Relevant Research Questions For Decentralised (Personal) Data Governance

Kurteva, A.K.; Pandit, Harshvardhan J.

Publication date
2023

Document Version
Final published version

Published in
CEUR Workshop Proceedings

Relevant Research Questions For Decentralised (Personal) Data Governance

Anelia Kurteva[1,*,†], Harshvardhan J. Pandit[2,†]

[1]Delft University of Technology, Delft, The Netherlands
[2]ADAPT Centre, Dublin City University, Dublin, Ireland

Abstract
Protecting and preserving individuals' personal data is a legal obligation set out by the European Union's General Data Protection Regulation (GDPR). However, the process of implementing data governance to support that, in a decentralised ecosystem, is still vague. Motivated by the need for lawful decentralised data processing, this paper outlines several relevant questions from legal, privacy and technology standpoints that need to be considered.

Keywords
Decentralisation, Data Governance, GDPR, Trust, Privacy, Semantic Web

1. Introduction

The rapid growth of the data economy and the privacy implications accompanying it have motivated a new paradigm shift towards decentralisation of data on the Web. This digital transformation aims to foster data sovereignty and interoperability and to empower individuals by allowing them to take back control over their personal data. Decentralised technology such as SOLID [1] has shown promising results and has slowly started to replace centralised digital infrastructures in several organisations across the European Union (EU). Motivated by the need for lawful decentralised data processing, this paper outlines several relevant questions from legal, privacy and technology standpoints that need to be considered.

1.1. How to describe and catalogue data for decentralised interoperability?

In a decentralised ecosystem, resources (e.g. data) should be described in a way that supports their interoperability by different services and machines to facilitate their discoverability, reuse and correct interpretation of their use policies. Utilising a consistent RDF vocabulary such as the Data Catalog Vocabulary (DCAT) [1], which is a W3C recommendation for describing datasets and online services in a machine-readable format, can be a starting point. DCAT's support for semantically representing resources, their role within a system, the associated agents and the

Trusting Decentralised Knowledge Graphs and Web Data (TrusDeKW) Workshop at ESWC 2023
*Corresponding author.
†These authors contributed equally.
a.kurteva@tudelft.nl (A. Kurteva); harshvardhan.pandit@dcu.ie (H. J. Pandit)

[1]https://www.w3.org/TR/vocab-dcat-2/

ability to classify them in catalogues based on themes can help structure decentralised data sharing and direct a service to a specific available resource that can be used for the specific purposes. To use any such vocabulary to describe a decentralised resource it should be clear what data exists in the resource and its availability in terms of what agent, when and for what purpose can use it and it what way. Mechanisms that automatically ensure resource's quality and completeness (e.g. specification of its availability) based on the set standard resource description format are needed as well.

1.2. How to establish, trust, and verify identities in a decentralised system?

Merely using the Web's domain-based identity may not be sufficient or even feasible in all cases. For example, cases where identities may always need to be *known* - such as a company's *legal identity* for accountability purposes, while in other contexts the identity may need to be *hidden* - such as to create a safe space for marginalised communities that use pseudonyms or identifiers instead of their real names. The issue of how to issue and manage identity useful for 'contextual identification' therefore also becomes an issue of trust to show or hide identities, to not misuse it, counter malpractices such as fraud - without surveillance or exposing sensitive information regarding private lives.

1.3. How to identify and ensure security of data and processing in decentralised systems?

Decentralised systems distribute the responsibilities for security mechanisms to be ensured and enforced across three levels. First for data - which could be encrypted, hidden from discovery, or be spread across locations. Second for data storage and transfer infrastructure, such as through encrypted communications or access control. Third in the *secure processing* of data, where involvement of multiple systems and actors establishes requirements for each actor to *identify* and *ensure* security of data and processing taking place *elsewhere* to avoid to detect lapses in security, such as failure to validate correctness or data breaches. While decentralisation reduces the scale for affected data, it increases the severity as all sensitive data relating to a context or individuals would be present within the single breached resource. Establishing accountability is a challenge under the current cybersecurity and legal frameworks due to lack of precedent and knowledge.

1.4. How to support individuals in making sense of decentralised data sharing?

It should be clear from the start if personal data (Art. 4(1)) and what specific categories of it (Art. 9) will be processed. In such cases, even in decentralised settings, at least one of GDPR's legal basis for data processing should be met to be legally compliant. For consent to be a legal base, it should be freely given, informed and unambiguous (Art. 7). Mediums for requesting consent online, such as web cookies, often present ambiguous information regarding the data processing, have deceptive design and dark patterns that are often invisible to a non-experts eye. User agents that correctly interpret GDPR's legal basis and their specific requirements and that have knowledge of common dark patterns can be used to filter out deceptive cookie banners

and consent requests. This will also minimise the information overload and consent fatigue of individuals and can help establish a level of trust in the agent and services that request access to data. Consent requests via cookies are usually accompanied with privacy policies that outline how data should be managed (i.e. gathered, used, disclosed). However, due to defining these rules in legal language and their length, they are often overlooked by data subjects. This is the case especially with non-experts who do not have the legal knowledge and patience to correctly interpret the privacy policy's content. To solve this, graph visualisations and UIs integrating them have emerged as a possible solution to ease individuals' comprehension [2].

1.5. How to establish responsibility and foster accountability across actors in decentralised settings?

Currently in centralised systems service providers are responsible for storing and processing individuals' data in a legally compliant way. In a decentralised system, data subjects are given control and ownership of their data, which can be a burden [3]. The responsibilities of agents should be clearly defined, agreed upon and described within each resources' metadata to establish accountability and transparency. For example, each resource can be catalogued and licensed (e.g. use of technology such as Data Licenses Clearance Center (DALICC) [4]). Machine-readable contracts [5]) and consent [6], outlining each actors' duties and responsibilities, can also be defined with specific service providers to minimise consent fatigue.

1.6. How to balance the legal obligations and responsibilities for decentralised actors?

Regulatory frameworks such as the GDPR are based on the conventional notions of centralised organisations collecting and managing data (as Controllers) that may utilise other actors to process it on their behalf (as Processors). The interpretation of such regulations towards decentralised solutions is unknown, which creates uncertainty, which ultimately hinders progress. Blindly charging forth with innovation may therefore end up not only harming the individuals involved, but also the service providers who want to develop new markets. A pragmatic and proactive solution therefore is to create solutions that function within the existing established boundaries of law, while developing new interpretations or legislations to facilitate further decentralisation. For example, there are no circumstances where users shouldering all the legal responsibilities of being a Controller have greater 'freedom', and instead will only face 'burdens' and 'exploitation' from lack of knowledge or willingness. Therefore a reasonable path forward is to identify mechanisms that either establish responsibilities, such as through model contracts for decentralised infrastructure and service providers, or to share responsibilities, such as through community bargaining and gatekeepers. While these happen, we should also engage with lawmakers and authorities to provide formal guidelines for the same and to develop future legislations. Of note, the European Union has already passed the Data Governance Act and has proposed Data Spaces that advance this conversation.

1.7. How to develop infrastructure and tools for decentralised systems?

In order to set up decentralised systems and services, an essential requirement is the availability of necessary infrastructure and tooling. For example, identity providers, data and processing associated resources - such as for storage, querying, computing, etc. - as well as specific tools for developers to create and users to consume and manage these resources. Before researching new methods to achieve the intended functionality, it is also necessary to enquire whether any of the existing tools and services can be reused or repurposed to provide all or some of the requirements. Where the market ecosystem has well established practices based on formal or de-factor standards, its reuse would be beneficial to increase the penetration and adoption of decentralised systems. For example, cloud technologies have reached the stage where they are widespread, are the subject of extensive standardisation, and have regulatory frameworks guiding responsible usage. Can we identify the "innovation" such existing technologies require to realise the decentralised vision and push market actors to developed these based on new markets and values? In parallel, existing infrastructure also has useful governance structures that can aid in resolving some of the pending issues with decentralisation. For example, rather than thinking of decentralisation as separation of independent nodes, we can establish decentralisation as communities where trust of services could be managed with gate-keeping or certification mechanisms such as that used within the app stores. For all the above, the existence of standards or common specifications is not a strict *necessity*, but will certainly accelerate development and adoption.

1.8. What is required to develop effective tools for automation in decentralised systems?

Automation requires machine-readable information, which also needs to be interoperable if it is to be shared between systems. While we are a community that propagates semantic interoperability to achieve decentralisation, the key question to ask ourselves is this: *"Can we ever reach an agreement to develop a standard?"* for any of the described topics here. While we have a variety of W3C recommendations as standards, and several tools and ontologies - often arising from large projects, we have neither seen their wider adoption and thus *effectiveness*, nor their acknowledgement as being *superior*. So the first requirement for the community is identifying what "standards" exist and what standards *should* exist - and from this creating a roadmap for achieving those. The second requirement is engaging with stakeholders to establish the minimum requirements agreeable to all, and codifying those as a standard to provide a guiding framework for interoperable solutions. The third requirement is then extending this standard with opinionated tools and methodologies to create operational services.

2. Conclusions

Decentralised systems are not identical replications spread over multiple locations, but instead facilitate diversity and variance while relying on commonality to communicate and inter-operate. Therefore, as long as we have a common vocabulary (e.g. Data Privacy Vocabulary

(DPV)[2] [7]) that can be semantically expressed and whose interpretation is *well-defined*, we can have decentralised solutions that act in a predictable manner while being free to perform with any technology or tools that they prefer to use. All of the above research questions that we have outlined should therefore be reframed to ask how to reach an agreement on communication of that topic between decentralised systems.

Acknowledgments

Anelia Kurteva is financially supported by the RePlanIT project funded by a Topsector Energy subsidy from the Ministry of Economic Affairs and Climate Policy in the Netherlands. The author thanks Ruud Balkenende and Alessandro Bozzon for their support and supervision. Harshvardhan J. Pandit's research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Center for AI-Driven Digital Content Technology at Dublin City University 13/RC/2106_P2. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulnaga, T. Berners-Lee, Solid: a platform for decentralized social applications based on linked data, MIT CSAIL & Qatar Computing Research Institute, Tech. Rep. (2016).

[2] C. Bless, L. Dötlinger, M. Kaltschmid, M. Reiter, A. Kurteva, A. J. Roa-Valverde, A. Fensel, Raising awareness of data sharing consent through knowledge graph visualisation, in: Further with Knowledge Graphs, IOS Press, 2021, pp. 44–57.

[3] H. J. Pandit, Making sense of Solid for data governance and GDPR, Information 14 (2023) 114.

[4] T. Pellegrini, V. Mireles, S. Steyskal, O. Panasiuk, A. Fensel, S. Kirrane, Automated rights clearance using semantic web technologies: The DALICC framework, Semantic Applications: Methodology, Technology, Corporate Use (2018) 203–218.

[5] A. Tauqeer, A. Kurteva, T. R. Chhetri, A. Ahmeti, A. Fensel, Automated gdpr contract compliance verification using knowledge graphs, Information 13 (2022) 447.

[6] A. Kurteva, T. R. Chhetri, H. J. Pandit, A. Fensel, Consent through the lens of semantics: State of the art survey and best practices, Semantic Web (2021) 1–27.

[7] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, et al., Creating a vocabulary for data privacy: The first-year report of data privacy vocabularies and controls community group (dpvcg), in: On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings, Springer, 2019, pp. 714–730.

[2] https://w3id.org/dpv