Achieving GDPR Compliance through Provenance: An Extended Model

Daniel P. Campagna[1], **Altigran S. da Silva**[2], **Vanessa Braganholo**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

[2]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brazil

danielcampagna@id.uff.br, alti@icomp.ufam.edu.br, vanessa@ic.uff.br

Abstract. *The approval of the General Data Protection Regulation (GDPR) brought a revolution in the way we treat data produced in digital media. The GDPR increases individuals' participation in the treatment of their data, and it also introduces technical challenges, whose failure can lead to a fine of 4% of the organization's annual revenue. Among many approaches that aim to contribute to the solutions of challenges introduced by GDPR, there is a research branch promoting the use of data provenance as a means to make transparent the increasingly complex workflows of systems. However, existing provenance models are not fully compliant with the GDPR. In this paper, we aim to contribute to the evolution of the GDPR data provenance model proposed by Ujcich et al.. We suggest eleven new changes that make the model more apparent and more compatible with the GDPR text. We also present two design patterns that should guide us in using these changes in real contexts.*

Resumo. *A aprovação do Regulamento Geral de Proteção de Dado (GDPR) trouxe uma revolução na maneira como tratamos os dados produzidos em meios digitais. A GDPR inclui uma maior participação dos indivíduos no tratamento dos seus dados e também introduz desafios técnicos cuja preterição pode levar a uma multa de 4% da receita anual da organização. Dentre muitas abordagens que buscam contribuir na solução dos desafios introduzidos pela GDPR, existe um ramo que tem promovido o uso de proveniência de dados como um meio para tornar transparente os passos cada vez mais complexos dos sistemas. No entanto, modelos de proveniência existentes não são completamente aderentes à GDPR. Neste artigo, buscamos contribuir com a evolução do modelo de proveniência de dados da GDPR proposto por Ujcich et al.. Ao final, sugerimos onze novas mudanças que tornam o modelo mais claro e mais compatível com o texto da GDPR, além de dois padrões de projeto que nos orientam em como usar essas mudanças em contextos reais.*

1. Introduction

In April 2016, the European Parliament introduced a Copernican revolution in the European Data Privacy law. The famous movement proposed by Immanuel Kant in philosophy, in the 18th century, shifted the manner we build knowledge by moving away from concepts embodied in the external objects towards the judgments that run in

an individual's mind. The General Data Protection Regulation (GDPR) [Council of European Union 2016] brings up "a similar revolution in European data protection law by seeking to shift its focus away from paper-based, bureaucratic requirements and towards compliance in practice, harmonization of the law, and individual empowerment." [Kuner 2012]. In practice, since the GDPR came into force on May 25th, 2018, European citizens (*subjects*) have the right to be informed of legal compliance when their data (*personal data*) are collected, used, stored, and shared. *Subjects* are also called upon to allow, deny, or restrict the use of their data, in such a manner that companies under GDPR Article 6 can only process data after procuring subject's consent.

A 2016 survey [Tankard 2016] has reported a lack of technologies to aid in evaluating whether *personal data* were processed and stored according to the owner's consent. These results, somehow, reflect our current scene. A recent survey [GDPR.EU 2019] from May 2019 – involving 716 small business leaders in Spain, the United Kingdom, France, and Ireland – reported that only 44% of organizations agree that their organization "describes its data processing activities in clear, plain language to data subjects" [GDPR.EU 2019, Art. 12]. In a word, at a 95% confidence level (+/-4% error margin), the survey authors report that millions of small businesses, only in European territories, are still ignoring part of the GDPR requirements. The penalties for organizations that do not comply with GDPR can be up to €20 million or 4% of their annual revenue.

In the literature, provenance mechanisms have been used to aid in meeting these requirements [Ujcich et al. 2018, Aldeco Perez and Moreau 2008, Martin et al. 2012, Bartolini et al. 2015]. According to the *Oxford English Dictionary*, provenance is "the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners." The World Wide Web Consortium (W3C) defines PROV, an extensible provenance model that uses three main concepts: *entities*, *activities*, and *agents* [Moreau and Missier 2013]. The model then proposes seven relations that can be used to link these concepts. The three concepts could be comprehended as vertices of a provenance graph, while the relations are edges that connect these vertices.

From the work that addresses the GDPR problem using provenance [Ujcich et al. 2018, Aldeco Perez and Moreau 2008, Martin et al. 2012, Bartolini et al. 2015], the model proposed by Ujcich et al. [2018] is the only one that is based on the GDPR law – others are based on previous versions of the law, or on country specific laws. Such model encodes GDPR semantics into subclasses of PROV components, using ontologies [Bartolini et al. 2015, Pandit and Lewis 2017]. The authors also propose three use cases leaning on an example used in prior work that involves collecting personal data for a retail shop. Despite all contributions, we found this model is missing an analysis concerning the law principles.

This paper addresses this gap by identifying and reporting limitations we find when comparing the GDPR data provenance model proposed by Ujcich et al. [2018] and some relevant articles of the law. Also, for each limitation, we propose extensions to the original model. Finally, we present a comprehensive practical case study using this new extended model.

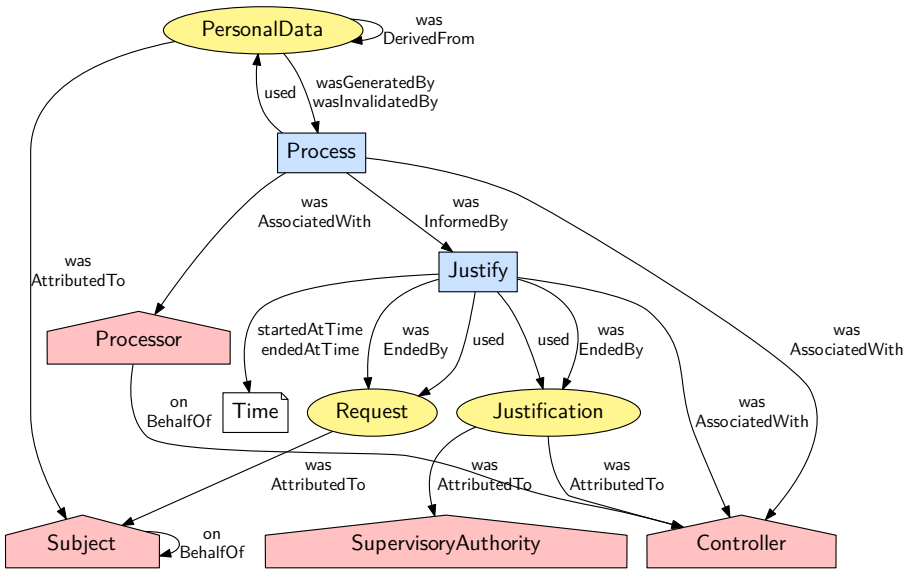

Figure 1. Ujcich et al.'s [2018] GDPR data provenance model with high-level classes.

The paper is organized as follows. Section 2 presents related work regarding the use of data provenance to cope with the GDPR requirements. Section 3 reviews the GDPR data provenance model proposed by Ujcich et al. [2018] . Section 4 discusses the limitations and introduces our extension to the GDPR data provenance model. Section 5 proposes two design patterns involving gathering and maintaining personal data for an on-line retail shop. Section 6 discusses future work and concludes this paper.

2. Related Work

Provenance is often used in e-science contexts to help in tasks related to the interpretation and understanding of results [Freire et al. 2008]. However, this need is also found in research that address problems involving data protection. For instance, Aldeco-Pérez and Moreau [2008] use provenance techniques to aid in auditing the United Kingdom's Data Protection (1998) Act. Martin et al. [2012] describe how provenance can help in tasks such as disclosing, access control, and data usage. Their context involves the earlier Data Protection Directive (1995) [Council of European Union 2016]. Moreover, other researchers explore provenance to solve different problems [Bonatti et al. 2017, Basin et al. 2018, Gjermundrød et al. 2016, Bier 2013].

Regarding the use of data provenance to address GDPR-related challenges, Ujcich et al. [2018] propose a GDPR data provenance model based on the data-processing components proposed in prior work [Bartolini et al. 2015, Pandit and Lewis 2017]. They use a subset of Bartolini et al. [2015] ontology, which represents knowledge about the obligations and rights that agents have among themselves. They also extend the proposal by Pandit and Lewis [2017] that introduces GDPRov. GDPRov extends the P-Plan ontology [Garijo and Gil 2013], which models expected workflow by using PROV's prov:Plan. Ujcich et al. 's model avoids using plans, as it requires pre-specification of the workflow. In exchange, the authors argue that more flexible specifications of how data can be used are possible. In general terms, their main contribution

is a translation of the rough text of the GDPR into a readable provenance language by recycling prior GDPR ontologies to map GDPR concepts into the W3C PROV model.

To deal with the practical issues of collecting provenance, there are whole-system provenance capture mechanisms such as Hi-Fi [Pohly et al. 2012], CamFlow [Pasquier et al. 2015], and Linux Provenance Module [Bates et al. 2015]. These mechanisms capture provenance information at the kernel-level for Linux-based operating systems. However, these systems are too fine-grained, which makes it difficult to use them to enforce data privacy policies.

There are also approaches not based on provenance data. Shastri et al. [2019] handle three traditional database management systems to ensure GDPR compliance. Pandit et al. [2019] demonstrate through use-cases that semantic representations of processes are useful towards maintaining ongoing GDPR compliance. Wang et al. [2019] propose a new paradigm for storing sensitive data. Their approach consists in encapsulating the personal data within a policy which governs its use. Therefore, to be accessed, a given piece of data needs to be submitted to a declassification step which checks whether its policy has been satisfied.

3. Background

Since this paper aims to analyze Ujcich et al.'s [2018] GDPR data provenance model (GDPR model), in this section we introduce an overview of this model. In their paper, the authors summarize six rights and obligations they identify in the GDPR text. Based on the text and recycling prior models, they propose their provenance model which is shown in Figure 1. This Figure graphically represents the high-level classes and relations of their proposed model. Table 1 contains a short explanation of each class of the model. They also propose several subclasses that enhance the model semantics towards the GDPR concepts.

For the purpose of the case study of this paper, we highlight the subclasses of Process and Justify. The remaining subclasses and a more detailed explanation are available in Ujcich et al.'s paper . Process has eight subclasses: Store, which represents an action in the past when some data were persisted; Disclose, which is useful to represent any data transmission between controllers and processors; Pseudonymize that represents "processing of personal data [so that it] can no longer be attributed to a specific data subject without the use of additional information". [Council of European Union 2016, Art. 4]. The remaining subclass are Collect, Retrieve, Combine, Erase, and Profile. Finally, the GDPR data provenance model expresses that any processing lawfulness should be informed by a *subject*'s Consent, or a *controller*'s Obligation, Interest or Authority [Council of European Union 2016, Art. 6]. In the model, these are all subclasses of Justify.

4. Extension to the GDPR Data Provenance Model

Now that we have the GDPR data provenance model [Ujcich et al. 2018] in place, it is possible to map the GDPR concepts into provenance components. However, while doing that, we found some limitations, for which we introduce extensions. In our efforts to propose extensions to this model, we have made a non-systematic read and interpretation and tried to fit the GDPR data provenance model components into each of

Component	Class	Explanation
Agent	Subject	An "identifiable natural person [...] who can be identified, directly or indirectly, in particular by reference to an identifier."
	Controller	An organization "which [...] determines the purposes and means of the processing of personal data."
	Processor	An organization "which processes personal data on behalf of the controller."
	Supervisory Authority	is "An independent public authority" [Council of European Union 2016, Arts. 4, 51-59] "that can monitor and enforce the application of" the GDPR and "handle complaints lodged by a data subject [...] and investigate." [Council of European Union 2016, Art. 57].
Activity	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."
	Justify	"The rationale that a controller uses in taking some action on personal data, which includes temporal notions of 'start' and 'end' times."
Entity	PersonalData	An "identifier [of a subject] such as a name, an identification number, location data, an on-line identifier or to one or more factors specific to the [...] identity of that natural person" [Council of European Union 2016, Art. 3].
	Request	A request sent from a Subject to a Controller for lawful processing.
	Justification	A justification for lawful processing.

Table 1. Classes of the PROV components proposed by Ujcich et al. [2018] .

the GDPR articles. In cases where it is not clear how a given GDPR article can fit the model, we discuss and propose extensions.

We found eleven limitations in which the model does not cover a concept. In this section, we present extensions that address all these limitations. However, we only discuss two of the proposed extensions (Sections 4.1 and 4.2) due to space constraints. The remaining are discussed in the project documentation page [1].

Tables 2-4 summarize all the limitations and extensions we propose. We classify all limitations and extensions into three types according to their impact on the model. The first set of extensions, which we call plumbing extensions, affects how the GDPR model works, and appends new elements (Table 2 summarizes them). The remaining are called porcelain (Table 3) and wallpaper (Table 4) extensions. Both intend to clarify and enrich this GDPR data provenance model by either suggesting new subclasses representing the semantic meanings of some part of the GDPR text or appending further information detailed in the law.

4.1. Provide the data subject with the period for which their data will be stored

Our first concern addresses the period for which each *personal data* will be stored. Both articles 13 and 14 [Council of European Union 2016, Arts 13, 14] prescribe which further information the *controllers* must provide to *subjects* whose *personal data* are collected. Into those, we found that the GDPR model does not cover the point a), which forces *controllers* to inform to *subjects* "the period for which [their] personal data will be stored [...]" [Council of European Union 2016, point a) of Article 12(2)]. Currently, the GDPR model represents this period using the startedAtTime and endedAtTime rela-

[1] https://dew-uff.github.io/gdpr-data-provenance-model/

GDPR Text	Limitation	Extension
"Where personal data relating to a data subject are collected from the data subject, the controller shall provide [...] the fact that the controller intends to transfer personal data to a third country or international organization [...]." [Council of European Union 2016, point f) of Art. 13(1)].	The GDPR model enables subjects to know whether the controller intends to transfer their data; however, the model does not provide a direct relationship between the controllers.	(i) A new self-relationship (wasAttributedTo) for Controller; and (ii) use Disclose class to represent this transfer.
"The controller shall [...] provide the data subject with [...] the period for which [her] personal data will be stored, or if that is not possible, the criteria used to determine that period" [Council of European Union 2016, point a) of Art. 13(2)].	Although it is possible to represent when a purpose ends, the authors of the GDPR model design patterns explicitly suggest that the obtention of that purpose-ending information does not occur at the beginning.	When *personal data* are obtained, both startedAtTime and endedAtTime relations should be created with their Time annotation, denoting the *Valid Time* [Ozsoyoglu and Snodgrass 1995], those *personal data* will be stored.
"Where the controller intends to further process [...] for a purpose other than that for which the personal data were collected, the controller shall [...]" reveal to data subject its intentions. [Council of European Union 2016, Art. 13(3)].	The current model does not inform whether a Justify is compatible with another one.	(i) A new self-relationship (wasCompatibleWith) for Justify; (ii) A new PseudonymizedData class, as a PersonalData subclass, which instances representing pseudonymized information from all person; and (iii) all those further processes informed by a *Justify* that wasCompatibleWith another must only use PseudonymizedData [Council of European Union 2016, point e) of Art. 6(4)].

Table 2. Plumbing extensions change the GDPR data provenance model high-level classes or relations by adding new elements or changing their use.

tions, which connects Justify entities to Time annotations, thus preserving the timing information of activity compliance. However, the text of this provenance model design suggests that those relations should be created either when its Justify entity is logically present in the database (*i.e.*, creating startedAtTime relation) or when a specific Request is made (*i.e.*, creating endedAtTime relation). This method to capture timing information is known in the literature by *Transaction Time* [Ozsoyoglu and Snodgrass 1995]. An example that enforces this understanding is the first design pattern introduced by Ujcich et al., which describes the use of this provenance model when a user "registers with and provides [her] personal data [...], along with [her] consent" [Ujcich et al. 2018, p. 5]. We notice that the GDPR model creates only the startedAtTime relation at this moment.

In order to represent the period that *personal data* will be stored, we understand

Limitation	Extension
The further processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes are often cited in the text of the law [Council of European Union 2016, Art. 5, 9, 14, 17, 21, 89].	A new ArchivingResearchingStatisticalPurposesJustify class, as sublclass of Justify, which represents those purposes.
Article 9 lists a set of purposes categories of personal data, which demands special justification to process. [Council of European Union 2016, Art. 9(1)]	We suggest the following new classes, all of them subclasses of PersonalData: RacialEthnicData, PoliticalOpinionsData, BeliefData, MembershipData, GeneticData, BiometricData, HealthData, SexLifeData, and SexualOrientationData.
Points b), d), e), and i) of Article 9 explain situations in which the special categories of personal data must not be prohibited [Council of European Union 2016, Art. 9(2)].	We suggest the following new classes: EmploymentSecurityObligation, as a subclass of LegalObligation (point b)); CommunityInterest (point b)) and PublicHealthInterest (point i)), both subclasses of PublicInterest; and PublicData, as a subclass of PersonalData, and Publish, as a subclass of Process (point e)).
Article 16 lays down rectification processes; however, the GDPR model does not propose a class representing that.	A new Rectification, subclass of Process, that represents a rectification process.
Article 21 defines the right of objection; however, the GDPR model does not propose a class to represent this right.	A new class ObjectionRequest, as a subclass of Request, that represents a subject's objection.

Table 3. Porcelain extensions are focused on improving the GDPR data provenance model understanding by providing new subclasses that better express some GDPR articles' points.

that it is enough to create both startedAtTime and endedAtTime relations, with their Time annotation denoting the *Valid Time* [Ozsoyoglu and Snodgrass 1995]. In a word, at the time the *personal data* are collected, along with its consents, *controllers* must inform to the GDPR data provenance model the real-time period this data could be stored. In terms of this provenance model, it means creating both startedAtTime and endedAtTime relations, at the time of data collection.

4.2. Reveal to data subject further processes

Another limitation we have found in this model is the representation of the data's use for a purpose other than the one for which they were collected. Article 5 establishes that "personal data shall be [...] collected for specified [...] purposes and not further processed in a manner that is incompatible with those purposes" [Council of European Union 2016, point b) of Art. 5(1)]. However, this same article sets out that some special types of further processing are not considered incompatible with those initial purposes, since these processes comply with Article 89. In other words, ensuring the principle of data minimization, special types of treatments can be processed regardless of the controller holding a specific justification. This brings us two challenges: (i) it must be clear how our model should represent these special type of further processing, and (ii) how to ensure that the *subject* is aware of these processes even though these processing use pseudonymized data.

We understand that it is possible to represent these treatments through the was-

GDPR Text	Limitation	Extension
"The data subject shall have the right to obtain from the controller [...] any available information as to their source." [Council of European Union 2016, point g) of Art. 15(1)].	The GDPR model does not present any information about the data source.	A new SourceExplanation entity, in which a PersonalData is associated with it, representing a document that describes the source of this data.
"The data subject shall have the right to obtain from the controller [...] the existence of automated decision-making [...] [and] meaningful information about the logic involved [...]." [Council of European Union 2016, point h) of Art. 15(1)].	The GDPR model does not present any that different information about the existence of automated decision-making.	A new AlgorithmExplanation entity, in which an automated Process is associated with it, representing a document that describes the logic involved in that process.
"Where personal data are transferred [...] the data subject shall have the right to be informed of the appropriate safeguards [...]." [Council of European Union 2016, Art. 15(2)] such as describe Art. 46, 47, and 49, also "the means by which to obtain a copy of them or where they have been made available" [Council of European Union 2016, point f) of Art. 13(1), 14(1)].	The GDPR model does not represent any information about the appropriate safeguards of the controller, by concerning their availability to lawfully transfer of personal data.	A new EvaluationOfAdequacy entity, in which a Controller is associated with it, representing a document that describes that required information.

Table 4. Wallpaper extensions present new additions to enrich the model with the information to be compliant with GDPR, although there are other means to present that information to the data owners.

InformedBy relation with an already given Justify activity. However, we see that this introduces ambiguity to the model, due to the representation of two different consents from a single Justify activity. Thus, we propose the wasCompatibleWith self-relation in the Justify activity. Article 6(4) prescribes the rules for valid compatibility between two Justify activities.

In addition, the *subject* should be aware "prior to that further processing" [Council of European Union 2016, Art. 13(3)]. The problem is that Article 89 imposes pseudonymization, which means that is not possible to derive the owner of the pseudonymized data anymore. In order to represent that limitation, we propose the new PseudonymizedData entity, which can not be *attributed to* a Subject. Each instance of this entity represents a pseudonymized property of all *subjects* that have that pseudonymized property. For example, suppose the *subjects* Alice and John, having the same blood type, have registered these data along with their consent in a clinical system. Later, after the *controller* pseudonymizes the blood-type data of all their clients, the provenance graph uses a single instance of PseudonymizedData to represent all pseudonymized blood-type data, including Alice and John's blood type. Following this solution, Alice (and John) are aware that her data could be in use, even though it is not possible to attribute whether the pseudonymized blood-type data belongs to her or to John. Finally, all those further processes should, now, use these PseudonymizedData entities.

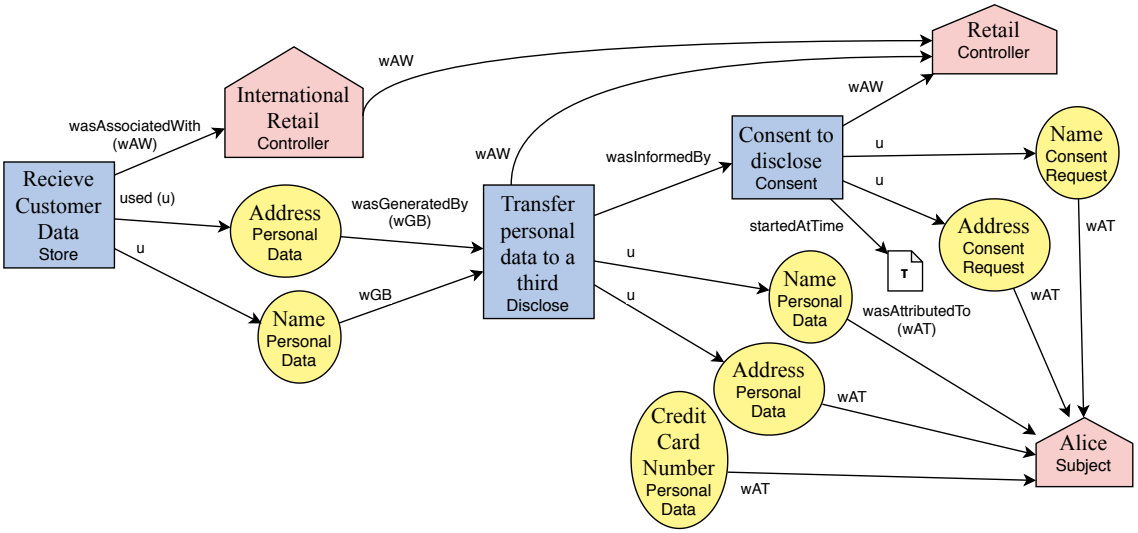

Figure 2. Holding Alice's consent, the retail shop intends to transfer Alice's data to an international retail shop controller. Note that Alice does not consent the controller to share her credit card number with a third party.

5. Design Patterns

Our goal in this section is to present design patterns to clarify *how* the extensions we proposed should be used in practice. Following the GDPR model authors' steps, we use the same running example based on the examples from prior work [Ujcich et al. 2018, Pandit and Lewis 2017, Basin et al. 2018] that involve collecting personal data for an on-line retail shop. Similarly, we assume a customer, Alice, interacts with the retailer by registering, making purchases, and subscribing to marketing information. For this work, we focus only on cases that require our extensions.

5.1. Transfer Data to an International Controller

After time τ, the retail shop intends to transfer Alice's data to an international retail *controller*. Note the association between both *controllers*, which makes it easier to identify the existence of the relation between them through wasAssociatedWith. Figure 2 shows the provenance graph generated from those activities.

Since we inherit the provenance model and did not discard any original component, we still represent a consent for personal data with purposes as a design pattern in the provenance graph by connecting Consent activities to ConsentRequest entities with the used relation. As shown in Figure 2, Alice does not consent to disclose her credit card information to a third party. Note that this use case could not be represented using only the original provenance model [Ujcich et al. 2018].

5.2. Further Process for Other Purposes

At the time τ, Alice registers with and provides her data to the retail shop, along with her consent. At the same time, Alice informs her wish that her data must be erased at the time $\tau + \delta$.

At the time $\tau + 1$, the retail shop intends to calculate the number of customers who share the same city using the address of all customers for statistical purposes.

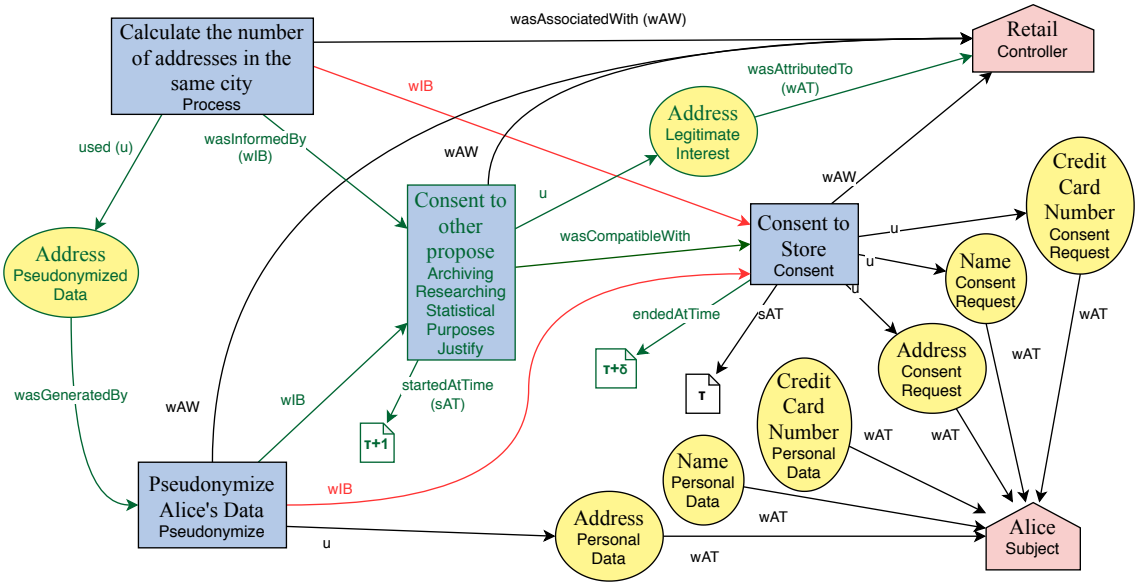

Figure 3. At the time $\tau + 1$, the retail shop intends to calculate the number of addresses in the same city by using personal data. This calculation uses pseudonymized data, and its purpose was not considered incompatible with Alice's consent for storing. Note that, before time $\tau + \delta$, Alice can be informed of the period in which her data will be stored. Note that we represent two graphs in this figure. The first one (black and red arrows) uses the original provenance model, The second one (black and green arrows) uses our extended model.

Thus, the retail shop uses customers' pseudonymized-addresses. We represent those pseudonymized addresses in the provenance graph through the PseudonymizedData entity, as shown in Figure 3. There must be only a single pseudonymized-address entity (regardless of the number of registered customers), representing all customers' pseudonymized-addresses generated by the Pseudonymize processes. This design pattern introduces ambiguity into relationships between data and data owners while enabling the subject to be aware of further processing. The latest design pattern presented in this example is the compatibility relationship between two different purposes. Figure 3 shows, in green and black, that the purpose of calculating the number of customers that share the same city wasCompatibleWith Alice's consent to store her data. In red and black, this Figure represents the use case without the proposed extensions. Note, however, that the provenance graph is more complete using our proposed extensions.

6. Conclusion and Future Work

In this paper, we introduce the problem of enforcing compliance with the GDPR from a data provenance perspective. We start by analyzing the GDPR data provenance model by Ujcich et al. [2018] ; after, we report our analysis comparing this model with some GDPR articles. In that effort, we found eleven points in the law text that are still not suitably represented by this model. We classify these limitations we have found into three types, each one expressing a degree of the impact they affect the model. For each of these limitations, we proposed an extension to the original model that addresses it. Finally, we present two design patterns that represent some of those extensions in the

provenance graph.

Although we have made a high-level effort to evolve this model by analyzing the GDPR text, in future work we also consider practical approaches. The path to helping organizations achieve true compliance should consider the high-level criticism aimed at improving a provenance model conjugated with practical measures that involve multidisciplinary and practical issues. Ujcich et al. [2018] anticipate some of these practical issues. They point out to the necessity to deal with metadata in provenance that also requires GDPR-compliance. They also note other problems, such as the need for inter-controllers audits; the mismatched provenance-granularity levels of; and the ensuring fraud-resistance in provenance collection mechanisms.

Finally, although the GDPR will always require some human activity [Basin et al. 2018], we believe that collaborative efforts can result in a mature solution that will minimize bureaucratic tasks and maximize the transparency of processes.

Acknowledgements. The authors would like to thank CAPES and CNPq for partially supporting this work, and João Felipe Pimentel and Maria Luiza Falci for their reviews and helpful comments.

References

Aldeco Perez, R. and Moreau, L. (2008). Provenance-based auditing of private data use. In *BCS International Academic Conference*.

Bartolini, C., Muthuri, R., and Santos, C. (2015). Using ontologies to model data protection requirements in workflows. In *JSAI International Symposium on Artificial Intelligence*, pages 233–248. Springer.

Basin, D., Debois, S., and Hildebrandt, T. (2018). On purpose and by necessity: compliance under the gdpr. In *International Conference on Financial Cryptography and Data Security*, pages 20–37. Springer.

Bates, A., Tian, D. J., Butler, K. R., and Moyer, T. (2015). Trustworthy whole-system provenance for the linux kernel. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 319–334.

Bier, C. (2013). How usage control and provenance tracking get together-a data protection perspective. In *2013 IEEE Security and Privacy Workshops*, pages 13–17. IEEE.

Bonatti, P., Kirrane, S., Polleres, A., and Wenning, R. (2017). Transparent personal data processing: The road ahead. In *International Conference on Computer Safety, Reliability, and Security*, pages 337–349. Springer.

Council of European Union (2016). Council regulation (EU) no 2016/679. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21.

Garijo, D. and Gil, Y. (2013). P-Plan: The P-Plan ontology. W3C recommendation, W3C. https://www.opmw.org/model/p-plan17092013/.

GDPR.EU (2019). 2019 GDPR Small Business Survey: Insights from European small business leaders one year into the General Data Protection Regula-

tion. https://gdpr.eu/wp-content/uploads/2019/05/2019-GDPR.EU-Small-Business-Survey.pdf.

Gjermundrød, H., Dionysiou, I., and Costa, K. (2016). privacytracker: a privacy-by-design gdpr-compliant framework with verifiable data traceability controls. In *International Conference on Web Engineering*, pages 3–15. Springer.

Kuner, C. (2012). The european commission's proposed data protection regulation: A copernican revolution in european data protection law. *Bloomberg BNA Privacy and Security Law Report (2012) February*, 6(2012):1–15.

Martin, A. P., Lyle, J., and Namiluko, C. (2012). Provenance as a security control. In *TaPP*.

Moreau, L. and Missier, P. (2013). PROV-dm: The PROV data model. W3C recommendation, W3C. http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Ozsoyoglu, G. and Snodgrass, R. T. (1995). Temporal and real-time databases: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):513–532.

Pandit, H. J. and Lewis, D. (2017). Modelling provenance for gdpr compliance using linked open data vocabularies. In *PrivOn@ ISWC*.

Pandit, H. J., O'Sullivan, D., and Lewis, D. (2019). Test-driven approach towards gdpr compliance. In Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 19–33, Cham. Springer International Publishing.

Pasquier, T. F.-M., Singh, J., Eyers, D., and Bacon, J. (2015). Camflow: Managed data-sharing for cloud services. *IEEE Transactions on Cloud Computing*, 5(3):472–484.

Pohly, D. J., McLaughlin, S., McDaniel, P., and Butler, K. (2012). Hi-fi: collecting high-fidelity whole-system provenance. In *Proceedings of the 28th Annual Computer Security Applications Conference on*, pages 259–268.

Shastri, S., Banakar, V., Wasserman, M., Kumar, A., and Chidambaram, V. (2019). Understanding and benchmarking the impact of gdpr on database systems. *arXiv preprint arXiv:1910.00728*.

Tankard, C. (2016). What the gdpr means for businesses. *Network Security*, 2016(6):5–8.

Ujcich, B. E., Bates, A., and Sanders, W. H. (2018). A provenance model for the european union general data protection regulation. In *International Provenance and Annotation Workshop*, pages 45–57. Springer.

Wang, L., Near, J. P., Somani, N., Gao, P., Low, A., Dao, D., and Song, D. (2019). Data capsule: A new paradigm for automatic compliance with data privacy regulations. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 3–23. Springer.