Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 262–268

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

Queryable Provenance Metadata For GDPR Compliance

Harshvardhan J. Pandit*, Declan O'Sullivan, Dave Lewis

ADAPT Centre, Trinity College Dublin, Dublin, Ireland

Abstract

Information associated with regulatory compliance is often siloed as legal documentation that is not suitable for querying or reuse. Utilising open standards and technologies to represent and query this information can facilitate interoperability between stakeholders and assist in the task of maintaining as well as demonstrating compliance. In this paper, we show how semantic web technologies can assist in representation and querying of compliance information related to the General Data Protection Regulation (GDPR), an European law governing the use of consent and personal data. We focus on the subset of obligations related to the use of consent and personal data, and represent the associated metadata using the previously published GDPRov ontology and GDPRtEXT resource. We present a proof-of-concept demonstration (available online) where information is queried to automatically populate the GDPR-readiness checklist published by Ireland's Data Protection Commissioner.

Keywords: GDPR; regulatory compliance; semantic web; provenance

1. Introduction

The General Data Protection Regulation (GDPR) [16] is an European data protection and privacy law that introduces several obligations and rights, whose compliance will require undertaking new practices for activities involving consent and personal data. With the large amount of fines, and the new set of required activities and practices, GDPR compliance has spurred a growth of innovation within the industry. This had led to the creation and offering of several approaches and tools, both commercial and non-commercial, for assisting with GDPR compliance.

While the use of technology in such solutions is innovative, the larger domain of regulatory compliance still focuses on information stored in static documents that do not use or provide any form of metadata. Such documents do not use a common vocabulary for representing technical aspects, nor are a form of linked data which can be used to semantically link related documents or regulations together. As a consequence, information associated with regulatory compliance cannot be easily queried or shared or represented.

* Corresponding author
E-mail address: harshvardhan.pandit@adaptcentre.ie

10.1016/j.procs.2018.09.026

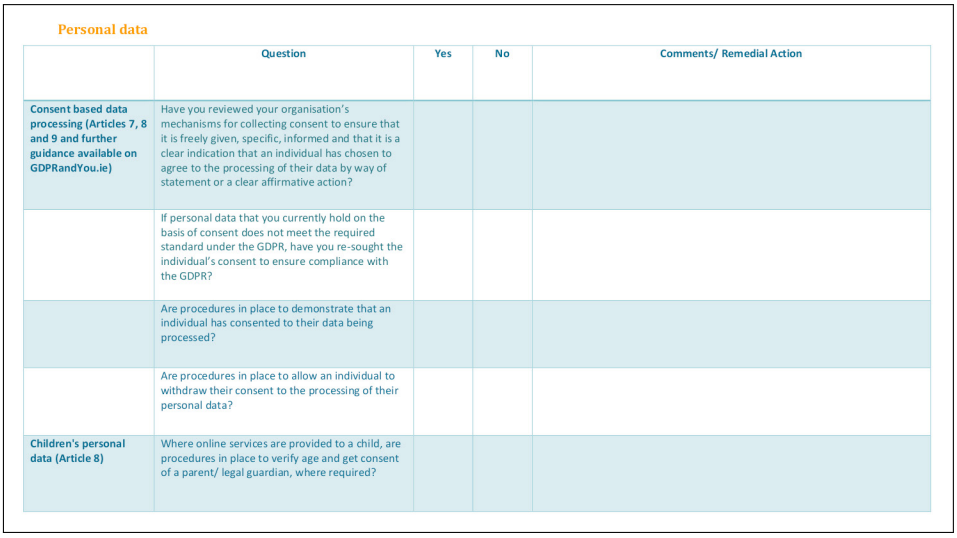
Provenance metadata forms an important part of compliance information. GDPR specifies several obligations over how personal data is collected, used, stored, and shared. It also provides obligations over consent, which is one of the legal basis for activities involving personal data. Compliance to such obligations invariably involve information related to the lifecycles of consent and personal data, which can be represented as provenance metadata. In addition, the provision of several rights provided by the GDPR also includes information that can be represented as provenance.

Through this paper, we present our work on modelling and querying provenance information related to GDPR compliance obligations using semantic web technologies. As the GDPR has not entered into effect at the time of this work, the demonstration of this research faces a lack of authoritative use-cases for compliance documentation. In light of this, we chose the GDPR readiness checklist published by Ireland's Data Protection Commissioner's office as an authoritative document for queries related to GDPR compliance. We present a proof-of-concept demonstration of our model for automating the retrieval of information related to these queries.

The work presented in this paper is based on previous published work addressing GDPR compliance and the use of semantic web technologies for representation of relevant metadata. GDPRov is a provenance ontology [14] for the representation of provenance information related to consent and data lifecycles. GDPRtEXT [13] provides a linked data version of the GDPR text as well as an ontology describing its various terms and concepts. Along with these, we have published our work on possible approaches towards representations for consent [7] and data sharing agreements [8]. An analysis of the interoperability model relevant to the GDPR was also published [15], which includes a discussion about the suitability of semantic web technologies for expressing metadata.

2. GDPR readiness checklist

Ireland's Data Protection Commissioner's office[1] launched a GDPR-specific website `www.GDPRandYou.ie` that provides guidance and resources to help individuals and organisations become more aware of their rights and responsibilities under the General Data Protection Regulation. One of the resources is a document[2] titled "Preparing Your Organisation for the GDPR – A Guide for SMEs", which we will refer to as "GDPR Readiness Checklist" for the purposes of our work. It provides a 'table' (see Fig. 1) containing various questions divided into contextual sections based on addressing certain GDPR articles and obligations. The docuemt is also provided online[3] as a webpage.



Personal data

	Question	Yes	No	Comments/ Remedial Action
Consent based data processing (Articles 7, 8 and 9 and further guidance available on GDPRandYou.ie)	Have you reviewed your organisation's mechanisms for collecting consent to ensure that it is freely given, specific, informed and that it is a clear indication that an individual has chosen to agree to the processing of their data by way of statement or a clear affirmative action?			
	If personal data that you currently hold on the basis of consent does not meet the required standard under the GDPR, have you re-sought the individual's consent to ensure compliance with the GDPR?			
	Are procedures in place to demonstrate that an individual has consented to their data being processed?			
	Are procedures in place to allow an individual to withdraw their consent to the processing of their personal data?			
Children's personal data (Article 8)	Where online services are provided to a child, are procedures in place to verify age and get consent of a parent/ legal guardian, where required?			

Fig. 1. GDPR Readiness Checklist document published by Ireland's Data Protection Commissioner (page 10)

[1] https://dataprotection.ie/
[2] http://gdprandyou.ie/wp-content/uploads/2017/12/A-Guide-to-help-SMEs-Prepare-for-the-GDPR.pdf
[3] http://openscience.adaptcentre.ie/GDPR-checklist-demo/demo/GDPR-readiness-checklist.html

The aim for this work was to demonstrate querying of information related to the document questions using semantic web technologies. For that, we first analysed document questions (Section 2.1) and represented them using SPARQL queries (Section 2.2). We then created an implementation model and tested it with an example use-case (Section 3.1) to generate a proof-of-concept demonstration (Section 3.2).

2.1. Analysis of Structure and Content

The document contains 13 pages (of relevance) with 63 questions divided into 9 sections. We categorise the questions into three categories - demonstrative, evaluative, and assistive - based on the requirements of information associated with them. Answers to demonstrative questions directly satisfy the question and do not need further actions or processing. Assistive questions can be directly evaluated, and therefore their answers contain information that assists in the evaluation. Evaluative questions have solutions that need to be evaluated based on further criteria not part of the question. They differ from assistive questions in that the information is complete and present, but cannot be evaluated for compliance without referring to additional actions or processing. We further distinguish between questions that contain or require provenance information and those that don't. Some of these questions could not be currently implemented due to a lack of metadata or additional information about how they would be implemented due to lack of real-world use-cases. Based on the analysis, we updated the GDPRov ontology with the required information from v0.4 to v0.6 to reflect the additional requirements. Our analysis of the document is available online[4].

2.2. Creation of SPARQL queries

Based on the analysis of the questions and their requirements, we modelled 33 SPARQL[5] queries to retrieve the related information using the GDPRov ontology to specify activities and entities related to consent and data, and GDPRtEXT to refer to specific terms and concepts within GDPR. The SPARQL queries are available online[6] with seperate files for each query and a common file containing the prefixes. An example query is presented in Listing 1, and retrieves information about steps and the processes along with the legal basis for their operation. The specified query G5 looks for steps that are part of a process and use some form of personal data. This is based on the modelling of information within GDPRov, that allows specifying of steps being part of a larger process (which GDPRov inherits from P-Plan). The argument for such a query is that a process would declare its legal basis, which the steps would inherit. This also allows a common step to be part of different processes with different legal justifications.

3. Implementation

3.1. Use-Case

We created an example use-case for our proof-of-concept demonstration based on an online shopping service that allows users to order products. Users can sign-up to the service to receive discounts and special offers (on the service). The service also serves ads to its users, which are generated by a Third Party. For the sign-up process, it collects consent and personal data such as name, address, email, and contact number. While ordering products, users are requested to provide sensitive information for transactions about their bank account or credit cards. The use-case defines certain additional subclasses such as *CustomerInfo* from *gdprov:PersonalData* for representing registered user information, and *gdprov:BankingInfo* from *SensitiveData* for representing the banking information. The use-case explicitly defines processes for handling various obligations and rights based on the terms provided by GDPRov. The use-case was generated using the Protégé[7] ontology editor, and uses a non-existant IRI[8]. It is availabe online as a single RDF file[9] that contains the use-case as well as GDPRov and GDPRtEXT ontologies as a self-sufficient dataset for

[4] http://openscience.adaptcentre.ie/GDPR-checklist-demo/demo/notes.html
[5] https://www.w3.org/TR/sparql11-query/
[6] https://opengogs.adaptcentre.ie/harsh/GDPR-readiness-checklist-usecase/src/master/sparql
[7] https://protege.stanford.edu/
[8] http://example.com/ontology/shoppingapp#
[9] http://openscience.adaptcentre.ie/GDPR-checklist-demo/demo/data.rdf

```
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gdprov:   <http://purl.org/adaptcentre/openscience/ontologies/gdprov#>
PREFIX gdprtext: <http://purl.org/adaptcentre/openscience/ontologies/GDPRtEXT#>

SELECT DISTINCT ?process ?legal WHERE {
  ?data a ?data_type .
  ?data_type rdfs:subClassOf gdprov:PersonalData .
  ?step a ?step_type .
  ?step_type rdfs:subClassOf gdprov:DataStep .
  ?step gdprov:usesData ?data .
  ?step gdprov:isPartOfProcess ?process .
  OPTIONAL { ?step gdprov:hasLegalBasis ?legal } .
  OPTIONAL { ?process gdprov:hasLegalBasis ?legal } .
} ORDER BY ?process
```

Listing 1: SPARQL query G5 - Legal basis for processing

querying. The Fact++[10] reasoner was used to compute additional facts about the use-case. Though not comprehensive, the use-case sufficiently provides an overview of how the SPARQL queries retrieve relevant information for answering the questions in the proof-of-concept demo.

3.2. Demonstration

Using the previously described SPARQL queries and use-case, we created an online[11] demo for automated querying of information related to the GDPR readiness checklist. The demo is intended to showcase how the static GDPR readiness checklist can be made more interactive and automated using semantic web technologies. It expresses the questions followed by the SPARQL query and its results. The results are retrieved on page refresh directly from our SPARQL endpoint[12] which contains the described use-case, which is based on the Openlink Virtuoso Open-Source Edition[13] triple-store. The demo uses YASQE[14] to represent the SPARQL queries as highlighted code. It uses YASR[15] to represent the results of queries in an interactive fashion. This allows the results to be a viewed as a HTML table (default view) or JSON, and allows exporting the results as a CSV file. The queries to the SPARQL endpoint and its responses are communicated as JSON documents. The source for the demo is available online[16] for introspection.

The results of each query contain the information associated with answering the relevant compliance questions. For the example query presented in this paper, which was G5 for legal obligations related to processing, the results express the steps and processes along with their legal obligations. This can be seen in Fig. 2 which shows five results of the query, of which three are processes that handle the various rights, which do not contain any legal basis. The remaining two are processes associated with the provision of the service, of which *OrderProcess* is considered to have legitimate interest and signing-up for an account (*NewUserSignUpProcess* is defined to be based on given consent.

[10] http://owl.cs.manchester.ac.uk/tools/fact/
[11] http://openscience.adaptcentre.ie/GDPR-checklist-demo/demo/
[12] http://openscience.adaptcentre.ie/sparql
[13] https://virtuoso.openlinksw.com/
[14] http://yasqe.yasgui.org/
[15] http://yasr.yasgui.org/
[16] https://opengogs.adaptcentre.ie/harsh/GDPR-readiness-checklist-usecase

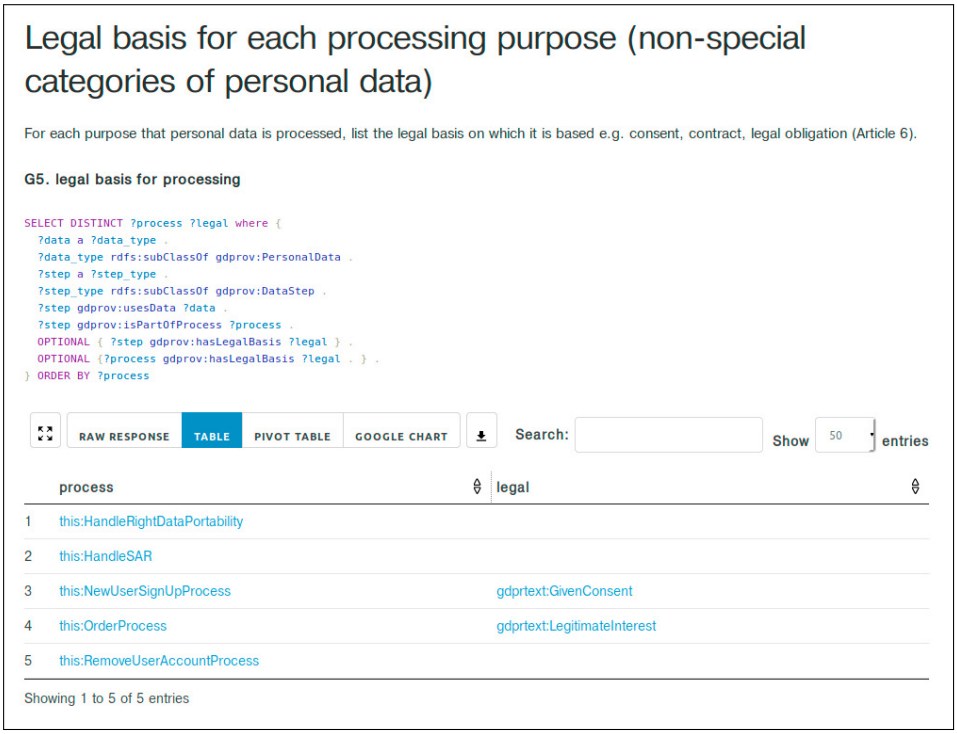


Fig. 2. Proof-of-concept demo for retrieving information related to query G5 in GDPR readiness checklist

4. Potential Applications

Automating Compliance Checks While regulatory compliance could be a periodic or continuous process within organisations, automating certain parts of the process can greatly aid in its efficient resolution, especially when the data is large in scale. Where it is not possible to automatically evaluate the compliance, the presence of information can be checked instead to ensure all the required metadata is present in the required form. An example of this in the proof-of-concept demo is the information provided for queries that could not be directly evaluated.

Documenting GDPR compliance A system for representing compliance related information in a structured and queryable format requires the creation of a compliance graph containing the appropriate metadata that can be retrieved using SPARQL queries. This information, and the associated SPARQL queries can then be used to generate documentation that highlights the specific steps and processes relevant to the obligation along with all necessary information. The documentation can itself be made more interactive, similar to the proof-of-concept demo presented, which enables investigation of related information in a simpler and consistent manner.

Impact Assessments, Continuous Evaluations, and Subject Access Requests GDPR stipulates that certain organisations may have to undertake Impact Assessments that evaluate the various consent and personal data related practices with a view towards compliance obligations. It also stipulates a process of continuous evaluation with respect to GDPR compliance, which involves a similar investigation of the organisation's consent and personal data related practices. A compliance system can assist in the retrieval of related information for such obligations similar to the proof-of-concept demo. Additionally, documenting the process of having carried out assessments is itself important to show due diligence when important changes are made to the organisation's practices.

Subject Access Requests (SAR) Data subject can retrieve information about their activities through the use of automated queries that simplify the provision of information by dynamically retrieving the required information and presenting itin a consistent and structured fashion to interactively access the required information.

5. Related Work

SPECIAL project The Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance (SPECIAL) project[17] is an European H2020 project that aims to provide a technical solution involving big-data innovation and privacy-aware data protection. Its contributions and publications are available online along with the publicly available deliverables[18] that describe their findings and reports to date. Their publication related to building a model for GDPR based on distributed ledgers [4] can allow for efficient information sharing between entities.
Impact Assessment This work provides a methodology and a template in the context of the GDPR for Data Protection Impact Assessment [2] and Privacy Impact Assessment [17].
Ontologies An initial work for describing GDPR obligations an ontology [1] addressed a draft version of the GDPR. It presented an OWL2 ontology describing the duties of data controllers for GDPR obligations. The work is described as preliminary with an explicit mention of intended changes and updates in the future.
Visualisation Approaches exist for interactive dashboards [3] that can show information flows for consent and personal data as well as provide features for the handling of various rights. Visualisation has also been used for representing contracts [6] and legal rules [18].
Smart Contracts Smart contracts for data sharing agreements between organisations [5] can be self-fulfilling and automated, which can fit well with the work described in this paper. The use of Artificial Intelligence techniques [9] towards supporting the compliance process can further aid in the management of such shared information.
Knowledge Graphs Recent work regarding creation of legal knowledge for multilingual services [12] can assist in the provision of compliance by design [11]. Such approaches enable efficient integration of technology into existing legal workflows.
Access Control A comprehensive and up-to-date survey [10] was recently published describing efficient and comprehensive access control using semantic web technologies in the areas of privacy, security, and policies published in the semantic web domain the various problems along with potential solutions and approaches.

6. Conclusion & Future Work

Through this paper, we demonstrated how semantic web technologies can assist in the representation and querying of information related to compliance towards General Data Protection Regulation (GDPR). We focused on obligations involving provenance metadata for consent and personal data, which we represented using the previously published GDPRov ontology and GDPRtEXT resource. Due to a lack of authoritative compliance use-cases, we presented the application of our work through the GDPR-readiness checklist published by Ireland's Data Protection Commissioner. We used SPARQL to represent the questions within the document, and presented a proof-of-concept demonstration of our model for automating the retrieval of information related to these queries. Our aim in undertaking this work was to sufficiently demonstrate the usefulness and maturity of the semantic web for representation of knowledge as well as for assisting in the compliance process. We also discussed the broader applications of our work for other aspects associated with GDPR as well as the regulatory compliance domain.

While we focus only on provenance metadata in the currently presented work, this is one aspect of information flows present in the GDPR information model. Our larger approach towards the GDPR [15] involves five information categories, which are provenance, data sharing agreements, consent, certification, and compliance. Our future work is primarily based on expanding our semantic web based approach towards the representation and incorporation of these information categories to create a knowledge-based system for GDPR compliance.

With the advent of the GDPR, it is expected that a significant number of information will be publicly available in relation to the practices surrounding consent and personal data. We aim to incorporate these as use-cases to both shape our work, as well as validate its applicability by demonstrating it over these use-cases.

[17] https://www.specialprivacy.eu/
[18] https://www.specialprivacy.eu/publications/public-deliverables

Acknowledgements

This work is supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] Bartolini, C., Muthuri, R., 2015. Reconciling Data Protection Rights and Obligations: An Ontology of the Forthcoming EU Regulation, in: Workshop on Language and Semantic Technology for Legal Domain.

[2] Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., Rost, M., 2016. A Process for Data Protection Impact Assessment Under the European General Data Protection Regulation, in: Privacy Technologies and Policy, Springer, Cham. pp. 21–37. doi:10.1007/978-3-319-44760-5_2.

[3] Bier, C., Kühne, K., Beyerer, J., 2016. PrivacyInsight: The Next Generation Privacy Dashboard, in: Privacy Technologies and Policy, Springer, Cham. pp. 135–152. doi:10.1007/978-3-319-44760-5_9.

[4] Bonatti, P., Kirrane, S., Polleres, A., Wenning, R., 2017. Transparent Personal Data Processing: The Road Ahead, in: Computer Safety, Reliability, and Security, Springer, Cham. pp. 337–349. URL: https://link.springer.com/chapter/10.1007/978-3-319-66284-8_28, doi:10.1007/978-3-319-66284-8_28.

[5] Corrales, M., Jurcys, P., Kousiouris, G., 2018. Smart Contracts and Smart Disclosure: Coding a GDPR Compliance Framework. SSRN Scholarly Paper ID 3121658. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=3121658.

[6] Esayas, S., Mahler, T., McGillivray, K., 2016. Is a Picture Worth a Thousand Terms? Visualising Contract Terms and Data Protection Requirements for Cloud Computing Users, in: Current Trends in Web Engineering, Springer, Cham. pp. 39–56. doi:10.1007/978-3-319-46963-8_4.

[7] Fatema, K., Hadziselimovic, E., Pandit, H.J., Debruyne, C., Lewis, D., O'Sullivan, D., 2017. Compliance through Informed Consent: Semantic Based Consent Permission and Data Management Model, in: Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn). URL: http://ceur-ws.org/Vol-1951/#paper-05.

[8] Hadziselimovic, E., Fatema, K., Pandit, H.J., Lewis, D., 2017. Linked Data Contracts to Support Data Protection and Data Ethics in the Sharing of Scientific Data, in: Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci), pp. 55–62. URL: http://ceur-ws.org/Vol-1931/#paper-08.

[9] Kingston, J., 2017. Using artificial intelligence to support compliance with the general data protection regulation. Artificial Intelligence and Law 25, 429–443. URL: https://link.springer.com/article/10.1007/s10506-017-9206-9, doi:10.1007/s10506-017-9206-9.

[10] Kirrane, S., Villata, S., d'Aquin, M., 2018. Privacy, security and policies: A review of problems and solutions with semantic web technologies. Semantic Web 9, 153–161. URL: https://content.iospress.com/articles/semantic-web/sw289, doi:10.3233/SW-180289.

[11] Mayer, W., Casanovas, P., Stumptner, M., 2017. Semantic Workflows in Law Enforcement Investigations and Legal Requirements, in: Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017).

[12] Montiel-Ponsoda, E., Rodríguez-Doncel, V., Gracia, J., 2017. Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe, in: Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017).

[13] Pandit, H.J., Fatema, K., O'Sullivan, D., Lewis, D., 2018a. GDPRtEXT - GDPR as a Linked Data Resource, in: 15th European Semantic Web Conference (in-press, Heraklion, Crete, Greece. URL: http://purl.org/ADAPT/pub/E18ESWC_GDPRtEXT.

[14] Pandit, H.J., Lewis, D., 2017. Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies, in: Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn). URL: http://ceur-ws.org/Vol-1951/#paper-06.

[15] Pandit, H.J., O'Sullivan, D., Lewis, D., 2018b. GDPR Data Interoperability Model, in: 23 rd EURAS Annual Standardisation Conference (in-press), Dublin, Ireland. URL: http://purl.org/ADAPT/pub/E18EURAS.

[16] Parliament, E., Council, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L119, 1–88. URL: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC.

[17] Reuben, J., Martucci, L.A., Fischer-Hübner, S., Packer, H.S., Hedbom, H., Moreau, L., 2016. Privacy Impact Assessment Template for Provenance, in: Availability, Reliability and Security (ARES), 2016 11th International Conference on, IEEE. pp. 653–660.

[18] Seppala, S., Ceci, M., Huang, H., O'Brien, L., Butler, T., 2017. SmaRT Visualisation of Legal Rules for Compliance, in: Proceedings of the 1st Workshop on Technologies for Regulatory Compliance co-located with the 30th International Conference on Legal Knowledge and Information Systems (JURIX 2017).