A SPECIAL vision for enabling Privacy through Linked Data

Authors (in alphabetical order):

- Piero Bonatti CerICT
- Sabrina Kirrane, WU (Wirschaftsuniversität Wien) - Vienna University of Economics and Business
- Axel Polleres, WU (Wirschaftsuniversität Wien) - Vienna University of Economics and Business
- Simon Steyskal, WU (Wirschaftsuniversität Wien) - Vienna University of Economics and Business
- Rigo Wenning, ERCIM

Abstract

This joint position statement by WU (Wirschaftsuniversität Wien) - Vienna University of Economics and Business (full W3C member), ERCIM (W3C Host), and CerICT outlines our vision for requirements and needed best practices to align and complement existing Linked Data vocabularies to deal with challenges arising through the new European General Data Protection (GDPR). We particularly focus on requirements regarding (i) consent management and (ii) keeping transparency records about personal data processing.

Background in Privacy, Linked Data and Web technologies

The Semantic Web and Linked Data activities initiated by W3C provide a rich toolbox of standards for interoperability and management of heterogeneous data from various domains. Consent records and transparency records for personal data usage stored as Linked Data additionally promise to seemlessly

1. integrate personal data and transparency records about interactions with different data processors from the user's perspective and interchange such records, as well as,
2. verify data usage is complying with given consent, if consented policies are described semantically using interchangeable vocabularies with a formal semantics.

Potential for improvement of interoperability

In using and applying the toolbox that Semantic Web and Linked Data Standards offer, we see the following main challenges:

- **Lack of integrated core vocabularies** for expressing personal data handling and consent.
- **Lack of formal semantics** to check compliance of personal data handling with given consent.

As secondary challenges (which we consider out of scope for the present position paper, but also relevant), we consider harnessing **standard architectures** to support these standard vocabularies and compliance checking natively in a scalable manner.

In the following, we sketch existing starting points for addressing both of the main challenges, that in our opinion deserve discussion at the workshop and further on in joint standardisation effort(s), as well as our plan to address the third challenge within the ="http://specialprivacy.eu">SPECIAL project. The SPECIAL (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance) project is a Research and Innovation Action funded under the H2020-ICT-2016-1 Big Data PPP call (Privacy-preserving Big Data technologies, ICT-18-2016).

Core Vocabularies

In order to keep transparency records about (i) consent collection, as well as (ii) personal data processing in accordandance with the GDPR, we deem it necessary to find agreement on how such records shall be recorded in an interchangeable manner. Additionally, in order to fulfill (iii) the data subjects' Right to Access to a copy of their personal data in an electronic format, also agreement on the format of this data would be desirable, where possible.

To this end, we propose to find agreement on *core Linked Data vocabularies*, or respectively, the agreement on best practices to reuse existing such vocabularies in order to model and exchange the relevant information about (i)-(iii). To our understanding, these vocabularies need to cover the core elements in Figure 1, related to data usage policies.

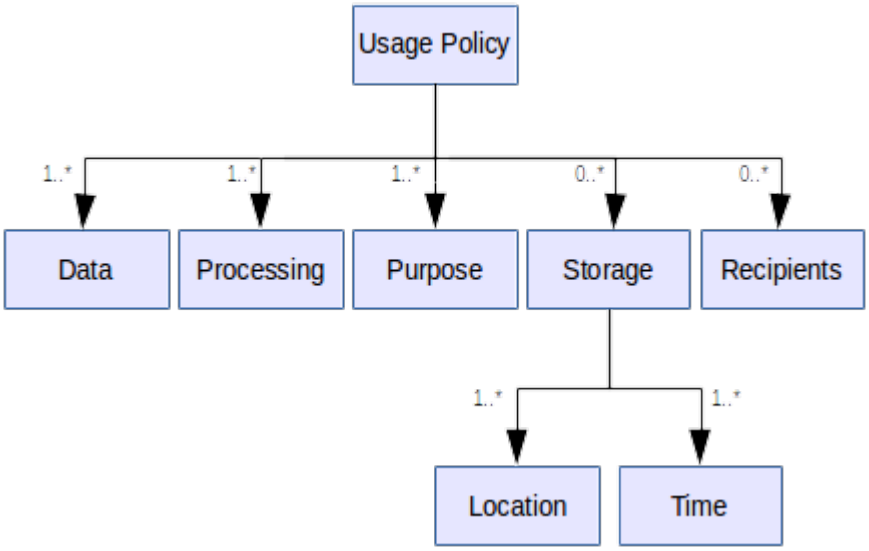

Figure 1: The minimum, core usage policy model (MCM) elements [SPECIAL D1.3].

That is, we need to enable modeling of the following key aspects:

1. **Data** describes the type of data processed. This including "static" personal data (such as identifiers or attributes by which an individual can be identified directly or indirectly, such as name, personal ID number, online identifiers such as nick- or username, home or work address, etc.), as well as "dynamic", context-dependent personal attributes or events related to an individual (e.g.: location of a person, heartrate at a certain point in time, participation at an event, visiting a Website, but also e.g. being assigned a certain IP address by a provider, etc.). Moreover, we assume that a taxonomic categorization of such data will be useful to classify such data.
2. **Processing** describes operations that are performed on personal data. Asprects related to processing include applying algorithms to the data, such as aggregation, anonymisation, as well as properties of the outcomes of the algorithm, such as anonymity metrics, as described in [SPECIAL D1.3].

3. **Purpose** specifies the objective of such processing, e.g. for marketing purposes.
4. **Storage** specifies where data are stored and for how long in the context of the specified processing.
5. **Recipients** specifies who is going to receive the results of data processing and, as a special case, whom data are shared with; that is, recipients need to be categorised in at least data subjects, data processors, data controllers, but also regulators, or â€“ as mentioned in [SPECIAL D1.3], business partners, etc.

Additionally, a core vocabulary will also need to express mechanism and concepts related to recording and controlling consent, including conditions defining "restriction of processing" (as per article 4 of the GDPR), which again could be tied to (possibly complex) policies. In [SPECIAL D6.3] we list and examplify a number of existing vocabularies that could in our opinion serve as starting point to model these aspects, such as for instance:

1. the FOAF, vCard and likewise schema.org offer vocabularies for modeling static personal data to describe persons, agents.
2. the DICOM ontology for healthcare metadata has dynamic attributes relevant to fitness and health, such as HeartRate; likewise, existing sich as the NeoGeo vocabulary, the GeoSPARQL vocabulary or the WGS84 Geo Positioning vocabulary ..
3. the P3P working group published an RDF vocabulary for expressing the concepts of P3P including a categorization of personal data and purposes.
4. the ODRL vocabulary provides means to model actions, prohibitions, and obligation to describe policies and consent.
5. the OWL Time Vocabulary provides means to describe the time and duraciton of processing, duration of consent, logging events.
6. finally, the PROV vocabulary provides an excellent starting point for modeling the provenance of consent and data processing actions.

Formal Semantics for Polcies and Automated Reasoning for Compliance Checking

In order to reason about whether (transparently stored) data processing records comply with data subjects' given consent policies. To this end we suggest to enable (ideally polynomial, scalable) OWL reasoning to decide whether the *authorized operations* specified by a data subject through their given consent, subsume the specific data processing records in the transpaency log. In [SPECIAL D2.1] we therefor specify *basic usabe policies* as OWL classes pf objects that model the aspects of the MCM (cf. Figure 1).

```
ObjectIntersectionOf(
   ObjectSomeValuesFrom(spl:hasData SomeDataCategory)
   ObjectSomeValuesFrom(spl:hasProcessing SomeProcessing)
   ObjectSomeValuesFrom(spl:hasPurpose SomePurpose)
   ObjectSomeValuesFrom(spl:hasRecipient SomeRecipient)
   ObjectSomeValuesFrom(spl:hasStorage SomeStorage)
)
```

where such a policy models that permission to *SomeProcessing* of *SomeDataCategory* for *SomePurpose* has been given to *SomeRecipient* in compliance with *SomeStorage* restrictions. Given that the respective attributes are described by respective class descriptions in OWL again, the usage policy adopted by the data controller (call it P_c) complies with the usage policy in the data subject's consent (call it P_s) if and only if all the authorizations in P_c are also authorized by P_s, that is, P_c complies with P_s if and only if $P_c \subset P_s$, which can be checked by an OWL reasoner. As a forthcoming task for formal reasoning, not only checking compliance of processing with consent, but also checking compliance with obligations and permissions

derived from the GDPR legislation itself may be necessary. To this end, we have collected first starting points on formalizing the general policies encoded in the GDPR itself as logical axioms in [SPECIAL D2.2]. We acknowledge that due to the deliberate ambiguity of legal text, such compliance checking can not be fully automated.

Standard architectures

Eventually, the SPECIAL project shall deliver a scalable reusable architecture which allows to log and process consent, transparency logs about personal data processing in a scalable manner and automatically check compliance as outlined above. We are currently discussing several architectural solutions for implementation of the transparency ledger [TELERISE] and deployment in different complementary use cases by Thomson Reuters, Deutsche Telekom AG, and Proximus.

Conclusions

Agreed core vocabularies to model aspects of consent and transparent personal date processing will be a key to a reusable architecture as we plan it in the SPECIAL project. We therefore look forward to present and discuss the aspects outlined above in the workshop. In order to reach agreement, out goal of the outcome of the workshop is the foundation of a W3C community group around standard vocabularies and linked-data based architectures to scalably maintain and interchange personal data processing and transparency records. The authors and their respective institutions are committed to participate in and contribute to such a community group.

References

[SPECIAL D1.3] P. A. Bonatti, S. Kirrane, R. Wenning *SPECIAL - Deliverable D1.3: Policy, transparency and compliance guidelines V1.*. 1 September 2017, URL: https://www.specialprivacy.eu/images/documents/SPECIAL_D1.3_M8_V1.0.pdf

[SPECIAL D6.3] A. Polleres, S. Kirrane, R. Wenning. *SPECIAL - Deliverable 6.3: Plan for community group and standardisation contribution.*. 30 September 2017, URL: https://www.specialprivacy.eu/images/documents/SPECIAL_D6.3_M9_V1.0.pdf

[SPECIAL D2.1] P.A. Bonatti, S. Kirrane, I. Petrova, L. Sauro, E. Schlehahn. *SPECIAL - Deliverable 2.1: Policy Language V1.*. 26 December 2017, URL: https://www.specialprivacy.eu/images/documents/SPECIAL_D2.1_M12_V1.0.pdf

[SPECIAL D2.2] P.A. Bonatti, S. Kirrane, I. Petrova, L. Sauro, C. Kerschbaum, E. Schlehahn, R. Wenning *SPECIAL - Deliverable D2.2: Formal representation of the legislation V1.*. 26 December 2017, URL: https://www.specialprivacy.eu/images/documents/SPECIAL_D2.2_M12_V1.0.pdf

[TELERISE] P.A. Bonatti, S. Kirrane, A. Polleres, and R. Wenning. *Transparent personal data processing: The road ahead*. In TELERISE: 3rd International Workshop on TEchnical and LEgal aspects of data pRIvacy and SEcurity @ SAFECOMP2017, volume 10489 of Lecture Notes in Computer Science (LNCS), pages 337--349, Trento, Italy, September 2017.

Last modified: Wed Feb 28 11:21:57 CET 2018