A Provenance Model for the European Union General Data Protection Regulation

Benjamin E. Ujcich[1,2] (✉), Adam Bates[3], and William H. Sanders[1,2]

[1] Department of Electrical and Computer Engineering
[2] Information Trust Institute
[3] Department of Computer Science
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
{ujcich2,batesa,whs}@illinois.edu

Abstract. The European Union (EU) General Data Protection Regulation (GDPR) has expanded data privacy regulations regarding personal data for over half a billion EU citizens. Given the regulation's effectively global scope and its significant penalties for non-compliance, systems that store or process personal data in increasingly complex workflows will need to demonstrate how data were generated and used. In this paper, we analyze the GDPR text to explicitly identify a set of central challenges for GDPR compliance for which data provenance is applicable; we introduce a data provenance model for representing GDPR workflows; and we present design patterns that demonstrate how data provenance can be used realistically to help in verifying GDPR compliance. We also discuss open questions about what will be practically necessary for a provenance-driven system to be suitable under the GDPR.

Keywords: data provenance, General Data Protection Regulation, GDPR, compliance, data processing, modeling, data usage, W3C PROV-DM

1 Introduction

The European Union (EU) General Data Protection Regulation (GDPR) [1], in effect from May 2018, has significantly expanded regulations about how organizations must store and process EU citizens' personal data while respecting citizens' privacy. The GDPR's effective scope is global: an organization offering services to EU citizens must comply with the regulation regardless of the organization's location, and personal data processing covered under the regulation must be compliant regardless of whether or not it takes place within the EU [1, Art. 3]. Furthermore, organizations that do not comply with the GDPR can be penalized up to €20 million or 4% of their annual revenue [1, Art. 83], which underscores the seriousness with which organizations need to take the need to assure authorities that they are complying.

A recent survey [2] of organizations affected by the GDPR found that over 50% believe that they will be penalized for GDPR noncompliance, and nearly 70% believe that the GDPR will increase their costs of doing business. The same

survey noted that analytic and reporting technologies were found to be critically necessary for demonstrating that personal data were stored and processed according to data subjects' (*i.e.,* citizens') consent.

Achieving GDPR compliance is not trivial [3]. Given that data subjects are now able to withhold consent on what and how data are processed, organizations must implement controls that track and manage their data [4]. However, "[organizations] are only now trying to find the data they should have been securing for years," suggesting that there is a large gap between theory and practice, as the GDPR protections have "not been incorporated into the operational reality of business" [5]. Hindering that process is the need to reconcile high-level legal notions of data protection with low-level technical notions of data usage (access) control in information security [3].

In this paper, we show how *data provenance* can aid greatly in complying with the GDPR's analytical and reporting requirements. By capturing how data have been processed and used (and by whom), data controllers and processors can use data provenance to reason about whether such data have been in compliance with the GDPR's clauses [6–8]. Provenance can help make the compliance process accountable: data controllers and processors can demonstrate to relevant authorities that they stored, processed, and shared data in a compliant manner. Subjects described in the personal data can request access to such data, assess whether such data were protected, and seek recourse if discrepancies arise.

Our contributions include: 1) explicit codification of where data provenance is applicable to the GDPR's concepts of rights and obligations from its text (Section 2.1); 2) adaptation of GDPR ontologies to map GDPR concepts to W3C PROV-DM [9] (Section 3); and 3) identification of provenance design patterns to describe common events in our model in order to answer compliance questions, enforce data usage control, and trace data origins (Section 4). We also discuss future research to achieve a provenance-aware system in practice (Section 5).

2 Background and Related Work

2.1 GDPR Background

The GDPR "[protects persons] with regard to the processing of personal data and ... relating to the free movement of personal data" by "[protecting] fundamental rights and freedoms" [1, Art. 1]. The regulation expands the earlier Data Protection Directive (DPD) [10], in effect in the EU since 1995, by expanding the scope of whose data are protected, what data are considered personally identifiable and thus protected, and which organizations must comply. As a result, it mandates "that organizations [must] know exactly what information they hold and where it is stored" [2]. Although the law does not prescribe particular mechanisms to ensure compliance, the law does necessitate thinking about such mechanisms at systems' design time rather than retroactively [2, 4].

The GDPR defines data *subjects* identified in the personal data, data *controllers* who decide how to store and process such data, and data *processors* who

Table 1. GDPR Concepts of Rights and Obligations as Applicable to Provenance.

Concept	Explanation	Provenance Applicability
Right to Consent [1, Arts. 6–8]	Controllers and processors can lawfully process personal data when subjects have given consent "for one or more specific purposes."	Provenance can model the personal data for which consent has been given, the purposes for which consent is lawful, and the extent to which derived data are affected.
Right to Withdrawal [1, Art. 7]	Subjects can withdraw consent regarding their personal data's use going forward but without affecting such data's past use.	Provenance can verify past compliance from before the withdrawal and prevent future use.
Right to Explanation [1, Arts. 12–15]	Subjects may ask controllers for explanations of how their data have been processed "using clear and plain language."	Provenance-aware systems can naturally provide such explanations by capturing past processing.
Right to Removal [1, Art. 17]	Controllers must inform processors if subjects wish to remove or erase their data.	Provenance can track when such removal requests were made, what data such requests affect, and to what extent derived data are affected.
Right to Portability [1, Art. 20]	Subjects can request their data from controllers or ask controllers to transmit their data to other controllers directly.	A common provenance model would allow each controller to link its respective provenance records with others' records.
Obligation of Minimality [1, Art. 25]	Controllers must not use any more data than necessary for a process.	Provenance can help analyze such data uses with respect to processes.

process such data on the controllers' behalf [1, Art. 4]. *Recipients* may receive such data as allowed by the subject's *consent*, which specifies how the personal data can be used. Controllers and processors are answerable to public *supervisory authorities* in demonstrating compliance.

For each GDPR concept that is a right of a subject or an obligation of a controller or processor, we summarize in Table 1 where data provenance can be applicable using the GDPR's text and where data provenance can help benefit all involved parties from technical and operational perspectives.

2.2 Related Work

The prior research most closely related to ours is that of Pandit and Lewis [8] and Bartolini *et al.* [3]. Both efforts develop GDPR ontologies to structure the regula-

tion's terminology and definitions. Pandit and Lewis [8] propose GDPRov, an extension of the P-Plan ontology that uses PROV's `prov:Plan` to model expected workflows. Rather than use plans that require pre-specification of workflows, we opted instead for creating relevant GDPR subclasses of PROV-DM agents, activities, and entities and encoding GDPR semantics into PROV-DM relations. Our model allows for more flexible specifications of how data can be used (*i.e.,* under consent for particular purposes while being legally valid for a period of time). Furthermore, our model focuses on temporal reasoning and online data usage control, whereas it is not clear how amenable GDPRov is to such reasoning or enforcement. The ontology of Bartolini *et al.* [3] represents knowledge about the rights and obligations that agents have among themselves. We find that a subset of that ontology is applicable in the data provenance context for annotating data, identifying justifications for data usage, and reasoning temporally about whether data were used lawfully. Bonatti *et al.* [7] propose transparent ledgers for GDPR compliance. Basin *et al.* [11] propose a data purpose approach for the GDPR by formally modeling business processes. Gjermundrød *et al.* [12] propose an XML-based GDPR data traceability system.

Aldeco-Pérez and Moreau [13] propose provenance-based auditing for regulatory compliance using the United Kingdom's Data Protection Act of 1998 as a case study. Their methodology proposes a way to capture questions that provenance ought to answer, to analyze the actors involved, and to apply the provenance capture. For using provenance as access control, Martin *et al.* [6] describe how provenance can help track personal data usage and disclosure with a high-level example of the earlier DPD [10]. Bier [14] finds that usage control and provenance tracking can support each other in a combined architecture via policy decision and enforcement points. Existing systems such as Linux Provenance Modules [15] and CamFlow [16] can collect provenance for auditing, access control, and information flow control for Linux-based operating systems.

3 GDPR Data Provenance Model

Motivated by data provenance's applicability to GDPR concepts as outlined in Table 1, we define a GDPR data provenance model based on the data-processing components of prior ontologies [3, 8]. Our model is controller-centric because the GDPR requires that controllers be able to demonstrate that their data processing is compliant, though we imagine that both controllers and processors will collect provenance data. Figure 1 graphically represents the GDPR data provenance model's high-level classes and their relations.

Tables 2, 3, and 4 explain the high-level classes shown in Figure 1 for Agent, Activity, and Entity W3C PROV-DM classes, respectively. Some high-level classes (*e.g.,* the Process activity) include subclasses (*e.g.,* the Combine activity) either because their notions are explicitly mentioned in the GDPR text or because they align with Bartolini *et al.*'s ontology for representing GDPR knowledge. We assigned more specific semantic meanings to several W3C PROV-DM relations; those meanings are summarized in Table 5.

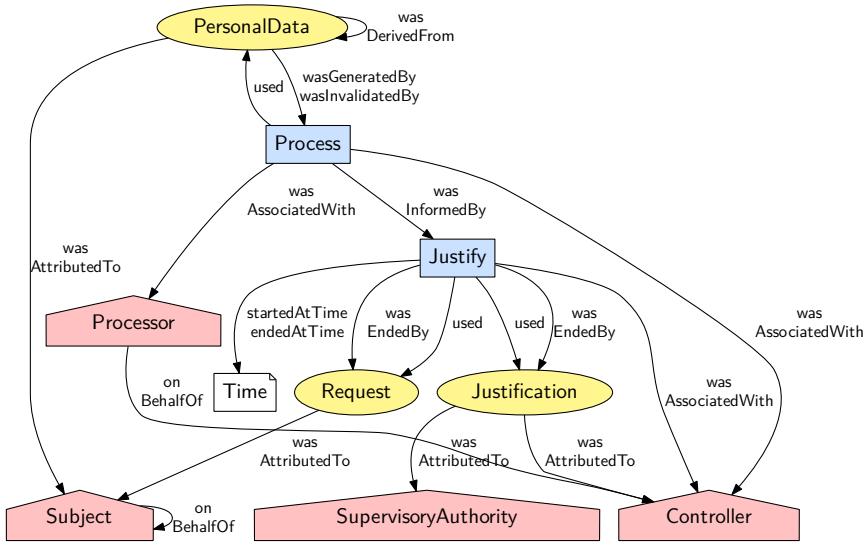

Fig. 1. GDPR data provenance model with high-level classes. House symbols represent agents (Table 2); rectangles represent activities (Table 3); ellipses represent entities (Table 4); arrows represent relations (Table 5); and notes represent other properties.

We found that the GDPR includes strong temporal notions throughout its text that affect whether processing is considered lawful. For instance, the notion of consent with respect to data usage may be valid only for a particular period of time. We use data provenance not only for capturing data derivations, but also for temporally reasoning about data usage, as we detail in Section 4.

4 Using the GDPR Data Provenance Model

Although the GDPR data provenance model describes *what* provenance to collect, it does not explain *how* to use such provenance. We present design patterns that modelers and practitioners can use to describe common events. We use a running example based on the examples from prior works [8, 11] that involve collecting personal data for a retail shop. We assume that a customer, Alice, interacts with the retailer by registering, making purchases, and subscribing to marketing information. We assume that each node and relation has a timestamp of its insertion into the graph so that we can perform temporal queries.

4.1 Design Patterns

Data Collection and Consent by a Subject At time τ, Alice registers with and provides her personal data to the retail shop, along with her consent. Figure 2 shows the provenance generated from these activities. Our design pattern decouples the personal data collected (PersonalData entities) from the subject's consent

Table 2. GDPR Data Provenance Model Agent Classes.

Class	Explanation and Subclasses
Subject	An "identifiable natural person ... who can be identified, directly or indirectly, in particular by reference to an identifier" [1, Art. 4]. *Subclasses*: Child subjects who cannot consent on their own and Parent subjects who can consent on their behalf [1, Art. 8].
Controller	An organization "which ... determines the purposes and means of the processing of personal data" [1, Art. 4]. *Subclasses*: EURecipient controllers (with country subclasses), NonEURecipient controllers (with country subclasses). (Data processing or transmission that leaves the EU is subject to additional regulations [1, Arts. 44–50].)
Processor	An organization "which processes personal data on behalf of the controller" [1, Art. 4].
Supervisory Authority	"An independent public authority" [1, Arts. 4, 51–59] that can "monitor and enforce the application of" the GDPR and "handle complaints lodged by a data subject ... and investigate" [1, Art. 57].

Table 3. GDPR Data Provenance Model Activity Classes.

Class	Explanation and Subclasses
Process	"Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means" [1, Art. 4]. *Subclasses*: Collect, Store, Retrieve, Combine, Disclose to another controller or processor via transmission; Erase to destroy personal data to fulfill the right to erasure [1, Art. 17]; Profile using "any form of automated processing ... to evaluate certain personal aspects relating to a natural person" [1, Art. 4]; or Pseudonymize by "processing of personal data [so that it] can no longer be attributed to a specific data subject without the use of additional information" [1, Art. 4].
Justify	The rationale that a controller uses in taking some action on personal data, which includes temporal notions of "start" and "end" times. *Subclasses*: a subject's Consent [1, Arts. 6–7]; a controller's Obligation, Interest, or Authority [1, Art. 6].

about such data (ConsentRequest entities), as personal data may be updated or rectified [1, Art. 16] independently of the giving of consent.

The GDPR specifies that processing is lawful when consent has been given "for one or more specific purposes" [1, Art. 6]. We represent this consent for personal data in relation to purposes as a design pattern in the provenance graph by mapping Consent activities to ConsentRequest entities with the used relation. As shown in Figure 2, Alice does not consent to use of her credit card

Table 4. GDPR Data Provenance Model Entity Classes.

Class	Explanation and Subclasses
PersonalData	An "identifier [of a subject] such as a name, an identification number, location data, an online identifier or to one or more factors specific to the ... identity of that natural person" [1, Art. 4]. *Subclasses*: DerivedData simplifies identification of data derived wholly or in part from PersonalData objects (by some Process).
Request	A request sent from a Subject to a Controller. *Subclasses*: ConsentRequest [1, Art. 6], WithdrawRequest [1, Art. 7], AccessRequest [1, Art. 15], CorrectionRequest [1, Art. 16], ErasureRequest [1, Art. 17], or a RestrictionRequest [1, Art. 18]
Justification	A justification (beyond a subject's consent) for lawful processing. *Subclasses*: LegalObligation "to which the controller is subject," a VitalInterest "of the data subject or of another natural person," a "performance of a task" in the PublicInterest, an OfficialAuthority "vested in the controller," a LegitimateInterest "pursued by the controller," or a Contract "to which the data subject is party" [1, Art. 6]

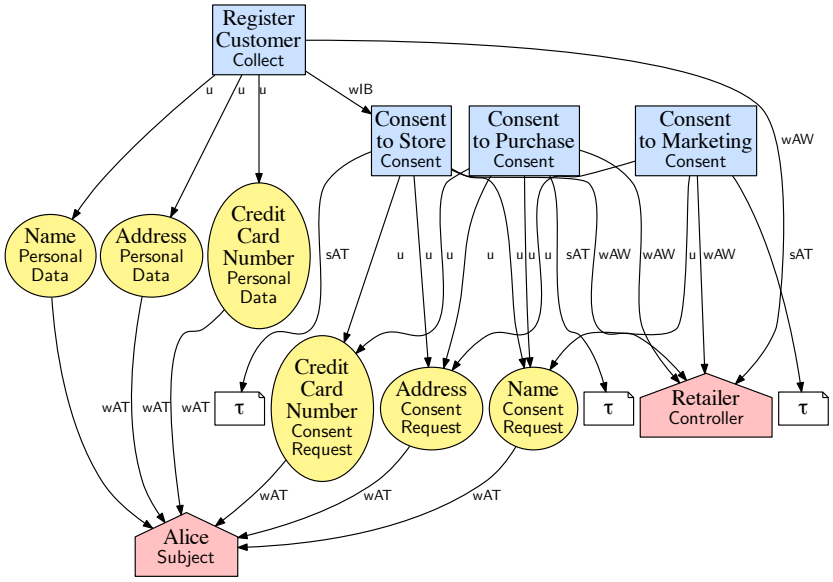

Fig. 2. Alice registers her personal data with a retail shop and consents to use of her data for storage, purchases, and marketing. Note that Alice does *not* consent to use of her credit card number for being shared for marketing purposes.

information for marketing, but she does allow it to be used for making purchases or for being stored by the retail shop (*e.g.*, to simplify future purchases).

Table 5. GDPR Data Provenance Model Relation Semantics.

From	Relation	To	Semantic Meaning
Process	wasInformedBy	Justify	Data processing actions under the GDPR require justification; we can reason about why data exist or why data were removed.
PersonalData	wasDerivedFrom	PersonalData	Data updates, such as corrections submitted by the subject as part of the right to rectification [1, Art. 16].
PersonalData	wasGeneratedBy or wasInvalidatedBy	Process	Personal data have lifespans. For instance, a subject may request that personal data be deleted. Both generation and invalidation require reasoning, so we use both relations.
Justify	used or wasEndedBy	Request or Justification	Justifications also have lifespans. For instance, a subject may withdraw his or her consent through a WithdrawRequest, which stops further data processing activities from using the Justify activity related to the withdraw request.
Justify	wasAssociated With	Controller	Justify activities are associated with controllers since controllers must keep such records for authorities; however, the information used to make the justification legal (*i.e.,* a Request or Justification entity) can be attributed directly to the source that produced it (*e.g.,* a Subject).

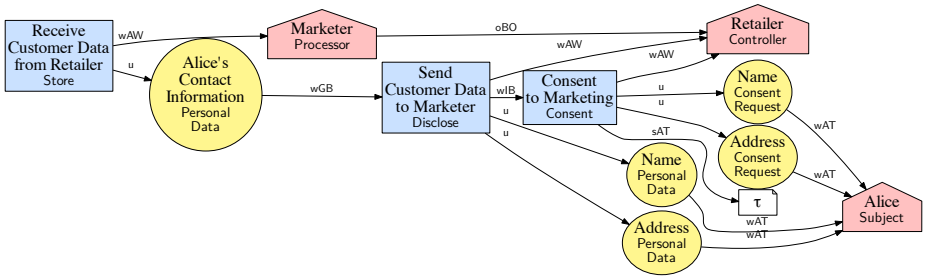

Fig. 3. The retail shop uses Alice's data for marketing purposes by employing a third-party marketer. The retail shop uses Alice's consent to receive marketing in allowing the processor to receive Alice's name and address. (For simplicity, portions of the provenance graph from Figure 2 that are not relevant are not shown.)

Data Transfers Among Controllers and Processors At time $\tau + 1$, suppose that the retail shop wishes to use a third-party marketing company to send marketing

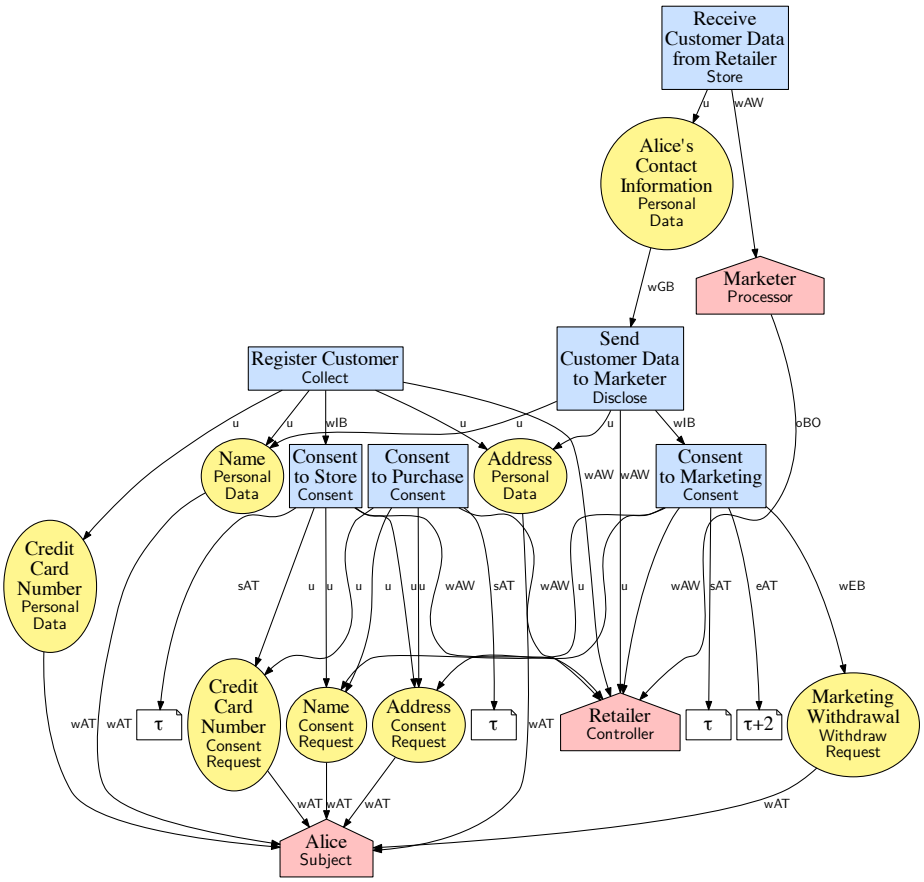

Fig. 4. Alice no longer wishes to receive marketing information, so she withdraws her consent for use of her personal data for marketing purposes. Future marketing activities that attempt to use the "Consent to Marketing" as justification will be noncompliant.

information to Alice. Figure 3 shows the provenance generated from the data transfer from the retail shop (the controller) to the marketing company (the processor). We model the transfer as a Disclose activity (by the controller) by which Alice's data are stored with a Store activity (by the processor).

Withdrawal by a Subject At time $\tau+2$, suppose that Alice no longer wishes to receive any further marketing from the retail shop. Figure 4 shows the provenance generated by her withdrawal of consent to receive marketing. We link the WithdrawRequest entity to the marketing Consent activity through the wasEndedBy relation to indicate that any prior Process that wasInformedBy the justification was valid, but that future uses will not be, after time $\tau + 2$. That ensures that

"withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal" [1, Art. 7].

4.2 Verifying Compliance

We can now use provenance to reason about several GDPR requirements, either at run time when a decision about data usage is being made (*i.e.,* access control) or after the fact during an audit. The choice of *when* to verify will depend on design decisions on what provenance information a controller or processor has the ability to access. We can answer compliance questions by querying a provenance graph such as the graph in Figure 4, as follows.

- *Was Alice's personal data used for marketing purposes after Alice withdrew her consent?* The "Send Customer Data to Marketer" activity was justified because it occurred during a time in which its justification activity, "Consent to Marketing," was valid (*i.e.,* after τ and before $\tau + 2$). If subsequent activities used the "Consent to Marketing" as justification after $\tau + 2$, then the controller would be noncompliant.
- *Who and what used Alice's address data?* One of the new operational and technical challenges with the GDPR is that of understanding where data "live" and what derived data are affected [2]. To answer this question, we start in Figure 4 at the PersonalData entity representing Alice's address and work backward from the relations. We find that her address was used during registration and was sent to and stored by the marketing processor as a bundled piece of contact information.
- *From the processor's perspective, under what usage conditions can Alice's address be used?* Processors are allowed to process data only if given the ability to do so by the controller. To answer this question, we start in Figure 4 at the "Receive Customer Data from Retailer" activity to find any paths in the graph that end at a Consent activity that, at the time of querying, have not yet ended. We find that the processor can use Alice's address to send marketing on behalf of the controller.

Our questions presented here are necessarily incomplete, but we find that provenance can be highly flexible in answering questions that subjects and supervisory authorities will have when controllers or processors are audited.

5 Discussion

Privacy Given that provenance collection includes metadata about *all* data processing activities, it introduces new privacy issues that will require that the metadata also be GDPR-compliant. That may require that PersonalData objects and Subject identifiers be stored as hashes of personal data and references to the personal data's actual locations rather than through embedding of the personal data in the provenance. We imagine that a *data protection officer* [1, Arts. 37–39]

will maintain access to the provenance graph for enforcing data usage control and for complying with audit requests. Subjects may be entitled to the portions of the controller's provenance graph related to their personal data [1, Arts. 12, 20]. Challenges arise, however, in ensuring a balance among the subject's fundamental rights [1, Art. 1], the privacy of the controller's own (proprietary) processes, and the privacy of other subjects so that releasing such data "shall not adversely affect the rights and freedoms of others" [1, Art. 20].

Standardization For inter-controller audits, we imagine that supervisory authorities will request provenance data from multiple controllers and processors so as to stitch together the relevant pieces of each's provenance graph. This will necessitate further standardization of 1) the granularity at which controllers and processors must collect provenance suitable for auditing; 2) the extent to which provenance collection mechanisms are built-in or retrofitted; and 3) tamper-proof and fraud-resistant provenance collection mechanisms.

Limitations Provenance collection and querying alone are not sufficient for meeting GDPR compliance, though we believe that automated provenance annotations will simplify much of the work involved in reasoning about data processing. The GDPR will always require some human activity to support reasoning about whether compliance was met or not [11]. Annotation of existing workflows and application processes (*e.g.,* reads and writes in databases) is generally a non-trivial and implementation-dependent process, though retrofitting of applications to collect provenance for information security [17] shows promise.

6 Conclusion

We outlined how data provenance can help with GDPR compliance by supporting reasoning about how data were collected, processed, and disseminated; reasoning about whether such collection and processing complied with subjects' intents; enforcing data usage control; and aiding auditing by authorities to check compliance. We presented a GDPR data provenance model and showed how our model can be used. Although many practical issues will need to be considered, we believe that provenance can reduce the burden on practitioners and make systems more accountable to the subjects from whom controllers collect data.

Acknowledgments

The authors would like to thank Jenny Applequist for her editorial assistance, the members of the PERFORM and STS research groups at the University of Illinois at Urbana-Champaign for their advice, and the anonymous reviewers for their helpful comments. This material is based upon work supported by the Maryland Procurement Office under Contract No. H98230-18-D-0007 and by the National Science Foundation under Grant Nos. CNS-1657534 and CNS-1750024. Any opinions, findings, and conclusions or recommendations expressed in this

material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)," in *Official Journal of the European Union*, vol. L 119, May 2016, pp. 1–88.
2. C. Tankard, "What the GDPR means for businesses," *Network Security*, vol. 2016, no. 6, pp. 5–8, 2016.
3. C. Bartolini, R. Muthuri, and C. Santos, "Using ontologies to model data protection requirements in workflows," in *Proceedings of New Frontiers in Artificial Intelligence*. Springer, 2017, pp. 233–248.
4. J. Vijayan, "6 ways to prepare for the EU's GDPR," *InformationWeek*, Sep. 2016.
5. W. Ashford, "Much GDPR prep is a waste of time, warns PwC," *ComputerWeekly*, Oct. 2017.
6. A. Martin, J. Lyle, and C. Namilkuo, "Provenance as a security control," in *Proceedings of the Theory and Practice of Provenance '12*. USENIX, 2012.
7. P. Bonatti, S. Kirrane, A. Polleres, and R. Wenning, "Transparent personal data processing: The road ahead," in *Proceedings of Computer Safety, Reliability, and Security*. Springer, 2017, pp. 337–349.
8. H. J. Pandit and D. Lewis, "Modelling provenance for GDPR compliance using linked open data vocabularies," in *Proceedings of Society, Privacy and the Semantic Web - Policy and Technology '17*, 2017.
9. World Wide Web Consortium, "PROV-DM: The PROV data model," https://www.w3.org/TR/prov-dm/, Apr. 2013.
10. Council of the European Union, "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 (Data Protection Directive)," in *Official Journal of the European Union*, vol. L 281, Nov. 1995, pp. 31–50.
11. D. Basin, S. Debois, and T. Hildebrandt, "On purpose and by necessity: Compliance under the GDPR," in *Proceedings of Financial Cryptography and Data Security '18*, Mar. 2018.
12. H. Gjermundrød, I. Dionysiou, and K. Costa, "privacyTracker: A privacy-by-design GDPR-compliant framework with verifiable data traceability controls," in *Proceedings of Current Trends in Web Engineering*. Springer, 2016, pp. 3–15.
13. R. Aldeco-Pérez and L. Moreau, "Provenance-based auditing of private data use," in *Proceedings of Visions of Computer Science '08*, 2008, pp. 141–152.
14. C. Bier, "How usage control and provenance tracking get together - A data protection perspective," in *Proceedings of IEEE 4th International Workshop on Data Usage Management*, May 2013, pp. 13–17.
15. A. Bates, D. Tian, K. R. B. Butler, and T. Moyer, "Trustworthy whole-system provenance for the Linux kernel," in *Proceedings of USENIX Security '15*, 2015, pp. 319–334.
16. T. Pasquier, J. Singh, D. Eyers, and J. Bacon, "CamFlow: Managed data-sharing for cloud services," *IEEE Transactions on Cloud Computing*, vol. 5, no. 3, pp. 472–484, July 2017.
17. A. Bates, W. U. Hassan, K. Butler, A. Dobra, B. Reaves, P. Cable, T. Moyer, and N. Schear, "Transparent web service auditing via network provenance functions," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 887–895.