The use of Decentralized and Semantic Web Technologies for Personal Data Protection and Interoperability

Mirko Zichichi[1,2,3], Víctor Rodríguez-Doncel[2], and Stefano Ferretti[3]

[1] Law, Science and Technology Joint Doctorate - Rights of Internet of Everything
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
vrodriguez@fi.upm.es
[3] Department of Computer Science and Engineering, University of Bologna, Italy
{mirko.zichichi2,s.ferretti}@unibo.it

Abstract. The enactment of the General Data Protection Regulation (GDPR) has been the response of the European Union to the growing data-driven economy backed up by the largest companies in the world. It provides the data protection and portability needed by individuals that "unconsciously" generate personal data for "free" services offered by providers that lack transparency on their use. Meanwhile, the rise of Distributed Ledger Technologies (DLTs) offers new possibilities for the management of general purpose data, hence being suitable for handling personal data in a trustless scenario. These decentralized technologies bring a new concept of contract called smart because of its ability to be self-executable. DLTs and smart contracts, together with the use of Semantic Web standards, allows the creation of a decentralized digital space controlled entirely by an individual, where his personal data can be stored and transacted.

Keywords: GDPR· personal data· distributed ledger technologies· smart contracts· semantic web

1 Introduction

With the introduction of the General Data Protection Regulation (GDPR) [5] in 2018, operations carried out regarding the management and the movement of personal data have radically changed. Data privacy of European Union's Citizen has been empowered through a series of rights that provide data protection and portability. The GDPR can be considered as a necessary response to the challenges posed by technological advancements brought mainly by large global companies that operates in a "not so new" data-driven economy, such as Google, Apple, Facebook and Amazon. It takes into account the need to protect personal data, which is increasingly public aware due to recent and frequent scandals on the abuse of personal information, such as the 2018 Cambridge Analytica revelations. A huge business, indeed, lies behind the trade of personal data and several companies make consistent profits operating in this sector. GDPR and

current literacy help the individual to understand how their personal data is often generate unconsciously and where, how or why the data is being collected, but still, further work is needed to let them develop the necessarily practical and interpretive skills [17]. More efforts are needed to reach both transparency and a balance between privacy and data sharing.

Even if GDPR requires data controllers, i.e. entities that collect and manage individuals' personal data, to release to their users the complete dataset they collected on them, upon request, there are currently no standards for this kind of requests and there is the tendency to hinder the progress of these, causing the entire process to become almost useless. These data controllers usually store this personal information in corporate databases, but they can become data providers to other parties if the individual agrees –and even if the individual does not agree they are obliged to act as data providers in extraordinary cases, e.g. national security. As of today, when these data transactions happen there is no transparency on the individual's data usage.

Meanwhile, between the many technologies that regards general-purpose data management and storage, Distributed Ledger Technologies (DLTs) are raising as powerful tools to avoid the control centralization. The current use of DLTs is in financial (i.e. cryptocurrencies) and data sharing scenarios. In both cases there are several parties that concur in handling some data, there is no complete trust among parties and often these ones compete to the data access/ownership. Such features suit perfectly with the process of moving the data sovereignty towards users and releasing them more influence over access control, while allowing anyone else to be able to consume this data with transparency. This can be made possible through smart contracts, the new concept of contract that brought a second blockchain revolution. The use of smart contracts grants to build Decentralized Applications (dApps) and Decentralized Autonomous Organizations (DAOs) that can realize novel important applications for social good [24].

The scope of this work concerns the design of methods and systems that support the right of individuals to the protection of their personal data and at the same favor its portability and economic exploitation and foster the social good. In the following sections the description of a proposal is made with the use of decentralized technologies and Semantic Web standards.

2 State of the Art

One of the most remarkable novelties in GDPR is the concept of *data portability*, which defines the right to have data directly transferred from one data provider to another making a step towards user-centric platforms of interrelated services [6]. This relates to the concept of data interoperability that embodies the complex network of users interaction based on personal data flow. To this scope, it is fundamental the use Semantic Web [2] standards, that bring structure to the meaningful contents of the Web by promoting common data formats and exchange protocols. The form of its most successful incarnation is Linked Data:

data published in a structured manner, in such a way that information can be found, gathered, classified, and enriched using annotation and query languages.

One of the most recent approach that involves the use of distributed technologies and Semantic Web integration in social networks is the Solid project [20]. Led by the creator of the Web Tim Berners-Lee, the project was born with the purpose of giving users their data sovereignty, letting them choose where their data resides and who is allowed to access and reuse it. Solid provides us a strong reference for our work because it uses Semantic Web technologies to decouple user data from the applications that use this data. Data is, indeed, stored in an online storage space called Pod, a Web-accessible storage service, which can either be deployed on personal servers or on public servers.

A great variety of solutions, instead, involve the use of DLTs for the management of general purpose data (for example applied to media contracts [12]) and personal data and few of them are in compliance with GDPR. A DLT is a software infrastructure maintained by a peer-to-peer network, where the network participants must reach a consensus on the states of transactions submitted to the distributed ledger, to make the transactions valid. The role of DLTs is to provide a trusted and decentralized ledger of data preserving immutability, traceability, transparency and pseudo-anonymity. The concept of DLT is the natural extension of the "blockchain" concept, because it includes those technological solutions that do not organize the data ledger as a linked list of blocks. The buzzwork blockchain has been presented together with the technology that has changed radically the vision that we have of the Internet, finance technologies, trust in communication and even digital democracy. It was made famous by Bitcoin [14], but the decentralized computation enabled by the Ethereum blockchain [3] enhanced the technology allowing the creation of a powerful tool: smart contracts. These contracts are self-managed structures that do not rely on a central control, thus eliminating the presence of single point of failures.

Related works for the management of Personal Data using DLTs include various proposals and many of them are also focused on GDPR [7, 4, 8], even in the corporate world.[4] However, these studies do not address DLTs and smart contract challenges, presenting only a conceptual approach. A more technical approach can be found in [22]. Meanwhile, Smart contract based data access control has been thoroughly studied in literature [21, 13] and still many scenarios are conceivable.

3 Moving Data Sovereignty Towards Users

By the description given so far, it is clear how there is a tendency to turn an individual to a simple source of data, concentrating the entire decision-making and operational power to the entity that make profit from that. It is still not possible to feasibly ensure to individuals the sovereignty of their personal data, nor the possibility of a appropriate data interoperability for data consumers. One

[4] https://wibson.org/

of the key problems is that individuals do not have control or even knowledge of the transfers that happen with their personal data. Data providers, indeed, store and maintain this data differently through several data silos, hampering their free exchange and economical exploitation. In this situation, individuals that may be good willing to offer their data for social good or that may simply make direct profit from it, do not have the power to do so. The main concern for individuals, then, is to invert this trend. The solution we propose, that supports the right of individuals to the protection of their personal data, data interoperability, economic exploitation and social good, is based on the following principles: i) to avoid the concentration of personal information and its opaque transfers it is needed a system that allows store and transact personal data in a controlled, transparent and non-centralized manner; ii) to favour personal data interoperability, hence to facilitate cross-domain application and services, a set of common languages and protocol must be used; iii) to let the individual define both high-level goals and fined-grained preferences for what regards the access to his data smart contracts can be used, since these allows to represent and reason with policies.

4 A proposal based on DLTs and Semantic Web

We present a solution that involves the use of Decentralized Technologies together with Semantic Web technologies to satisfy the principles posed in the previous section. In order to go further, it is needed to specify the types of personal data that concern this solution (but also the general case [17]). Personal data is defined as any piece of information that can identify or be identifiable to a natural person. Digital personal data, in particular, is generated by the interaction of a user with a software or a hardware in form of numbers, characters, symbols, images, sounds, electromagnetic waves, bits, etc. [11]. Then this digital personal data can be divided in:

- Data that users give to systems - This type of data is given in input by an individual to the system he is using, including self-tracking information, social networks sites data (including videos, pictures, texts and tweets), emails and videos. It is important to notice the case of social networks sites because these have increased the potential to collect personal data that individuals consciously give to systems.
- Data that systems extract from users - This type of data is extracted by a system from its user and is collected on behalf of other entities (that may be the same social network sites of the previous case). The reasons could be different, such as surveillance, harvesting people's activities or structurally required (e.g. a Mobile Service Provider must know its users' location in order to redirect telephone calls). It is necessary to stress the fact that the entity that controls the system that generated this data can claim its control [10].
- Data that systems process on behalf of users - This type of data is generated by inferring on data of previous types. This kind of data is processed by

data processors in order to obtain more meaningful information regarding individuals.

4.1 Data that users give to (Smart Transportation) systems

A novel system architecture has been presented that allows to create, store and share personal data generated by users in a Smart Transportation System [25]. The use of distributed ledgers and related technologies can serve as the basis to build novel smart services and to promote social good for what concerns individuals' personal data. Users within the Transportation Network produce various kind of data coming from their vehicle or smartphone, that actually concerns the same features exploited by social networks, e.g. user's location and activities. An approach may be to use a composition of different services, such as:

- IOTA DLT [18], to provide a level of data traceability, verifiability and immutability;
- IPFS to store data [1] ;
- Zero Knowledge Proof [9] algorithms to guarantee the necessary privacy;
- Ethereum Smart Contracts to control data access.

The shift from current centralized technologies consists in the fact that the user completely controls access on data he generated. This infrastructure gives the opportunity to build a Data Marketplace based on the personal data individuals decide to sell or to set access rules in order to provide data for the social good.

4.2 Data that systems extract and process from users

The more the data is centralized in "silos" (not communicating between each other), the more individuals lose control over their personal data information. Conversely, in this work we propose the use of a unique digital space for each data subject, where data flow is ruled and data providers and consumers can meet to transact. A possible solution to this can be achieved through decentralization and shared standards.

In particular, the use of DLTs to represent and transact with personal data would grant data validation and access control, as well as no central point of failure, immutability and most importantly traceability. Moreover, it is possible to use decentralized file systems, that allow continuous data availability. These properties are necessary in order to associate each individual to the digital space that will contain his personal data and that will be used to attend the requests of data providers and data consumers. Crucial is the use of smart contracts, since they provide a new paradigm where unmodifiable instructions are executed in an unambiguous manner during a transaction between two parts. Without the presence of a third party, then, a user may completely control the access to his personal data, being sure that his decisions on how and when to access his data are always observed. Every process is completely traced and permanently stored

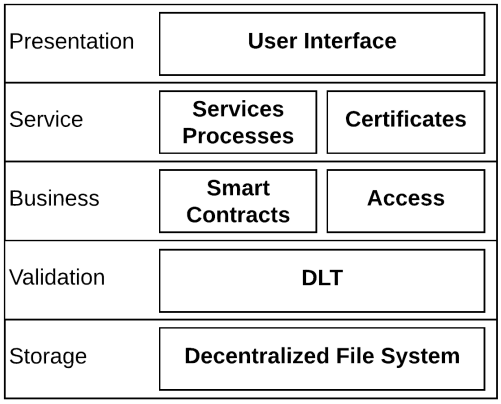

Fig. 1. Layered Architecture

in the blockchain. For what concerns the expression of legal requirements and privacy preferences, and the compliance with GDPR, smart contracts unintelligibleness (i.e. how their instructions, expressed in a programming language, become a contract) still needs deep investigation.

However, our interest is on the decentralization aspects of these contracts. And most of all, how the user's data flow can be regulated. The interesting aspect is that smart contracts, i.e. algorithms executed by a machine, allow two parties, e.g. data subject and provider or data provider and consumer, to reach an agreement in the process of the data flow. Nevertheless, smart contracts algorithms should be described to satisfy at least the following reasoning tasks: (i) determine if a policy satisfies the legal requirements; (ii) determine if a data request can be satisfied according to the individual's preferences and the legal constraints. On the other hand interoperability can be best achieved if a network of ontologies is used to model the personal data life-cycle and their actors.

4.3 Semantic web based policies

The smart contracts must be thus represented in a language that favours reasoning and a language that eases interoperability. Fortunately, the W3C has published over the last twenty years a set of specifications to describe resources which simultaneously addresses these two design goals: those of the semantic web.

Whereas these specifications were born to represent data in the web, their use has gone beyond and today many applications run totally offline but using the semantic web specifications. In the most spread paradigm, information is represented using RDF (Resource Description Framework). In this framework, resources are identified with URIs and described with collections of triples. The

precise meaning of each resource can be formally established with OWL ontologies. An ontology is a formal representation of knowledge through a set of concepts and a set of relations between these concepts, within a specific domain. Through the use of these ontologies it is possible to convey the meaning of data, hence to facilitate cross-domain applications and services. Ontologies in these scenarios effectively act as data models, eventually complemented with RDF Shapes[5], which further impose restrictions on the data that are easy to be evaluated.

Whereas new ontologies can be created whenever necessary, there is a set of *de facto* standard ontologies which should be reused whenever possible. For example, there are ontologies to describe the basic personal contact information, such as vCard[6], to describe basic geographical information[7] or to represent computer policies[8] or contracts[19]. Other vocabularies and ontologies have recently appeared in the domain of privacy and data protection [16][15].

The two advantages of 'interoperability' and 'reasoning' can be now well illustrated: first because the aforementioned ontologies are recommended by the W3C and thus universally understood. Second, reasoning with the information represented using these data models is easy because they are mapped in a formal language. An individual may want to say: whenever I am in the province of Lombardy, I want my data not to be transacted. If properly connected to other datasets, the system knowing that the individual is in Milano will infer that the individual is also in Lombardy and should not tranfer the data.

Zero Knowledge Proof While Semantic web technologies can be considered transversely to layers in the architecture of Figure 1, the Certificate component lies on top of the components that constitute the infrastructure pillars, and right under the user interface. That is because it serves as the middle layer that protects individual's privacy over his data. The use of "suitable" cryptographic techniques, such as Zero Knowledge Proof, may allow to prove that an individual possesses a certain property without revealing his data. For instance, using Zero Knowledge Proof of Location [23] is possible to prove that an individual finds himself in a certain zone without revealing his exact location.

5 Vision and conclusions

5.1 Vision

After having explained how a unique digital space can be built, it is fundamental to explain its use. The main idea is that this infrastructure can lead personal data flow towards a "safe" place where the individual can enforce his rights. There are different actors behind the successful implementation of this vision.

[5] https://www.w3.org/TR/shacl/
[6] https://www.w3.org/TR/vcard-rdf/
[7] https://www.w3.org/2003/01/geo/
[8] https://www.w3.org/TR/odrl-vocab/

First of all, the individual is obviously favoured because he assumes the full control over such digital structure. Then, all the actors behind the decentralized structure are incentivized by the use of the technology specification itself, e.g. monetary retribution. Finally, the main actors who use the space both to provide and gather data, i.e. data providers and consumers, are the one to which focus on. In particular, GDPR requires data providers to release personal data to data subjects, but this does not implies the use of the digital space. The use of common standards provided by Semantic web is a necessary incentive, but not sufficient. Hence both providers and consumers must be incentivized by the data market that generates behind the digital space. This is matter of investigation, in particular regarding the complex structure that the system may assume.

5.2 Conclusion

GDPR has brought an important evolution for what concerns individuals' personal data protection, providing them rights to contrast data abuse. However, the data flow that occurs behind the scenes between data providers and consumers and the creation of data "silos" prevent the execution of transparent processes at the eye of data subjects. A possible approach may be the one where each individual maintain his personal digital space in which his personal data is stored and transacted. This can be achieved through decentralized technologies in the form of DLTs, decentralized file systems and smart contracts that provide transparency and data access control, and through semantic web technologies in the form of linked data that provide data portability. Further methods may protect the privacy of individuals, e.g. Zero Knowledge Proof, while new methodologies for the analysis of such systems may bring to light new GDPR interaction models, e.g. to understand possible actors and manners to infer data.

Acknowledgment

This work has been partially funded by the EU H2020 MSCA project LAST-JD-RIOE with grant agreement No 814177 and PROTECT with grant agreement No 813497.

References

1. Benet, J.: Ipfs-content addressed, versioned, p2p file system. arXiv preprint arXiv:1407.3561 (2014)
2. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284**(5), 28–37 (2001)
3. Buterin, V., et al.: Ethereum white paper (2013), `https://github.com/ethereum/wiki/wiki/White-Paper`
4. Chen, Y., Xie, H., Lv, K., Wei, S., Hu, C.: Deplest: A blockchain-based privacy-preserving distributed database toward user behaviors in social networks. Information Sciences **501**, 100 – 117 (2019). https://doi.org/https://doi.org/10.1016/j.ins.2019.05.092

5. Council of European Union: Regulation (eu) 2016/679 - directive 95/46
6. De Hert, P., Papakonstantinou, V., Malgieri, G., Beslay, L., Sanchez, I.: The right to data portability in the gdpr: Towards user-centric interoperability of digital services. Computer Law & Security Review **34**(2), 193–203 (2018)
7. Faber, B., Michelet, G., Weidmann, N., Mukkamala, R., Vatrapu, R.: Bpdims: A blockchain-based personal data and identity management system. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. pp. 6855–6864. Hawaii International Conference on System Sciences (HICSS), United States (2019). https://doi.org/10125/60121
8. Farshid, S., Reitz, A., Roßbach, P.: Design of a forgetting blockchain: A possible way to accomplish gdpr compatibility. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)
9. Feige, U., Fiat, A., Shamir, A.: Zero-knowledge proofs of identity. Journal of cryptology **1**(2) (1988)
10. Hearn, A.: Structuring feeling: Web 2.0, online ranking and rating, and the digital'reputation'economy. ephemera: theory & politics in organization **10** (2010)
11. Kitchin, R.: The data revolution: Big data, open data, data infrastructures and their consequences. Sage (2014)
12. Kudumakis, P., Wilmering, T., Sandler, M., Rodríguez-Doncel, V., Boch, L., Delgado, J.: The challenge: From mpeg intellectual property rights ontologies to smart contracts and blockchains. Signal Processing Magazine **37**(2) (2019)
13. Maesa, D.D.F., Mori, P., Ricci, L.: Blockchain based access control. In: IFIP international conference on distributed applications and interoperable systems. pp. 206–220. Springer (2017)
14. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2009), `http://www.bitcoin.org/bitcoin.pdf`
15. Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.: Pronto: Privacy ontology for legal reasoning. In: International Conference on Electronic Government and the Information Systems Perspective. pp. 139–152. Springer (2018)
16. Pandit, H.J., O'Sullivan, D., Lewis, D.: An ontology design pattern for describing personal data in privacy policies. In: WOP at ISWC. pp. 29–39 (2018)
17. Pangrazio, L., Selwyn, N.: 'personal data literacies': A critical literacies approach to enhancing understandings of personal digital data. New Media & Society **21**(2), 419–437 (2019)
18. Popov, S.: The tangle. cit. on p. 131 (2016)
19. Rodríguez-Doncel, V., Delgado, J., Llorente, S., Rodríguez, E., Boch, L.: Overview of the mpeg-21 media contract ontology. Semantic Web **7**(3), 311–332 (2016)
20. Sambra, A.V., Mansour, E., Hawke, S., Zereba, M., Greco, N., Ghanem, A., Zagidulin, D., Aboulnaga, A., Berners-Lee, T.: Solid : A platform for decentralized social applications based on linked data (2016)
21. Siris, V.A., Dimopoulos, D., Fotiou, N., Voulgaris, S., Polyzos, G.C.: Interledger smart contracts for decentralized authorization to constrained things. arXiv preprint arXiv:1905.01671 (2019)
22. Truong, N.B., Sun, K., Lee, G.M., Guo, Y.: Gdpr-compliant personal data management: A blockchain-based solution. arXiv preprint arXiv:1904.03038 (2019)
23. Wolberger, L., Fedyukovych, V.: Zero knowledge proof of location. `https://platin.io/yellowpaper` (2018)
24. Zichichi, M., Contu, M., Ferretti, S., D'Angelo, G.: Likestarter: a Smart-contract based social DAO for crowdfunding. In: Proc. of the 2st Workshop on Cryptocurrencies and Blockchains for Distributed Systems (2019)

25. Zichichi, M., Ferretti, S., D'Angelo, G.: A distributed ledger based infrastructure for smart transportation system and social good. In: IEEE Consumer Communications and Networking Conference (CCNC) (2020)