A European Framework for Regulating Data and Metadata Markets

Pompeu Casanovas[1, 2], Víctor Rodríguez-Doncel[3], Cristiana Santos[1], Asunción Gómez-Pérez[3]

[1]Institute of Law and Technology, Universitat Autònoma de Barcelona
Yaiza Cabedo for pompeu.casanovas@uab.cat; cristiana.teixeirasantos@gmail.com
[2] Data to Decisions Cooperative Research Centre, Deakin University, Geelong, Australia
[3]Ontology Engineering Group, Universidad Politécnica de Madrid
{vrodriguez, asun}@fi.upm.es

Abstract. In this paper we examine the possibilities offered by the EU legal framework to set and regulate a data and meta-data market. It is our contention that a policy and legally-driven market could benefit from analytical concepts —meta-rule of law, semantic web regulatory models, legal ontologies— to reduce privacy and data protection risks. We introduce a general and integrated framework, and provide examples of existing privacy ontologies and of the practical use of linked data.

Keywords: privacy, meta-rule of law, metadata markets, ontologies

1 Introduction

In this position paper we introduce some notions related to the concepts of privacy and data protection recently embraced by the EU in the General Data Protection Reform (GDPR),[1] notions not explicitly used in the legal texts. However, it is our contention that *meta-rule of law*, *semi-automated regulations*, *Semantic Web Regulatory Models* (SWRM), *Ontologies*, among others, are analytically related and could be useful to make effective the principles and rights which have been included in the Reform within tight deadlines.[2] A Regulatory Model (RM) is the specific normative suite encased by computational devices built up to monitor a regulatory system encompassing hard law, soft law, governance and ethics. When using Semantic Web (SW) tools, RMs turn into SWRMs, as introduced by Casanovas [1].

Since its early stages, legal scholars have pointed out the Copernican legal turn of GDPR compared to the previous situation [2][3]. The set of principles contained in the enacted legislation are not new, representing the natural follow-up of citizens' protections contained in the EU Charter of Fundamental Rights (2000). But, contrary

[1] http://ec.europa.eu/justice/data-protection/reform/index_en.htm
[2] The Regulation shall be applied from 25/05/2018 and the Directive should be transposed into national law before 06/05/2018.

to US prevailing opinions, personal data protection has emerged in Europe as a *fundamental right* [4]. Thus, principles of transparency, data minimisation, proportionality, purpose limitation, consent, accountability, data security, and the rights of access, correction, erasure, etc. can be enforced through economic sanctions, and monitoring instruments. Moreover, the creation of a European Data Protection Board "reveals substantial legal incentive to create a new industry with regard to application of the Regulation" [26]. However, in spite of their broad scope and the progressive attitude shown by EU legal drafters, the Article 29 Working Party, and the EU Court of Justice (e.g. Judgment C-131/12), the environment of the Web of Data raise some concerns about its implementation. It is worth mentioning that (i) legal regulations of sectors such as banking and finances are too complex and detailed to be easily handled without the participation and consent of mighty stakeholders; (ii) policies and ISOs represent another dimension to be taken into account, as it is not always clear when, why, how and to whom they apply in the global markets; (iii) data and metadata are still poorly regulated; (iv) technical protocols and standards (e.g. W3C Recommendations, Oasis standards) are not mandatory; (v) Linked Open Data (LOD) scenarios [27] entail new regulatory challenges. The semantic approach is being aligned with statistical differential privacy methods to prevent attacks against high-dimensional micro-data [28] [30].

Legal scholars, computer and social scientists have identified the main elements of this new regulatory framework —including the results of fifteen years of research on legal XML, Legal RuleML and legal ontologies [5]; linked data publication and consumption [6], copyright related terms [7], licensing [8], patents [9], privacy risks [10], and the emergence of data and metadata markets [11]. Stemming from a practical point of view, a general framework is needed to regulate such a market and the exercise of rights.

Section 2 presents the meta-rule of law and a possible policy-driven data market. Section 3 illustrates the particularities of the linked datasets with respect to privacy and data protection. Section 4 presents examples and open challenges, and Section 5 concludes the paper.

2 Policy and legally-driven integrated Metadata and Data Markets: The Meta-rule of Law

The EU Market requires a complex balance between roles of EU and national controllers, as each EU country sets its specific market and finance controls.

Due to this situation and to the financial crisis, the Commission issued the *European Market Infrastructure Regulation*[3] (EU 648/2016) (EMIR) to enhance the market infrastructure resilience and promoting financial stability. The *European Securities and Markets Authority*[4] *(ESMA),* supervisory body created to harmonise the banking sector and financial market, and other EU control bodies, such as the Body of

[3] http://ec.europa.eu/finance/financial-markets/index_en.htm
[4] http://ec.europa.eu/finance/general-policy/committees/index_en.htm

European Regulators for Electronic Communications (BEREC) are struggling to make EU national countries to comply with security measures.

Fig. 1 depicts a general integrated framework to develop this Digital Single Market, showing two different poles: (i) official supervision bodies and civil society stakeholders; (ii) protective principles and values of the rule of law, and the layer of meta-rule of law embracing such principles through data and metadata semi-automated processing. This figure reproduces schematically the metadata workflow through individuals, organizations and institutions within the market. Civil society stakeholders —organizations representing the interests of companies and holdings, and those representing consumers at national and European level— are situated on the left. The opposite column, on the right side of the figure, refers to institutional bodies of the national states and of the EU (with supervisory, ruling or monitoring roles). For the sake of simplicity, we have selected only few of them, those with regulatory (and sanctioning or rewarding) powers over market players, and institutions set to protect personal data and privacy, enacting the ethical and information principles contained into GDPR or, at national level, into the DP Acts. The idea of distinguishing a meta-rule of law intends to go further, linking all the elements at play in the metadata workflow [14]. The central layers in Fig. 1 plot knowledge acquisition, knowledge representation, reasoning, and evaluation as necessary procedural stages of the DSM social ecosystem by contextualizing the whole process within the institutional requirements laid down through the complex network of entangled rules, protections, warrants and decisions at stake. This is the sense of setting a meta-rule of law as analytical device to implement the guarantees of the rule of law on digital environments.

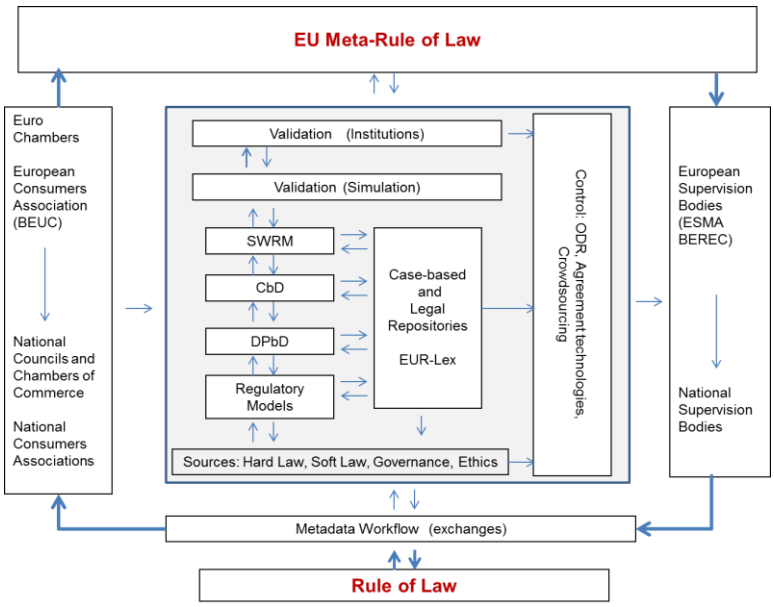

Fig. 1 Policy and legally-driven data and meta-data markets. Acronyms: Data Protection by Design [DPbD), Compliance by Design [CbD], Semantic Web Regulatory Models [SWRM].

Allegedly, all stakeholders and institutions are under the provisions of the rule of law. Legislation and case-law have been traditionally defined by legal scholars as *the binding sources of law*. With the emergence of the internet and the digital world, many legal scholars such as Lessig pioneered the view that not only traditional legal instruments but programming, and technological languages might shape human or artificial behaviours through all types of codes, protocols and standards [12]. The "identity layer" or "metasystem layer" of the Internet, defines "a protocol to enable a kind of virtual wallet of credentials". The web of data focuses both on linking data and on end-users, thus, on *personalization* of knowledge. Legal components address issues concerning knowledge and management of regulations, ethics, governance and the law, *alike* [1]. Both approaches (legal and computational) converge on almost every aspect of the identity layer and the valuation of data, protocols and behaviour of citizens, consumers, companies, and administrations on the web. A shared identity ecosystem is still under development, both in EU and in USA (NIST) [25].

3 Privacy and Data Protection in Data Markets

This section deepens on how technology is in its way to provide elements towards the management of privacy and data protection in a semi-automated manner during data exchanges. Section 3.1 describes the attempts to represent key aspects of data protection and privacy in a structured machine-readable form. Section 3.2 proposes a dataset labelling with respect to their privacy level. The first milestone is the description of datasets in terms of their privacy level, answering the question: *does a certain dataset contain personal identifiable information* (PII)*?* A negative answer is important to be asserted for data consumers to confidently make use of the traded resource. A positive answer helps a data processor to comply with the data protection norms. Section 3.3 identifies potential risks that technology cannot address.

3.1 Computer ontologies representing privacy

Table 1 identifies some of the ontologies in the domain of privacy and data protection along with their modelling objective. These ontologies are very often thoroughly designed considering theoretical aspects, but in practice are difficult to be learnt or used in practical settings.

Ontology	***Modelling objective***
LegLOPDontology [17]	Privacy of users of location-based services
OntoPrivacyontology [18]	Supporting a tool allowing to query legislative data
Neurona Ontologies [19]	Data protection compliance to offer reports regarding the correct application of security measures to data files containing personal data for administrations and organizations
Privacy by Design ontology framework [20]	Implementation of data protection measures prior to the determination of the means of processing
European healthgrids [21]	Ontologies for privacy compliance on European healthgrids
Data Protection Requirements	Specifies data protection legal requirements for business

in Workflows [22]	processes legal compliance

Table 1. Privacy-related ontologies and their purpose

The *LegLOPD* ontology aimed at the preservation of privacy of users in location-based services. It modelled concepts from the Spanish data protection law. The core concept in the ontology is *private data*. The *OntoPrivacy* ontology modelled the concepts of the Italian Personal Data Protection Code; a bottom-up approach was used as the lexicon was the basis to build the ontology. *OntoPrivacy* has been created to support a tool that allows querying the functional profile of legislative data. The application-oriented *Neurona* ontologies modelled the knowledge for the development of data protection compliance to offer reports regarding the correct application of security measures to data files containing personal data. Its design is based on a (i) Data Protection Knowledge Ontology, which contains the core concepts of the system; and a (ii) Data Protection Reasoning Ontology, to assess data protection compliance. Regarding the *Privacy by Design (PbD) ontology framework* consists of nine base ontologies, eight domain ontologies and four application specific ontologies; it requires that data protection measures be implemented prior to the means of processing being determined. *Privacy compliance and enforcement on European healthgrids* through ontologies expresses the legal norms using to enforce access control policies regarding the sharing of medical data between different healthcare organizations in Europe. Finally, the ontology to model *Data Protection Requirements in Workflows* uses an ontology to extend notations to specify data protection legal requirements with which business processes must comply with. This approach highlights the new duties of data controllers, the auditors, and the DPAs and fosters the transition of IT-based systems, services/tools and businesses to comply with the GDPR. Other ontologies (e.g. *PrivOnto*, on privacy policies) are being developed and will be added to the list [29].

3.2 Annotating the Privacy Level of Datasets

Datasets may be annotated in order to describe their legal status with respect to privacy. As of today, there is little use of metadata properties to annotate the privacy status of a dataset [23]. While the adoption of the Dublin Core *license* property has gained massive widespread to declare the license of a resource, no equivalent property is being used to specify if a database contains or not personal data or any other related information. The Dublin Core *rights* property might be used to such end, and specific elements have been proposed such as *ldr:hasPersonalData* or *ldr:hasPrivacyLevel*.[5] But beyond these simple properties, there is an evident shortage of elements to qualify the dataset in terms of data protection and privacy. The following metadata items might be necessary: (i) date of expiry of the consent; (ii) purpose which has been consented by the user; (iii) countries where the dataset can live.

The existing legal framework suggests the addition of new properties, like the specific country where the personal data file has been registered, the privacy level of the dataset (for example in Spain three different levels are defined), or the different security measures that should be taken. Once these properties have been defined, a further specification of privacy-related information can be made using the ontologies

[5] The ldr prefix stands for the Linked Data Rights vocabulary. http://purl.oclc.org/NET/ldr/ns#

described in the section 3.2. These properties might also be part of a reviewed version of the DCAT application profile for data portals in Europe[6].

3.3. Challenges and risks

It is worth to notice that securing interoperability and a common format for data exchange is a necessary but not sufficient condition for a digital regulated market compliant with the requirements of EU GDPR. The path towards a semi-automated management of privacy and data protection is not a straightforward one, and some risks can be identified [6]:

1. *Personal data publicly available in social networks, or illegally leaked datasets due to security breaches*. User generated content from social networks (Facebook, YouTube or LinkedIn, etc.), represents roughly 50% of the Linked Open Data cloud [24] —large parts of it constituting personal data records.
2. *Spamming or other bad purposes*: many personal data records contain e-mail addresses which can be spammed in a more sophisticated manner.
3. *Indirect identification and re-identification of identifiable personal information*. Integrating personal data from distinct sources of available linked data (even from apparently innocuous or anonymized resources), may trigger indirect identification and re-identification.
4. *Profiling of individuals*. Big data analytics may foster the integration of data to create and reuse personal profiles.
5. *Security risks*. When large databases with personal information are created, security breaches grow faster than the adopted protected measures (the use of encryption keys by unauthorised persons may disclose personal identifiers).
6. *Onerous duties for publishers*. We may bring into the discussion the liability of the publisher releasing anonymised data into the public domain without the capability of controlling its access (duty to ensure that no-one can be identified from the data).

A data and metadata market and citizens' rights would be enhanced by a legally and policy-driven framework for daily economic exchanges. E.g. How could we protect financial transactions? Is a bank operator entitled to manipulate, combine, or even sell the end-user geo-located metadata of transactions performed in ATMs or trough mobile money? These geolocations give information about our close environment, the places we visit, etc. Based on geolocation, a software agent is able to infer structured information to build users' profiles. Metadata, end-user routes, behavioural patterns, and personal options can be inferred. When using the mobile phone for transactions, both geolocational and transactional data are generated. The smartphone combines communication data and social media data. Credit card metadata is used for identification purposes too. The formalisation of policies and the proposed meta-rule of law would contribute to mitigate these risks in a legally-framed, controlled, and monitored machine-machine communication for Digital Financial Services (DFS). A semi-automated networked market would know and foresee in advance (i) the different national regulations; (ii) the different scenarios

[6] https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v1

across countries (e.g., mobile/plastic card personal metadata can be treated differently depending on the country); (iii) the legal value of all possible trading moves. Financial inclusion, integrity, and anti-fraud and money laundering policies could benefit from this perspective too [31].

General provisions, exceptions and conditions for particular cases can be better managed through SWRM, which in addition: (i) have the property to coordinate the agency of different related powers (Legal Enforcement Agencies, DP agencies, supervisory commissions, local and national bodies...), (ii) can facilitate the common interoperability of norms and concepts, (iii) may offer a general framework to coordinate the legal actions to be taken before, during and after the transaction.

5 Conclusions

This paper outlines challenges related to the regulation of a data and meta-data market considering the EU legal framework. Assuming the implementation of a meta-rule of law, we described a framework where a policy-driven market can help enhancing the management of personal data reducing the privacy risks. These policies will need a better qualification of the data in terms of privacy, and the existing resources and needs sketched in this paper should be refined. Semi-automated regulation could manage these and many other cases saving time, efforts, money, and avoiding further legal problems. That is to say, again: agency through SWRMs would turn to be more effective, secure and protective, because what it really counts is aligning the *legal value* of actions with the *real needs and decisions* of individuals, companies and organizations.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (projects TIN2013-46238-C4-2-R, RTC- 2014-2946-7, DER2012-39492-C02-01), the EU 520250-1-2011-1-IT-ERA MUNDUS-EMJD, and the CRC Data2Decisions Australian Programme. Louis de Koker, Yaiza Cabedo, and three anonymous reviewers provided useful insights for this position paper.

References

[1] Casanovas, P., Semantic Web Regulatory Models. Why Ethics matter? Philosophy & Technology 28 (1) pp. 33-55 (2015)
[2] Gawith, S., Lens, R., de Hart, P. (Eds.) Reforming European Data Protection Law, Springer, Dordrecht (2015)
[3] Gawith, R. Lens, P. De Hart (Eds.). Data Protection on the Move. Current Developments in ICT and Privacy/Data Protection, Springer Verilog, Dordrecht (2016)
[4] Gonzalez-Foster, G. The Emergence of Personal Data Protection as a Fundamental Right of the EU, Springer, Dordrecht (2015)
[5] Casanovas, P., Pal Mirani, M., Perini, S., et al. (2016). Special Issue on the Semantic Web for the Legal Domain, Guest Editors Editorial: The Next Step". Semantic Web Journal 7 (2), pp. 1-13 (2016)

[6] Rodríguez-Doncel, V., Santos, C., Casanovas, P. et al., Legal aspects of linked data – The European framework, CLSR: The International Journal of Technology Law and Practice (2016)
[7] Rodríguez-Doncel, V., Santos, C., Casanovas, P. et al., A Linked Term Bank of Copyright-Related Terms, in A. Retool (Ed.) Legal Knowledge and Information Systems, pp.91-99, IOS Press (2015)
[8] Rodríguez-Doncel, V., Suárez-Figueroa, M. C., Gómez-Pérez, A., Poveda-Villalón, M. License Linked Data Resources Pattern. In Proceedings of the 4th WOP. CEUR Workshop Proceedings 1188 (2013)
[9] Ramakrishna, S., Paschke, A process for knowledge transformation and knowledge representation of patent law. In International Workshop on Rules and Rule Markup Languages for the Semantic Web, Bikakis et al. (Eds.), RuleML 2014, LNCS 8620, pp. 311–328, Heidelberg, Springer (2014)
[10] Pape S, Serna-Olvera J, Tesfay WB. Why Open Data May Threaten Your Privacy, inWorkshop on Privacy and Inference, PRINF 2015 (2015)
[11] Steyskal, S., Kirrane, S. If You Can't Enforce It, Contract It: Enforceability in Policy-Driven (Linked) Data Markets, in Filipowska, A., et al. (Eds.), International Conference on Semantic Systems - SEMANTiCS 2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15). CEUR Workshop Proceedings 1481, pp. 63-66 (2015)
[12] Lessig, L. Code and Other Laws of Cyberspace, NY: Basic Books (1999), Code v2 (2006)
[13] Pagallo, U. The Laws of Robots. Crimes, Contracts, and Torts. Berlin, Dartmouth: Springer (2013)
[14] Casanovas, P. Conceptualisation of Rights and Meta-rule of Law for the Web of Data, Democracia Digital e Governo Eletrônico (Santa Caterina, Brasil) 10 (2), pp. 18-41 (2015)
[15] Cabedo, Y. OTC regulatory reform: risks of the clearing obligation from a competition perspective, LSE Risks and Regulations 31 (2016)
[16] Koops, B.J., Hoepman, J.H., Leenes, R. Open-source intelligence and privacy by design, CLSR 29, pp. 676-688 (2013)
[17] Mitre, H.A., González-Tablas, A.I., Ramos, et al., A legal ontology to support privacy preservation in location-based services. In: Meersman, R., et al. (Eds.) On the Move to Meaningful Internet Systems: OTM 2006 W, LNCS, vol. 4278, pp. 1755–1764, Springer Berlin Heidelberg (2006)
[18] Cappelli, A., Lenzi, V.B., Sprugnoli, R., et al., Modelization of domain concepts extracted from the Italian privacy legislation. In Proceedings of the 7th Int. W. on Computational Semantics (2007)
[19] Casellas, N., Nieto, J.E., Roig et al., Ontological semantics for data privacy compliance: The Neurona project. In Proceedings of the Intelligent Privacy Management Symposium. pp. 34–38 (2010)
[20] Kost, M., Freytag, J.C., Kargl, F. et al., Privacy verification using ontologies. In Proceedings of the Sixth International Conference on Availability, Reliability and Security. pp. 627–632 (2011)
[21] Rahmouni, H. B., Solomonides T., Casassa M. et al., Privacy compliance in european healthgrid domains: An ontology-based approach. In 22nd IEEE Symposium on CBMS, pp. 1-8 (2009)
[22] Bartolini C., Muthuri R., Santos C.: Using Ontologies to Model Data Protection Requirements, in Workflows Ninth International Workshop on Juris-informatics, JURISIN (2015)
[23] Rodríguez-Doncel, V., Gómez-Pérez, A., Mihindukulasooriya, N.: Rights declaration in Linked Data, in Proc. of the 3rd Int. W. on Consuming Linked Data, O. Hartig et al. (Eds) CEUR vol. 1034 (2013)
[24] Schmachtenberg, M., Bizer, C., Paulheim, H. Adoption of the Linked Data Best Practices in Different Topical Domains. The Semantic Web – ISWC 2014 vol. 8796 LNCS pp 245-260 (2014)
[25] Grassi, P.A., Nadeau, E.M., Galluzzo, R.J. et al., Attribute Metadata, NIST Internal Report (2016)
[26] De Hert, P., Papakonstantinou, V., The new General Data Protection Regulation: Still a sound system for the protection of individuals? CLSR 32, pp. 179-194 (2016)
[27] Boella G, et al.: Linking Legal Open Data: Breaking the Accessibility and Language Barrier in European Legislation and Case Law, in Proc. of the 15th ICAIL (2015)
[28] Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy, Foundations and Trends, in Theoretical Computer Science 9 (3-4) pp. 211-407 (2014)
[29] Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Sadeh, N., Reidenberg, J. PrivOnto: a Semantic Framework for the Analysis of Privacy Policies, Semantic Web Journal (under review) (2016)
[30] Narayanan, A., Shmatikov, V. Robust de-anonymization of large sparse datasets, in 2008 IEEE Symposium on Security and Privacy (sp 2008), pp. 111-125, IEEE (2008)
[31] De Koker, L., Jentzsch, N. Financial inclusion and financial integrity: Aligned incentives?. World development 44 pp. 267-280 (2013).