Data & Policy (2020), 2: e4, 1–10
doi:10.1017/dap.2020.1

COMMENTARY

Data protection by design: Building the foundations of trustworthy data sharing

Sophie Stalla-Bourdillon[1], Gefion Thuermer[2]*, Johanna Walker[2], Laura Carmichael[1] and Elena Simperl[2]

[1]Law School, University of Southampton, Southampton, United Kingdom
[2]Electronics & Computer Science, University of Southampton, Southampton, United Kingdom
*Corresponding author. Email: gefion.thuermer@soton.ac.uk

(Received 20 August 2019; revised 14 November 2019; accepted 15 January 2020)

Keywords: data-driven innovation; data protection by design; data trusts; General Data Protection Regulation; organizational DPbD process

Abstract

Data trusts have been conceived as a mechanism to enable the sharing of data across entities where other formats, such as open data or commercial agreements, are not appropriate, and make data sharing both easier and more scalable. By our definition, a data trust is a legal, technical, and organizational structure for enabling the sharing of data for a variety of purposes. The concept of the "data trust" requires further disambiguation from other facilitating structures such as data collaboratives. Irrespective of the terminology used, attempting to create trust in order to facilitate data sharing, and create benefit to individuals, groups of individuals, or society at large, requires at a minimum a process-based mechanism, that is, a workflow that should have a trustworthiness-by-design approach at its core. Data protection by design should be a key component of such an approach.

Policy Significance Statement

There is an emerging consensus that safe data-sharing environments are crucial to encourage data flows between actors and accelerate innovation. These safe data-sharing environments have sometimes been described as data trusts. In this article, we suggest that the key to prevent and minimize risks for individuals in the context of data sharing is that all parties involved in data sharing follow a common workflow comprising three phases. Focusing on workflows and processes rather than legal forms is the most effective way to ensure that data-related practices can be considered trustworthy.

Introduction

Data protection by design (DPbD) was recently introduced into law via Article 25 of the General Data Protection Regulation (GDPR). The requirement of DPbD builds upon research and applied work conducted in the field since the end of the 1990s (Cavoukian, 2009). Article 25(1) places a legal obligation on controllers[1] to "*implement appropriate organisational and technical measures […] designed to implement data-protection principles […] in order to meet the requirements of this Regulation and*

[1] The following legal definition of controller is provided by Article 4(7) of the GDPR: "*'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law*".

protect the rights of data subjects." DPbD therefore plays a key role in enabling and demonstrating compliance with the GDPR.

In this article, we address the question of how the requirements of DPbD should shape the development of data trusts (this concept is explored in more detail below). We will argue that both technical and organizational requirements are foundational to ensuring trustworthy data sharing. We further insist on the necessity of starting with organizational measures and creating a DPbD process, which are prerequisites to the selection of appropriate technical measures.

In order to strengthen our claim, we also draw on our experience as interdisciplinary members of Data Pitch[2]—an open innovation program—to inform our proposed approach. Data Pitch aims to bring together data providers (i.e., corporate and public sector organizations) to share data with successful program applicants (i.e., startups and small and medium enterprises [SMEs]) to reuse for innovation purposes. The project launched in January 2017 and will end in December 2019. It is funded by the European Union's Horizon 2020 Research and Innovation Programme.[3]

Data Trusts

Data trusts have been conceived as a mechanism to enable the sharing of data across entities where other formats, such as open data or commercial agreements, are not appropriate and make data sharing easier, more scalable (Hall and Pesenti, 2017), and mutually beneficial for members (Lawrence, 2016). Although the form and purposes of data trusts are currently a topic of much discussion (e.g., Alsaad et al., 2019; Hardinges, 2018; O'Hara, 2019; Wylie and McDonald, 2018), a broadly accepted definition has not yet emerged. This is in part because data trusts may be of benefit in data-driven innovation, as well as many other situations such as personal or health data management (Lawrence, 2016) and security, safety, and efficiency, like in the Internet of Food Things project.[4] The concept of the "data trust" requires further disambiguation from other facilitating structures such as data collaboratives (Susha et al., 2017). Furthermore, the use of data trusts as an internal data-sharing methodology, as it is established by firms such as Truata,[5] has created further ambivalence around the term.

By our definition, a data trust is a legal, technical, and organizational structure for enabling the sharing of data (Walker et al., 2019); they can assist with the exchange of data for a variety of purposes, one of which is to help solve business or societal problems. In that, they differ from data collaboratives, which have the distinct goal to solve societal problems through collaboration between organizations from diverse sectors (Verhulst et al., 2015). For data trusts, as well as related structures such as data collaboratives, the design, development, and utilization of robust mechanisms for responsible data sharing are crucial to engender trust and ultimately drive forward data-driven innovation and achieve their organizational goals, regardless whether these are social or economic.

The need for increased data sharing

Data-driven innovation is regarded as a new "growth area" for the global economy (Organisation for Economic Co-operation and Development [OECD], 2015). Given data-driven innovation is contingent upon "*the use of data and analytics to improve or foster new products, processes, organisational methods and markets*" (OECD, 2015), it is vital that interested parties have lawful access and rights to (re)use vast amounts of robust data where necessary and appropriate. It is therefore unsurprising that a key obstacle to the growth of data-driven innovation is a lack of data sharing (Mehonic, 2018; Skelton, 2018)—also

[2] https://datapitch.eu/

[3] For more information about Data Pitch, visit the project website at https://datapitch.eu/ (last accessed on May 10, 2019).

[4] For more information about the Internet of Food Things project, see https://www.foodchain.ac.uk/ (last accessed on May 10, 2019).

[5] For more information about Truata, see https://www.truata.com/ (last accessed on May 10, 2019).

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

referred to as the "data-pooling problem" (Mattioli, 2017). For instance, a deficiency of training datasets has led to the failure of multiple private and public machine learning initiatives (Mehonic, 2018).

Alongside economic benefits, innovation enabled through greater data sharing also provides many societal and ecological benefits. For instance, data-driven innovation may lead to improved customer service experiences, such as through the extended use of chat-bots, and better diagnosis or more efficient provision of care in health services. When data sharing is used to increase efficiencies in industry, it does not only save costs for businesses, but can also reduce emissions and energy consumption, and thereby improve air quality and public health, or even help tackle climate change. In the public sector, data sharing could help to improve road safety, traffic flows, or maintenance, making for a safer public environment. More direct benefits for citizens are found in new products that improve individual control of personal data, and increase organizations' compliance with the GDPR (Thuermer et al., 2019).

There are numerous reasons why organizations may be reticent to share data for innovation purposes, including concerns over privacy, data quality, free-riding, competition, reputation, and proprietary issues (Mattioli, 2017). Data trusts are proposed as one approach that could encourage increased data sharing and reusage within a wider data-driven innovation strategy[6]; especially for personal and anonymized data (Edwards, 2004; Reed and Ng, 2019).

Sharing personal data

The GDPR applies only to information pertaining to an identified or identifiable natural person. In many instances of data sharing, however, as has been shown by Data Pitch, the data that are shared are, or could become personal data. With sensitive sectors such as healthcare and research increasingly utilizing artificial intelligence, this is only likely to increase (Lawrence, 2016).

Personal data should be processed only where necessary, proportionate, and lawful. As a starting point, organizations should consider whether types of data-sharing activities with lower risk of reidentification to data subjects are most appropriate in the given circumstances, for example, sharing anonymized data[7] or fully synthetic data[8] rather than personal data. However, the use of anonymized data is not always suitable, in particular, where more granular individual-level data are required (e.g., for some patient-based studies). Furthermore, while the use of fully synthetic data may minimize the risk of reidentification substantially, it may not be an option in all instances as "*the truthfulness of the data is lost*" (Surendra and Mohan, 2017). The quality of synthetic data also varies, as it is dependent on the standards of its underlying generation practices (UK Government Department for Digital, Culture, Media, & Sport, 2018).

In all other cases, organizations should minimize the risk of reidentification as far as possible, for example, by deidentifying personal data through other anonymization techniques that remove and/or mute certain personal-identifying features to protect the privacy of individuals. Such deidentified data may be rendered anonymous or pseudonymous—note that the latter remains personal data and therefore falls under the scope of the GDPR. For instance, differential privacy[9] is "one of the strongest" (Stalla-Bourdillon, 2019a) anonymization techniques, and may be employed by an organization to publish aggregate data while retaining individual-level data internally.

[6] For further elements of such a strategy, see, for example, the British Academy and The Royal Society (2017) report, which focuses on the need for "a renewed governance framework" and stewardship body for trustworthy data sharing.

[7] Recital 26 of the GDPR defines anonymous information as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*" For further guidance on anonymisation practices, see Anonymisation: Managing Data Protection Risk Code of Practice; Information Commissioner's Office (ICO) (2012) and Elliot et al. (2016).

[8] Note that there are three main types of synthetic data identified by Surendra and Mohan (2017): (i) fully synthetic data—that is, data are "completely artificially generated"; (ii) partially synthetic data—that is, some original values are masked by artificial data; and (iii) hybrid synthetic data—that is, "*[f]or each record of original data a nearest record in the synthetic data is chosen and both are combined to form hybrid data*" (Surendra and Mohan, 2017).

[9] For more information on differential privacy, see Dwork (2006).

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

Once again, while deidentified personal data may preserve privacy, in some cases, it may significantly reduce the utility of data, for example, by decreasing the accuracy of linkability across datasets. The words of guidance issued by the UK Government Department for Digital, Culture, Media, and Sport (2018) make clear that: "*[i]t is important to remember that pseudonymising or anonymising data does not make it automatically appropriate to use. It is possible to make incorrect inferences and develop potentially intrusive or damaging policies based on less identifiable data.*" It is therefore essential that organizations strike the correct balance between preserving the privacy of data subjects while also maintaining a sufficient level of data utility—there is no one-size-fits-all approach. Organizations need to ensure that robust organizational and technical measures are in place, which ultimately remain suited to the specific context and purpose of the processing activity in question.

Given the broad definition of personal data, it is imperative that those designing, developing, and utilizing data trusts remain compliant with the GDPR. Due to the key role of DPbD in enabling and demonstrating compliance with the GDPR, it is vital that we further explore how the requirement of DPbD does or should impact upon the construction of data trusts. As O'Hara (2019) argues, the purpose of data trusts is to "support trustworthy data processing," which is achieved by applying constraints that go beyond the law. This requires determining what the law actually mandates and adding to its prescription.

Data Protection by Design

Despite the concept of privacy-by-design being well established in principle, its technical implementation has been rather limited thus far (Hansen, 2016; Tsormpatzoudi et al., 2016). Given Article 25 of the GDPR now directly places a legal obligation on controllers to practice DPbD, there is a real incentive for its widespread implementation. Especially, as pursuant to Article 83(4), any infringements of Article 25 may result in "*administrative fines up to 10 000 000 EUR, or in the case of an undertaking, up to 2 % of the total worldwide annual turnover of the preceding financial year, whichever is higher.*"

Organizational as well as technical measures

It still remains difficult to find practical DPbD guidance that provides extensive coverage of both the organizational and technical dimensions mandated by Article 25. When DPbD is presented and explained, the focus is often set on its technological dimension (Wiese Schartum, 2016)—the engineering of data protection principles through design strategies and privacy enhancing techniques (Danezis et al., 2015; Deng et al., 2011). Less emphasized is that the requirement also has a vital organizational dimension—that is, Article 25(1) places a legal obligation on controllers to "*implement appropriate organisational and technical measures […] to protect the rights of data subjects.*" For instance, organizational measures may refer to the adoption of particular procedures and the selection of particular individuals to decide and action various aspects of data processing, including the type of privacy-enhancing technologies (PETs) to be employed across the data sharing and reusage lifecycle (The Royal Society, 2019).

Seven core data-protection principles

This organizational dimension of DPbD implies a particular workflow, that is, a series of accountable decisions and actions taken by responsible individuals with appropriate expertise prior to the commencement of the data processing activities under consideration. Note that an organization may also choose to automate some of these decisions for reasons of scalability; in that case, the accountable decisions by individuals concern the design of the automation.

The main nodes of this workflow echo the seven core data-protection principles at the heart of the GDPR and directly referred to by Article 25(1): (a) "lawfulness, fairness, and transparency"; (b) "purpose limitation"; (c) "data minimization"; (d) "accuracy"; (e) "storage limitation"; (f) "integrity and confidentiality"; and (g) "accountability." These data protection principles are outlined in GDPR Article 5 and impose high-level restrictions upon how personal data should be collected and used, how data quality

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

should be ensured and maintained, and how personal data should be protected. These principles are particularly important when data are not only processed internally, but also shared between organizations.

DPbD workflow

Essentially, before any processing starts, the data controller should put in place technical and organizational measures in order to facilitate compliance with the data protection principles as listed in Article 5. Article 25 thus refers to Article 5. The basic structure for a DPbD workflow—comprising eight nodes I–VIII—can be derived from Article 5 of the GDPR as follows:

I. Define your *purpose* for sharing data in this instance. (See Article 5(1)(b)—"purpose limitation.")
II. Identify your *legal basis* for sharing data in this instance. (See Article 5(1)(b)—"purpose limitation.")
III. Determine which data are *necessary* for your specific purpose. Ensure that you *reduce*: (a) any *nonessential processing* activities and (b) the *amount of data* required—for example, mask or hide direct identifiers that are not required for processing in this instance. *If you can anonymize data, just do it!* (See Article 5(1)(c)—"data minimization.")
IV. Set a *data retention period* in relation to the purpose. (See Article 5(1)(e)—"storage limitation.")
V. Ensure the data to be shared are *accurate.* (See Article 5(1)(d)—"accuracy.")
VI. Verify that the processing is *fair.* (See Article 5(1)(a)—"lawfulness, fairness, and transparency.")
VII. Ensure the data are *not altered or disclosed without permission*—for example, define who is eligible to access data—and the processing is *confidential.* (See Article 5(1)(f)—"integrity and confidentiality.")
VIII. Ensure the processing is *transparent* and *monitored*, for example, by logging activities so that you can know what is happening with the data (and ultimately demonstrate compliance). *Best practice: assess risk before initiating processing.* (See Article 5(1)(a)—"lawfulness, fairness, and transparency" and Article 5(2)—"accountability.")

If data trusts are the mechanism through which data sharing will be enabled in the future, it is therefore clear that they should embed a DPbD workflow, and thereby be underpinned by organizational and technical measures as defined by GDPR Article 25.

Two lessons learnt from Data Pitch

After familiarization with the DPbD workflow in principle, the next step toward trustworthy data sharing is determining how to carry out this DPbD workflow in practice. From our experience with Data Pitch, we raise two key organizational lessons learnt for successful implementation of a DPbD workflow.

Strong engagement across business functions for responsible data sharing and reusage

Responsible data sharing can be viewed as a chain of decisions and actions.[10] For instance, a company may consider: why it may wish to share data; what kind of entity might be eligible to access the data; what the purpose of data sharing is; what authority it has to share the data; and how it might ensure that the data sharing is compliant. It is extremely unlikely these decisions and actions will be taken by one person alone. Such decision-making needs strong engagement across business functions—from security experts and data scientists to data protection officers and business strategists.[11] Senior-level support is crucial to overcome ambiguities in the decision-making process.

[10] For instance, Bunting and Lansdell (2019) examine how to design "decision-making processes for data trusts."
[11] Tsormpatzoudi et al. (2016) also highlight the importance of an interdisciplinary approach for effective DPbD implementation.

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

An agreed process for accountable decision-making

It is vital that there is a process in place where organizational and technical measures are selected to uphold the seven core data-protection principles across the lifecycle of the data processing activity (e.g., over the course of an open innovation program). These organizational and technical measures must be appropriate, that is, well-suited to the specific context and purpose of the data processing activity in question.

Embedding a DPbD Approach Within Data Trusts

Therefore, we argue that the effective entrenchment of DPbD within the construction of data trusts requires (at least)

1. Cognizance of the minimum legal requirements for DPbD—including both its organizational and technical dimensions—as mandated by Article 25 together with its accompanying DPbD workflow located in Article 5.
2. An organizational DPbD process that addresses (at minimum) the legal requirements for DPbD across the entire data trust lifecycle (i.e., from initial plans for creating a data trust to a data trust in operation).
3. Strong, cross-functional business engagement that brings the required expertise to successfully shape, execute, and appraise the DPbD process.

Given that we have already examined both points (1) and (3), we will now turn our attention to what an organizational DPbD process for data trusts is likely to involve. Note that we are only able to signpost some key aspects of a DPbD process to act as a point of reference for data trusts—there is no one-size-fits all approach. A DPbD process must always take into account the specific context and purpose of the data sharing and reusage activities in question.

Scenario

A few organizations are interested in working together to form a new data trust. This data trust would be centered around the creation of a data pool so as to improve their current levels of innovation activities. This data pool would involve each organization sharing their data with authorized members of the data pool, that is, the other organizations and (potentially) third parties. A significant amount of these datasets are likely to be personal or anonymized.

Three-layer approach

As there is no agreed configuration for data trusts, we represent data trusts through three core layers that feature in many data-sharing ecosystems. These three core layers comprise: (a) the data layer—where interested parties make plans to create a data pool; (b) the access layer—where pooled data are made discoverable through a data trust; and (c) the process layer—where pooled data are approved for (re-) usage via the data trust. Note that data may be stored centrally (e.g., all datasets will be held by the data trust) or disparately (e.g., individual datasets will be held by different parties).

The data layer: preparation of data sources

DPbD should be embedded into the plans for the new data trust through the following process:

1. Ensure that all potential members are aware of the legal requirements for DPbD (in particular Article 25 and Article 5)—and the overarching DPbD process for the data trust. Recognize any gaps in knowledge—and provide further training and guidance where necessary.
2. Identify the appropriate persons across all organizations that have the authority and required expertise to decide and action on the pooled data.

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

3. Provide clear guidelines for reviewing data in the planned data pool, including guidance on: how to assess whether data can be understood as personal data and high risk processing.
4. Apply standardized procedures for the removal of unnecessary personal data. The data minimization principle should directly impact the way datasets are redacted and presented. For instance, direct identifiers should be stripped away as often and as early as possible to minimize the personal data contained in datasets.

The access layer: discovery of pooled data

The datasets within the planned data pool should then be made discoverable to authorized parties through metadata. DPbD should be embedded into the access layer of the new data trust through the following process:

5. Define who is eligible to access the pooled data, and place limitations on who accesses the data, and why. These boundaries are defined around the purpose of the data trust itself, but also include a clear distinction between the raw data and metadata.
6. Provide standardized access through centralized technical solutions, underpinned by monitoring and auditing processes, or provide governance processes to manage peer-to-peer direct sharing that enable auditing.

The process layer: approval of pooled data (re)usage

The (re)usage of datasets within the data pool should be managed by the data trust, which should be in the position to make informed decisions about whether (or not) to permit data sharing with interested parties. DPbD should be embedded into the process layer of the new data trust through the following process:

7. Control data usage through standardized risk assessments. Once the processing purpose and data sources are confirmed, there should be an assessment of the intended versus allowed use of the data, to guarantee in particular the lawfulness and fairness of processing and ultimately the impact upon the rights and freedoms of data subjects. Such an assessment should be done in context of the intended use, and therefore renewed each time a new purpose is suggested. Once again, risk assessment is the key for accountability. Importantly, risk assessment should be iterative—it should start as early as the pooling phase and be reviewed at the inception of the reusage phase.
8. Ensure that data are tailored to queries. Queries that are interested in aggregates should only be responded to with aggregate data. Where raw data are required, this should be limited to the necessary attributes. Traditional techniques based on extract, transform, load should be reconsidered as they tend to create unnecessary movements of data. The potential for PETs, such as differential privacy, should be fully explored at this stage.[12]

Examples from the Data Pitch program

Guidance

Data Pitch provided guidance on the key legal and privacy aspects of data sharing and reusage of (closed) data for a variety of purposes that can be understood by nonlegal specialists through their key resources: The Legal and Privacy Toolkit (2017, 2018, 2019). Privacy and data protection is a key focus for the Legal and Privacy Toolkit, including (a) strategy for pseudonymization and anonymization; (b) guidance on the data spectrum; (c) high-risk processing; and (d) data flow mapping as one method which organizations sharing and/or reusing data can use, to support and demonstrate legal compliance with the GDPR and other applicable laws.

[12] For instance, Stalla-Bourdillon (2019b) provides an overview of some DPbD methods for data analytics projects.

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

Contracts and oversight

All organizations formally taking part in the Data Pitch program signed a bilateral, asynchronous contract. The Data Pitch consortium supported all the organizations to interpret and instill best practices throughout their involvement. The consortium required data providers and SMEs to provide information about the data they intended to share and/or reuse as part of the program via a Data Provider Questionnaire or Self-Sourced Data Record, and made risk-minimizing suggestions. For instance, where data providers proposed to supply data that had been subject to pseudonymization processes ahead of reuse by the participating SMEs, the Data Pitch consortium could oversee and recommend the implementation of best practice safeguards on a case-by-case basis.

Data ethics

Compliance with data ethics is complementary to the Legal and Privacy Toolkit, for example, it was obligatory for SMEs to sign an Ethics Statement in order to participate in the program.

Training

Training related to the Legal and Privacy Toolkit was provided to participating SMEs via workshops and webinar in order to further promote legal and ethical awareness. The provision of more interactive forms of dissemination was also important for improved engagement, such as group tasks based on data sharing and reusage scenarios and interactive legal decision-trees.

Data access

Data Pitch provided access to metadata through a dedicated platform that enabled applicants to explore the available datasets without exposing the actual data. Once contracts were signed, a direct exchange of data between the data holders and participants took place. In the majority of cases, the data were shared to the SME's infrastructure or to the commercial cloud (paid for by the SME); in a smaller number of cases, the data remained in the provider's infrastructure or was accessed on the commercial cloud paid for by the provider.[13]

Data protection impact assessments

Data Pitch required some data users to evaluate their use of the data through a data protection impact assessment where necessary. This ensured that the purpose of data use was sufficiently reflected upon, and any risks to data subjects were addressed before processing commenced.

Conclusion

Data trusts, as legal, technical, and organizational structures to enable the sharing of data, are conceived as an important tool to engender trust as part of a wider response to data sharing barriers that may impede data-driven innovation. Given the likelihood that the data to be shared may be personal data or could become personal data (e.g., through purpose or result of use, reidentification), it is vital that data trusts embed DPbD through the implementation of appropriate organizational and technical measures that uphold the seven core data-protection principles at the heart of the GDPR. The DPbD workflow defined by Article 5 is therefore key to the effective implementation of the appropriate organizational and technical safeguards that lead to trustworthy data sharing.

There is an opportunity for data trusts to lead the way with the practical implementation of DPbD by giving equal attention to its organizational and technical dimensions. Strong engagement across business functions will be critical for the creation and adoption of well-considered processes that embed DPbD.

[13] https://datapitch.eu/wp-content/uploads/2018/08/D2.3_v5.pdf

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

Acknowledgment. An earlier version of this article was made available for the Data for Policy conference 2019.[14]

Funding Statement. Data Pitch is funded by the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement 732506.

Competing Interests. The authors declare no competing interests exist.

Authorship Contributions. S.S.-B.: conceptualization, writing—original draft, writing—review and editing; G.T., J.W.: investigation, writing—review and editing; L.C.: writing—original draft, writing—review and editing; E.S.: writing—review and editing, funding acquisition.

Data Availability Statement. Details about the Data Pitch program used in the case study are available at https://datapitch.eu/.

Abbreviations

DPbD	data protection by design
GDPR	General Data Protection Regulation
PETs	privacy-enhancing technologies
SME	small and medium enterprises

References

Alsaad A, **O'Hara K and Carr L** (2019) *Institutional Repositories as a Data Trust Infrastructure.* In Proceedings of Web Science 2019, 30 June–3 July 2019. Boston, MA: ACM.

British Academy & The Royal Society (2017) Data Management and Use: Governance in the 21st Century. *Report.* Available at https://royalsociety.org/.

Bunting M and Lansdell S (2019) Designing Decision-Making Processes for Data Trusts: Lessons from Three Pilots. *Report.* Available at https://theodi.org/.

Cavoukian A (2009) *Privacy by Design: The 7 Foundational Principles*. Canada: Information and Privacy Commissioner of Ontario.

Danezis G, **Domingo-Ferrer J**, **Hansen M**, **Hoepman J-H, Le Métayer D, Tirtea R and Schiffner S** (2015) Privacy and Data Protection by Design—From Policy to Engineering. *European Union Agency for Network and Information Security (ENISA) Report.* Available at https://www.enisa.europa.eu/.

Deng M, **Wuyts K**, **Scandariato R**, **Preneel B and Joosen W** (2011) A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering* **16**(1), 3–32.

Dwork C (2006) *Differential Privacy.* In Proceedings of International Colloquium on Automata, Languages, and Programming (ICALP), pp. 1–12.

Edwards L (2004) The problem with privacy: A modest proposal. *International Review of Law, Computers & Technology* **18**(3), 263–294.

Elliot M, **Mackey E**, **O'Hara K and Tudor C** (2016) The Anonymisation Decision-Making Framework. *UK Anonymisation Network (UKAN).* Available at https://ukanon.net/.

Hall W and Pesenti J (2017) Growing the Artificial Intelligence Industry in the UK. *Independent Review.* Available at https://www.gov.uk/.

Hansen M (2016) Data protection by design and by Default à la European General Data Protection Regulation. In Lehmann A, Whitehouse. D. Fischer-Hübner, S. Fritsch, L. and Raab, C. (eds.), *Privacy and Identity Management. Facing up to Next Steps*. Cham, Switzerland: Springer, pp. 27–38.

Hardinges J (2018) What is a Data Trust? *Open Data Institute Blog.* Available at https://theodi.org.

Information Commissioner's Office (ICO). (2012) Anonymisation: Managing Data Protection Risk Code of Practice. Available at https://ico.org.uk/.

Lawrence N (2016) Data Trusts Could Allay our Privacy Fears. *The Guardian.* Available at https://www.theguardian.com/uk.

Mattioli M (2017) The data-pooling problem. *Berkeley Technology Law Journal* **32**(1), 179–236.

Mehonic A (2018) Can Data Trusts be the Backbone of our Future AI Ecosystem? *The Alan Turing Institute Blog.* Available at https://www.turing.ac.uk/.

O'Hara K (2019) Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship. *Web Science Institute White Paper.* Available at https://eprints.soton.ac.uk/.

[14] https://zenodo.org/record/3079895

https://doi.org/10.1017/dap.2020.1 Published online by Cambridge University Press

Organisation for Economic Co-operation and Development (OECD) (2015) Data-Driven Innovation: Big Data for Growth and Well-Being. *Report.* Available at https://www.oecd.org/.

Reed C and Ng I (2019) Data Trusts as an AI Governance Mechanism: Response to the Singapore Personal Data Protection Commission. Available at https://www.ssrn.com/.

Skelton SK (2018) New Forms of Governance Needed to Safely and Ethically Unlock Value of Data. *ComputerWeekly.com.* Available at https://www.computerweekly.com.

Stalla-Bourdillon S (2019a) Anonymising personal data: Where do we stand now? *Privacy & Data Protection* **19**(4), 3–5.

Stalla-Bourdillon S (2019b) Data protection by design and data analytics: Can we have both? *Privacy & Data Protection* **19**(5), 8–10.

Stalla-Bourdillon S and Carmichael L (2018) Legal and Privacy Tookit v2. With Contribution from Zhang P. Available at https://datapitch.eu/.

Stalla-Bourdillon S, and Knight A (2017) Legal and Privacy Tookit v1. Available at https://datapitch.eu/.

Stalla-Bourdillon S, and Carmichael L (2019) Legal and privacy aspects of transnational, cross-sector data sharing in open innovation. With Contribution from Zhang P. Available at https://datapitch.eu/wp-content/uploads/2020/01/D3.9.pdf.

Surendra H, and Mohan HS (2017) A review of synthetic data generation methods for privacy preserving data publishing. *International Journal of Scientific & Technology Research* **6**(3), 95–101.

Susha I, **Janssen M and Verhulst S** (2017) *Data Collaboratives as a New Frontier of Cross Sector Partnerships in the Age of Open Data: Taxonomy Development.* In Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 2691–2700.

The Royal Society (2019) Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis. *Report.* Available at https://royalsociety.org/.

Thuermer G, **Walker JC and Simperl E** (2019) Data Sharing Toolkit. Available at https://datapitch.eu/.

Tsormpatzoudi P, **Berendt B and Coudert F** (2016) Privacy by design: From research and policy to practice—The Challenge of multi-disciplinarity. In Berendt B, Engel T, Ikonomou D, Le Métayer D and Schiffner S (eds.), *Privacy Technologies and Policy.* Cham, Switzerland: Springer, pp. 199–212.

UK Government Department for Digital Culture Media & Sport (2018) Guidance: 3. Use Data that is Proportionate to the User Need—How to Implement Principle 3 of the Data Ethics Framework for the Public Sector. Available at https://www.gov.uk/.

Verhulst S Sangokoya D and The GovLab. (2015) Data Collaboratives: Exchanging Data to Improve People's Lives. Available at https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a.

Walker JC, **Simperl E**, **Stalla-Bourdillon S and O'Hara K** (2019) *Decision Making Processes for Data Sharing: A Framework for Data Trusts.* ACM WomENcourage 2019: Celebration of Women in Computing, Rome, Italy. Available at https://eprints.soton.ac.uk/434736/.

Wiese Schartum D (2016) Making privacy by design operative. *International Journal of Law and Information Technology* **24**(2), 151–175.

Wylie B and McDonald S (2018) What is a Data Trust? *Centre for International Governance Innovation (CIGI).* Available at https://www.cigionline.org/.

Cite this article: Stalla-Bourdillon, S. Thuermer, G. Walker, J. Carmichael, L. and Simperl, E. 2020. Data protection by design: Building the foundations of trustworthy data sharing. *Data & Policy* 2: e4, 1–10, doi:10.1017/dap.2020.1