Personal Linked Data: A Solution to Manage User's Privacy on the Web

Patricia Serrano-Alvarado, Emmanuel Desmontils

Personal Linked Data: A Solution to Manage User's Privacy on the Web

Patricia Serrano-Alvarado and Emmanuel Desmontils

LINA, Université de Nantes,
2, Rue de la Houssinière BP 92208
44322 Nantes Cedex 03 France
Name.LastName@univ-nantes.fr
http://www.lina.univ-nantes.fr/

Abstract. While using modern applications, personal digital data is spread over hundreds of servers all around the world and users have very poor control over these data. To tackle this issue, based on the semantic Web, we are developing a framework, named Privacy-Lookout, to allow people to be on the lookout for transgressions of their personal data privacy. Concretely, we propose to construct a *personal linked data view* of individuals to organize and semantically enrich the meta information of their personal data existing in the Web. The mean idea is to allow users to know if the information the Web posses about them respects their privacy principles. This paper introduces the first ideas of such approach.

Keywords: Data privacy control, Linked Data, Semantic Web

1 Motivation and objectives

Nowadays, all aspects of individuals' life, either private, professional or public, are massively digitized. Personal digital data is spread over hundreds of servers all around the world and users have absolutely no control over their data. In general, servers collect, process, share and trade personal data with or without the explicit consent of concerned persons. Collected data is stored forever and users forget or are unaware of the huge traces the digital world has about them. Fortunately, more and more people is being conscious about the importance of safeguarding their privacy and controlling their personal data while using modern applications. But lot of work left to do, meanly among young or misinformed people and over all general tools allowing people to control their digital footprint need to be developed [7,20].

For the European Union, protecting personal data is a fundamental right and the free flow of personal data is considered a common good. Aiming at reinforcing individuals' rights, the European Commission recently proposed to update and modernize the European Privacy Guidelines [10,15]. Among others, these principles mention "personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes", "such data should not be processed, unless the data subject gives his explicit consent" and "data subjects should have the right that their personal data are erased and no longer processed, where the data are no longer necessary in relation to the purposes for which the data are collected...". Hence, European guidelines support individuals' concerns about their data privacy. And it is clear that individuals should be able to protect their personal data to reuse it and to control somehow their usage, storage and sharing.

Several current projects and tools go in that direction like MesInfos[1], MiData[2] and Give Me My Data[3] that allow people to obtain and use their personal data (retained by companies) in a way that is portable. In general, Web applications provide users more or less clear and easy to use tools to manage their personal data. For instance the dashboard of Google[4], where users can manage a tiny portion of the iceberg peak of the data that google retains about them. But a general approach allowing users to know and manage the privacy of their data does not exist.

In order to tackle this problem, the objective of this abstract is to discuss a general approach, based on the *semantic Web*, that empowers users with the ability of possessing and managing a detailed *view* of their data privacy in the Web. The questions we aim at answering with this approach are: According to their rights, how users can

[1] http://fing.org/?MesInfos-Experimenting-the-Sharing
[2] http://www.thebigdatainsightgroup.com/site/article/big-data-notes-002-midata
[3] http://givememydata.com/
[4] https://www.google.com/dashboard/

define their personal wishes about the privacy policies for their data? How users can know which are the current privacy policies of their data existing in the Web? How users can detect misuse, misinformation, inaccuracy or transgression of data protection rules of their data in the Web automatically? How users can give a feedback or a report to data storers/servers when data privacy problems arise?

Answering these questions is a great challenge because of several reasons. On the one hand, lot of information about privacy policies is missing or existing in heterogeneous formats. Besides, privacy policies are hard to locate automatically. On the other hand, there is no global solution for individuals to define their personal privacy policies, i.e., their needs and wishes about the privacy policies of their data. Consequently, it is hard to define an automatic verification of the consistency between the privacy policies governing individuals' data and their privacy wishes.

This paper is organized as follows. Section 2 introduces the approach we propose to allow people to be on the lookout for transgressions of their personal data privacy. Section 3 shows first ideas for an implementation of such approach and Section 4 concludes and gives future work.

2 Privacy-Lookout: our general approach

Semantic Web [4,3,16,12] helps understanding the resources existing on the Web and facilitates the automatization of any process. Semantics provides well-defined meanings that can be used to better understand data. For instance, a picture can be *enriched* with a description of the person appearing in it, the person who took the picture, the date when it was taken, the place, etc. Other important semantic information can be attached like the privacy rules over the picture, e.g., the license under which it may be reused.

RDF (Resource Description Framework[5]) is a W3C recommendation that proposes concepts that allow to describe data semantics. A description is a triple composed of a subject, a predicate and an object. To have proper RDF triples it is necessary to use unique identifiers (i.e., URIs). To be useful, at least the subject of the triple should be uniquely identified. To simplify the understanding of triples they can be seen as an instance identifier, a property name, and a property value. For example, the RDF triple describing who took a picture can be represented simply by:

```
Patty creator stephPicture
```

The person who took the picture is identified by http://fooenterprise.com/patty/. Concerning stephPicture, hundreds of objects named stephPicture may exist in the world. That is why it is important to have unique identifiers for objects too. In our description, stephPicture is identified by the following URI http://foopersonalsite.net/steph/pictures/stephpicture.jpg. Concerning the property name, to automatically allow processes to understand what the property is about, it should come from a well-defined vocabulary that frequently is organized in *ontologies*. A commonly used ontology to describe persons is FOAF [8] and the ontology we use in our triple is Dublin Core[6] [19,13], which was defined to describe published documents. The URI identifying the property creator as proposed by the Dublin Core ontology is `http://purl.org/dc/terms/creator`.

If we continue with our example, the semantic information concerning the data privacy of stephPicture can be described by the triple:

```
stephPicture license CC_BY_NC_ND_1
```

Here, license is a property taken from the Dublin Core ontology and identified by the URI `http://purl.org/dc/terms/license`. The property value is a Creative Commons (CC) license specifying that the picture should not be modified nor used for commercial purposes. The ontology used for this license is taken from SemanticCopyright[7] and the corresponding URI is `http://www.semanticcopyright.org/files/basic-owl/html-doc/instances/basic-CC-BY-NC-ND_1.html`.

Figure 1 illustrates our example[8]. Identifiers come from different domains (foopersonalsite.net, fooenterprise.com, purl.org, etc.), properties come from different ontologies (*dc* for Dublic Core, *foaf* for FOAF and *ex* for our current example), and the triples describing data semantics *link* them constructing a meaningful graph. Links among domains raise the *Linked Data*.

[5] `http://www.w3.org/RDF/`

[6] `http://dublincore.org/documents/dces/`

[7] `http://semanticcopyright.org/`

[8] The complete description of this example can be found at `http://pagesperso.lina.univ-nantes.fr/~serrano-p/stephpicture.htm`

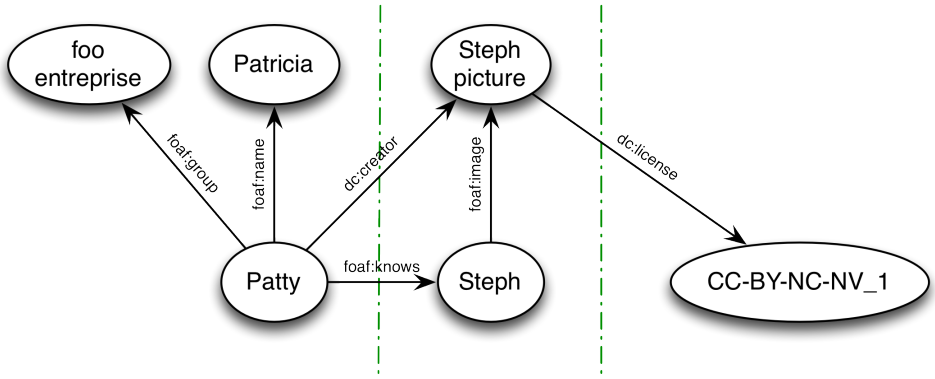

Fig. 1. A simple example of Linked Data

The Linked Data [5,9] is the evolution of a global information space of linked documents (i.e., the Web) to one where both documents and data are linked. It is a set of best practices for publishing and connecting structured data on the Web. The increasingly adoption of the Linked Data best practices has lead to the extension of the Web with a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programs, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews. This Web of Data enables new types of application, we adopt it to empower people with tools to better protect their data privacy.

In our example, Steph's picture provides enough semantic information to protect his privacy, i.e., the license of the picture is clearly identified, Steph knows who took the picture, and knows that he appears in this picture. If he wants to suppress the picture he can do it because it is stored in his personal Web page. If he wants to change the picture's policy he can do it as well. If he detects that somebody does not respect the picture's license he can undertake legal procedures. So, Steph has total control on his picture. But what about all other Steph's data that is out his control? How can he know which is the data concerning him that may affect his privacy and that is spread all over the digital world? How can he control (at least know) his privacy rights concerning this data?

Based on the semantic Web, we are developing a framework, named Privacy-Lookout, to allow people to be on the lookout for transgressions of their personal data privacy. Concretely, we propose to construct a *personal linked data view* of individuals that will organize and semantically enrich the meta information of their personal data existing in the Web. Such meta information will focus on who stores these information, for what purposes, under which privacy policies, etc. The mean idea is to allow users to know if the information the Web posses about them respects their privacy principles. If they figure out their privacy is being transgressed, supported by existing regulations, they will be able to react and make their privacy rights be enforced.

Figure 2 shows the Privacy-Lookout process. Roughly speaking, we have an Entity that can be an individual or, to generalize, a legal Entity. From a set of queries submitted to the Web and to the Linked Data spaces, Privacy-Lookout obtains the meta information of scattered data may concern the specified Entity. These meta information is filtered and semantically enriched to obtain a right set of meta information that concerns the Entity. Then, the corresponding privacy policies are detected (if they exist), semantically enriched and linked to right data. This is what we call the *resource and privacy detection* process that produces the Entity's linked data view.

The *resource analysis* process considers a knowledge base (or corpus) composed of the Entity's known resources and the privacy policies the Entity wishes to protect its data. This corpus will be the data privacy principles of the Entity that should be well structured and semantically specified. The mean objective of the resources analysis process is to construct the Entity's semantic corpus and to confront it with the Entity's linked data view. This confrontation may be defined to detect any kind of privacy problem related with Entity's data, for instance detecting misuse, misinformation, inaccuracy or transgression of its data protection rules.

Next section discuses in more detail Figure 2 and some ideas of implementation.

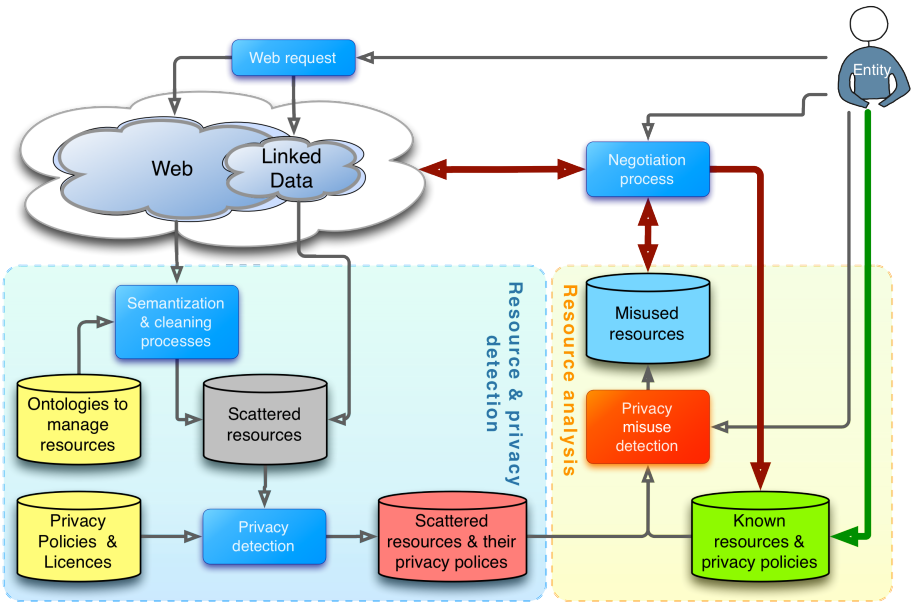

Fig. 2. Process of Privacy-Lookout

3 Towards a semi-automatic process

Privacy-Lookout needs several complementary process, some of them concern technical issues but other are open research problems.

In the resources and privacy detection process, to obtain scattered data from the Web, all available APIs can be used. APIs of well known search engines like Google and Yahoo can be used but also APIs of Web services like Graph API[9] of Facebook or the twitter API[10]. The Linked Data can be interrogated through semantic web engines like Sig.ma[11] and OpenLinkDataExplorer[12], and well known stores like DBpedia [2] and Freebase [6] can be also directly accessed. Indeed, all kind of API that allow to locate Entity's resources in the Web and Linked Data spaces are useful. The challenge here is to deal efficiently with the heterogeneity of sources.

To detect right Entity's data, a filtering process is made based on the corpus of the Entity. For instance, the Entity's corpus may contain a picture where the subject is the Entity. This picture will be the base of a comparison to decide if a multimedia resource corresponds to the Entity.

Then, from right obtained resources, the personal linked data is constructed (e.g., RDF Triple store) for the concerned Entity by adding semantics to every resource. The more semantics has a resource the more it is meaningful and exploitable. Web resources in general have no semantic information, nevertheless in some cases, some semantics is attached to the (X)HTML Web page containing the resource through *microdata* (like with Schema.org[13]) or *microformats*. To add semantics to resources, pertinent ontologies can be used (RDF-Wordnet [18], DBPedia [2], Dublin Core [19], Yago [17], OpenCYC [14], etc.) but in most cases, it is necessary to use a concept extractor like Open Calais[14]). Web results are merged with the results obtained from the Linked Data to build a uniform and homogeneous data set of scattered resources as shown in Figure 2.

The cornerstone process that completes the personal linked data view is *privacy detection*. This process detects the privacy conditions regulating Entity's resources. It is a complex process because there is no consensus about how to specify privacy policies, and over all, not all resources have a privacy policy attached. If a resource is attached to a privacy policy, it can be done in several ways: implicit, semi-implicit or explicit.

[9] `https://developers.facebook.com/docs/reference/api/`
[10] `https://dev.twitter.com/docs/api/1.1`
[11] `http://www.w3.org/2001/sw/wiki/Sig.ma`
[12] `http://www.w3.org/2001/sw/wiki/OpenLinkDataExplorer`
[13] `http://schema.org/`
[14] `http://www.opencalais.com/`

- Implicit. An implicit privacy policy is when the domain of the resource is clearly related to well-known privacy conditions. For instance, all pages of Wikipedia are under a Creative Common License CC-by-sa [21]) and all data published in Facebook are under the Facebook's license[15].
- Explicit. For us, an explicit privacy policy is the one that is specified semantically. It can be done by using an ontology in the Linked Data space, like the ontology of Creative Common [1][16]), or microdata/microformat for Web pages.
- Semi-implicit. It corresponds to a standard sentence in natural language like "©John Doe" or images for licenses like OBbL[17], OL[18], CC-by[19], CC-0 [20], etc.

Then, based on dedicated ontologies for privacy (e.g. SemanticCopyright[21] or CopyrightOnto [11]), each resource is attached to the detected privacy policy. As a result, we obtain the personal linked data view enriched with corresponding privacy policies or licenses.

In the resource analysis process, the personal linked data view is confronted to the Entity's data privacy principles. These principles are constructed following the wishes of the Entity and based on privacy regulations. They will be described also using the semantic Web techniques. If the Entity knows its resources (or some of them), it attaches the privacy policies it considers should govern them. Constructing and organizing the Entity's corpus (semi)automatically is challenging. Currently there is no approach allowing individuals to specify globally their data privacy policies. For instance, it should be possible for an Entity to specify that all pictures where he appears should be at most under a particular privacy policy, and pictures where its children appear under another one, more restrictive. Thus, it should be able to define privacy rules to express privacy by data type. But if some particularities should be considered, the Entity should manually refine the privacy for specific data.

The *privacy misuse detection* process will detect resources that do not preserve (intentionally or not) Entity's privacy. Defining the misuse detection process is also challenging, it should detect and analyze the coherence between the personal linked data view and the Entity's data privacy principles. The goal is to arise several kinds of problems like misuse, misinformation, inaccuracy or data privacy transgressions. Techniques to manage equivalence of privacy policies should be employed. Covering rules will be necessary also to detect for example that there is no problem if a more restrictive privacy policy currently regulates a resource compared to the Entity's data privacy policy. Finally, once all privacy problems are detected, the Entity can enter into a negotiation process with corresponding resource managers.

4 Conclusion and ongoing work

This paper introduced the first ideas of Privacy-Lookout, a framework to empower individuals with a tool to be aware of their data privacy. The proposed approach aims at being general but for the instance we focus on the Web of Data. Hence, supported by the semantic Web and the Linked Data, we are exploring the feasibility of our ideas.

The general process of our approach is made in two steps, first a local linked data view of resources concerning a person (or to generalize an Entity) is constructed, then it is confronted to the Entity's wishes to protect its data. Three challenging issues are faced, the first one is constructing efficiently an homogeneous and semantically rich Entity's linked data view, the second one is constructing the global Entity's data privacy policies and the third one is detecting and analyzing inconsistencies over these two data sets.

[15] https://www.facebook.com/about/privacy/
[16] http://creativecommons.org/ns
[17] http://opendatacommons.org/licenses/odbl/
[18] http://www.data.gouv.fr/Licence-Ouverte-Open-Licence
[19] http://creativecommons.org/licenses/by/3.0/
[20] http://creativecommons.org/publicdomain/zero/1.0/
[21] http://semanticcopyright.org/index.php/ontology

References

1. Abelson, H., Adiba, B., Linksvayer, M., Yergler, N.: ccREL: The Creative Commons Rights Expression Language (mar 2008), http://wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007), http://dbpedia.org/
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American pp. 35–43 (May 2001)
4. Berners-Lee, T.: Semantic Web Road Map (1998), http://www.w3.org/DesignIssues/Semantic.html
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data-The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
6. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM (2008)
7. Bouguettaya, A., Eltoweissy, M.: Privacy on the Web: facts, challenges, and solutions. Security Privacy, IEEE 1(6), 40–49 (2003)
8. Brickley, D., Miller, L.: FOAF vocabulary specification 0.98. Namespace Document 9 (2010), http://xmlns.com/foaf/spec/
9. Cyganiak, R., Jentzsch, A.: Linking open data cloud diagram. LOD Community (2011), http://lod-cloud.net/
10. European Commission: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (2012), http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf
11. Garcia González, R.: A semantic web approach to digital rights management. Universitat Pompeu Fabra (2006), http://rhizomik.net/html/ontologies/copyrightonto/
12. Hendler, J., Berners-Lee, T.: From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence 174(2), 156–161 (2010)
13. Initiative, D.C.M., et al.: Dublin Core Metadata Element Set, version 1.1 (2008), http://dublincore.org/
14. Matuszek, C., Cabral, J., Witbrock, M., Deoliveira, J.: An introduction to the syntax and content of Cyc. In: Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. vol. 3864, pp. 44–49 (2006), http://www.cyc.com/platform/opencyc
15. Reding, V.: The EU Data Protection Reform 2012: Making Europe the Standard Setter for Modern Data Protection Rules in the Digital Age 5 (Jan 2012), http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/26&format=PDF
16. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. Intelligent Systems, IEEE 21(3), 96–101 (2006)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th international World Wide Web conference (WWW'2007). pp. 697–706. ACM Press, New York, NY, USA (2007), http://www.mpi-inf.mpg.de/yago-naga/yago/
18. Van Assem, M., Gangemi, A., Schreiber, G.: Conversion of WordNet to a standard RDF/OWL representation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. pp. 237–242 (2006), http://www.w3.org/2006/03/wn/wn20/
19. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin Core Metadata for Resource Discovery. Internet Engineering Task Force RFC 2413, 222 (1998), http://dublincore.org/
20. Weitzner, D.J.: Beyond secrecy: New privacy protection strategies for open information spaces. Internet Computing, IEEE 11(5), 94–96 (2007)
21. Wikipédia: Wikipédia:Citation et réutilisation du contenu de Wikipédia — Wikipédia, l'encyclopédie libre (2013), http://fr.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Citation_et_r%C3%A9utilisation_du_contenu_de_Wikip%C3%A9dia&oldid=89798151, [En ligne; Page disponible le 3-mai-2013]