

Machine Interpretable Expression of Compliance

Dave Raggett <dsr at w3 dot org>, W3C,
Rigo Wenning <rigo at w3 dot org>, W3C

Copyright © 2012 Dave Raggett, Rigo Wenning

This work contains ideas that were conducted as part of the [PrimeLife project](#) with funding from the European Union's 7th Framework Programme. The work reported is experimental and the examples shown are fictitious, and taken from a working demonstrator. In a slightly different form, this work was already submitted to an earlier [W3C Workshop on Privacy and data usage control](#)

Introduction

W3C began as early as 1996 to think about technologic remedies to the privacy issues created by the Web. It started with PICS for privacy. The W3C Platform for Privacy Preferences (P3P) 1.0 was published as a W3C Recommendation in July 2002 [1]. It defines a machine interpretable format for websites to express their privacy practices. A revised format (P3P 1.1) was published as a W3C Note in November 2006, but failed to reach Recommendation status [2]. Since 2004, W3C has been involved with Privacy Research Projects in the European Commission's 7th Framework programme.[3][4] And now W3C is working on defining a Tracking Preference Expression Specification TPE[5] and a Tracking Compliance and Scope Specification (TCS)[6].

During the work on the TPE and the TCS the question came up what it means to be compliant and how to express compliance regimes. Having tokens representing complex concepts in long human readable documents is not scalable. The Privacy Policies on Web sites and their pages of legal language did not help to improve the feeling of the users of the Web.

But how can a Service actually describe why a certain data item is needed, why it is not harmful to send it, what they have done to take the privacy worries of end users into account, convince them to use the service. Privacy is more and more in the center of marketing strategy as not addressing the topic will keep users away from services. Cloud computing, where nobody really knows where the data actually is, where the complexity of the system is leading to lack of understanding that in turn nurtures doubts about misuse of collected data. The air of transparency is the best remedy to counter fear, uncertainty and doubt about the unknown. This was the use case for P3P. This use case is higher on the agenda than ever.

To counter fear of the unknown, in summary, P3P described the business name and address responsible for the website, the dispute resolution procedures, the means (if any) for users to access personal data collected by the website, the kinds of data collected, the purposes it will be used for, the data retention policy, and the recipients of the data. It lacked a way to give a short notice to the user on how the different facts expressed in a machine readable form are relating to each other. Software was supposed to explain this and that failed. Thus the need for a new approach.

P3P supports a notice and consent model of privacy, where websites describe their privacy policies and users can review the policy and decide whether to walk away or to proceed to interact with the site, and by so doing indicate their consent to that policy.

Rather than expecting users to review the privacy policy for each website that they visit, a P3P enabled web browser performs an automatic comparison of the user's recorded preferences with the website's policy, and only alerts the user if there is a mismatch.

With the work on DNT and its exception mechanism, this has taken a new turn. The Tracking Preference Expression Specification will contain an API to ask the user for his permission to personalize content and thus collect personal data. In Europe, the service needs some kind of consent to continue storing information client side. But how would one convince the user to accept the data collection, how do we reassure the user? Again, 22 pages of legalese haven't done the trick in the past and they won't do the trick here. The P3P statement vocabulary contains a lot of the semantics actually needed to have an internationalized interface to tell people what is collected and what the service intends to do with the collected data. This can be taken as a basis and extended to cater for the new needs. This paper tries to go first steps in that direction.

The PrimeLife Dashboard

With increasing public awareness of the amount of information being collected by websites, it seems timely to consider new approaches covering more than just cookies, whilst enabling a practical treatment of the user interface for expressing privacy preferences.

To investigate this, a Firefox extension was developed to look at the issues involved. This had to support:

- a. auto-generation of a human readable version of the policy
- b. automatic comparison of the user preferences with the policy
- c. automatic generation of a human readable report on any mismatches
- d. user interface for viewing and changing user preferences

The scope was taken as the data that websites can collect from HTTP request headers during a session. This includes the IP address, cookies, the user agent header, information on user preferences for language and data formats, the requested URL, the date and time of day, and more.

To simplify the user interface for preferences, a subset of P3P was chosen. This has the following object model:

- The URI for the site's full (human readable) policy
- The URI for instructions that users can follow to request or decline to have their data used for a particular purpose (optional)
- The name of the business responsible for the website
- The set of categories of collected data as defined by P3P 1.1
- The set of purposes collected data can be used for as defined by P3P 1.1
- The set of recipient types as defined by P3P 1.1
- The data retention policy type as defined by P3P 1.1

Note this uses P3P's data categories rather than the taxonomy of data items. This was found to be a much better fit to the needs for describing the kinds of data collected from HTTP requests.

The simple object model allows the preferences user interface to be provided as a set of grouped checkboxes, as shown below:



Generic Privacy Preferences

Adjust the following choices to your taste:

The following types information may be collected about me:

- ☒ Name, address, phone number, or other physical contact information
- ☒ Email address or other online contact information
- ☒ Website login IDs and other identifiers (excluding government IDs and financial account numbers)
- ☒ Information about your purchases, including payment methods
- ☒ Financial information such as accounts, balances, and transaction history
- ☒ Information about the computer you are using, such as its hardware, software, or Internet address
- ☒ Which pages you visited on this web site and how long you stayed at each page
- ☒ Activities you engaged in at this web site, such as your searches and transactions
- ☒ Information about social and economic categories that might apply to you, such as your gender, age, income, or where you are from
- ☒ Messages you send to us or post on this site, such as email, bulletin board postings, or chat room conversations
- ☒ Cookies and mechanisms that perform similar functions
- ☒ Which groups you might be a member of such as religious organizations, trade unions, and political parties
- ☒ Health information such as information about your medical condition or your interest in

Cancel

Update

Accessing the policy and generating a human readable version

To reach a website, the user can type in a URL, follow a bookmark, or follow a link from another site, e.g. on the results page from query on a search engine like Google. The browser extension intercepts the Firefox location change event and cancels the HTTP request before it is sent. The extension then sends an HTTP HEAD request to the website's root. The response is examined to find a reference to the site's generic privacy policy. This is represented as an HTTP Link header (analogous to the HTML link element), e.g.

```
Link: <http://localhost/w3c/policy.json>;
      rel="http://primelife.eu/generic-privacy-policy"
```

This header is easy to add to pages generated via PHP. The URI for the policy is then dereferenced to obtain the policy itself. Note P3P 1.0 defined a P3P HTTP header rather than using the generic Link header. This is something that could be considered if and when this work is brought into the standards track.

The object model for policies is decoupled from the on-the-wire transfer format, but from a practical point of view it was easiest to implement the transfer format with JSON [3]. Here is an example policy in JSON:

```
{
  "fullURI": null,
  "optURI": null,
  "name": "ACME widgets online inc.",
  "purposes": [ "current", "admin", "tailoring", "individual-analysis" ],
  "recipients": [ "ours", "delivery", "same" ],
  "retention": "business-practices",
  "categories": [ "computer", "navigation", "interactive" ]
}
```

Generating a human readable version of the privacy policy

The P3P 1.1 specification includes suggested text for each element in the taxonomy. This was copied into JavaScript and used to generate a human readable version of the policy. Here is an example:

PrimeLife Policy Preferences Editor

Generic Privacy Policy for ACME widgets online inc.

We may collect the following types of information about you:

- Information about the computer you are using, such as its hardware, software, or Internet address
- Which pages you visited on this web site and how long you stayed at each page
- Activities you engaged in at this web site, such as your searches and transactions

The ways your information may be used:

- To provide the service you requested
- To perform web site and system administration
- To customize the site for your current visit only
- To do research and analysis that uses information about you

With whom we may share your information:

- Companies that help us fulfill your requests (for example, shipping a product to you), but these companies must not use your information for any other purpose
- Delivery companies that help us fulfill your requests and who may also use your information in other ways
- Companies that have privacy policies similar to ours

How long we may keep your information:

- Our full privacy policy explains how long we keep your information

Further information

- Link to full privacy policy not provided

Return to view of preferences warning

The same text was also used for constructing a dialog summarising the mismatch between the user's preferences and the website's policy, for example:



Generic Privacy Preferences Warning

Each time your browser sends a request to a website, some information is disclosed in the HTTP headers. This includes your browser's external IP address, information about your browser and operating system, and your language preferences. The IP address may provide information about your location and your identity.

This website's generic policy conflicts with your preferences in the following ways:

1. The website says it would like to use your information
 - *To do research and analysis that uses information about you*
2. The website says it would like to share your information with
 - *Companies that have privacy policies similar to ours*
3. The website says it would like to keep your information for a longer period
 - *Our full privacy policy explains how long we keep your information*
4. The website doesn't provide a link to its full privacy policy

Please select between the following actions:

Cancel load

Load page this time

Always load this page

Edit preferences

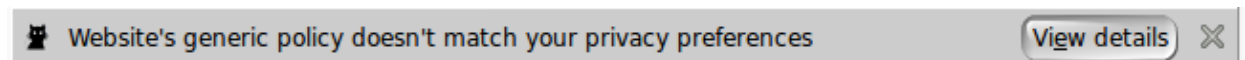
View policy

If the site's policy matched the user's preferences, or the user decided to override the mismatch, the browser extension then proceeds to relaunch the HTTP request for the original URL.

The Firefox notification bar is shown when a site is found to lack a privacy policy.



The Firefox notification bar is shown when a mismatch is found.



Clicking "View details" brings up the warning dialog shown earlier.

A local SQLite database was used to capture the user's preferences, and to cache the policy for sites as a performance optimization.

Anonymising Proxies

The act of making an HTTP HEAD request on a website's root discloses the browser's external IP address. This can be avoided by routing the request through an HTTP proxy. This could be configured via a user preference.

Summary and suggestions for further work

This paper has described a fresh take on P3P that goes beyond the limitations of compact policies, whilst still enabling a simple user interface for setting preferences. The object model lends itself to the use of JSON as a policy transfer format. The restricted semantics for a machine readable policy covering data collected in HTTP requests, is supplemented by a link to the site's full human readable policy. The proposal starts to think about how to integrate the ontology created by P3P statement vocabulary into the HTML5 and javascript interactions to allow for higher transparency and suggests further work in that area.

A further consideration is the privacy policy for other kinds of personal information collected by websites, for example, credentials coupled to a user's public or partial identity. Can the P3P taxonomies be extended to support these?

P3P and the approach described in this paper are couched in legal terms relevant to the obligations extended by websites to their users. Websites also have the challenge of operationalizing privacy policies when it comes to controlling access and usages of personal data in the website's backend. This suggests the need for transforming privacy policies into data handling policies. The PrimeLife project is looking at extending the XACML access control language to cover data handling policies, see H5.3.2 [4].

Widespread support for machine readable privacy policies is likely to involve a legislative mandate with measures in place to ensure that sites conform to the policies they disclose. However, this would only apply to the countries with the corresponding laws. A way is needed to allow the browser to verify the jurisdiction a given website is subject to. This could take the form of digital certificates issued by national agencies.

A separate issue is many people aren't sufficiently motivated to set privacy preferences. One reason is the desire to just get to the website in question without having to bother with reviewing the policy. Another is a lack of knowledge sufficient for an informed decision. This points the way to the use of independent third parties for help with setting privacy preferences, and for monitoring the data handling practices of websites. Some progress has been made with the latter in terms of a browser extension (Privacy Dashboard) that tracks what information is collected by the websites you visit, together with a means to set your preferences on a site by site basis [5].

Further reading

- [1] <http://www.w3.org/TR/2002/REC-P3P-20020416/>
- [2] <http://www.w3.org/TR/2006/NOTE-P3P11-20061113/>
- [3] <http://www.json.org/>
- [4] <http://www.primelife.eu/results/documents/activity-5-policies>
- [5] <http://www.primelife.eu/results/opensource/76-dashboard>