Position Paper

Building Privacy into the Semantic Web: An Ontology Needed Now

Anya Kim
Center for High Assurance Computer Systems
Naval Research Lab
Washington, DC 20375
1-202-767-6698
anya@itd.nrl.navy.mil

Lance J. Hoffman
Dept. of Computer Science
The George Washington University
Washington, DC 20052
1-202-994-4955
hoffman@seas.gwu.edu

C. Dianne Martin
Dept. of Computer Science
The George Washington University
Washington, DC 20052
1-202-994-3394
diannem@seas.gwu.edu

ABSTRACT

The Semantic Web can bring more meaning and structure to the Web, but it also has the potential to profoundly affect how personal information is collected and used. We point out that achieving privacy requires trust and security as well a standard method of exchanging privacy policies. We suggest that privacy be built into the Semantic Web now, and stress the need for a privacy ontology.

1. INTRODUCTION

Both in our everyday life and online, we are forced to make privacy decisions. We sometimes unknowingly give up some elements of our privacy, while at other times we consciously do so. When we make a decision to disclose personal information, we do so either because it is convenient, we trust the other party, or because we believe the information to be insignificant. What we don't realize is that the seemingly innocuous bits of information, relatively harmless by themselves, can divulge a lot of information about us when put together. So how does this affect us when we talk about the Semantic Web?

The methods of the current Web that collect privacy information all too often use cookies inappropriately, capture unneeded IP addresses, even track the URLs of the sites we visit as in [12], or require input of unnecessary personal information [13]. To address privacy concerns on the Web, attempts have been made such as TRUSTe [15] and, more recently, the W3C's Platform for Privacy Preferences (P3P) Project [9]. Unfortunately, privacy has been an afterthought in building the Web (and, indeed, in building the Internet).

While security and trust have been discussed to some extent with regards to the Semantic Web, the concept of privacy has to a large extent been overlooked. In this paper, we address privacy issues related to the Semantic Web, the effect the Semantic Web will have on managing personal information, and possible solutions. We suggest that rather than attempting to add suitable privacy measures as an afterthought, privacy issues should be addressed before the Semantic Web gains popularity.

The potentials of doing this right and pitfalls of ignoring privacy issues are described numerous places in the literature [1, 2, 3, 6, 14]

2. PRIVACY ON THE SEMANTIC WEB

While enabling machines (agents) to understand data and exchange it may create limitless applications, the Semantic Web can also create greater risk in the way personal information is managed. As an example, agents will be able to exchange data about you that will allow them to compile databases of sites you visit, items purchased online, hobbies, etc. Users will shy away from using a Semantic Web that enables them to do more yet subjects them to unexpected assaults on their privacy.

The most appropriate mechanism available today on the Web for describing privacy policies is P3P which uses XML to display policies in a standard, machine-readable format [9, 10]. There are several other issues to consider in protecting personal information on the (Semantic) Web. For example, assume a patient needs to have a prescription filled by a doctor. The patient's agent can contact the doctor's agent to request a prescription. The doctor's agent might then look at the patient's medical records, verify that the medication is almost exhausted, and issue a prescription renewal form (digitally signed and encrypted). The patient's agent then would send this to the patient's preferred pharmacist's agent, who after having verified the information and checked its records, would arrange shipment of the medication to the patient's home address with an agent that handles physical delivery (for example, Federal Express, United Parcel Service, etc.). The physical delivery agent in turn may contact the patient's agent to schedule a convenient delivery time. Some financial transaction is also likely to be included with each data exchange.

Semantic Web Workshop 2002, Hawaii USA

The total solution here involves not only privacy protection, but trust and security as well: the four parties involved need to form a trusted relationship with the parties that they exchange information with, and security protocols need to be used to protect the confidentiality of any patient information and financial data that travels over the Web (as well as to ensure integrity and non-repudiation). In the above scenario, we want to be sure that only the minimal/pertinent information required to perform each process is exchanged between agents. The pharmacist need not know the nature of the illness or the patient's entire medical history, and the physical delivery agent requires only the shipping address and contact information to ship the package. Indeed, some services such as iPrivacy [5] provide this "blind delivery service" today.

To the delight of some and the dismay of others, privacy may soon become an explicitly negotiable commodity, perhaps with some baseline limits on what is not negotiable. Certain applications may allow a bargaining process where a user is enticed to give up some degree of his or her privacy in exchange for more privileges, or a discount on fees [7]. However, the above is only possible if users and agents can communicate incentives and changes to privacy preferences to each other in a common language.

The P3P specification supports a base set of data elements that are used to describe the type of data that are frequently collected from Web sites and referenced in the sites' privacy policies. However, it allows sites to declare additional data elements by publishing their own schemas [10]. While this approach may be suitable for the current Web, on the Semantic Web where agents gather information from various sources and exchange and compile this information to fulfill the user's mission, inter-understandability of this information and how it is used is a concern. Suppose a patient's agent is on the Semantic Web, comparing pharmacists. If privacy is of a concern to the patient, the agent can compare privacy policies of the pharmacists. However, if pharmacists create their own data elements to describe their privacy policies regarding the prescription that's been ordered, how will the patient's agent be able to compare policies received from different sources? The following example shows a definition for a 'drug' data element and 'medication' data element in P3P syntax. Even though the data elements are named differently, they refer to the same piece of data, the prescription that the user will order.

```
<DATA-DEF name="drug.name"
    short-description="Name of drug"/>
<DATA-DEF name="drug.rxnum"
    short-description="Rx number"/>


<DATA-DEF name="medication.name"
    short-description="Nameof medication"/>
<DATA-DEF name="medication.no"
    short-description="Unique id of
    medication"/>
```

The multi-interaction in the Semantic Web will require that agents communicate with one another using an agreed-upon language to protect the privacy of their users.

Therefore, the most important piece of a privacy-respecting Semantic Web is a privacy ontology that enables agents to exchange privacy-related information using a common language. Even today, the speedy adoption of P3P is hindered because of the difficulties in arriving at an agreed-upon vocabulary. This was the same problem encountered by the Platform for Internet Content Selection (PICS) specifications developed by the W3C in the late 1990's for content labeling [8, 11]. Until the recent creation by the Internet Content Rating Association (ICRA) of a common vocabulary to specify certain types of content, the development of many templates for content control based upon different value systems was not possible [4]. The privacy ontology should be able to clearly define the various dimensions of privacy (e.g. privacy of personal behavior vs. privacy of communications), and contain enough parameters and index terms to enable specification of a privacy policy in a standard machine-understandable format. It should be descriptive enough to specify the highest known standards of data protection and privacy (expand the standard vocabulary and base data elements of P3P). It should also not only allow specification of a user's privacy preferences to a Web site, but also allow users (through their agents) to collect and store (in a specified format) certain information about the Web sites and other agents they interact with.

Currently, policies are created by Web sites regarding collection of information about the user accessing the site or page. When semantic agents are roaming the Web looking for information, we may see a shift in privacy exchanges: users create privacy metatags that describe the desired privacy level of data on their own web sites. These tags will explain to the agent how different elements in the web page can be accessed. For example if user Joe Smith posts his resume on his Web page, the section containing address information can be labeled as <current/>, while the section containing authored papers can be labeled <individual-decision/>[1] . Now, when semantic agents browse Joe Smith's page, they can automatically understand how he wants the data elements to be handled by agents. The agents can either agree to this and browse his web page, or disagree and be shown nothing. This gives owners of personal information greater control over its dissemination than currently available.

The Semantic Web will bring about new challenges for privacy management. Careful consideration of the issues now will result in a privacy-respecting Semantic Web that allows users to employ agents to carry out sophisticated tasks while being confident that personal information is being managed in their desired fashion.

[1] For the purpose of this example, we have borrowed tags from the Purpose element of the P3P specification [10]. The specification states that <current/> information may be used for completion of the activity for which it was provided. <individual-decision/> information may be used to determine the habits or interests of the user, combined with other data to make a decision that directly affects that individual.

3. REFERENCES

[1] Center for Democracy and Technology (CDT) Briefing Book on Privacy Legislation - 2000, http://www.cdt.org/privacy/plif.shtml.

[2] Etzioni, A. The Limits of Privacy. Basic Books, New York (2000).

[3] Hoffman L. J. (ed.). Building in Big Brother. Springer-Verlag, New York Berlin Heidelberg (1995).

[4] Internet Content Rating Association site, http://www.icra.org.

[5] iPrivacy site, http://www.iprivacy.com/.

[6] Landler, M. Fine-Tuning for Privacy, Hong Kong Plans Digital ID. The New York Times. http://www.nytimes.com/2002/02/18/technology/18KONG.html. (Feb. 18 2002).

[7] Laudon, K.C.: Markets and Privacy. Communications of the ACM 39:9 (1996), 92-104.

[8] Martin, C. D., and Reagle, J. M. A Technical Alternative to Government Regulation and Censorship: Content Advisory Systems for the Internet. Cardozo Arts & Entertainment Law Journal, 15:2, 409-427.

[9] Platform for Privacy Preferences (P3P) Project site, http://www.w3.org/P3P/.

[10] The P3P Specification 1.0, http://www.w3.org/TR/2001/WD-P3P-20010928/.

[11] Platform for Internet Content Selection (PICS) site, http://www.w3.org/PICS/.

[12] Privacy.net site, http://www.privacy.net.

[13] PrivacyRight.com, Control of Personal Information, White Paper. (May 2001), http://www.privacyright.com.

[14] Rosen, J. The Unwanted Gaze: The Destruction of Privacy in America. Random House, New York (2000).

[15] TRUSTe site, http://www.truste.com.