The Need for Data Sharing Agreements in Data Management

George Konstantinidis

University of Southampton, UK
g.konstantinidis@soton.ac.uk

Abstract. There is an evidently growing legal, cultural and technological need for tools and models that allow users to express their own intentions and consent over the usage of their personal data and information. Service providers and institutions that manage personal data rely on specifying monolithic "Terms and Conditions" written in natural language and enforced in an ad-hoc manner, by presenting users with top-down, coarse-grained, opt-in/out options. We advocate the need for users to describe their personal contract of data usage in a formal, machineprocessable language. Semantic Web technologies can have a central role in this approach by providing the formal tools and languages required. Expressing data sharing intentions, consent and data usage agreements in a technical way enables the development of algorithms that automatically respect a user's policy. This helps organisations increase technological capabilities, abide by legal requirements, and avoid ad-hoc processes, thus saving engineering resources.

1 The need for machine processable data sharing agreements

In this paper we advocate the need for machine processable data sharing agreements which we believe will prove particularly valuable to the health data management domain.

Traditional data privacy approaches such as privacy[4], k-anonymity[11], or l-diversity[8], provide top-down mechanisms to specify global, authoritative, coarse-grained privacy policies. These problems as studied in the Web, A.I., Databases, and Knowledge Representation communities thus far have focused on ensuring confidentiality of individuals' identity while releasing data, by either masking, suppressing, altering or inserting noise to data in order to achieve de-indentification. On the other hand, access control approaches, including the most common role-based [10] and attribute-based [5] approaches, aim to maintain confidentiality of private information by completely disallowing access to certain data, based on roles/attributes of users and purposes.

All these technologies consider protection against a non-trusted party and as such they aim to directly limit access to data, or hide the identity of the individual, rather than provide a contract of access that can be used also for later or indirect use by a relatively trusted party.

The latter idea has been limitedly studied in the context of privacy languages for the World Wide Web, with most notable the now deprecated example of the P3P language [13]. P3P alloweded website owners to specify coarse-grained policies, using predefined options, regarding usage of the clients' data; web clients/browsers would also specify their preferences in a similar way and automatically accept or reject visiting a page depending on a match between the policies. Such coarse-grained predefined options, and the fact that these languages mostly provide "accept/reject" policies and don't offer flexibility for partial access has led almost to an abandonment of machine processable efforts to specify access and usage contracts, leaving data owners and service providers to rely exclusively on legal, natural language, "Terms of Use" agreements.

We claim the need to develop the theory, algorithms and implementations of expressing, supporting and managing fine-grained intentions, or contracts of data access and data usage, as well as personal user consent. To achieve that, we believe that systems need to pursue the following objectives:

1. Create a bottom-up setting where individuals and organisations can create data sharing policies backed-up by a formal machine-processable technical language. This is in contrast to the classic service-client data sharing model where clients commit their data to a service provider after being presented with coarse-grained "Terms and Conditions" written in natural language, and on which the clients get only a few and simple opt-in/opt-out options.
2. Enable rich and more dynamic expressions of access and usage policies. Current access control systems are based on a predefined set of static rules in order to prohibit access to certain fields of a data repository; these access control rules cannot express sharing contracts such as a policy that prohibits a data item to be shared depending on its history of sharing, or on the amount of data already given to the data requester.
3. Develop algorithms that will allow data requesters to obtain the maximal result set of the query that still abides by the contracts of the data owners.
4. Support the goal of data auditing [3] which is to determine if private information was disclosed in answering queries, as well as the goal of accountability [15] which aims to understand the responsibilities of different parties in data processing. Data sharing agreements would be central in data auditing and accountability scenarios, since they inform exactly how to (or not to) use a particular dataset.

2 The Role of the Semantic Web

We believe that semantic web technologies are inherently suitable to serve the role of providing the common shared vocabularies for data sharing intentions and agreements, together with the algorithmic machinery that is needed to process these agreements.

Building on top of more general schemas such as FOAF, and schema.org that can be used to describe persons and personal data, or DICOM [1] for modeling healthcare and medical imaging metadata, there are several ongoing approaches using knowledge graphs to express aspects of data sharing agreements. The Open Digital Rights Language (ODRL) [12] is a policy expression language that models content, services, actions, prohibitions, and obligations. PROV-O [14] models provenance information generated in different systems and under different contexts. More interestingly, the SPECIAL [2] provides a vocabulary for expressing consent together with data processing workflows which take such consent into account, while in [7] the authors develop an ontology that models privacy policies described in actual medical research data sharing agreements.

These knowledge graphs are great steps towards a vision where users or parties encode their preferences and intentions of data usage in a machine processable way and data processing algorithms automatically respect these preferences. In order to achieve this, the developed vocabularies have to be backed by the development of generic and re-applicable algorithms; possibly borrowing from data integration [6] or ontology based query answering [9].

References

1. Digital imaging and communications in medicine, https://www.dicomstandard.org/
2. The special usage policy language, https://aic.ai.wu.ac.at/qadlod/policyLanguage/
3. Agrawal, R., Bayardo, R., Faloutsos, C., Kiernan, J., Rantzau, R., Srikant, R.: Auditing compliance with a hippocratic database. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. pp. 516–527. VLDB '04 (2004)
4. Dwork, C.: Differential privacy: A survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) Theory and Applications of Models of Computation. pp. 1–19. Springer Berlin Heidelberg (2008)
5. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM conference on Computer and communications security. pp. 89–98. Acm (2006)
6. Konstantinidis, G., Ambite, J.L.: Scalable query rewriting: a graph-based approach. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 97–108. Athens, Greece (2011)
7. Li, M., Samani, R.: Dsap: Data sharing agreement privacy ontology. In: Semantic Web Applications and Tools for Healthcare and Life Sciences (2018)
8. Machanavajjhala, A., Venkitasubramaniam, M., Kifer, D., Gehrke, J.: l-diversity: Privacy beyond k-anonymity. In: ICDE (2006)
9. Perez-Urbina, H., Rodrıguez-Dıaz, E., Grove, M., Konstantinidis, G., Sirin, E.: Evaluation of query rewriting approaches for owl 2. In: Joint Workshop on Scalable and High-Performance Semantic Web Systems (SSWS+ HPCSW 2012). p. 32

10. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based access control models. Computer 29(2), 38–47 (1996)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
12. W3C: Odrl information model 2.2, https://www.w3.org/TR/odrl-model/
13. W3C: The platform for privacy preferences 1.0, https://www.w3.org/TR/P3P/
14. W3C: Prov-o: The prov ontology, https://www.w3.org/TR/prov-o/
15. Weitzner, D.J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., Sussman, G.J.: Information accountability. Comm. of the ACM 51(6), 82 (2008)