Limiting risks to personal privacy using the semantic web

S.Atkinson[1], C.Johnson[2] and A.D.Phippen[1]

[1.] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2.] University of Plymouth, Plymouth, United Kingdom
e-mail: shirley.atkinson@plymouth.ac.uk

Abstract

Collecting information about an individual can be carried out with ease and without the individual being aware. Problems are faced when the information is collected, combined and used in an abusive fashion. This research explores the potential for harm posed by the combination of Internet technologies and the release of personal information to vulnerable individuals. A prototype semantic web application has been designed to reduce the risk to the individual by acting upon the release of personal information.

Keywords

Personal Privacy, Domestic Abuse, Teenagers, Semantic Web

1. Introduction

Modern living introduces an unprecedented onslaught in regard to dissemination about information about us as individuals. Information about our location, name and address, date of birth, preferences and dislikes are all easily transmitted around the globe through Internet protocols. New devices become Internet enabled to share in this web of transmission in the name of making our lives easier, for example our mobile phones can access WAP pages or their location can be tracked through the Internet.

One of the hindrances to this transmission is the interoperability of personal data, in that not all formats are recognisable in all places. However, work is being done in this area. The Semantic Web proposes that all data should be marked in the same way, linked through to the concepts it represents, so that machines can interpret the data as humans would.

The Semantic Web is not yet close to achieving its vision, however, if it does, this has serious implications for the transmission of our personal data. Business views personal data as a commodity and marketing desires drive forward the collection of as much data as possible. Recommender systems that encourage people to spend more money with a particular retailer require large amounts of personal data to build their profiles.

This paper introduces the background influences to privacy from the perspective of the individual in their normal day to day life before introducing the Semantic Web, with a high level overview of the elements that fit together to create it. The research

carried out so far is introduced along with the prototype developed out of the findings. This paper presents an evaluation of the prototype before finishing with the conclusion.

2. Influences

Privacy has progressed from the original concepts of natural privacy, where moats and drawbridges provided a barrier to unwanted intrusion to the more modern concepts surrounding the control of personal information. Controlling personal information has been proposed by both Clarke (1999) and Tavani (2007) as only one of a number of elements to privacy.

Raab (2004) describes privacy as being highly subjective. An individual's perception of a privacy invasive situation is based on the context created by a combination of a complex mixture of circumstances. These circumstances include the elements of technology, legitimate activity, perceived imbalances of power and the social context (Hine and Eve, 1998).

Personal privacy is affected by the interaction of many differing factors that range from external influences upon an individual to more internalised behavioural issues. Legal controls provide elements of protection for personal data and constraints for commerce; personal data is collected and manipulated by commercial activity; government policy dictates the collection, maintenance and distribution of publicly available personal information and influences educational campaigns which seek to highlight concerns and influence behaviour; and technology provides varying levels of protection under different circumstances.

The global nature of the Internet brings complexities in terms of jurisdiction. An individual interacting with a websites that collects and stores personal information can involve a number of countries and their privacy laws, as General Motors found to their cost when attempting to create an online telephone directory (Windley, 2005).

Increasingly it is easy to collect and store personal information both in an explicit fashion and unobtrusively (Cranor and Garfinkel, 2005). Consumer preferences, location or information are now inferred and gathered by either sensors or clever software. This implicit gathering of data is much easier than explicit requests to the user to provide specific information. Powerful recommender systems designed to encourage the consumers to spend more money with the company require large quantities of personal data to be effective. Amazon.com is a good illustration of this with the user being shown books and related products that have been selected based on their previous interests.

Online accessibility has become a goal for many public bodies. Public records are now available online for very little cost. The Office of National Statistics provides a fully searchable database for all birth, marriage and death registrations post 1984. Registration data prior to that date are in image format (ONS, 2005). The Land Registry provides information about how owns which property and the mortgage

details associated with it (Land Registry, 2005). Medical records are now available through the NHS net to those medical professionals who need access (NPFIT, 2005).

Advances in database technology, sharing of information combined with DNA testing has an influence in the fight against crime. DNA is now routinely taken as part of police enquiries for the National DNA database (Davies, 2000) and surveillance of public spaces through CCTV is commonplace (Garfinkel, 2000). Biometric approaches are considered for incorporation into identity cards and passports.

Privacy Enhancing Technologies (PETs) seek to redress privacy concerns by making use of solutions such as anonymisation (HISPEC, 2002) to hide activity or location; or making use of control based access to detect intrusion (Lecomte et al, 2005). Animated user friendly visual aids such as the Privacy Bird (Byers et al, 2004) or Hector the Protector (Microsoft, 2005) whilst visually appealing, have their limitations.

3. Semantic Web

The Semantic Web began as a vision provided by Berners-Lee (2000) where content would be annotated so that other machines would be able to manipulate the data to gain an understanding of the context. This approach would lead to all data being made available, irrespective of format. For example, contents within a sound file would be accessible.

The goal for the Semantic Web is to streamline the interchange of data (Passin, 2005) to make it interoperable. It will seek to combine the many different ways that people interact with the web, for example: web pages, newsgroups, email, ubiquitous devices and mobile phones (Fensel et al, 2003).

Figure 1 represents a high level overview of how the Semantic Web layers will be used. The author creates an RDF document which contains triples that express what the data is. These triples link to concepts from an agreed standard set of terms, an ontology that is expressed using OWL. The data is stored either as pages on a web server or created as a self-describing web service using OWL-S a standard. Consumers instruct software agents to search for information using the same agreed domain ontology. The information is then gathered from the heterogeneous sources and combined, thus presenting the consumer with the information relevant to achieving their goal.

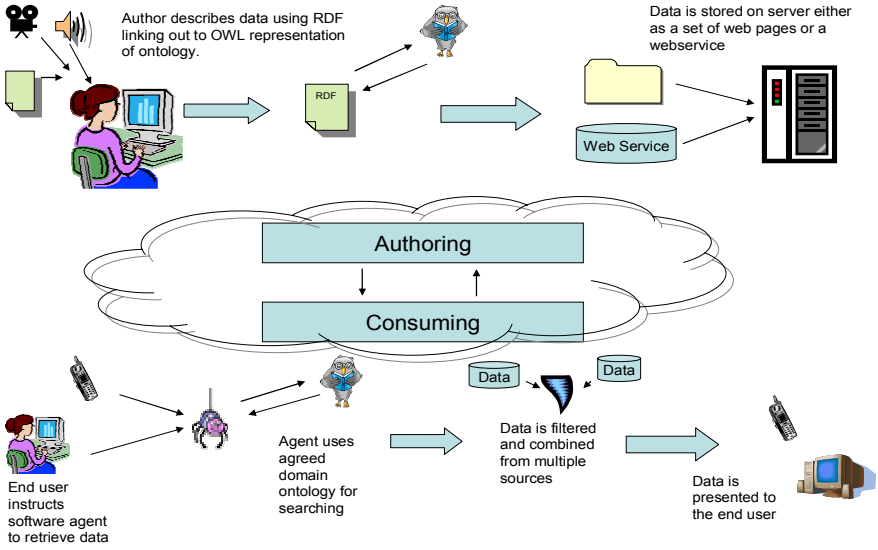

Figure 1: Representation of Semantic Web Activity

With the emerging design for the Semantic Web, privacy together with the use and control of personal information is not a specific individual part of the infrastructure. It is often represented within the element of Trust (Guerrra, 2003) along with identity and security. Golbeck (2003) outlines how the current focus for trust in the Semantic Web is on trustworthiness and credence given to data on the Internet. Other work concentrates on the manipulation and representation of privacy policies (W3C, 2005) or making links between individuals to determine whether to trust them or not (FOAF, 2006).

4. Research

The use of the Semantic Web to combine and infer lends itself to concerns when involving personal information. The move towards interoperability of all data would appear to remove the rudimentary protections that might have been in place when data could not be combined, it was harder and more costly to gather the information together into a profile.

Vulnerability is the perceived risk of mental or physical harm and has been strongly linked to the disclosure of personal information (Dinev and Hart, 2004). Solove (2003) proposed that technologists are building an "architecture of vulnerability" whereby individuals are placed at risk and yet are powerless do reduce those risks.

Mitigating these vulnerabilities involves an analysis of the risks likely to be faced by individuals. Risk management is a complex area involving the highly subjective assessment of risk. Raab and Bennet (1998) propose that risk is best assessed by using a combination of expert knowledge with objective calculation of risk along with subjective views of individuals.

This research therefore focuses on the intersection between the fields of personal data and Internet connectivity. An examination of the potential for harm through abuse of personal data by other individuals is made, combined with the influences that PETs have when combined with risk control where risks to harm are identified and managed.

The issues for privacy for Domestic Abuse Survivors (Survivors) and Teenagers were explored through the use of semi-structured interviews, focus groups, workshops and online questionnaires.

4.1. Findings

Concerns were raised about how easily personal information was divulged through mobile phones, emails, social networking websites, public records and third party databases. The risks manifested themselves through women being tracked to safe houses and refuges; harassment and stalking.

83% of the teenagers interviewed divulged personal information with 27% expressing concern about it. Coping mechanisms were employed where requests for personal information were considered to be excessive. These included either ignoring the request or if the request was mandatory, giving false information. Blocking mechanisms were frequently used if there was unwanted contact.

4.2. Prototype

A set of use cases were developed based on the findings and took three different perspectives: the end user, a support worker and an organisation responsible for others. The use cases illustrate how the actors would want technology to protect their personal information.

A Vulnerability Assessment Framework (VAF) and a Taxonomy of Threat were also developed for use within the prototype software. The taxonomy of threat detailed the relationships between the different types of threat and potential consequences to personal privacy. The VAF provides a framework to measure potential vulnerability to an individual, so that the required protection levels could be calculated.
The prototype took the format of an Internet Explorer 7 Browser plug in which had the objectives to:

- Identify where risks lay.
- Identifying where individual's or refuges are located through the use of postcode identification and mapping software
- Personal information given out by individuals
- About themselves
- Or about other people.

The plug in shows up as an additional toolbar at the top of the browser and is shown in figure 2.

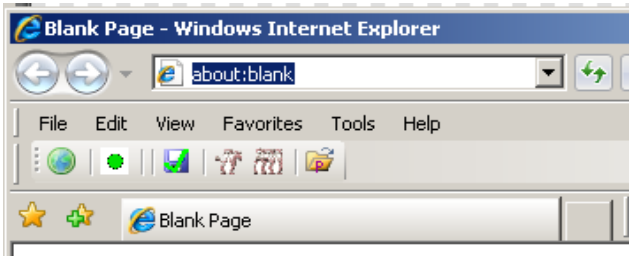

Figure 2: Prototype toolbar

A search is made for instances of the address, the postcode and the names given in the settings. A right click of the mouse button will allow the user to visit the URL where the information has been collected. In addition, the prototype displays to individuals whether a web page being visited poses a threat. The circle changes to red or yellow depending upon the assessment of the threat and the relevant fields in the web form are also highlighted in the relevant warning colour.

An analysis of where information has been divulged is shown to the individual. Details are saved about where information is given out, to which URL and what was filled in. It also illustrates details of where other people's names have been given out (the colleague bar at the end of the bar chart in figure 2).

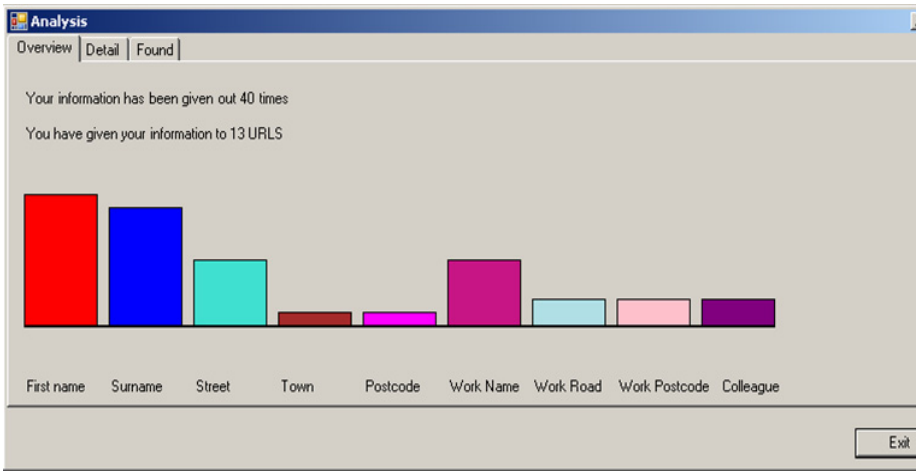

Figure 3: Analysis of information divulged.

5. Evaluation

Evaluation by the target user groups is in progress, however early feedback indicates that the prototype has some encouraging effects. One teenager remarked

> *"Yeah, this would make me think a bit more about what I put out and where! It's nice to know what I've said".*

A manager for a refuge felt that the software would be useful not only to determine where information was being given out by residents within the refuge but also to act upon the behaviour of the residents in encouraging them to think more deeply about the consequences of their actions.

6. Conclusion

Early feedback for the prototype are that it is useful in encouraging people to think more about where they wish to give out their information, rather than acting in a restrictive manner. Evaluation continues to determine if controls that influence the giving out of personal information can reduce the potential risks for harm to an individual.

7. References

Berners-Lee, T. with Fischetti, M. (2000), *Weaving the Web*. Texere Publishing. London.

Byers, S., Cranor, L., Kormann D and McDaniel, P. (2004), "Searching for Privacy: Design and Implementation of a P3P-Enabled Search Engine", In *Proceedings of the 2004 Workshop on Privacy Enhancing Technologies (PET 2004)*, Toronto, Canada

Clarke, R. (1999), "Introduction to Dataveillance and Information Privacy, and Definitions of Terms", Xmamx Consultancy Pty Ltd, http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html#Intro (Accessed 17 May 2007)

Cranor, L.F. and Garfinkel, S. (2005), *Security and Usability*, O'Reilly, USA

Davies, S. (2000), *The Death of Privacy: A Personal View*, BBC, Video Cassette

Dinev, T. and Hart, P. (2004), "Internet Privacy Concerns and their Antecedents - Measurement Validity and a Regression Model", *Behaviour and Information Technology*, Vol 23, No 6, pp 413-422

Fensel, D. Hendler, J, Lieberman, H. Wahlster, W. (Eds), (2003), *Spinning the Semantic Web*, The MIT Press, Cambridge, Massachusets

FOAF, (2006), "The Friend of a friend Project", http://www.foaf-project.org. (Accessed 9 March 2006)

Garfinkel, S. (2000), *Database Nation*, O'Reilly Associates, Sebastopol, CA

Golbeck, J., Parsia, B., Hendler, J. (2003), "Trust Networks on the Semantic Web", In *Proceedings of Cooperative Intelligent Agents 2003*, Helsinki, Finland, August 2003 http://www.mindswap.org/papers/CIA03.pdf (Accessed 9 March 2006)

Guerra, A.G., Zizzo, D.J., Dutton, W.H and Peltu, M. (2003), "Economics of Trust in the Information Economy: Issues of Identity, Privacy and Security", *Oxford Internet Institute, Research Report No 1*, April 2003. http://www.oii.ox.ac.uk/resources/publications/RR1.pdf (Accessed 9 March 2006)

Hine, C., Eve, J. (1998), "Privacy in the marketplace : Social construction of privacy", *The Information society* , Vol. 14, No 4, pp253-262

HiSPEC, (2002), "Privacy Enhancing Technologies: State of the Art Review", http://www.hispec.org.uk/public_documents/7_1PETreview3.pdf, (Accessed 12 March 2006)

Land Registry, (2005), "Land Registry Online", http://www.landregisteronline.gov.uk/lro/landing.htm, (Accessed 12 March 2006)

Lecomte, J., Clarke, N.M., Furnell. S.M. (2005), "Artificial Imposter Profiling for Keystroke Analysis on a Mobile Handset", In *Proceedings of 5th International Network Conference*, University of the Aegean and University of Plymouth

Microsoft, (2006), "Hector the Protector", http://www.microsoft.com/nz/athome/security/children/hector.mspx, (Accessed 12 March 2006)

NPFIT, (2005), "NHS Care Records Service", National Programme for IT, http://www.connectingforhealth.nhs.uk/delivery/programmes/nhscrs, (Accessed 12 March 2006)

ONS, (2005), "General Register Office", , http://www.gro.gov.uk/gro/content (Accessed 17 May 2006)

Passin, T.B. (2005), *Explorers guide to the semantic web*, Manning, Greenwich, USA

Raab, C.D and Bennett, C.J. (1998), "Distribution of Privacy Risks: Who Needs Protection", *Information Society*, Vol 14, No 4, pp263-274

Raab, C.D. (2004), "The Future of Privacy Protection", *Cyber Trust and Crime Prevention Project*, http://www.foresight.gov.uk/Previous_Projects/Cyber_Trust_and_Crime_Prevention/Reports_and_Publications/The_Future_of_Privacy_Protection/The_Future_of_Privacy_Protection.html (Accessed 17 May 2007)

Solove, D.J. (2003), "Identity Theft, Privacy, and the Architecture of Vulnerability", *Hastings Law Journal*, Vol 54 No 1227, pp1232

Tavani, H.T. (2007), *Ethics and Technology* 2nd Edition, Wiley, USA

W3C. (2001), "Semantic Web Activity Statement", www.w3.org/2001/sw/activity.html (Accessed 2 March 2006)

Windley, P. (2005), *Digital Identity*, O'Reilly, USA

As the users are very privacy conscious such a service has to take care of providing privacy while delivering a service. Saltzer and Schroeder define the term "privacy" that it "… denotes a socially defined ability of an individual (or organization) to determine whether, when, and to whom personal (or organizational) information is to be released" (Saltzer and Schroeder, 2004). Ross Anderson describes privacy as "ability and/or right to protect our personal secrets, the ability and/or right to prevent invading our personal space" (Anderson, 2001). In the Lategan & Olivier paper (Lategan and Olivier, 2002) it was expressed that: "The privacy of information used on the Internet is a very real and important issue. Many users have concerns about the security of private information supplied to organisations on the Internet, and rightfully so, as tales of compromised information abounds.".

Privacy policies are being used more and more to promise the security of an individual's private information ..." (Lategan and Olivier, 2002). In order to achieve this the "Chinese Wall" approach is proposed which is based on a trusted middleman to allow push services based recommendation without sacrificing privacy. By doing so the organization which wants to offer recommendations can select a user group entirely based on their interests, their location and the available temporal information of the user without knowing the user personally. This way offers anonymity to the user but allows selecting a matching target audience. The vital requirement is that the user trusts the middleman (acting as the "Chinese Wall") and that the information provider is able to get his message through to potential clients.

2. In "whom" we trust

The perception of trust and privacy varies with every user and the individual experience in using online services. An interesting fact is that people have developed a distrust towards online services which has been caused by illegal activities like phising, identity theft and the suspicion that "somebody" does something with their data.

In the general perception it seems that users feel saver in the "offline" world then in the "online" world. This interesting fact has to be considered when introducing a service like Multi Dimensional Personalisation. Another interesting fact is that a study has shown that even if an internet user describes himself as privacy concerned they give out more information about themselves as they initially wanted to do (Berendt et al, 2005). So the MDP could protect users from themselves.

As this service works across the borders from the "online world" to the "offline world" it might get affected by the privacy and trust concerns of the users. As the offline world is the everyday environment in which everybody is used to live, most people do not longer see that in this world the same risks are there.

2.1. Online vs. Offline world

In the online world there will be the same or similar services available as in the offline world. By the "bad" reputation the internet has gained recently there is this "distrust" towards online transactions whereas it seems that there is a higher level of