Using the Concept of Topic Maps for Enhancing Information Privacy in Social Networking Applications

Stefan Weiss

Institut für Wirtschaftsinformatik
Johann Wolfgang Goethe-Universität
Gräfstraße 78
60054 Frankfurt am Main
stefan.weiss@m-lehrstuhl.de

Abstract: The growth of online social networking applications (SNA) and the economic potential that might be generated by their underlining business models is getting broad attention recently. Public discussions about the information privacy for SNA users have been revived as a result of some privacy-invasive activities taking place such as identity theft, stalking, discrimination, distortion of facts, embarrassment, and many others. One of the problems why these threats to information privacy are not merely theoretical in nature is the apparent loss of control over the personal information provided and the lack of transparency over its usage on the Internet. This paper suggests a privacy-enhancing technology (PET) that adapts the semantic technique concept of a topic map for the use with SNAs. Topic maps in such a setting would provide the user with more transparency over the status of his information privacy when using SNAs and would provide means to exercise choices on particular data sets. The paper explains how a topic map concept would be adapted as PET, which privacy principles it would cover, and how it would generally solve some of the issues of information privacy in online social networks.

1 Introduction

The exponential growth of the membership of online social networks such as MySpace, Facebook, YouTube, Flickr, LinkedIn, Xing, and others in the last few years is observed with great interest by different disciplines such as economics, social sciences, law, and technology. Besides the economic potential such new business models may generate, it is the users' behaviour on these sites and particularly the degree of personal information users provide publicly to enrich their social networking profiles that is getting a lot of attention – a fact that also revives the discussions about information privacy on the Internet.

A recent position paper by the European Network and Information Security Agency outlines the most important security and privacy threats to users and providers of social networking sites and offers policy and technical recommendations to address them [En07]. When reading the technical recommendations, though, it is apparent that explicit technical solutions do not address the individual's information privacy from a holistic perspective but rather tries to firefight current threats.

This paper suggests a specific technical concept through which a user of social networking sites could control and manage his privacy preferences of a number of sites simultaneously. The semantic technique called topic maps would create more transparency of the personal data that is provided and would give a mechanism to place controls within the solution to limit the personal data processing to its original purpose. The proposed solution diverts from the traditional viewpoint of a privacy-enhancing technology but as prior privacy research and the dynamics of open Web environments have shown, the pre-eminent goal for privacy research and PET development is likely to shift from access protection, anonymity and unlinkability type of solutions to privacy safeguarding measures that enable greater transparency and that directly attach context and purpose limitation to the personally identifiable data itself [We07].

The rest of this paper is organized as follows. Chapter 2 illustrates the problems that need to be addressed when trying to enable privacy in social networking applications. Chapter 3 points out the limitation of a selection of current solutions and why they are not appropriate when using social networking sites. Chapter 4 suggests using the semantic technique of topic maps to address the described problems and Chapter 5 explains the background and major attributes of topic maps. Chapter 6 visualizes a sample case on how to manage privacy preferences by using the concept of topic map. Chapter 7 lists some remaining challenges in the context of the semantic web and the use of topic maps and Chapter 8 gives the conclusion.

2 Illustrating the Problem

Privacy in our physical world is about using some form of control mechanism to create our own personal and private space. Think about the curtains on your window. If you want some privacy in your home and you don't want onlookers to disturb your evening with your family, you close the curtains.

Information privacy on the Internet is quite a different story. The Web is evolving with lots of open spaces and new modern architectures that have windows without curtains, not even curtain rails. Especially with the increasing use of social networking applications on the Web, it becomes apparent how much knowledge on and about individuals is accumulated and exposed openly. Web protagonists argue that privacy is dead and that people don't care about their information privacy anymore anyway. They argue that most users of social networking applications provide their own sometimes very personal details by free choice.

However, the moment privacy-invasive activities on the Internet become known and it becomes apparent that the misuse of openly available personal data can cause financial harm, reputational damage, embarrassment or discriminatory acts to the individual, the calls for more information privacy online get louder.

In order to enhance the information privacy of Internet users, a number of requirements need to be addressed and solved. This paper attempts to focus at this point only on the most obvious problems inherent in the open nature of the Web and especially in the use of online social networks while others are mentioned as remaining challenges (Ch. 7).

2.1 Limited structure and control

One of the most important aspects to assure information privacy on the Web is that control mechanisms can be implemented to actually control the use of personal data for its specified purpose. Jonathan Zittrain sees this point in an article published in the *Stanford Law Review* [Zi00] by saying that the problems of privacy and copyright being exactly the same. With both, there is a bit of “our” data that “we’ve” lost control over. In order to implement any kind of control, however, the data to be controlled, the data handling procedures, any rules, and the parties involved need to be known and transparent in a controllable structure. Unfortunately, the Web – and even more so Web applications such as SNAs – is low on structure and control.

Control models used today originated in confined and trusted environments with clear perimeters where controls could be placed. These models, however, are not appropriate for the enforcement of privacy requirements in the open world of the Web. Current privacy-enabling solutions mainly focus on data access and authorization restrictions or on reducing the identifiability of an individual by any data traces left behind. Besides the fact that these controls are undermined by the open nature of the Web and the unaware user (see section 2.2) releasing most of his personal data without restrictions, there are currently no automated controls in place to assure the purpose binding of personal data usage on the Web, one key privacy principle apparent in most privacy and data protection laws and regulations.

Similar to the security concepts for Web 2.0 applications [Da07], the information privacy controls in social networking applications need to be enforced at the data level. Currently, personal data produced through Web applications can be categorized with XML and rules can be attached to it by using the Resource Description Framework (RDF). When it comes to controlling the purpose binding aspect of the personal data, though, it gets difficult in an environment where not all involved parties are transparent and where the personal data spreads uncontrollably.

2.2 Increasing technical complexity

The increasing complexity of web application technology and the introduction of mashable applications in the so-called Web 2.0 environment are not helpful in providing the needed structure. In fact, the Web evolves without minding the necessity for control over personally identifiable data. This is especially apparent in online social networks. Personal data provided by the user to one application may spread to a number of other applications without further control over its usage simply by mashing applications.

Technology experts speaking at the Web 2.0 Expo in Berlin in November 2007 have focused very much on expanding the social graph of an individual user's relationships by further connecting and combining personal profiles [Re07] and designing for a web of data [Co07] without much attention on users' privacy considerations. In these discussions, user privacy was seen at best by providing some additional features like profile view- and access settings and by stressing the development work around more convenient identity management systems such as OpenID. However, intertwining social network applications, their content and ultimately the user's personal data increases the complexity for assuring clear accountability.

2.3 Unawareness of the user about exposure and lack of transparency

When using social networking applications, a user may want to decide on a case-by-case basis if he wants to provide a certain set of personal information about himself in a specified context and if he only wants to provide it for a specific purpose and for a specific data receiver. However, in order to make those decisions, the landscape of his social connections to others, the so-called 'social graph', and the respective personal data he has released to whom, for which purpose, and in what context needs to be completely transparent and controllable over time – a requirement not solved today.

Recent privacy studies on Facebook reveal the fact that most users are unaware of specific risks of privacy-invasive activities and have no idea to what degree their online profile and the personal information connected to it is visible and exposed to others [AG06]. And most users leave their privacy default settings in place because they are overwhelmed by the number of choices they would have to make at the time of data collection. Subsequently, it is impossible to make decisions on information privacy preferences and purpose limitations in different contexts over time without the complete and transparent picture of someone's social graph and spread of personal data.

3 Limitations of Current Solutions

A range of privacy-enhancing technologies try to address some of the problems described above. However, they have certain limitations when it comes to their use in social networking applications.

Ian Goldberg argued in 2002 that most privacy-enhancing technologies to date have been concerned with the privacy of identity [Go02]. Protecting the identity information of an individual is an important aspect of an individual's information privacy. However, it does not address the purpose limitation and control of all other types of personally identifiable data – an aspect that is especially important with the use of SNAs where personal data is openly provided.

The Platform for Privacy Preferences (P3P) provides a technical way to achieve personal choice, informed consent, and a commitment from providers about the use of data [Li03]. P3P is able to turn data handling practices of a Web site provider in machine-readable policies. Web site users can then match their predefined privacy preferences against the P3P policy of the Web site provider. As such, P3P addresses the translation of data handling rules into an automated control – an important step towards enhancing information privacy. However, as discussed earlier, the appropriate privacy controls in social networking applications need to be at the data level. P3P in its current version does not require the Web site user to categorize or tag the data provided with respective usage purposes and only covers a one-on-one matching of general privacy preferences against a general data handling policy.

In order to establish a stricter form of control over specific personal data, the architecture of so-called 'Hippocratic databases' is an interesting option for giving structure and control mechanisms needed for privacy-enhanced Web applications. They limit disclosure of specified data by matching the query about wanting to use the data with a pre-set privacy policy. Hippocratic databases are based on the idea of having one unified database architecture [Ag02] which is also its major limitation for a use within a Web architecture. The Hippocratic database concept actually has boundaries where controls can be set up but the logic in this solution is described in database languages and does not scale to the Semantic Web [Ni04].

What is needed is a technique that puts context or more meaning to the personal data provided by the user to one or more social networking sites. Semantic Web concepts may provide such means. Gottfried Vossen describes the goal of the Semantic Web as making the Web more accessible to computers and thereby enabling new applications for humans [Vo07]. Why not researching the way that semantic technologies and concepts can do this for privacy-enhancing tools and applications?

4 Addressing Information Privacy in Social Networks by using the concept of Topic Maps

Discussions around creating a policy-aware web address different ways of using semantic techniques for data categorization and programming rules into machine-readable policies. Yet, the increasing complexity of data linkage in social networks makes it difficult to address the different associations of a set of personal data, its usage purpose, and the context it was provided in. Tim Berners-Lee sees the increasing need of general logic frameworks that can establish accountable systems supporting an entity in being aware of the provenance of information and responsible for its disposition [Be06].

Topic maps represent such a general logic framework even though used for different purposes so far. Similar to how Google structures data to improve its search engine capabilities and then uses the data search results for different purposes, the user of social networking applications could use a topic map-type of application to structure his personal data and the information privacy preferences attached to each data set. The idea is to give the user access to one central repository where his personal data and the various associations to specific purposes, applications, etc. are stored. One can think of an index of his personal data provided over time in various contexts.

The main goal of using the semantic technique of a topic map at this point would be to provide a basis for reasoning decisions matching intended data usage with specified privacy preferences and their relationships to subsets of personal data. The topic map in such a privacy-enabling application would provide contextually relevant information. Since the topic map concept can only address the privacy requirements of providing transparency and accountability, it could also be labelled as a "transparency-enhancing technology" (TET). With such a solution, an individual should be able to access his own 'privacy topic map' giving him transparency over the different data use cases (topics), how they associate with different SNAs and mashed applications in different contexts (associations), and which data sets have been fed into these (occurrences).

In other settings, it has been suggested to use topic maps for example to catalogue business rules, the applications that use them, and the documents that mention them to support impact analysis and change management [BD05] – a similar approach suggested here but for the purpose of managing privacy preferences. The difference to a P3P approach is that policies or rules would be set at the data level and they could be carried forward to other applications. The proposed privacy-enabling topic map concept would be able to represent the following:

- forms of data usage, purpose, and context;
- shifting relationships between those purposes over time; and
- assignment of personal data to SNAs.

5 How Topic Maps work

The topic map standard ISO/IEC 13250 is an international standard that defines a way of encoding information subjects and the relationships that exist between and among them. It provides a mechanism for organizing unstructured information on the Web. As such they constitute an enabling technology for knowledge management. Dubbed "the GPS of the information universe", topic maps are also destined to provide powerful new ways of navigating large and interconnected corpora [Pe00].

Topic maps go further than what Resource Description Framework (RDF) can do. They are not only making statements about particular resources but they rather capture the relationship between particular resources and representing them as “knowledge”. That is were they become interesting for the use in enhancing the information privacy of SNA users. Topic maps enable the description of complex relationships, whether it will be in information, knowledge, process or social engineering [On07].

One of the major advantages that topic maps have versus other semantic techniques to represent and tag data is what is called reification. To reify something is to regard something as a real thing[1]. With reification, topic maps cannot only be used for representing data sets and their relationships to other data sets but they can be used to reify an association such as a specific data purpose to the original data owner or to the date the data was created – a concept that establishes a control mechanism for privacy assurance purposes.

And finally, one of the major advantages of a topic map versus other semantic techniques is the ability that its structures can be interchanged between applications. Especially when considering Google’s announcement of Open Social’s common set of APIs [Go07], interchangeable platforms will be key for their acceptance in social networking applications.

6 Sample Case “Managing Privacy Preferences”

Starting with a P3P process, the following sample case attempts to show where a topic map concept could extend a P3P solution and where it could address the transparency and accountability requirement of information privacy. Figure 1 shows a simple case of representing a web site provider’s personal data handling practices in form of a P3P privacy policy. The data subject or user of a web site in this case pre-sets his privacy preferences in the same P3P taxonomy as the provider’s policy. When the user visits the particular website, his preferences are compared with the website’s privacy policy. The result of the matching is given to the user in form of a notice.

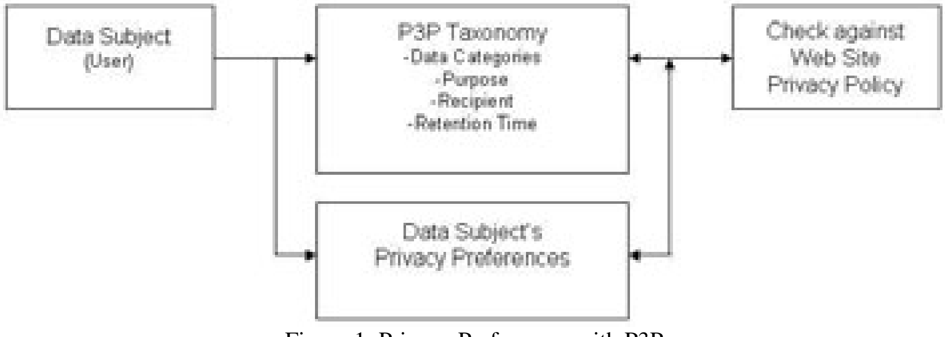

Figure 1: Privacy Preferences with P3P

[1] Merriam-Webster Dictionary

Besides the fact that the user may have a difficulty in setting his high level privacy preferences correctly and probably leaving the privacy settings of the browser in its default setting, there is no mechanism provided that tells the user which type of personal data he provided to whom and for what purpose it was intended or being used. Also, the proper execution of the stated P3P policy and the proper handling of the user's data cannot be checked.

Figure 2 extends the P3P process representing the use of a topic map concept where privacy preferences would be set on the data level and related to various applications.

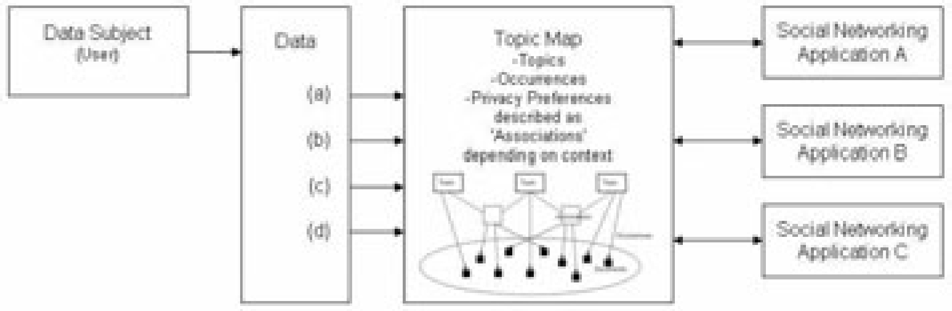

Figure 2: Privacy Preferences with Topic Map

The user would be asked to tag each provided data set with specific preferences and group it in a specific context, for a defined purpose and for the use in one or more social networking applications. The technique would be similar to the tagging functionalities of social tagging sites such as digg or del.icio.us. Tagging categories are easy to attach to data and the kind of metadata is easily transferred to various applications. SNA users are used to such tagging functionalities and, therefore, the topic map type of functionalities would provide an easy way of self-control for their own personal data. The resulting 'privacy topic map' could be visualized either in simple lists or in a mindmap type of graphical element grouping data sets and their attached privacy preferences either by purpose, context, data category, or by application.

Rules such as data handling procedures of the different website providers defined in their P3P policies could now be linked to the 'privacy topic map'. As such, the topic map concept would represent an extension of existing P3P policies and would provide a lot more transparency and accountability to the proper data handling procedures. Of course, such an application would require the use of the same taxonomy in all involved systems but there are various research projects currently underway that try to come up with standardized policy languages and semantic ontologies. Researchers and practicioners in the W3C Policy Languages Interest Group (PLING) for example also want to evaluate the augmentation of existing policy and governance frameworks by semantics specific to privacy and data protection (see mailing list archives on www.w3.org/Policy/pling/).

Figure 3 gives an example for a topic map on a high level where privacy preferences would be categorized according to the required topic map taxonomy. The ‘staying in touch with friends’ can be one topic. Its association with certain types of business networking activities could be that only non-sensitive personal data can be shared with individuals that belong to a business networking application. On the other hand, the occurrence of friends being on pictures from the last birthday party (a meaning the user would attach via a tag to each photo) may not be linked to the business networking topic so that these pictures do not get shared with business type of contacts.

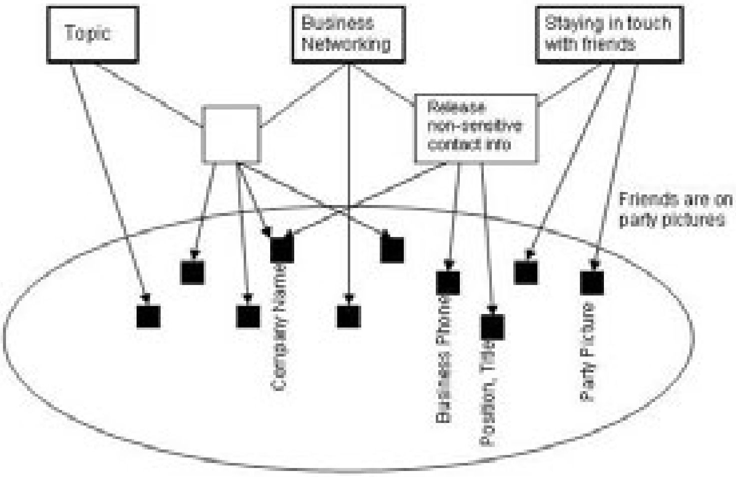
Topic
Business Networking
Staying in touch with friends
Release non-sensitive contact info
Friends are on party pictures
Company Name
Business Phone
Position, Title
Party Picture

Figure 3: High Level Topic Map for Online Social Networking

Each line depicting an association would carry along privacy preferences stating a rule for the purpose the data can be used for, the owner of the data and to what parties the data can be released or not. The use of privacy preferences in such a way would structure preferences in an aggregation, similar to what Jonathan Zittrain suggests adopting from the music industry’s structure where the use of aggregation of preferences is applied to the problem of ill-informed (or simply disinterested) customers [Zi00].

The topic map would encapsulate the privacy preferences scheme, capturing the relationships between the information and the grouping of the dos and don’ts of data processing. The topic map code would become an identifier for the context-driven privacy preferences. Therefore, the respective personal data could be processed according to a set of rules taking in consideration the different contexts.

Additionally, the reification functionality of the topic map would enable the application to reify the fact that the usage purpose for the respective data set was truly set by the original data owner. This gets more important in a Web 2.0 environment where, for example, a phone number of an individual could have been provided by someone other than the actual "owner" of the phone number. The topic map application would automatically check or reify if the originator of the content such as the phone number is in fact the same individual who has set the privacy preferences for the usage of the phone number.

7 Remaining Challenges

Semantic web technologies are relatively new concepts that have to gain ground but as illustrated here, it is worthwhile to evaluate how they could support parts of the privacy requirements. The concept of topic maps and its proper technical implementation for a privacy-enhanced SNA still needs to be carefully tested. However, the proposed solution mainly attempts to initiate a discussion among researchers and developers to extend current solutions with semantic techniques. In order to express privacy requirements in a web environment and especially for social networking applications, these technologies could be enhanced for utilising tagging and semantic metadata. The appropriate method for integrating these techniques into the current Web architecture and for enabling an interchangeable format across different applications for privacy preferences remains to be developed. In general, though, there is the notion that the current Web architecture needs to be extended to support transparency and accountability [WA07] and the concept of topic maps might be able to address at least those requirements.

Other semantic techniques such as microformats, structured tagging, or content labelling need to be researched on their fit for addressing privacy requirements. The same holds true for the interoperability of different web policy languages and their application for privacy issues. The W3C interest group on policy languages (PLING), for example, is discussing issues arising when different policy languages such as OASIS XACML (eXtensible Access Control Markup Language), IETF Common Policy, and P3P are used together. Other research underway currently addresses the interlinkability of semantic data (see for example SIOC, sioc-project.org) – an issue that also needs to address how privacy preferences can be imported and exported among mashed applications used more often in the social networking environment.

It does remain a challenge to start researching the adaptation of semantic technology concepts for building new classes of privacy-enhancing technologies for the Web while these concepts themselves are still in their infancy. Additionally, it is not clear today if these concepts will be scalable and cost-productive.

Before applications such as the described adoption of a topic map concept and other web techniques such as microformats and tagging functionalities get accepted as appropriate means to address information privacy requirements, the relevant privacy preferences for each personal data set in a specified context do not only have to be turned into a machine-readable format but also in what Wahlster and Dengel [Wa06] call a 'machine-understandable' description. One can think of a research project similar to 'Zonetag Photos from Yahoo!' where camera phones automatically attach tags with information on the location of the photo with the picture. In a privacy-related project similar to Zonetag, the privacy preferences information would be attached to each data set automatically according to a pre-set machine-readable and context-understandable policy or privacy profile of the user.

Finally, the world of Web platforms in the context of developing social networking applications is changing rapidly and it remains a challenge to make privacy a top agenda item for designers, developers, and providers of social networking applications while at the same time not loosing sight of new legal, social and technical developments.

8 Conclusion

The proposal to use semantic techniques such as the concept of topic maps for enabling a greater degree of information privacy and transparency when using social networking applications addresses the limited structure and control on the data level currently existing in these web applications, tries to reduce the technical complexity by using standardized and scalable technologies that web users are familiar with, and looks at solving the existing lack of transparency over the status of the user's information privacy.

Topic maps are all about structure and could represent and visualize the personally identifiable data of a user provided to various social networking platforms. The sample case for managing privacy preferences presented in this paper shows how such a structure or logical framework could enable the implementation of controls on the personal data level, thus, establishing the necessary accountability for authorized data usage. Topic maps are fairly simple constructs and are scalable in semantic web environments. Adapting and implementing them for enhancing the information privacy in SNAs would at least be able to address the provision of full transparency over the personal data being processed and would provide means to exercise choices on particular data sets.

The discussion among security and privacy experts on applying the concept of topic maps to the field of privacy-enhancing technologies should be expected to be very diverse and interesting. Yet, it is the time now to expand the scope of privacy-enhancing technologies from identity-related solutions to broader privacy principles such as enabling the purpose limitation and transparency needs while at the same time making use of new technical developments of the Semantic Web.

References

[AG06] Acquisti, A., Gross, R., Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, PET 2006.

[Ag02] Agrawal R., Kiernan J., Srikant R., Xu Y., Hippocratic Databases, 28th International Conference on Very Large Data Bases (VLDB), August 2002.

[BD05] Battle, L., Degler, D., Can Topic Maps describe context for enterprise-wide applications?, Extreme Markup Languages 2003®, Montréal, Québec.

[Be05] Bellovin, S., Clark, D., Perrig, A., Song, D. (Eds), Report of NSF Workshop on A Clean-Slate Design for the Next-Generation Secure Internet, GENI Design Document 05-05, July 2005.

[Be06] Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., Hendler, J., N3 Logic: A Logic for the Web (2006), http://dig.csail.mit.edu/2006/Papers/TPLP/n3logic-tplp.pdf.

[Co07] Coates, T., Yahoo! Tech Development, Designing for a Web of Data, Web 2.0 Expo Berlin, November 6, 2007.

[Da07] Davidson, M., Yoran, E., Enterprise Security for Web 2.0, IEEE Computer Society Magazine, IT Systems Perspectives, Pages 117-119, November 2007.

[En07] ENISA, European Network and Information Security Agency, Position Paper No.1, Security Issues and Recommendation for Online Social Networks, October 2007.

[Go02] Goldberg, I., Privacy-enhancing technologies for the Internet, II: Five years later, 2002.

[Go07] Google Press Release, Google Launches OpenSocial to Spread Social Applications across the Web, November 1, 2007, http://www.google.com/intl/en/press/pressrel/opensocial.html.

[Li03] Lindskog, H. and S., Website Privacy with P3P, Chapter 5 – Platform for Privacy Preferences Project, Pages 55f., Wiley Publishing, 2003.

[Ni04] Nivargi, P., A Generic Privacy Model for Data Access Using Semantic Web Technologies, Arizona State University, December 2004.

[On07] Ontopia Website, Description of Topic Maps, Topic Maps: The GPS of the Web, http://www.ontopia.net/topicmaps/what.html.

[Pe00] Pepper, S., The TAO of Topic Maps – Finding the way in the age of infoglut, Ontopia AS, 2000, http://www.ontopia.net/topicmaps/materials/tao.html.

[Re07] Recordon, D., Six Apart, Opening the Social Graph, Web 2.0 Expo Berlin, November 6, 2007.

[Vo07] Vossen, G. and Hagemann, S., Unleashing Web 2.0: from concepts to creativity, Morgan Kaufmann Publishers, p. 335, 2007.

[Wa06] Wahlster, W. and Dengel, A. (2006), Web 3.0: Convergence of Web 2.0 and the Semantic Web, Technology Radar Feature Paper Edition II/2006, Deutsche Telekom Laboratories, 2006.

[We07] Weiss, S., The Need for a Paradigm Shift in Addressing Privacy Risks in Social Networking Applications, Post-Proceedings: The Future of Identity in the Information Society ; Third International Summer School organized by IFIP WG 9.2, 9.6/11.7, 11.6 in cooperation with FIDIS Network of Excellence, Karlstad (Sweden) 2007.

[WA07] Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., Sussman, G.: Information Accountability, Computer Science and Artificial Intelligence Laboratory Technical Report, MIT, June 13, 2007.

[Zi00] Zittrain, J., What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Era of Trusted Privication, February 2000, Stanford Law Review, Vol. 52, Available at SSRN: http://ssrn.com/abstract=214468 or DOI: 10.2139/ssrn.214468.