Prywatność z protokołem P3P w transakcjach online

Igor Margasiński, Krzysztof Szczypiorki
Instytut Telekomunikacji PW
e-mail: igor@margasinski.com, krzysztof@szczypiorski.com

Streszczenie

W referacie podjęto dyskusję na temat perspektyw użycia nowego protokołu – P3P (*Platform for Privacy Preferences*) – będącego obecnie jednym z najważniejszych przedsięwzięć standaryzacyjnych z dziedziny prywatności w systemie WWW oraz w handlu elektronicznym. Przeprowadzono analizę możliwości i ograniczeń płynących z tworzenia automatycznych polityk w oparciu o język XML (*Extensible Markup Language*). Zaprezentowano zarys działania systemu P3P oraz dokonano konfrontacji celów osiąganych przez P3P z obecnymi oczekiwaniami w zakresie prywatności online. Poza nakreśleniem możliwych obszarów zastosowań i praktycznych korzyści, dokonano krytycznej analizy i rozważono, czy protokół opracowany i promowany przez W3C (*World Wide Web Consortium*) może być skuteczną odpowiedzią na obecne zagrożenia prywatności. Wskazano, że P3P nie rozwiązuje wszystkich obaw związanych z prywatnością WWW – w szczególności nie może samodzielnie zapewnić prywatności.

(...) I know what is wrong,
And I know what is right.
And I'd die for the truth
In My Secret Life.

Leonard Cohen - *In My Secret Life*

1. Wprowadzenie

W szybkim tempie powiększają się obszary życia, w których podstawą działania staje się wymiana **informacji**. Dla jednostek ludzkich są to między innymi: nauka, poszukiwanie znajomości, produktów, kontakty towarzyskie, zakupy, komunikacja z bankiem... Globalnie, współczesna gospodarka już dziś opiera się często wprost na **wiedzy**. Te przemiany społeczne, często szumnie określane jako początek nowej epoki cywilizacyjnej – formowania społeczeństwa informacyjnego – oznaczają, że informacja staje się podstawowym towarem.

Powszechność komunikacji poprzez Internet oznacza, że wiele wartości, w tym między innymi tzw. zacisze domowe, nie spełnia już należycie nadanej im roli ochrony prywatności. Prywatność, będąca podstawowym prawem każdego człowieka, wymaga obecnie nowego postrzegania – potrzebna jest weryfikacja prywatności z uwagi na nowe środowiska i obecne tam nowe zagrożenia. Potrzebne jest także poszukiwanie nowych środków ochrony prywatności.

W zależności od obiektu, który posługuje się danymi informacjami, informacje te mogą mieć różne znaczenie i wagę. Dla firm marketingowych, badających upodobania konsumentów, informacja jest cennym towarem. W obliczu terroryzmu informacja staje się potężnym narzędziem, które wraz ze skutecznym zarządzaniem wiedzą może być wykorzystane w różnych intencjach, a w skrajnych przypadkach może służyć do prowadzenia wojny informacyjnej.

Na początku roku 2001 świat obiegła wiadomość o luce w systemie teleinformatycznym jednej z firm zajmujących się handlem danymi [25]. Organizacja, której obroty ze sprzedaży danych osobowych i profili firmom ubezpieczeniowym oraz agencji FBI w roku 2000 wyniosły prawie 600 milionów dolarów, udostępniała przez kilka tygodni zawartość swoich baz danych każdej osobie potrafiącej posługiwać się przeglądarką internetową. Ujawniona luka stała się pretekstem do dyskusji o dopuszczalności i granicach takiej działalności. Przy okazji okazało się, że np. pewna kobieta została zwolniona z pracy, ponieważ pracodawca otrzymał informację o niechlubnej przeszłości pracownicy, rzekomo okradającej sklepy i skazanej za handel środkami odurzającymi. Nigdy nienotowana kobieta stała się ofiarą błędu silnego narzędzia. Jeden ze specjalistów, badający bezpieczeństwo systemu, odkrył także, że według analizowanych baz danych zmarł w roku 1976.

1.1. Furtki do prywatności

Przeglądając strony internetowe nie pozostajemy anonimowi [21]. Zdarzenia odwiedzin oraz poszczególne **kroki nawigacji są rejestrowane przez serwery WWW**, z którymi następuje połączenie. Do typowych danych gromadzonych przez serwery należą: adres IP klienta, czas nadejścia żądania, adres URI

przeglądanych zasobów, czas przesłania, nazwa klienta, informacje o błędach, informacje o przeglądarce użytkownika, oraz w przypadku, gdy przejście do strony nastąpiło przez odnośnik – adres URL poprzednio odwiedzanej strony (*refer link*). Wypełniając formularze przekazujemy serwerom WWW dalsze informacje.

Poszczególne odwołania użytkowników (w bezstanowym protokole HTTP [11], [5]) wiązane są ze sobą najczęściej za pomocą mechanizmu zarządzania stanem HTTP – *Cookies* [17]. Serwery zapisują w ten sposób, na stacji użytkownika, dowolne informacje (najczęściej kodowane w tylko sobie znany sposób), aby przy kolejnych odwołaniach móc je odczytywać. W ten sposób możliwe jest **rejestrowanie ścieżek nawigacji** poszczególnych użytkowników. Serwery mogą odczytywać tylko zapis *Cookies* dokonany przez siebie. Jednak zastosowanie *Cookies* może być znacznie silniejsze. Pomimo wspomnianego ograniczenia, możliwe jest globalne śledzenie nawigacji pomiędzy wieloma serwisami. Tu z pomocą przychodzą popularne bannery. W dziedzinie systemu WWW nośniki reklamy służą nie tylko do promowania określonych produktów – przekazywania użytkownikom informacji reklamowych i odnośników – ale również są przekaźnikiem informacji w drugim kierunku, czyli do organizacji umieszczającej bannery (takiej jak *DoubleClick*). Te informacje to wskazówki o zainteresowaniach użytkowników. Na stronach WWW możliwe jest osadzanie elementów pochodzących od trzecich stron. Firmy specjalizujące się w reklamie WWW często stosują bannery, opierając treść reklam na szczegółowych informacjach o internautach. W sieciach zrzeszających strony współpracujące ze sobą w tworzeniu profili użytkowników wyróżnić można centralny węzeł, z którego pochodzą bannery, czyli AdSerwer. Odwołując się do jednej ze stron z takiej sieci, w sposób niewidoczny, odwołujemy się również do AdServer'a poprzez pobranie bannera. Temu odwołaniu towarzyszy na ogół zapis i odczyt odpowiednio preparowanych *Cookies*. Dla osoby postronnej są to niezrozumiałe ciągi znaków. Dla AdSerwera to cenne informacje umożliwiające szczegółowe odtwarzanie ścieżek nawigacji poszczególnych użytkowników [20].

Obszerne zbiory danych poddawane są następnie **automatycznym procesom odkrywania wiedzy**. Przy przetwarzaniu danych o użytkownikach systemu WWW stosowane są głównie rozwiązania wywodzące się z teorii zbiorów rozmytych (*fuzzy sets*), sztucznych sieci neuronowych (*artificial neural networks*), algorytmów genetycznych (*genetic algorithms*) oraz z teorii zbiorów przybliżonych (*rough set theory*). Teoria zbiorów rozmytych określa zasady postępowania ze zjawiskiem niepewności. Sieci neuronowe są szeroko stosowane do modelowania złożonych funkcji, umożliwiają uczenie się oraz uogólnianie. Z kolei algorytmy genetyczne służą głównie do zapewniania wydajnego wyszukiwania i optymalizacji wyników, natomiast teoria zbiorów przybliżonych związana jest bezpośrednio z wydobywaniem wiedzy. W wydobywaniu wiedzy służącej do profilowania użytkowników systemu WWW, wyróżniane są na ogół trzy etapy: przetwarzanie wstępne, odkrywanie wzorców oraz analizowanie wzorców.

Faza przetwarzania wstępnego opiera się na przekształcaniu danych o aktywności, treści stron WWW oraz ich strukturze i wzajemnych powiązaniach, na formę abstraktu określającego ciąg działań dokonywanych przez użytkowników [8]. Następnie otrzymany ciąg zdarzeń dzielony jest na sesje. W przypadku, gdy użytkownik dokonuje zapytań do serwerów spoza sieci profilującej, trudne jest określenie, kiedy użytkownik opuścił daną stronę. Przyjmuje się wtedy umowną granicę trzydziestu minut jako czas bezczynności użytkownika, po upływie którego sesja uznawana jest za zakończoną. Bardziej inwazyjna metoda to zastosowanie mechanizmów rozszerzających standard HTML, takich jak JavaScript, czy ActiveX, do wymuszania komunikacji w określonych interwałach z serwerem WWW (np. sekwencyjne pobieranie plików graficznych).

Identyfikator sesji często zapisywany jest bezpośrednio w adresach URI (*Uniform Resource Identifier*) lub w *Cookies*. Na ogół na podstawie żądań klienta HTTP można wnioskować o przeglądanych treściach. Zdarza się jednak, że same adresy URI nie zawierają wszystkich potrzebnych informacji i konieczna jest dodatkowa komunikacja z serwerem WWW. Poważną przeszkodą jest również szeroko rozpowszechniony mechanizm schowków internetowych (funkcje *cache*'u). Wprowadza on zaburzenia w rejestracji aktywności użytkowników. Najskuteczniejszym rozwiązaniem jest tu również zastosowanie elementów wprowadzających dynamikę po stronie klienta, wspomagających śledzenie (np. *JavaScript*). Mniej skuteczne jest stosowanie mechanizmu weryfikacji pola ostatnio odwiedzanej strony (*refer link*) z plików logów.

Na etapie **odkrywania wzorców** stosowane są algorytmy znane z takich dziedzin jak statystyka matematyczna (*Statistics*), wydobywanie wiedzy (*Data Mining*), uczenie się maszyn (*Machine Learning*) i rozpoznawanie obrazów (*Pattern Recognition*).

Techniki statystyczne należą do najpopularniejszych metod wydobywania wiedzy o odwiedzinach serwerów WWW. W czasie badania danych pochodzących z poszczególnych sesji przeprowadzane są różne analizy statystyczne, wykorzystujące częstość (rozkład prawdopodobieństwa), średnią, wyznaczanie mediany

(wartości środkowej rozkładu), itp. Analizowanymi danymi są: odwiedzane strony, czas zapoznawania się ze stronami, czas i długość ścieżek nawigacji użytkowników. Wyniki tych działań pozwalają określić, które zasoby cieszą się popularnością poszczególnych użytkowników.

W celu ustalenia związków pomiędzy najczęściej odwiedzanymi stronami w obrębie sesji stosuje się algorytmy odkrywania reguł asocjacyjnych [1]. Dzięki zastosowaniu tych technik – często algorytmu *a priori* [2] lub jego wariantów – możliwe jest np. odnalezienie korelacji między tym, że użytkownik interesuje się hodowlą zwierząt, a faktem, że śledzi wydarzenia polityczne.

Kolejną stosowaną metodą jest klasyfikacja ([9], [29], [14]). Pozwala ona na przyporządkowanie profili użytkowników do poszczególnych klas i kategorii. Potrzebne jest tu dokonanie odpowiedniego wyboru czynników najlepiej opisujących poszczególne kategorie. Klasyfikacja realizowana jest zazwyczaj poprzez wykorzystanie mechanizmów uczenia indukcyjnego. Szczególnie przydatne okazują się algorytmy oparte na indukcji drzew decyzyjnych, naiwnym klasyfikatorze bayesowskim, k-najbliższym klasyfikatorze. Dzięki zastosowaniu klasyfikacji możemy dowiedzieć się, że np. ponad 40% użytkowników dokonujących zakupów produktów sportowych mieści się w grupie wiekowej 25-35 lat i pochodzi z dużych miast. W przypadku, gdy zachodzi potrzeba przewidzenia kolejnych kroków użytkownika stosowane są metody odkrywania wzorców ([3], [27]) oraz analiza zależności przyczynowych. W tym przypadku budowany jest model odzwierciedlający ważne zależności pomiędzy różnorodnymi zmiennymi. Istnieje kilka metod, które pozwalają na wyznaczenie modelu zwyczajów przeglądania stron. Techniki te to głównie ukryte modele Markowa i sieci bayesowskie.

Analiza wzorców stanowi ostatni krok w procesie odkrywania wiedzy o profilach użytkowników WWW. Celem tej fazy jest odrzucenie nieinteresujących wzorców z otrzymanego wcześniej wzoru. Osiągane jest to głównie poprzez mechanizmy zapytań (*knowledge query mechanism*) opartych na języku SQL lub poprzez zastosowanie działań OLAP (*Online Analytical Processing*). W efekcie uzyskiwane są profile użytkowników zawierające, poza danymi osobowymi, szczegółowe informacje o ich zainteresowaniach, zwyczajach, itp.

Łatwość w pełni zautomatyzowanego i globalnego pozyskiwania oraz przetwarzania danych o osobach korzystających z ogólnoświatowej sieci, to zjawisko odstraszające użytkowników od poważnych zastosowań Internetu. Według badań opinii publicznej [16] większość konsumentów nie ufa przedsiębiorstwom przetwarzającym ich prywatne dane. Użytkownicy oczekują rozwiązań zapewniających ochronę ich prywatności w sieci publicznej. Ponad połowa ankietowanych przyznaje, że dopiero w przypadku pewności, co do przestrzegania przez firmę polityki prywatności, poleciłaby ich usługi znajomym lub rodzinie. Ponad 90% konsumentów zaznacza, że korzystaliby częściej z zakupów internetowych, gdyby polityka prywatności sklepu była poddana odpowiedniej kontroli. Wyniki badań przeprowadzonych w Kanadzie [10] są jeszcze bardziej wyraziste. Około 83% użytkowników podało, że nie korzysta ze sklepów internetowych, ponieważ nie jest dla nich jasne, co dzieje się z ich prywatnymi danymi, komu i do jakich celów są przekazywane, oraz przez kogo ich zainteresowania i zwyczaje są obserwowane.

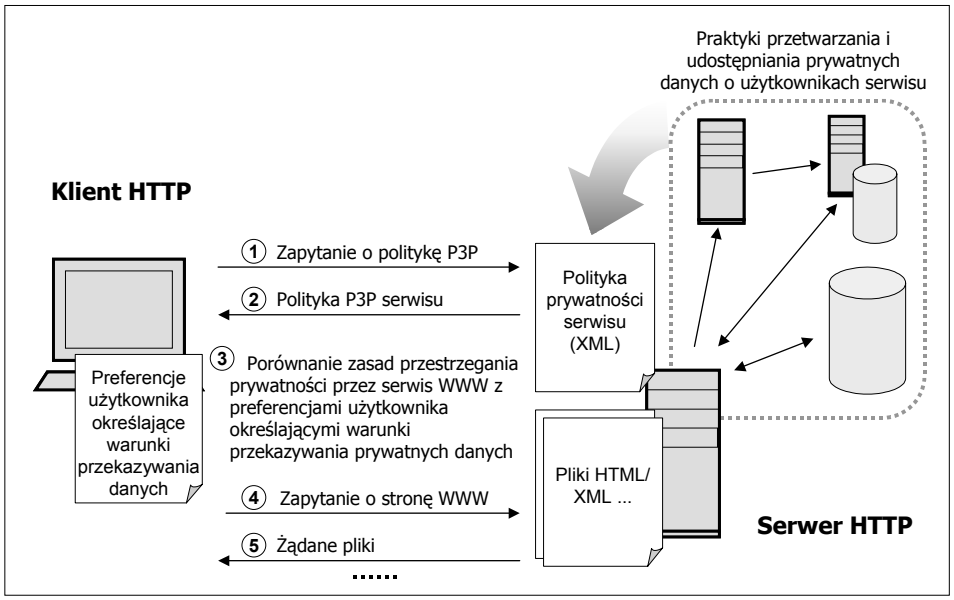
1.2. Zarys technologii P3P

Naruszenia prywatności internatów, agresywne i nie podlegające praktycznie żadnej kontroli, spowodowały narodziny różnego rodzaju inicjatyw i wpłynęły na tworzenie mechanizmów ochrony prywatności. Zgodnie z duchem otwartego rozwoju Internetu, w ostatnich latach spontanicznie pojawiło się szereg koncepcji i implementacji rozwiązań zapewniających prywatność oraz anonimowość [21]. Główne kierunki to wprowadzanie serwerów pośredniczących ([4], [20]), sieci wielu węzłów pośredniczących ([12], [13], [22], [24], [28]) w oparciu o ideę Mixnet-u Davida Chauma [7], sieci P2P zapewniające anonimowość (*Peer-to-Peer*) [26], a także aplikacje filtrujące zapis *Cookies* i niepożądane dodatki do stron WWW. Brak ochrony prywatności użytkowników sieci Internet i wynikająca z tego powszechność nadużyć oraz chaos panujący wśród dostępnych rozwiązań spowodowały, że podjęte zostały także prace standaryzacyjne, mające na celu regulację zasad przestrzegania prywatności. W dalszej części artykułu będziemy zajmować się właśnie tym oficjalnym kierunkiem.

W roku 1997 World Wide Web Consortium rozpoczęło prace nad projektem **Platformy Preferencji Prywatności – P3P – Platform for Privacy Preferences** [19]. W roku 2002 protokół P3P 1.0 ([18]) został przedstawiony w formie oficjalnych zaleceń W3C. Choć w większości przypadków nie zdajemy sobie z tego sprawy, już dziś na co dzień mamy do czynienia z P3P. Na przykład, przy domyślnych ustawieniach przeglądarki *Internet Explorer 6*, zapis *Cookies* od stron trzecich, nie wspierających podstawowych elementów P3P, jest odrzucany. Obecnie w technologię P3P wyposażonych jest ok. 2000 serwisów WWW. Trwają dalsze prace nad wersją 1.1 protokołu oraz opracowania koncepcji P3P 2.0.

Protokół P3P jest standardem zapisu i przekazywania polityk prywatności serwisów WWW w jednolity sposób. Polityka, w formie plików XML (*Extensible Markup Language*) [30], może być następnie automatycznie interpretowana przez przeglądarkę użytkownika. Z punktu widzenia użytkownika, agent P3P automatycznie pobiera i odczytuje zapis polityki danego serwisu WWW. Przeglądarka wyposażona w technologię P3P może zweryfikować politykę prywatności serwisu i poinformować użytkownika o praktykach tam opisanych (np. w języku narodowym użytkownika). Przeglądarka może również porównać politykę serwisu z preferencjami bezpieczeństwa, wybranymi przez użytkownika. Na tej podstawie automatycznie dokonywane są decyzje dotyczące dalszej komunikacji z serwerem. Agent P3P może być częścią składową przeglądarki, stanowić plug-in, może być zewnętrznym programem lub np. modułem elektronicznego portfela, czy niezależną aplikacją stosowaną przez dostawcę usług internetowych. Popularne przeglądarki posiadają już wbudowany moduł P3P. Rysunek 1 przedstawia poglądowy schemat przebiegu komunikacji w relacji przeglądarka – serwer WWW z zastosowaniem protokołu P3P.

.



Rysunek 1. Poglądowy schemat działania protokołu P3P

2. Polityka P3P

Polityka P3P serwerów WWW powinna odzwierciedlać rzeczywiste praktyki przetwarzania danych użytkowników. Technologia P3P nie zapewnia jednak, że warunek ten w rzeczywistości będzie spełniony. Przełożenie praktyk na standard P3P poprzedzane jest zazwyczaj dokumentacją określającą, jakie dane i do jakich celów są gromadzone. Wzorcowo dokumentacja oparta jest na przeprowadzonym wcześniej audycie stron internetowych firmy. Najważniejsze informacje zawarte w polityce P3P to:

- dane identyfikacyjne organizacji,
- identyfikator pliku lub zasobu,
- opis zasobu,
- informacje, czy dane są gromadzone przy pobieraniu zasobu przez użytkownika, a jeśli tak, to jakiego rodzaju są to dane (anonimowe, pseudoanonimowe czy dane osobowe),
- czy stosowany jest mechanizm *Cookies*, a jeśli tak, to o jakim identyfikatorze oraz od kogo pochodzi zapis,
- gdzie dane są przechowywane,
- do jakich celów dane są wykorzystywane,
- jakim stronom dane są przekazywane,
- czy użytkownik ma możliwość akceptowania praktyk (schematy *opt-in, opt-out*).

Dodatkowo, w przypadku, gdy serwis wykorzystuje mechanizm *Cookies* lub pośredniczy w jego dostarczaniu, w zapisie polityki P3P umieszczane są następujące informacje:

- nazwa (identyfikator) Cookie,
- opis,
- czy zapis jest usuwany po zakończeniu sesji,

- czy Cookie pochodzi bezpośrednio od docelowego serwera WWW, czy od trzeciej strony,
- jakie dane są gromadzone (anonimowe, pseudoanonimowe lub dane osobowe),
- czy dane są kiedykolwiek wiązane z danymi osobowymi,
- od jakiej strony pochodzi zapis,
- do jakich celów uzyskane dane są wykorzystywane,
- jakim stronom dane są udostępnianie,
- czy użytkownik ma możliwość akceptacji zapisu Cookie (schematy opt-in, opt-out).

Serwis WWW może posiadać oczywiście wiele polityk P3P. Poszczególne podstrony serwisu, różniące się podejmowanymi w ich obrębie praktykami, mogą mieć przyporządkowane różne polityki P3P. Liczba polityk P3P jest najczęściej wynikiem wyważenia, ponieważ zarówno duża ich liczba, jak i mała mają wady i zalety. Mała liczba polityk (w szczególności jedna) oznacza przyjęcie zasad bardziej ogólnych, które muszą pokrywać się z praktykami z zakresu większej liczby działów serwisu. Zaletą takiego podejścia jest prostota tworzenia i modyfikacji polityki P3P oraz mniejsze ryzyko związane z naruszeniem prawa, polegającym na błędnym przyporządkowaniu polityki. Wadą jest konieczność wyspecyfikowania polityk o charakterze bardziej inwazyjnym w odniesieniu nawet do stron gdzie takie praktyki nie występują. Jest tak, np. gdy jedną polityką objęte są strony o zróżnicowanych praktykach przetwarzania prywatnych danych. Może to działać odstraszająco na użytkowników. Duża liczba polityk oznacza stworzenie oddzielnych plików P3P do większej liczby działów serwisu. Jest to podejście umożliwiające bardziej szczegółowe sprecyzowanie podejmowanych praktyk w zależności od poszczególnych obszarów serwisu. Zaletą tego rozwiązania jest umożliwienie użytkownikowi zapoznania się z praktykami serwisu w sposób bardziej dokładny. Może to oznaczać, że użytkownicy przywiązujący większą wagę do prywatności nie będą unikać szerszego zakresu serwisu - w szczególności całej witryny. Wadą jest komplikacja i praktyczne trudności w implementacji i aktualizacji dużej liczby polityk.

2.1. Struktura polityki P3P

Serwery WWW posiadające pojedynczą politykę prywatności P3P udostępniają dwa pliki XML: **plik odwołań** (*Policy Reference File*) oraz **plik polityki**.

Plik odwołań to pierwszy plik, jaki przeglądarka internetowa wyposażona w technologię P3P odczytuje. Plik ten stanowi mapę określającą, gdzie znajduje się plik (bądź pliki) polityki prywatności. Zawiera również informacje o skojarzeniach poszczególnych polityk prywatności z zasobami, do których się odnoszą, a także wskazanie okresu ważności. Zasobami, które identyfikowane są poprzez adresy URI, mogą być całe katalogi, pojedyncze strony WWW, pliki Cookie, itp. Pliki PRF mogą być umieszczane na trzy sposoby:

- tzw. dobrze znana lokalizacja (*well-known location*) – umieszczenie pliku w katalogu: /w3c i nazwanie go p3p.xml – rozwiązanie najprostsze,
- poprzez nagłówek HTTP (*HTTP header approach*) – w nagłówku odpowiedzi serwera HTTP zawierana jest informacja o umiejscowieniu pliku PRF,
- poprzez znaczniki HTML (*HTML link tag approach*) - każda ze stron WWW posiadająca politykę P3P zawiera odpowiedni znacznik HTML stanowiący odnośnik do pliku PRF.

Plik odwołań PRF wskazuje, gdzie umieszczona jest polityka prywatności. Postać pliku może być następująca:

```
<?xml version="1.0"?>
<POLICIES xmlns="http://www.w3.org/2002/01/P3Pv1">
  <!-- Expiry information for this policy -->
  <EXPIRY max-age="86400"/>
  <POLICY
    xml:lang="pl">
    <ENTITY>
         <DATA-GROUP>
         </DATA-GROUP>
    </ENTITY>
    <ACCESS><nonident/></ACCESS>
    <STATEMENT>
      <EXTENSION optional="yes">
         <GROUP-INFO
         xmlns="http://www.software.ibm.com/P3P/editor/extension-1.0.html"
         name="Access log information"/>
      </EXTENSION>
    <CONSEQUENCE>
         Our Web server collects access logs containing this information.
    </CONSEQUENCE>
    <PURPOSE>
         <admin/><current/><develop/>
```

```
    </PURPOSE>
    <RECIPIENT><ours/></RECIPIENT>
    <RETENTION><indefinitely/></RETENTION>
    <DATA-GROUP>
         <DATA ref="#dynamic.clickstream"/>
         <DATA ref="#dynamic.http"/>
    </DATA-GROUP>
  </STATEMENT>
</POLICY>
</POLICIES>
```

Do tworzenie poprawnych plików P3P można obecnie zastosować gotowe narzędzia generujące politykę P3P na podstawie parametrów wprowadzanych przez użytkownika. Na uwagę zasługuje bezpłatne narzędzie *IBM P3P Policy Editor*. Ułatwia ono stworzenie szczegółowej polityki, a także automatyczne przetłumaczenie jej do postaci kompaktowej (patrz punkt 2.3 *Skrócona forma zapisu*). Dodatkowo możliwe jest sprawdzenie poprawności sporządzonej polityki za pomocą narzędzia *W3C's P3P Policy Validator*, udostępnianego na stronach *World Wide Web Consortium*.

2.2. Najważniejsze elementy

Polityka zapisana w formacie P3P zawiera przede wszystkim następujące elementy:

- **dane identyfikacyjne** organizacji (*entity*) – adres oraz inne informacje potrzebne do nawiązania kontaktu z organizacją,
- **politykę prywatności** (*policy*) – podstawowe informacje o pojedynczej polityce prywatności takie jak nazwa (*name*), adres URI odpowiedniej polityki w języku naturalnym (*discuri*), adres pod którym podane są instrukcje w jaki sposób użytkownik może wyrazić zgodę lub niezgodę na praktyki zdefiniowane w polityce (*opturi*) w schemacie *opt-in* lub *opt-out*,
- zasady **dostępu** (*access*) – dostęp użytkowników do ich danych osobowych (np. umożliwia poprawienie danych adresowych),
- **sprawy sporne** (*disputes*) – informacje o nadrzędnych zasadach lub o innych prawach, którym podlegają praktyki serwisu,
- **sposoby rekompensaty** (*remedies*) – zawarte w elemencie *disputes* i określające sposoby rekompensaty w przypadku naruszenia zasad zdefiniowanych w polityce,
- **cele** gromadzenia danych (*purpose*),
- zbiór **odbiorców** (*recipient*),
- warunki przechowywania danych (*retention*).

Warto zwrócić uwagę na trzy ostatnie pozycje, za pomocą których głównie różnicowany jest stopień inwazyjności polityki. Możliwe **cele** gromadzenia danych podzielono na 11 kategorii: od najmniej inwazyjnych (takich jak przeprowadzenie procesu, do jakiego zostały dostarczone – kategoria *current*, czy potrzeb administracji serwerem – kategoria *admin*) do celów gromadzenia informacji, które mogą być wykorzystywane do określania zwyczajów lub zainteresowań odwiedzających. Najbardziej inwazyjne kategorie zawierają informację o możliwości łączenia profili użytkowników z danymi osobowymi, w celach bezpośredniego wpływania na te osoby (np. kategoria: *individual-decision*). Możliwe jest również podanie innej kategorii. Element **odbiorcy** został podzielenie na 6 kategorii. Poza docelowym serwerem WWW (kategoria: *ours*), strony trzecie zostały zróżnicowane względem tego jak postępują z otrzymanymi danymi. W znaczniku *retention* – **warunki przechowywania** możliwe jest podanie jednej z 5 kategorii. Np. podana w przykładzie kategoria *indefinitely* – brak polityki przechowywania – oznacza, że informacje przechowywane są przez czas nieokreślony. Inny przykład to kategoria *business-practices* informująca, że organizacja przechowująca dane w celach marketingowych musi określić termin ich niszczenia. Przedstawione kategorie sprawdzają się przy zastosowaniu do typowych funkcji serwisów WWW. W niektórych specyficznych przypadkach jednak mogą okazać się niewystarczające.

2.3. Skrócona forma zapisu

Standard P3P wprowadza również możliwość zapisu polityki prywatności serwisu w formie skróconej, czyli tzw. polityki kompaktowej. Choć jest to element opcjonalny, większość agentów obecnie bazuje, w znacznym stopniu właśnie na polityce kompaktowej. Złożoność opisu, jaki możemy zastosować w polityce kompaktowej jest mniejsza i dotyczy jedynie praktyk związanych z zastosowaniem mechanizmu *Cookies*. Polityka w formie kompaktowej to ciąg trzyliterowych słów oddzielonych spacjami. Dostępne słowa podzielone zostały na kategorie. Politykę w formie kompaktowej przekazuje serwer klientowi w nagłówku HTTP, np. po zapytaniu klienta o stronę abc.html:

```
GET /abc.html HTTP/1.1
Host: np_jakis_serwer.pl
Accept: */* Accept-Language: en, pl
User-Agent: WonderBrowser/1.0
```

serwer może poinformować klienta, w skróconej formie, o polityce prywatności dotyczącej tej strony, poprzez odpowiedź:

```
HTTP/1.1 200 OK
P3P: policyref="http://np_jakis_serwer.pl/P3P/PolicyReferences.xml",
   CP="NON DSP ADM DEV PSD OUR IND STP PHY PRE"
Content-Type: text/html
Content-Length: 1234
Server: WonderServer/1.0

...dane...
```

Każdy z trzyliterowych skrótów odzwierciedla, zgodnie ze specyfikacją P3P, odpowiednie praktyki. Wymienienie takiego skrótu w polityce kompaktowej oznacza zadeklarowanie określonej praktyki przetwarzania danych. Dostępne skróty przyporządkowane są do poszczególnych elementów polityki P3P, takich jak np.: cele przetwarzania danych, czy zbiór odbiorców. Prześledźmy, co oznacza przedstawiony w przykładzie zapis (Tabela 1):

Dostęp	**NON** brak		
Sprawy sporne	**DSP** sprawy sporne są określone w pełnej polityce P3P		
Rekompensata			
Brak identyfikacji			
Cele	**ADM** do administracji systemem i serwerem WWW	**DEV** do badań i rozwoju	**PSD** do celów marketingowych w oparciu o pseudo-anonimową identyfikację
Odbiorcy	**OUR** odbiorcą jest wyłącznie firma		
Warunki przechowywania	**IND** wraz z informacjami identyfikującymi	**STP** dla celów stanowych	
Kategorie	**PHY** informacje o potrzebne do nawiązania fizycznej komunikacji	**PRE** informacje o preferencjach	

Tabela 1. Interpretacja przykładowej polityki kompaktowej P3P

Fakt opierania działania przeglądarek głównie na polityce kompaktowej, ograniczającej deklarację polityki do stosowania mechanizmu *Cookies*, oznacza, że obecnie głównie w tym zakresie praktycznie realizowany jest protokół P3P. Pełna polityka prywatności jest na ogół tylko prezentowana użytkownikowi. Istnieją jednak już odpowiednie *plug-iny* umożliwiające korzystanie z protokołu w pełni.

3. Perspektywy

Poważna luka w systemie WWW związana z brakiem mechanizmów ochrony prywatności spowodowała, że oczekiwania dotyczące pierwszych prac standaryzacyjnych z tej dziedziny są wysokie. Wraz z pojawieniem się wiadomości o powstawaniu protokołu P3P, niejednokrotnie dochodziło do poważnej krytyki rozwiązania. Pojawiło się wiele opinii, że P3P jest protokołem słabym i nieskutecznym w zapewnianiu prywatności. Nieporozumienia wynikały głównie z niezrozumienia zamierzonej przez twórców roli protokołu. Protokół P3P realizuje dwa główne cele. Po pierwsze umożliwia serwerom WWW zaprezentowanie swoich praktyk przetwarzania danych użytkowników w standaryzowany sposób. Po drugie umożliwia użytkownikom odwiedzającym serwisy WWW poznanie deklaracji: jakie dane o nich będą gromadzone, jak będą użyte oraz w jaki sposób użytkownik może wyrazić zgodę lub niezgodę na poszczególne działania. W praktyce okazuje się, że

cele te są na ogół skutecznie realizowane, choć w niektórych zastosowaniach sztywno zdefiniowane kategorie okazują się niewystarczające. Co jest jednak zaskakujące, sami twórcy P3P przyznają, że nie należy zaliczać protokołu do technologii rozszerzających prywatność (PET – *Privacy Enhancing Technologies*). P3P to platforma preferencji prywatności, czyli platforma pozwalająca na zaprezentowanie komunikującym się ze sobą stronom swoich preferencji z zakresu prywatności. Zatem rozwiązanie należy traktować jako pomocniczy element w technikach zapewniających prywatność. Należy pamiętać, że sam protokół P3P w żaden sposób nie podwyższa bezpieczeństwa i pełni jedynie rolę informacyjną. Zgubne jest więc opieranie ochrony prywatności jedynie na P3P. Co jest również istotne, protokół P3P nie zawiera mechanizmów pozwalających na stwierdzenie, czy praktyki deklarowane w polityce P3P są zgodne z rzeczywistością. Po wprowadzeniu do jednej z najpopularniejszych przeglądarek internetowych domyślnych ustawień wymagających polityk kompaktowych P3P do akceptacji *Cookies* trzecich stron, wiele serwisów zastosowało technologie P3P z konieczności. W sieci Internet można obecnie odnaleźć wiele gotowych *recept* konfiguracji P3P, które *obchodzą* nowy sposób działania przeglądarek i pozwalają na działanie serwisu w praktycznie niezmieniony sposób.

Rozwój mechanizmów rozszerzających prywatność w systemie WWW to włączanie technologii P3P jako jednego z komponentów – platformy komunikacyjnej – do pełnego rozwiązania, zapewniającego prywatność. Rozwój technologii P3P to wprowadzanie mechanizmów kontroli korelacji rzeczywistych praktyk z deklarowanymi. Proponowane obecnie rozwiązania to systemy zarządzania prywatnością (*Privacy Manager*), składające się z dwóch podstawowych komponentów:

- serwer – definiujący politykę prywatności, przypisujący politykę do poszczególnych zasobów, przygotowujący procedury audytu oraz zapewniający narzędzia raportujące,
- monitory – będące elementami pośrednimi pomiędzy serwerem systemu zarządzania prywatnością, a środowiskiem aplikacyjnym.

Systemy zarządzania prywatnością mogą znacznie podwyższyć jakość praktyk serwisów. Mogą okazać się narzędziem znacznie upraszczającym implementacje i przestrzeganie określonej polityki. Z punku widzenia użytkownika nadal nie pozwalają na zyskanie pewności co do rzeczywistego losu udostępnianych świadomie i nieświadomie danych. Semantyka polityki P3P nie zawsze może odzwierciedlić rzeczywiste praktyki. Kategoryzacja praktyk ma *płaską* strukturę. Brakuje również możliwości wprowadzania wyrażeń warunkowych. Wydaje się jednak, że najważniejszym problemem pozostaje wiarygodność publikowanych polityk i ich związek z rzeczywistością. Nieuczciwa organizacja może posługiwać się przez pewien okres polityką prywatności P3P, informującą o nieinwazyjnych praktykach. Następnie może usunąć zapis P3P i wyprzeć się wcześniejszych deklaracji. Użytkownicy, którzy udostępnili swoje prywatne dane takiej organizacji, pozostają bezbronni. Obecnie rozważa się zastosowanie technik podpisów cyfrowych oraz certyfikatów X.509 do zapewnienia uwierzytelnienia i niezaprzeczalności. Być może sprawi to, że protokół P3P zacznie być stosowany z przekonaniem.

4. Podsumowanie

Protokół P3P jest odpowiedzią na obawy internautów dotyczące gromadzenia i przetwarzania ich prywatnych danych przez serwisy WWW. Technologia nie ogranicza bezpośrednio nadużyć, ale pozwala na przekazanie praktyk serwisów użytkownikom odwiedzającym strony WWW. Ujednolicona postać zapisu polityki prywatności jest dobrym narzędziem regulującym sposób informowania konsumentów o ustalonych zasadach przetwarzania ich danych. P3P to narzędzie pozwalające na ustalenie kompromisu pomiędzy ekonomicznymi potrzebami firm, a prawem użytkowników do prywatności i kontrolowania danych o sobie w sieci Internet. W przypadkach, gdy strony gromadzące dane nie są wiarygodne, rozwiązanie okazuje się mało przydatne, a zaufanie do protokołu może być szkodliwe.

5. Literatura

[1] Agrawal, R., Imielinski, T., Swami A. Mining Association Rules Between Sets of Items in Large Databases. Materiały: ACM SIGMOD Conference on Management of Data, Washington DC, USA, May 1993

[2] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. Materiały: 20th VLDB Conference, str. 487-499, Santiago, Chile, 1994

[3] Agrawal, R., Srikant, R. Mining Sequential Patterns. Materiały: 11th Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995

[4] Anonymizer (`http://anonymizer.com`).

[5] Berners-Lee, T., Fielding, R., Frystyk, H. Hypertext Transfer Protocol – HTTP/1.0. RFC 1945, 1996

[6] Catledge L.D., Pitkow J.E. Characterizing Browsing Strategies in the World Wide Web. Materiały: 3rd Int'l World Wide Web Conference, 1995

[7] Chaum, D. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. Communications of the ACM Vol. 24 no 2, str. 84 – 88, 1981

[8] Cooley, R., Mobasher, B., Srivastava, J. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. Materiały: 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), Newport Beach, California, November 1997
[9] Fayyad, U., Piatetsky-Shapiro, G., Smyth P. From data mining to knowledge discovery: An overview. Materiały: ACM KDD, 1994
[10] Ferneyhough, C. Online Security and Privacy Concerns on the Increase in Canada, Ipsos-Reid, 2001
[11] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee T. HyperText Transfer Protocol – HTTP/1.1. RFC 2616, 1999
[12] Goldberg, I., Shostack, A. Freedom Network 1.0 Architecture and Protocols. Zero-Knowledge Systems. White Paper 1999
[13] Goldschlag, D. M., Reed, M. G., Syverson, P. F. Onion Routing for Anonymous and Private Internet Connections. Communications of the ACM Vol. 42 no 2 , str. 39-41, 1999
[14] Hartigan, J. Clustering Algorithms. John Wiley, 1975
[15] Jamtgaard, L. P3P Implementation Guide. The Internet Education Foundation, 2003
[16] Krane, D., Light, L., Gravitch D. Privacy On and Off the Internet: What Consumers Want. Harris Interactive (2002)
[17] Kristol, D., Montulli L. HTTP State Management Mechanism. RFC 2965, October 2000
[18] Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marchall, M. The Platform for Preferences 1.0 (P3P 1.0) Specification, W3C Recommendation, April 2002
[19] Cranor, L. Web Privacy with P3P. O'Reilly & Associates, September 2002.
[20] Margasiński, I. Zapewnianie anonimowości przy przeglądaniu stron WWW. Materiały: 18. Krajowe Sympozjum Telekomunikacji – KST'2002, Bydgoszcz, 2002
[21] Margasiński, I., Szczypiorski, K. Web Privacy: an Essential Part of Electronic Commerce. Materiały: 3rd International Interdisciplinary Conference on Electronic Commerce "ECOM-03", Gdańsk, 2003, str. 65-72
[22] Margasiński, I., Szczypiorski, K.: VAST: Versatile Anonymous System for Web Users. Materiały: The Tenth International Multi-Conference on Advanced Computer Systems ACS'2003, Międzyzdroje, 2003
[23] Presler-Marshall, M. The Platform for Privacy Preferences 1.0 Deployment Guide, W3C Note, May 2001.
[24] Reiter, M.K., Rubin, A.D.: Crowds: Anonymity for Web Transactions. ACM Transactions on Information and System Security, str. 66-92, 1998
[25] Scheeres, J. What They (Don't) Know About You. Wired, May 2001
[26] Six/Four System (`http://www.hacktivismo.com/projects`)
[27] Srikant, R., Agrawal, R. Mining Sequential Patterns: Generalizations and Performance Improvements. Materiały: 5th Int'l Conference on Extending Database Technology (EDBT), Avignon, France, March 1996.
[28] Syverson, P. F., Goldschlag, D. M., Reed, M. G.: Anonymous Connections and Onion Routing. IEEE Symposium on Security and Privacy, 1998
[29] Weiss, S.M., Kulikowski, C.A. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann, San Mateo, CA, 1991
[30] Yergeau, F., Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. Extensible Markup Language (XML) 1.0 (Third Edition), W3C Recommendation February 2004

Artykuł recenzowany