Web Site Privacy with P3P®

Helena Lindskog
Stefan Lindskog
Web Site Privacy
with P3P®
Web Site Privacy with P3P®

Helena Lindskog
Stefan Lindskog
To our wonderful children Caroline, Sofia, David and Johanna
## Contents

**Acknowledgments** xiii

**Chapter 1  Introduction to Privacy** 1
- Privacy Awareness 1
- The Right to Be Left Alone 4
- Means for Privacy
  - Anonymization 6
  - Pseudonymity 7
  - Unlinkability 8
  - Unobservability 9
- The Origin Server Perspective 10
- When the Origin Server Meets the Privacy-Aware User 11
- Platform for Privacy Preferences 13
- Trust 14
- What’s Ahead 14

**Chapter 2  Internet Security** 17
- Terminology and Definitions 17
- Vulnerabilities and Threats 18
  - Vulnerabilities 18
  - Threats 20
- Security Policy 21
- Protection Mechanisms 21
  - Authentication Systems 22
  - Access Control 22
  - Cryptographic Systems 23
    - Usage of Cryptographic Systems 23
  - Auditing 26
  - Firewalls 26
  - Intrusion Detection Systems 27
  - Anti-Malware Software 27
  - Vulnerability Scanners 28
- Summary 28
- Additional Reading 28
Chapter 7  Five Steps to Creating a Privacy Policy 77
Step 1—Create a Written Privacy Policy for the Site 77
Step 2—Decide Which Policies Apply to Which Pages 78
Step 3—Create P3P Policies 79
Step 4—Create a P3P Policy Reference File 82
Step 5—Validate the Policies 83
Summary 84
Additional Reading 84

Chapter 8  Privacy Policy in English 85
Information in an Online Privacy Policy 85
What Type of Personal Data Is Collected? 86
Why Is Personal Data Collected? 87
How Is the Collected Personal Data Used? 88
Is the Collected Personal Data Redistributed or Shared with Other Organizations? 88
How Is Collected Personal Data Protected? 89
How Can I Access My Own Personal Data? 90
Whom Do I Contact with Questions about the Privacy Policy? 90
Summary 90

Chapter 9  Privacy Policy Using P3P 91
Create a Reference File 91
Create the Policy File 94
Create the Surrounding Tags 94
Entity Information 95
Access Information 95
Disputes 97
Statements 99
The DATA-GROUP 100
The CONSEQUENCE Tag 109
The NONIDENTIFIABLE Tag 109
The PURPOSE Tag 110
The RECIPIENT Tag 114
The RETENTION Tag 116
P3P and Multiple-Language Support 117
The Legal Perspective on Policies 119
Summary 120
Additional Reading 120
Chapter 10  Cookies and P3P  
   Cookies Revisited  
   P3P and Cookies  
   Cookie Filtering  
   Full Policies  
   Create a Reference File  
   Create the Policy File  
   Compact Policies  
   Example of a Third-party Cookie That Uses a  
   Compact Policy  
   The Implications of the Three-Letter Combinations  
   Legislation  
   Cookie Alternatives  
   Session Cookies  
   Ad hoc sessions  
   Login sessions  
   Persistent Cookies  
   Cookie Policy Receipts—A Suggestion  
   Summary  
   Additional Reading  

Chapter 11  User-Agents and Other P3P Tools  
   Policies  
   Client Side  
   Viewers  
   User-Agents  
   Intermediary Servers  
   P3P for Data Conveyance  
   Existing P3P Tools  
   Validators  
   User-Agents and Viewers  
   The JRC User-Agent  
   Summary  
   Additional Reading  

Chapter 12  P3P and the Mobile Internet  
   Mobile Internet—The Vision  
   Mobile Internet Architecture  
   The User Device  
   The Proxy/Gateway  
   The Service Provider
Appendix A  An XML Tutorial  183

Background  183
General XML Concepts  184
  Tags and Attributes  184
  Document Type Definition (DTD)  186
  Prolog  187
  Processing Instructions  188
  Namespaces  188
Other Related Recommendations  189
Additional Reading  190

Appendix B  Taking Charge of Profile Information Conveyance  191

Introduction  191
Privacy  192
Mobile Internet Architecture and Services  194
  WAP 1.2.1 Architecture  194
  WAP 2.0  194
  Personal Trusted Devices  195
  User-Agent Profiles  196
  Location-Based Services  197
  Context-Aware Services  197
Privacy Risks  198
  Risks Factors  198
  Exposed Data  199
  Spam  200
Privacy-Enhancing Technologies  201
  Basic Concepts  201
  Platform for Privacy Preference Project (P3P)  201
    P3P Agreement  201
    A P3P User Agent  202
The PiMI Prototype  202
  Minimal Profile Conveyance  202
  Overview  203
  Results and Suggestions  205
  Use Case  205
For very valuable advising and proofreading, we would like to thank: Andreas Ljunggren, Mikael Nilsson, Giles Hogben, Jörgen Sigvardsson, Johan Hjelm, Magnus Johnard, and Fredric Palmgren.

*Stephen Kenny*
privacy in Internet environments is not only about technology and legislation, but also about straightforwardness, in both directions. It is about making sure that sites are up-front with the people using their services, and it is about providing the end user with the choice of whether to share information with the origin server.

This book is about how to use technology in the service of end users.

Privacy Awareness

The struggle for respect for privacy has always been an ongoing battle. In 1890, after the press had published some personal information about the Warren family, Louis Brandeis and Samuel Warren wrote “The Right to Privacy,” where they defined privacy as a fundamental right of the individual that should not be tampered with.

Our opinions of and feelings toward privacy vary for a number of factors. The first is our age. Our respect for authorities varies with the time we were born. People born in the 1940s and 1950s appear to have less respect for authorities than do people born in the 1920s and 1930s; those of us born in the 1960s grew up in shock over the controversial behavior of grown-ups.
The rights of “normal” individuals are one thing—what about the rights of those not like us? Lobotomy was used until the end of the 1970s in most countries, and sterilization of the mentally challenged is still discussed. However slowly, a common understanding of the rights of people who are different from ourselves in some aspect is developing over the world, and the younger we are, the more natural this understanding is to us.

Another interesting difference is how the new generation, which never knew a world without email and SMS, views privacy. Our two daughters would never dream of chatting using their own names. They each have a personal email address that uses the family domain, which they use with their close friends, and one public address, where they have given themselves nicknames like skate_girl_160@hotmail.com. How the concept of pseudonymization got into their minds, we do not know, but for some reason we think the next generation will claim their right to privacy with a lot more enthusiasm than the rest of us.

A second factor that makes us different from each other is our cultural background and nationality. In Sweden, we have a concept of openness that influences the way we all think. We have not participated in a war since the beginning of the nineteenth century, so military supervision has not been a routine part of our lives. According to Swedish law, the salary of any citizen who has a job in the public sector, such as a school-teacher, is public. The salaries are printed in a book that anyone can buy. On the other hand, due to the lack of military supervision, there is little chance that privacy laws are broken for the sake of public safety. In countries that have been exposed to severe threats, like Germany during the 1950s to 1970s and the United States in 2001, the understanding that police and military authorities may have to bend the rights of individuals to stop real danger is much greater.

In some countries there is distrust of those same authorities. Those who live in countries that have suppressed their people, or at least some groups of their people, have a greater suspicion of their government’s data collection activities. For instance, many Germans may have experienced first-hand crimes committed by the Nazi as well as the East German regimes.

Our cultural background, our experiences, and those of our parents and grandparents thus affect our views of privacy. What about religion and gender? Do Christian women have a different view of privacy than

1 Short Message Service - text messages received to your mobile device.
Muslim men? Well, of course. Our social status is also a factor. Even if we find two Finnish women, aged 43, both of whom had fathers who fought in World War II, are Protestants, and have the same income, they may well feel very different from each other.

This brings us to one other basic definition of privacy, made by Alan Westin in 1967:

*Privacy is the claim of individuals, groups, and institutions to determine for themselves how, when, and to what extent information about them is communicated to others.*

In other words, privacy is about self-determination. As individuals, we need to decide for ourselves what we want. We do not want to be treated as a group of tourists pushed around Versailles on the assumption that we all wish to see the bed of Louis XIV instead of having a closer look at the paintings of Marie Antoinette.

We may have very good reasons for wanting privacy. We may be trying to avoid divorce attorneys or the Mafia. We may have no more specific reasons than simply to exercise our right to be left alone. Note that this would not be a chapter about privacy if this factor was not mentioned: We may also be criminals trying to hide illegal deeds. Regardless of whether this illegal deed is distribution of copyrighted material, burglary, or terrorism, the fact remains: Providing individuals with rights, such as freedom of speech, the right to an attorney, or the right to privacy, might be advantageous to a criminal. A phone can be used for criminal purposes. So can a car, a computer, and the Internet.

We are not going to dwell on the issue of privacy versus crime for long because many scientific papers have been written on the subject, but there are two things we would like you to bear in mind. The first is that if we really want to send a message from, say, Australia to Seattle using the Internet, without anybody noticing, there are about 50 ways to do it, and none of them involves the use of P3P. The second is that our feelings about privacy may well depend on all the factors we previously stated.

Even if we, as individuals, decide that we do not think this privacy stuff is very interesting, and even if we do not care if people want to collect our IP numbers, there is still the business perspective. Valued customers who visit our business Web sites care about privacy, and will abandon our sites if we do not show that we care as well. We should respect the rights of individuals to determine for themselves whether they want private information communicated to others. If you want
your customer’s respect, you should convey the following message to your customer:

“I, your local friendly service provider, respect your right to self-determination.”

The Right to Be Left Alone

Have you ever received a spam email? Our first experience with it was a few years ago when our then nine-year-old daughter’s mailbox was filled with ads for sex sites. We were completely outraged. We tried to put a stop to it any way we could, including contacting the ISP and, silly enough, sending “unsubscribe” messages to those sending the email. We know better now, of course. Sending an unsubscribe message has the opposite effect because it proves to the originator of the offending message that there is someone receiving it at the other end.

There are three kinds of privacy, all of which have bearing on the use of Internet:

- **Personal privacy against moral offense**, meaning that we should not be exposed to information that offends our moral senses
- **Territorial privacy against trespassing**, meaning that people should not trespass on our property
- **Informational privacy against gossip**, meaning that we decide about data that can be tied to us as individuals

When our daughter received these messages, it was a violation to all three kinds of privacy. First, someone obviously had collected her email address and sold it, implying disrespect for her informational privacy. Second, they sent email messages that she did not want, showing disrespect for her territorial privacy because it is fair to call our inboxes our territory. After all, they are part of our computers. We should decide for ourselves what should end up there. Third, the links that were sent contained offensive content—fortunately she was too young to follow them—and were thus an intrusion to her personal privacy.

When we discuss privacy issues, we should keep these differences in mind. There are ways to handle all three, but in this book, we focus on informational privacy, which is, in most cases, the starting point for most personal and territorial intrusions. Surveys have shown that exposure to unsolicited email and SMS messages is the greatest privacy
concern for most people. Thus, this message should also be conveyed to your customers:

“I will not share identifiable information about you with others.”

**Means for Privacy**

In order to create applications for the privacy-aware user, we need to think like that user. With that in mind, it is now time to introduce our users: Hans and Greta. Hans is the careless sort of person who has never given privacy two thoughts. He will freely give out information, without ever considering that it could be used for something else. He will just shake his head at the fourth Viagra spam he got this week and delete it. Greta wants to stay away from the spotlight, and if she thinks that a service will store or process her information for purposes other than what she intended, she will not use it. She would rather walk to the ATM to get cash for groceries than use her credit and bonus cards where she runs the risk of having information about her purchase stored somewhere for the future.

This book is dedicated to Greta. So why do we need Hans? Even though a user like Hans has no interest in maintaining privacy, he must still have access to all the services he wishes. When we design sites for the privacy-aware user, like Greta, we must consider the ease of use for users like Hans as well, in addition to the legislation of the country where that lives.

Greta wants to go to a Web site that sells the particular book she wants to purchase. She does not want anyone to know what she reads or buys. Let’s take a look at her options.

There are four kinds of privacy-enhancing technologies (PETs) defined by the Common Criteria\(^2\) group:

1. Anonymization
2. Pseudonymization
3. Unlinkability
4. Unobservability

---

\(^2\) The Common Criteria represents the outcome of a series of efforts to develop criteria for evaluation of IT security that are broadly useful within the international community. See [http://www.commoncriteri.org/](http://www.commoncriteri.org/).
Anonymization

The Common Criteria (CC) definition of anonymity is as follows:

*Anonymity requires that other users or subjects are unable to determine the identity of a user bound to a subject or operation.*

Anonymization, as illustrated in Figure 1.1, implies that the destination will not know who Greta is at all. If Greta uses a technology of this kind, she will be completely anonymous to the origin server.

What is interesting to note, though, is that even though people want anonymity, few are willing to pay for it. In 1998, a Canadian company called ZeroKnowledge did an implementation of the Chaum white papers, which introduced a technology called mix nets that allows anonymous transmissions to take place. It called the service the Freedom Network. In October 2001, the network was closed down—mainly for financial reasons.

The end user will or might have to pay for anonymity in several ways:

- The cost that the company will charge Greta to use the service
- The cost of overhead when it comes to deciding which service to use and why it is good—such as whether to trust the ISP’s services or to select a third party
- The cost of round-trip times—that is, the number of extra seconds that Greta will have to wait to get to the requested URL
There is another not-so-irrelevant problem in the particular case of Greta wanting to buy books. She can be anonymous when surfing around looking for them, but how will she pay for them without having her identity revealed? Well, it is possible. There are technologies for digital cash transmissions that are based on the same ideas as cash cards, but they have three drawbacks. One is that they are complex and rarely used. Another is that they are not secure—that is, money can be lost, just as you can lose money by misplacing your wallet, which means that the concept should be used only for smaller amounts. The main reason, though, is that banks may use these transactions to preserve data about which user issued which check. All this means the following:

- A lot of overhead
- Difficulty and uncertainty when it comes to getting back lost money
- A willingness to trust the bank with the customer’s privacy

Let’s suppose that Greta wants anonymity so badly that she will still take the overhead of using an anonymous routing protocol to reach the origin server and will use digital cash for the payment. How is the vendor supposed to ship the goods to her if she does not want to give away her identity?

**Pseudonymity**

The CC definition says the following:

_Pseudonymity requires that a set of users and/or subjects are unable to determine the identity of a user bound to a subject or operation, but that this user is still accountable for its actions._

The general idea of pseudonymization is that you use an identity that cannot be tied to you as a person—that is, an alias or a nickname. It could also be a serial number. Pseudonyms are used, for example, when providing location data from a telecom operator to a service provider; see Appendix D. There are several ways of doing this, but in general the location is passed on with a number that will identify the user throughout the session.

Thus, the pseudonym is a recurring ID that is indirectly tied to the user. The matter of accountability is relevant according to the CC definition;
the pseudonym is reversible, if it is really necessary for reasons of crime or nonrepudiation. This obviously does not include the case where the user pseudonymizes herself (Figure 1.2). Cases can fall outside the scope of the CC definition, and we would prefer lifting the issue of accountability out of the definition.

Pseudonymization can also take place at the receiving side. Pseudonymizing databases to be able to store data longer than allowed by law is a common method, used when building statistical databases, for example.

The pseudonymity concept is what teenagers use when they use weird Hotmail addresses to communicate with people they do not know. Pseudonymity also occurs when you create an account with your favorite Internet chess club, calling yourself “Bobby Fischer II” or when you play Quake as “The Mega Wizard.” It is often used in statistical databases. Origin servers that want to keep data for statistical purposes can pseudonymize this data before storing it.

**Unlinkability**

The third category defined by the CC says the following:

*Unlinkability requires that users and/or subjects are unable to determine whether the same user caused certain specific operations in the system.*
Figure 1.3  Unlinkability.

In this particular case, the user is part of a group that has access to the service, but no one can determine which one in the group actually performed the specific operation. You can see that the group that Greta is part of sent the request to the server, but not that she was the one who did it. To benefit from this, Greta would need to join a network of users that collectively use a privacy protocol that implies that no one can determine from which particular person in the network the action took place.

**Unobservability**

Finally, the last category of privacy-enhancing technologies is unobservability. The CC definition says the following:

> Unobservability requires that users and/or subjects cannot determine whether an operation is being performed.

This is about getting the message through without anybody noticing that a message was even transmitted. A common method is steganography, where a message is hidden inside something else. This could be a picture, downloaded from the Web, an audio file, or even a simple text message. If you were the kind of kid who liked playing detective or spy, you have probably tried the last category. You often see this kind of technology used in Hollywood movies, where a character receives an email message with no meaning, then the letters suddenly change places, and a secret message is shown.
Unobservability is about hiding from the rest of the world, not the receiver, which is why it has no bearing on Greta’s particular problem.

The Origin Server Perspective

We are still stuck with a problem. Greta wants access to a service. Somewhere there is an origin server that wants to provide her with this service, but neither one can reach the other. The reason why they cannot is simple: lack of trust. If Greta had trusted services on the Internet in the first place, there would not be a problem.

Let’s take a look at the bookstore, which we will call werespectyou.com. This is a nice little company that tries hard to meet the needs and wishes of its customers. At this company work two Web designers named Harry and Sally, who have learned in the past few years that personalization is important. Thus, they provide users like Hans with a number of features to improve his experience when he visits their site. They remember the books he has bought and send him email messages in which they state that because he bought a book about monkeys last time he was there, he might be interested in a new, particularly interesting book about gorillas. He has made settings in which he stated that he is interested in sports and music; thus, they can tailor the site when he reaches it to make sure that books concerning these subjects are the first he sees whenever he logs on. They have placed cookies on his client, so that he barely even needs to log on. Any time he wants to buy something from the site, his name and address are automatically filled out.
for him on the order form. In order to do this, werespectyou.com has stored a lot of data about Hans. This data is also used for statistical purposes so that the company can build customer profiles and make strategic decisions about what books to have in stock in the future.

werespectyou.com also has a system administrator named Karen who has a hard time chasing hackers out of the system. She keeps log files of all the activities, so if Trudy the intruder breaks in or makes an attempt to, Karen will be ready to catch her. This activity takes place in the basement of the company, where the servers are located. When Karen first installed the Web server, she used the standard logging facilities, but as the company has grown, the number of attempted intrusions has grown to several a week, and now she has created several Perl scripts that log everything. Those log files are sometimes cleaned, but some of them remain on DAT tapes in a closet. That way, she knows that she can retrieve those files to track an intruder if she comes back.

In essence, werespectyou.com has the following needs:

- Collect data about the buyer, in order to fulfill the request
- Collect data about all users to create statistical reports
- Keep log files about activities in order to maintain an acceptable level of security
- Profile the user, in order to enhance the user’s experience at the Web site

When the Origin Server Meets the Privacy-Aware User

Three extreme categories of people reach werespectyou.com: Hans, Greta, and Trudy the intruder. Karen handles Trudy. Harry and Sally have already created customized applications for Hans. Now it is time to take care of Greta’s needs. Before we do that, we need to take a closer look at which one of the previously listed needs of the company is absolutely necessary.

It is necessary to collect data about the buyer in order to fulfill the request, but is it necessary to store this data? Well, it is useful but not necessary. If werespectyou.com wants Greta as a customer, it should consider retaining Greta’s data only at the time of the purchase, then erasing it.

Statistics are important; however, there are ways to store data that is anonymized or at least pseudonymized that can fulfill that purpose.