

[eric]

This aspect is part of the user participation section and the answer is typically the data subject (or someone on his behalf) decides about the sharing of the data. While the mechanisms may be standardized for managing these permissions (see, for example, OAuth) the actually list of entities that can have access is obviously different from user to user.

In addition, it is worth pointing out that some protocols involve multi-parties (like WebRTC) and the question of who has access to what data is an architecture decision with significant privacy implications. The available security mechanisms may also help to reduce the number of parties that have access to data elements.

Relevant questions from the privacy guideline document are:

> *How do the protocol's security considerations prevent surveillance, including eavesdropping and traffic analysis?*
>
> *How do the protocol's security considerations prevent or mitigate stored data compromise?*

In addition the section about data minimization also discusses this aspect:

> *b. Data. What information does the protocol expose about individuals, their devices, and/or their device usage (other than the identifiers discussed in (a))? To what extent is this information linked to the identities of the individuals? How does the protocol combine personal data with the identifiers discussed in (a)?*
>
> *c. Observers. Which information discussed in (a) and (b) is exposed to each other protocol entity (i.e., recipients, intermediaries, and enablers)? Are there ways for protocol implementers to choose to limit the information shared with each entity? Are there operational controls available to limit the information shared with each entity?*