Privacy protection in government mashups

Janice Warner\textsuperscript{a,*} and Soon Ae Chun\textsuperscript{b}
\textsuperscript{a}Georgian Court University, 900 Lakewood Ave, Lakewood, NJ 08701, USA
\textsuperscript{b}City University of New York, College of Staten Island, Staten Island, NY 10304, USA

Abstract. The Web 2.0 technologies allow dynamic content creation using syndications or mashups, extracted from diverse data sources, including government enterprise data. As a primary source of citizen data, the US government has the obligation not only to make public data available for citizen access as stated in the Freedom of Information Act, but also to protect the privacy of individual citizen’s records as stated in the Privacy Act. In a mashup, a third party mashup Web application provider requests the individual’s data from the government agencies through Web services. Since the data is public data and not necessarily provided through electronic interactions, individual citizens may not be able to express fine-grained privacy policies on how data may be used. In addition, the government agency’s privacy policy is very coarse grained, and the relative sensitivity of individual information is not considered. We discuss the opportunities and issues associated with the programmable web and mashups, provide a Privacy Protection Model for Mashup Applications, using a mashup related multi-dimensional privacy protection space and present policy recommendations to complement the technological solutions. The model and recommendations include deployment of a personal privacy policy network, a distributed system over which citizens can publish their individual privacy policies. These policies are accessible by all web service providers to be consulted in real time by data providers including government agencies for the purposes of automated privacy protection reasoning concerning data release.

Keywords: E-government mashup services, citizen privacy, regulatory privacy policy, personal privacy policy network, Web 2.0

1. Introduction

Mashups, a term originated with the pop music industry, is the use of a combination of web services to create a new application. It is used creatively by amateurs and experts alike using publicly available data and other resources available on the Web. Business enterprises have begun to use the technique to meet their needs. We envision that governments will begin doing so as well. Currently, some public governmental data is available via application programming interfaces (APIs). Our research addresses how the privacy of individuals whose information is part of the public record could be protected while not discouraging the general development of mashup applications or precluding the advantages of an open, programmable web. We begin by looking at the privacy landscape in general and then the mashup opportunity and how it impacts upon privacy.

1.1. Privacy landscape in today’s networked society

Privacy was identified as a fundamental human right in the Universal Declaration of Human Rights adopted by the United Nations in 1999. However, the definition of privacy is still unclear. It has been identified as a fundamental issue of concern for Internet users but surveys of their behavior have shown that they will readily provide private information if they perceive a benefit in doing so [7]. We believe...
that a user should have the right to control the way personally identifiable information (PII) is obtained, processed, distributed, shared and used by any other entity. To that effect, several technological solutions have been defined and deployed. Most famous is P3P [8], a protocol for exchanging structured data about privacy policies and preferences. However, having a mechanism in place does not guarantee that it will be used properly or that the policies will be enforced. Therefore, governmental policies and legislation are needed if privacy is really to be preserved.

The E-Government Act of 2002 sets requirements on government agencies regarding how they assure the privacy of personal information in government information systems. It requires that US federal websites include a privacy notice that states what information is to be collected, why it is being collected, its intended use, what notice or opportunities for consent are available to individuals regarding what is to be collected and how it will be shared, and how it will be secured. An example of privacy policy can be seen at the Oregon state e-government site.\(^1\) This information must be available on websites in machine-readable format. Other legislation (P.L 109–115) signed into law at the end of 2005 prohibits federal websites from collecting data about visitors to those websites [28].

To date, privacy solutions have typically addressed collection of data by web site operators from electronic visitors to these web sites. Some believe that is too narrow a concern. A study by Tufekci [26] showed that although users were concerned about privacy, they wanted to control their privacy through the use of nicknames and visibility, not limit the release of information. Thus, they did not generally limit the personal information (besides telephone numbers) disclosed. They saw value in releasing and sharing information. The results were in line with Altman’s theory of privacy [7] as an optimization between competing pressures for disclosure and withdrawal. In addition, the people in the study were not well aware of or were unconcerned by persistence of data, the later use or discovery of information about an individual not envisioned at the time of submission. Based on the study, Tufekci believes that we have come back full circle to the grassroots village surveillance situation where everyone potentially knows everyone else’s business and past history. What needs to be built are mechanisms for building boundaries between what is kept private and what is public through different levels of visibility both over time and space. This has become more difficult because once information is released, an individual has lost some control over the audience who will have access to that information. As we will show, this is exactly the concern associated with mashups because the data used in the service is not typically directly collected by the mashup provider.

While we are generally concerned with privacy issues, we are particularly concerned with data revealed in unexpected circumstances and more specifically when information is revealed through the combination of data from a variety of sources. This mashup privacy problem is discussed in Section 1.2. While we view the ability to create mashups as an opportunity, enabling a more flexible and useful World Wide Web, we believe limits need to be placed on usage of PII in mashups. We present technical safeguards in Section 3. However, since technological safeguards are almost never sufficient, we discuss policy considerations in Section 4. Finally, we conclude in Section 5 which includes a discussion of our continuing work and a summary of recommendations.

1.2. Mashup opportunities and risks

Mashups are new content created by blending or mashing of data (i.e. application results, video, data, and pictures) from two or more sources using Web services available on the Internet. The term “mash-up” originally came from the music industry relating to the combination of various sound tracks. Since

around 2005, it has appeared as the term used when Web services are combined initially for entertainment purposes. Its use in more formal undertakings was described in an article in the journal “Nature” [29]. A mashup is often created through accessing to one or more Application Programming Interfaces (APIs). These interfaces contain basic information on how to electronically use Web services. Mashups have extended the utility of data, allowing enormous variations. Previously, the data produced and posted by a Web application were “locked” in a data owner’s site and the users were passive “information sinks” [12]. In order to get the Web data or content in a form usable by another application, data consumers had to develop special purpose programs called “screen scrapers” which process web pages. Today, because of Web 2.0 and APIs, mashup editors allow source data to be dynamically available through Web services in real time. Given the variety of content available on the Web, one can mix-and-match multiple data sources in unexpected ways to suit countless purposes. Specifically, mashups and the interfaces and programs developed to support them facilitate:

- Dynamic content composition across organizational or national boundaries,
- Combining of disparate data sources including data, pictures, video and sound, and
- Creativity of presentation.

Most current mashups typically combine results from one or more databases with a map service. For example, maps overlayed with crime data can be created such that crime data taken from a corrections database is combined with a map application such as Google maps via use of address information that is transferred from one to the other. Besides map mashups, there are photo mashups, as well as shopping, video, game, crime, sports, celebrity and government mashups just to name a few. There are close to 4000 mashup links on http://www.programmableweb.com/mashups/directory as of January 2009. Mashups are also being used in business. An IBM case study [11] describes the combination of weather forecasts, google maps and the inventory data used to tweak inventory levels for products whose sale depends upon the weather.

Mashups are easy to develop by end users. Several general purpose mashup editors exist including Yahoo! Pipes, Microsoft Popfly, and Google Mashup Editor. They automate the connection of data sources and tools for operations such as filtering and searching. In addition, drag and drop Web Widgets (or Gadgets) allow portable chunks of code to be embedded and executed within any Web page. For example, several Real Simple Syndication (RSS) feeds are used to create new content using Yahoo!Pipes. Figure 1 shows the basic mashup architecture.

Public data is readily available through local, state, county or provincial and federal or national government sources in many countries. Types of data include census data, labor data, tax data, commerce data, transportation data, local law enforcement data, and many more. For example, the British government recently exposed through APIs public, but not personal, data including neighborhood statistics, transportation, education, environment and health care data.  

Government mashups can benefit constituents by rapidly developing and disseminating combined information that may result in new knowledge of benefit to them. For instance, information on government legislation on gun control can be pulled from many levels of government Web services and combined with news feeds to give users a complete view of the issue.

On the negative side, however, the ease of creating mashups can potentially result in the revelation of more information than one might like, creating possible breaches of individual privacy. One example [15, 24] envisions a mashup of data Amazon.com wishlists used to create a profile of a person’s interests.

\[^2\]http://www.showusabetterway.co.uk/call/data.html.
Parts of the profile, however, could be embarrassing to the individual or cause suspicion. For example, the profile might indicate that the person has an interest in subversive books, which might trigger a watch list by government agencies. The mashup described aggregated the data and plotted the exact location of the profiled people on a map.

Another example of existing governmental mashups is the consolidation and presentation of detailed information about members of the current US Congress. It may be acceptable to reveal detailed information about such public figures which is arguably the price paid for office. However, if similar mashups were created for access to data on any individual citizen, there would be serious privacy implications.

Availability of private information on-line may be exacerbated over time as more e-government initiatives are introduced. These initiatives make data or content available through Web services for the purpose of citizen convenience. Just like on-line shopping, it is more convenient for many constituents to use a government Web service rather than to visit a government office. Some of the current Web services include car registration, on-line application for social services such as food stamps, and on-line information on a variety of topics.

Some data made available on-line are not collected via the Web. Public records are increasingly being made available on-line. These records include court proceedings, bankruptcies, real estate records, occupational licenses, security and exchange commission filings, political contributions, births and deaths. Several third party on-line companies, such as Intelius provide fee-based mashup services that consolidate queries made to the various public record databases provided by state governments in the
form of a background check. Others use the publicly available data to provide a mashup such as a 
mapping of convicted sex offenders to neighborhoods. ³

The development of mashups using governmental public data could be a threat to privacy for the people 
on whom data is collected (i.e., anyone resident in the municipality, state or country) who makes any 
transaction that is considered part of the public record. The basic question of whether making public 
records available over the Internet is a violation of privacy is not addressed here. Instead, we consider 
only the issues associated with releasing data collected by government agencies, whether it be public 
record data or other data, for the purposes of allowing others to create mashups. Our main concern is 
the right of individuals to control information about themselves. Besides privacy, a government entity 
needs to be concerned with integrity, not only of their data, but also of other data sources.

In an e-government mashup, data is pulled from government sources, mixed with other data sources 
and manipulated to provide alternate viewpoints of the data which could include important discoveries 
about the population covered by the data set. The mashup provider might be a government agency or 
organization or any other web service provider. For data that is part of the public record and available 
on-line, third party mashup providers can combine the public data with other data whether or not a user 
would like his or her data to be manipulated or not. An example of a third party mashup that uses 
both governmental and public sources of data is in www.openmass.org. It monitors and presents in an 
integrated fashion various sources of data about what is happening in the Massachusetts state legislature. 
Currently, there is no safeguard on public record information. The US Office of Management and 
Budget (OMB) has recently published APIs for access to the federal contracts and assistance databases 
at www.fedspending.org. Rhode Island was the first state government to offer an API to its databases at 
www.govTracker.org. Sunlight labs⁴ provide an API to a database containing data about Congress and 
it converts IDs so that data about the same person from different sources can be mapped. It requires 
authentication via registered application ID.

Governments could choose to make it harder or easier to query and retrieve electronic information 
from the public record. For the protection of those whose data it is, it should not be easy to pull personal 
data from a government service even if the data is public domain data, except for purposes deemed to be 
of benefit to the public at large. Thus the API to the public domain data could require as much effort as 
going to the seat of government and asking to see a public record. Controlled access interfaces, however, 
could be established for organizations that promise to use the data in a manner in line with public policy 
and the privacy desires of the people associated with the data.

Government agencies might also want to push data from one-government service to another to link 
access to services. This can help coordinate national and local programs as well as external sources 
of aid and benefits and provide one-stop services which have been identified as a high priority for 
users. For example, mashups could be used to coordinate relief in response to an emergency. Based 
on user or neighborhood characteristics collected and made available through a Web Service, various 
relief organizations could be pushed data on the situation. Coordination of response could thus be 
automated. This would also benefit individual citizens. A mashup service will allow citizens to access 
a one-stop service on a portal, alleviating them from interacting with each individual agency separately, 
by transaction service compositions. In [22], an example of a one-stop service for a pregnant mother 
who needs food stamps, foster care and health care is provided.

⁴http://services.sunlightlabs.com/api/.
Portals provide some of the functionality necessary. In particular, at www.GovBenefits.org, users can fill out a survey and then get a list of services that might help them. However, it is anonymous and the users must then investigate each of the services themselves. Random attempts by the authors often gave rise to a list of over thirty potential services. The time required to investigate and apply for each is not insignificant. Mashups could provide easier access to services by allowing users not only to search for but to automatically apply for those that match their characteristics.

The main concern with mashups is that data should not be pushed to services in which the citizen is not interested, or to which they do not want to disclose private information. For example, the social security number used in one transaction service for a government agency should not be pushed to another service (whether it is another government or commercial service) that may cause damages in privacy or other areas. In other words, the dataflow between internal mashup processes should be controlled to adhere to the privacy requirements.

Our view of mashup services is broad. They consist of web-service access to data from multiple on-line sources. This data may have been collected on or off the Internet itself. It can be designed for public or closed group consumption. We believe mashups will provide constituents with many benefits including more efficient access to services. The availability of data will also help in answering important research questions for science, health, and social policy. However, knowledge as power can have a dark side if it is used maliciously to profile individuals and deny them services or discriminate against them. Because of the negative potential, individuals should have some control over personally identifiable information (PII). This control needs technology components to automate the setting, monitoring and exchanging of policies and supported by regulation and laws as incentive to implement and support the technology components.

2. Technological safeguards for privacy protecting mashups

Our privacy protection model protects PII associated with data from government sources mashed up with any other data. It provides fine grained access control based on content type, recipient, and the recipient’s intended use of the data. Data that government agencies could release but should protect include those that represent the public record and those collected outside the public record. Such data might be captured when a citizen tries to obtain services such as education or health benefits. As discussed in the introduction, data in the public record is not protected from being seen but could be protected from easy access for purposes that might be undesirable for the individual. Therefore, the privacy protection model proposed here governs an interface to data under the jurisdiction of a government body, irrespective of whether it is in the public record or not. We build on those approaches proposed in [4], concerning policy-based fine-grained access control for protecting multi-level classified information. Like the platform for privacy protection, P3P [3], our model provides a mechanism for policies to be specified and exchanged. However, we additionally propose enforcement and reasoning mechanisms so that compliance can be verified.

2.1. Privacy policy specification

A government agency has to observe privacy policies in conformance with legal mandates such as the Privacy Act [1] or agency-specific rules. These are referred to as Regulatory Privacy Policies (RPP). Privacy preferences, specified by individuals and called Personal Privacy Policies (PPP), are published to repositories and made accessible using Uniform Resource Identifiers (URIs) to the data providers. Of
Fig. 2. Overall Interaction among Mashup providers, Government data providers and citizens to protect the private data used in mashup services.

Concern are data providers who have information deemed to be of a private nature. This would include name, identification number, financial information and/or any other information that an individual may want to keep private from the general public. Data providers would consult the URIs on behalf of their constituents when making a decision concerning release of data to a mashup provider. For an E-government mashup service, a government agency as a data provider would take into consideration both the Regulatory Privacy Policies (RPP) as well as the Personal Privacy Policies (PPP), providing fine-grained privacy protection. In addition, the use of PPP allows for notification to the data subjects of policy violations by third party mashup providers. Finally, the PPP repositories can be used to make the mashup provider aware of the personal privacy policies that impact the implementation of its service with an indication of why a specific data access is denied.

Mash-up providers are also expected to have published their own privacy policies, called Mashup Privacy Policies (MPP). Data providers can consult these policies before releasing data. These policies contain an indication of the purpose of the mashup. Data providers could choose not to release data for use in mashup services when the purpose is not acceptable to the individuals whose privacy is being protected. Figure 2 shows the overall interactions among mashup providers, government data providers, and citizens.

PPP, RPP, and MPP are uniformly specified based on five aspects of the data and mashup service: DataTypes (DT), Linking Parameters (LP), Operations (OP), ProviderType (PT) and MashupPurpose (MP). The uniformity enables a common enforcement component be applied to all policy types. A privacy protection space is formed by combining specific values specified for these five parameters. It defines the dataset that either is allowed to be used or is prohibited from use in a mashup. When a parameter is specified with a “+” sign, it means that only the values listed are covered by the protection space. When a parameter is specified with a “−” sign, it means that all values other than those listed are covered by the protection space. Upon protection space instances, constraints are placed authorizing or
restricting release of data when the parameters are met. Individuals may have multiple protection spaces thus allowing private data to be shared under certain protection spaces and not in others.

2.1.1. Data types

The data types, DT, describe the individual personal data items that should be protected. They are one or more uniform resource identifiers (URIs) which reference resource descriptions written in XML using the resource description framework (RDF). For example, we use The Friend of a Friend (FOAF) schema that describe people and things with associated objects (i.e., images) forming decentralized linked information system. The FOAF documents contain not only the basic information about people, organizations, or groups, such as names, addresses, but also how they are related to things, such as “attends meetings” or “is depicted with photos” etc.

A DT to be protected for privacy is specified with a tuple \( (E, URI, Sign) \) where \( E \) represents the protected element, \( URI \) denotes the optional resource URIs where the definition of \( E \) can be located, and \( Sign \) to show the inclusion or exclusion in the protection space. For example, a person’s name would be designated as protected using the following DT definition: \( DT = (\text{foaf:name}[\text{rdf:resource=http://xmlns.com/foaf/0.1/#term name}] \text{ sign} = “+”) \). When a DT is specified, all equivalent DTs and all subclasses of the DT are also included in the protection space.

2.1.2. Linking parameters

The Linking Parameters, LP, are data types which may or may not be used in mashup services as the key data type, i.e. uniquely identifying piece of information, on which data from multiple sources is combined. As data types, they are also described by URIs using publicly available schemas or ontologies. The main purpose of including LP in the protection space is to ensure the integrity of the combination of data linked from different sources. For example, a person with a very common name such as “Lee” might be concerned if their name were used as the key to link data from different sources. The results might include data that is not really associated with them. Therefore, they might want to prevent any mashups created from data linked on the basis of a last name. Thus, they would set up a protection space where \( LP = (\text{foaf:name}[\text{rdf:resource=http://xmlns.com/foaf/0.1/#term surname}] \text{ sign} = “−”) \).

2.1.3. Operations

Mashup Operations, OP, are those that the mashup providers may perform on a set of data types linked from multiple sources. These include Combine, an operation that takes data from multiple sources and displays it together; Represent that takes data from multiple sources and displays a representation of it; the Overlay operation that links data and displays them on top of each other such as via a map or graph; the Sequence operation that links data from different sources in a temporal sequence; and the Personalize operation that links data with content in which a person is expected to be interested based on an analysis of the data values.

2.1.4. Provider type

Provider Type, PT, is a tricky aspect to define for protection spaces. It is included as a proxy for “trust” in the provider to use the data only for stated purposes. This is similar to how trust is implemented in P3P where some web sites are allowed to leave and access cookies and others are not. Alternatively, PT can be specified via industry codes such as those standardized in the North American Industry Classification System (NAICS). The NAICS codes are six digit codes where the first two digits indicate a broad

industry such as agriculture, health or finance. Each additional digit indicates a more specific industry within a larger industry. Figure 3 shows a piece of the NAICS classification scheme for the Finance and Insurance Industry.

2.1.5. Mashup privacy protection spaces

At the highest level, at least one mashup protection space applies to a particular data set. This represents the basic policy of the government agency that has control over the data set. For example, property sale transactions are available on-line. The data for these transactions are in the public domain. However, suppose that the state restricts on-line access only by pre-certified members of the real-estate market (i.e. PT + = real-estate) which may include companies that participate in the market such as brokers, those that provide information about the market such as real-estate reporters, or economic analysts who study the housing and building markets. In addition, this group may only have the data if they use addresses only for mapping without releasing names of those involved. The purpose would be “mapping” with a positive attribute. No other purpose is allowed.

As another example of specifying personal privacy policies, constituents may want to prevent mashup providers from combining their address information with income information. However, they may be willing to allow government agencies to create a map overlay of the information for the purpose of surveying income bands in an area. This policy could be represented by \{DT + = (income, address), PT + = government-agency, P + = surveying, OP + = overlay, LP + = (map, satellite image)\}.

2.2. Personal Privacy Policy (PPP) network

Citizens may participate in protecting their privacy from unwanted exposure of their government data via mashups by publishing their Personal Privacy Policies (PPP) in a preference repository. Technically,
these policies can be used to govern all release of their PII whether to a mashup provider or to any other individual or organization that makes a request.

This repository is assumed to be widely known and accessible to data providers, through URIs. Data providers can, in fact, ask individuals for their preferences when they collect data. They can then inform the individuals of the repository and submit their privacy policies for them. In order for the concept to gain traction, government agencies are choice organizations for establishing PPP repositories and collect individuals’ data – because of the scope and size of their collection. These repositories form a virtual network, the PPP network.

The PPP can be maintained by the citizens by accessing the PPP network and by performing updates on their personal privacy policies. Whenever users provide information to a government organization, they should be issued with a certificate or credential such that they can access the privacy preferences repository and specify their personal privacy policies for external use of the data by mashup providers. This credential could be electronic or a smart card or both [13]. The smart card would serve both as an identity card and a store of information that gives the user authority to access information about themselves and change policy governing that information. It can also include links to identification IDs associated with various services.

The PPP network will allow citizens to have more control over their own private data, through direct participation in protecting the private data. This participatory privacy protection also accommodates a high degree of individual differences in privacy, and may foster the level of trust in government agencies. It also simplifies the requirements on individuals. They can specify their preferences once for all known as well as unknown potential uses of their data.

2.3. Mashup privacy policy enforcement and reasoning

Today some data providers will only allow access to their data to pre-certified mashup providers to whom they have given a credential, key or access login. This paradigm of data release is limited in a large scale environment, such as the Web and Internet. Our proposed model opens up the process by allowing mashup providers to indicate the business they are in (i.e., their provider type), the data types that will be used, and the purpose of their mashup.

Specifically, an interactive dialogue occurs between the mashup provider and the data provider where the mashup provider indicates the parameters associated with themselves and their service, the data provider consults individual privacy preferences in the PPP network and either denies access or returns the requested data subject to any restrictions discovered in the repository.

When government data is requested for release, the government as data/content providers should conform to the regulatory privacy policies (RPP), enforcing the no-disclosure without consent rule of the US Privacy Act with a dozen exceptions related to PII shown in the list below [5]:

- Intra-agency need-to-know exemption authorizes the intragency disclosure of a record for necessary official purposes.
- Required Freedom of Information Act (FOIA) disclosure exemption states that the data should be released to FOIA requests unless it is exempted from the FOIA rules.
- Routine use exception states that data release or sharing is allowed if used in obvious routine tasks, e.g., federal tax payer information collected by the federal tax administration is disclosed to state tax officials for state tax administration.
- A record can be disclosed to the Bureau of Census for the purposes of planning or carrying out a census or survey.
A record can be disclosed to the National Archives and Administrations for the purpose of archiving or the evaluation of archiving.

A record can be released to support law enforcement activities.

Disclosure to the Comptroller General, or any of his authorized representatives is permitted for the performance of the duties of the General Accounting Office.

A data record should be released for the normal course of court proceedings, including court-ordered discovery.

The Debt Collection Act authorizes agencies to disclose bad debt information to credit bureaus.

These RPP policies are specified using the mashup privacy space model and can be enforced in the context of mashup providers. However, the enforcement needs to consider the PPP where the personal privacy policies state that some data types can not be used for a certain purposes and can not be combined with other data sources, etc.

There are three different cases to consider in the enforcement of the mashup privacy policies that includes both RPP and PPP.

Case 1: RPP and PPP are in agreement – apply either policy in disclosure decision

Case 2: RPP prohibits but PPP permits disclosure. The tighter policy should apply but the government agency could justify disclosure since the PPP permits it.

Case 3: When the RPP requires the record disclosure, but the PPP states preference for non-disclosure, the enforcement follows the RPP but data source should notify the data subject of the data release. In this case, the government’s RPP takes precedence but the citizen is notified of the data release, and the third party mashup provider is also notified of the PPP expressing the individual’s sensitivity of the private information on the particular record.

In case 3, the notification to the citizen/individuals of the disclosure against his privacy preference alerts the citizen with the information such as to whom the data was released and for what purpose the data may be used, and why the data is required by law to be released. The notification to the mashup provider of the PPP that states the citizen’s preference of data not to be released for a particular purpose may make the mashup provider aware of the sensitivity the citizen feels about the data use.

2.4. Architecture and implementation

Our implementation builds upon existing languages and structures for policy specification and communication as in [2,6,8,9,16,17,19]. The data request by the mashup service provider is made through Web services. The PPP and RPP policies are specified with P3P statements in XML, and can be placed in URIs (e.g. http://mygov.gov/policy1.xml). The policies are a collection of policy statements where each policy statement specifies the name of the policy, the URI of the natural language description of the policy, expiry date, the provider information, purpose, the intended recipients of the data, data types, etc., as schematically shown below.

```xml
<POLICIES xmlns="http://www.w3.org/2002/01/P3Pv1">
  <POLICY name="policy1" discuri="http://mygov.gov/disc1" .... </POLICY>
  <POLICY name="policy2" discuri="http://mygov.gov/disc2" .... </POLICY>
  <POLICY name="policy3" discuri="http://mygov.gov/disc3" .... </POLICY>
</POLICIES>
```
To accommodate the mashup-related privacy protection, our work involves identifying the tags that need to be extended to specify intended usage of the data (purpose), operations, and linking parameters.

The policy file containing the policy statements is referenced by a set of services where the policies specified in the URIs are applicable. For instance, the Web service definition can contain a P2P policy URI, as shown in the fragment of a Web service WSDL definition below. The Web service requested by a third party mashup provider for this particular service is subject to the privacy policy stated in the URI http://mygov.gov/policy.xml file and its relevant privacy statement.

```
<wsdl:definitions>
  ...
  <service name="MyService" interface="myns:Interface"
    p3p11:p3p="http://mygov.gov/policy.xml\#policy3">
    <endpoint name="Endpoint1" binding="myns:binding">
      <soap:address location="http://ws.example.org/myservice" />
    </endpoint>
  </service>
</wsdl:definitions>
```

Similarly, the Personal Privacy Policies can be specified and stored within the PPP network. The mashup provider’s request is made in the Web service invocation using the SOAP format. The requested Web service has relevant policy statements stored in URIs for the PPP and the RPP policies, and tries to identify the relevant data subjects (relevant individuals) whose PPP is also consulted for evaluating the data disclosure as described in Section 2.3.

The secure mashup architecture shown in Fig. 4 supports two basic functions. The first is an interactive data access process that ensures data is not released in conflict with privacy policies of both the government agency and the individual associated with the data. The second is an off-line management process that helps to ensure that data access is provided as efficiently as possible.

### 2.5. Interactive privacy preserving data access process

In step 1, the mashup provider (developer) requests content using APIs and Web services. This step starts with the mashup provider discovering suitable content provider(s), such as a government agency (e.g. IRS)\(^6\) to whom a data access request is made. The request includes the mashup provider’s own privacy policy statements, such as the purpose and requested data items, intended target audience, etc. Once the request arrives, step 2 begins. In step two, the privacy protection engine that is implemented in a government agency looks up its own regulatory and organizational policies (RPP) as well as the citizen’s (whose data is requested) PPPs. When the requested data usage policy agrees with the policies of the data provider as well as the data owner’s personal policies, the data is retrieved and returned. When there are conflicts or disagreements on the data access or usage policies stated in the organizational policies or data owner’s personal policies, the request is transformed to a privacy-aware request that applies the most stringent privacy policies, and the request result is sent over to the mashup providers. In addition, a data disclosure notification is sent to the individual (data subject), in case the RPP mandates data disclosure despite the personal privacy preferences not to disclose.

\(^6\)The discovery of the content provider is out of scope of this paper.
2.6. Managing privacy policies

Logically, privacy protection spaces and policies are easy to imagine. Creating them from the various combinations of protection requests is less straightforward. For example, the specification of purposes can vary widely. To prevent that, we use a limited set of purposes, given as a choice in a drop down list in a form-based interface. The initial set considers a limited list of purposes such as “marketing”, “analysis”, “aggregation”. However, our plan is to develop a taxonomy of purposes which may be selected by users.

Another issue is that the specification and management of policy parameters is not easy even when the data is static. Here, we assume the data is constantly changing with new data being constantly generated and with user preferences on existing data also changing. Therefore, a system is needed to derive applicable privacy protection spaces whenever a change in individual privacy preferences occurs.

The data provider is responsible for knowing where individual privacy preferences are housed and ensuring that they are linked to the relevant repositories. Since real-time access to these repositories for hundreds or even thousands of individuals would be prohibitive, privacy protection spaces should be pre-calculated and pre-defined. These spaces are then updated only when data source providers are notified that a privacy preference change has been made.
3. Policy recommendations

This section focuses on legislation within the US. However, similar policy might be employed by other government entities. Legislation concerning privacy has focused on health related information via the Health Insurance Portability and Accountability Act (HIPAA) of 1995 and education information via the Family Educational Rights and Privacy Act. Various standards and regulations also apply to corporations. For example, the Gramm-Leach Bliley Act (GLBA) requires the financial industry to have a privacy policy that describes how and when nonpublic personal information is shared with other institutions. Customers must acknowledge the policy and have the option to opt-out. In addition, protection tools must exist to prevent unauthorized access to information. However, enforcement is left in the hands of the US federal government and there is no individual right of action, thus eliminating the possibility of individual lawsuits [23].

The e-government Act of 2002 introduced by Turner and Lieberman sets requirements on government agencies regarding data use and storage. It requires US federal websites to include a privacy notice that addresses information that is collected, the intended use of the information. It also requires the policy to state how and when users will have opportunities to give or retract their consent to use personal data and how and when users will be notified of use. The policy must also state whether it will be shared and how it will be secured. Finally, it requires that policies must be in machine-readable format. This requirement is usually met by privacy statements in P3P format. Federal websites were further prohibited from collecting data about visitors to those websites in a law initiated by Knollenberg and signed in late 2005. Several representatives and Senators including Markey and Bean have introduced more general legislation designed to regulate the conduct of information brokers.

What has not yet been addressed is privacy issues for data collected outside of direct interaction with individuals at a website. Individuals should be able to set a general policy to govern all data no matter how it is collected and by whom. Governments can start this process through initiation of personal policy registries that are open and accessible via a web service through an open application interface. This would aid not only in the implementation of the mashup privacy protection model described here but also but in protection of personal identifying information in general.

4. Conclusions

Mashups introduce new privacy challenges for individuals, because data might be combined and released by web services with which the user has no interaction. We introduced a model for addressing the privacy challenges in government mashups. It allows a user, as a data owner, to describe their privacy preferences as Personal Privacy Policies (PPP), government agencies, as data providers, to specify Regulatory Privacy Policy (RPP), and mashup service provider to specify their privacy policy (MPP). These privacy policies are expressed in terms of data types, provider types and mashup purposes. The policy statements are defined with an extensive use of semantic and social Web technologies such as existing ontologies.

The proposed technology solution includes a PPP network where citizens can register their personal privacy preferences, and a Privacy Enforcement engine that interprets PPP, RPP and MPP before releasing individual’s data requested by third party applications such as mashups. Our contribution is to make

---

7House of Representatives Bill 1080 and 3140.
specification of privacy policy fine-grained so that data is usable for as wide an audience as possible, for situations where data sharing is for the public good, while keeping data private from the general public.

We recommend that national governments consider legislation for the protection of personally identifiable information. In support of this legislation, such governments should also develop and deploy repositories to house personal privacy policies on data collected by governments regardless of whether the collection process is via a Web interaction or through more traditional means. These repositories must be publicly accessible not only to individuals so that they can review and update their policies but also to any web service provider who collects and makes available data on individuals. We wish to promote such a distributed system so as not to preclude data from being used in ways that will benefit society at large but will offer some protection to individuals at a level in which they are comfortable.

Acknowledgements

The work is partially funded by Research Foundation of City University of New York under PSC-CUNY Research Award Program Cycle 38 awarded to the second author.

References


