Department of Computer Science

Personal Data Management for Privacy Engineering: An Abstract Personal Data Lifecycle Model

Majed Alshammari and Andrew Simpson

CS-RR-17-02

Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford, OX1 3QD

Personal Data Management for Privacy Engineering: An Abstract Personal Data Lifecycle Model

Majed Alshammari and Andrew Simpson
Department of Computer Science, University of Oxford
Wolfson Building, Parks Road,
Oxford OX1 3QD, UK
Email: firstname.secondname@cs.ox.ac.uk

***Abstract*—It is well understood that processing personal data without effective data management models may lead to privacy violations. Such concerns have motivated the development of privacy-preserving systems and legal frameworks such as the EU General Data Protection Regulation. However, there is a disconnect between policy-makers and engineers with respect to the meaning of privacy. In addition, it is challenging to establish that a system complies with its privacy requirements, to provide technical assurances, and to meet data subjects' expectations. In the spirit of engineering privacy, we propose an abstract personal data lifecycle (APDL) model to support the management of personal data. The APDL model represents data processing activities in a way that is amenable to analysis using an appropriate privacy risk management model. As such, it helps facilitate the identification of potentially harmful data processing activities; it also has the potential to demonstrate compliance with legal frameworks and standards.**

I. Introduction

Privacy is typically articulated at a high level of abstraction. Thus, its concrete manifestations are ambiguous both to those concerned with data protection and to those responsible for developing and maintaining systems [1], [2]. Further, incorporating privacy requirements into the early stages of the development process requires an appropriate interpretation of legal, social and political concerns [3]. These challenges lead to a disconnect between policy-makers and software engineers with regards to understanding the meaning of privacy, its related concepts, and the ways in which systems can be developed to comply with legal frameworks and standards, as well as to meet data subjects' expectations [4]. Therefore, there is a need for generalised techniques that support the effective translation of abstract privacy principles, models and mechanisms into implementable requirements [1], [4].

The dominant approach to embedding privacy into the early stages of the design process is Privacy by Design (PbD) [5], which is built upon a set of principles that aim to identify and mitigate potential privacy risks and meet regulatory compliance requirements [4]. However, the principles of PbD are given at a high level of abstraction, which leads to challenges with regards to translation into engineering activities [3]. Data minimisation has been proposed as a necessary and foundational step to engineer systems in line with the principles of PbD [3] — but applying the principle of data minimisation is a challenge in itself.

To achieve the aim of the PbD, detailed privacy impact and risk assessments need to be conducted with the aim of identifying and addressing potential privacy risks [6]. A Privacy Impact Assessment (PIA) is a process that identifies and mitigates the impact of an initiative on privacy with stakeholders' participation [7]. Specifically, a PIA provides non-technical guidelines for stakeholders on identifying high level privacy requirements; however, it does not provide guidelines on translating these into technical system requirements [2]. In order for a PIA to be holistic and effective, it needs to be complemented by an appropriate privacy risk management model. It also needs to be complemented by a sufficiently robust model that supports the identification of potential privacy risks in a proactive, comprehensive and concrete manner. The representation of such a model tends to be relatively straightforward, capturing possible states and possible changes in these states brought about by operations [8].

The first step towards bridging the gap between policy-makers and software engineers involves providing a common language for privacy engineering that considers protection, manageability and traceability of personal data. Such a common language expresses privacy concerns and expectations of multiple stakeholders. Often, legal frameworks and standards are given at a high level of abstraction without relying on rigorous models that explicitly specify privacy-related concepts [9]: types and sensitivity of personal data; the purposes for, and the manner in which, this data is processed; involved actors; and assigned roles and responsibilities. Thus, an abstract data model plays a crucial role in providing a privacy-aware data lifecycle model in the context of data protection. Furthermore, such a model serves as a stepping stone for the translation of high level privacy requirements into system requirements by defining a foundation for contextual analysis. This includes identifying key concepts of privacy, and associated properties and relationships.

The abstract data lifecycle model (ADLM) [10] was developed to serve as a generic data lifecycle model for data-centric domains. As such, we have chosen the ADLM as a starting point for our contribution. Crucially, it is a generic model that can be used as a means to classify, compare and relate other data lifecycle models, as well as to provide the basis to devise new data lifecycle models [10].

We present an Abstract Personal Data Lifecycle (APDL)

model that represents the main stages, associated activities and involved actors of the personal data lifecycle. It helps facilitate the management and traceability of the flow of personal data, as well as the identification of data processing activities that may lead to privacy violations or harms in a concrete and comprehensive manner. Furthermore, it has the potential to help demonstrate privacy compliance with legal frameworks and standards. Finally, it has the potential to underpin a conceptual framework for privacy engineering with the aim of helping stakeholders reason about design decisions and ground discussion in a common terminology.

II. Foundations

In the context of data-centric domains, data undergoes a variety of actions — including creation, use, publication and destruction — by several actors for various purposes. These actions in combination constitute a data lifecycle. It is understandable that each domain is concerned with a specific type of data and each data lifecycle model has its own specific focus in relation to a domain of interest. Most importantly, they all consider the same item of interest — data, which is a "living thing" that moves though various stages during its lifecycle; it is at the heart of these systems [10]. In the context of data protection, personal data often moves through various stages that are governed by laws, regulations or standard principles, such as collection, retention, usage, disclosure and destruction. Accordingly, personal data should be at the heart of methods, techniques and tools that systematically and proactively identify and address privacy concerns at the early stages of the design process. This, in turn, demands a shift to data-centric software engineering practice in such contexts.

The Abstract Data Lifecycle Model (ADLM) [10] was derived from specific instances of data lifecycle models using a bottom-up approach to ensure broad coverage and wide applicability [10]. For each domain, a list of models was analysed in terms of their lifecycle phases, features, roles, actor features and metadata features. By analysing, comparing and contrasting these models, the ADLM was derived as an abstract data lifecycle model for data-centric domains. It establishes five areas of classification: lifecycle phases, features, roles, actor features, and metadata features. The ADLM provides a means to classify, compare and relate other data lifecycle models, and provides the basis to develop new lifecycle models for other data-centric systems and domains [10].

We illustrate only relevant parts of the ADLM: lifecycle phases and roles. The former generalise all possible stages, steps or processes in the analysed lifecycle models. The ADLM consists of the following lifecycle phases: ontology development, planning, creation, archiving, refinement, publication, access, external use, feedback and termination [10]. The latter, along with the identified lifecycle phases and features, help describe and classify lifecycle models. The ADLM considers the following roles: ontology designers, data creators, metadata creators, administrators and end users.

Engineers, with the help of domain experts, can translate high-level privacy requirements into technical system requirements by defining a foundation for contextual analysis [2]. This necessitates identifying key concepts for describing relevant privacy aspects, associated properties and relationships. Such an analysis needs to be based on well-defined modelling languages and vocabularies [2]. Conceptual modelling has been previously used for several purposes, including ontology modelling and data modelling [11], which both help in defining an abstract personal data lifecycle model.

III. Case Study

We now introduce the ePetition system, the aim of which is to implement the European Citizens' Initiative (ECI) [12], which we shall use as an illustrative case study. The ECI is used to support a formal request — provided by organisers — to an authority for submitting a proposal for a legal act. In particular, it enables EU citizens from a number of EU Member States to invite the European Commission to propose a legal act on issues where it has competence to legislate. The main purpose of the ePetition system is to verify and certify the number of valid signatures that support a certain initiative. In order for signatories to support a specific initiative, they need to provide 'identifying' personal data — such as full names, dates of birth and unique identifier numbers — which is typically retained in databases. In compliance with applicable regulations, data controllers are required to apply appropriate security measures to protect the collected personal data, and ensure that it is only used for the specified purposes and retained only as long as necessary to achieve these purposes.

The main steps of preparing and launching an initiative are as follows. The first step involves setting up a citizens' committee of at least seven EU citizens. All of the committee's members need to be permanent residents or citizens of the EU Member States and old enough to vote in elections to the European Parliament. This committee acts in its capacity as the official organiser of the initiative and is responsible for preparing and managing the initiative. Secondly, the organisers need to prepare an initiative and register it with the European Commission. In order to register an initiative, the organisers need to specify the title, the subject matter, its objectives, the committee members' personal data, and provide an email address and telephone number for the representative and their substitute. The organisers also need to find a hosting provider when signatures are intended to be collected electronically by an online collection system — either using an instance of the open source software that is provided by the European Commission and hosting it at its site or by developing their own collection system and using a hosting service provider. For both, organisers need to obtain a certificate from the competent national authority to verify its compliance with minimum technical requirements [13]. Then, the certificate should be posted in the online collection system. Next, individuals, who act as signatories, are able to submit their personal data and their statements of support. To give their support for the initiative, signatories need to provide the specified personal data, such as full names, permanent residence, date of birth and nationality. However, in some Member States,

such as France and Spain, personal identification numbers are required. It is important to ensure that duplicate signatures by the same individual are avoided [14]. Having reached the required number of signatures, organisers should send this personal data to relevant competent national authorities to verify this data and certify the number of valid statements of support. Having received all certificates from competent national authorities, organisers should submit the initiative by sending these certificates to the European Commission.

In accordance with the EU Data Protection Directive [15] and the Regulation (EU) No. 211/2011 on the Citizens' Initiative [14], organisers and competent national authorities act as data controllers. In particular, organisers are required to notify the Data Protection Authority in the EU Member State where the personal data will be processed. They are also required to apply appropriate measures to protect personal data in compliance with the Directive and relevant regulations. This includes that personal data must be "adequate, relevant and not excessive" in relation to the purpose of supporting the initiative and verifying the statements of support. Accordingly, the organisers and the competent national authorities must ensure that collected personal data is not used for purposes other than those specified for supporting the initiative and verifying the statement of support respectively. In addition, the data controllers must destroy all statements of support and any copies one month after submitting the initiative to the Commission or issuing the certificate respectively as per [14].

IV. The APDL model

The APDL is an abstract model that represents personal data in terms of states and operations. It identifies a set of stages through which personal data moves during its lifetime and indicates the order and depth in which these activities can occur. The APDL has the potential to be used as a means to classify, compare and relate other personal data lifecycle models, as well as to be used as the basis for defining new personal data lifecycle models for various domains. We will use and adapt features of the ADLM as points of reference for analysis.

A. Lifecycle stages

As there are obligations and limitations on the stages of the personal data lifecycle and associated activities, our analysis of the ADLM has to consider such concerns. Some stages will be combined — generalised — according to their characteristics and associated activities, while others will be defined — specialised — to limit associated activities to particular privacy principles. Those stages not relevant to personal data will be discarded.

It is essential to adopt a set of universal privacy principles that can be applied in a variety of contexts in various jurisdictions. As an example, the Fair Information Practice Principles (FIPPs) were developed as core principles of the Code of Fair Information Practice [16]. Subsequently, other guidelines and principles have been developed by a variety of organisations to codify the FIPPs with the aim of protecting the privacy of individuals and ensuring that personal data flow across borders is appropriate, such as [17] and [18]. In 2006, at the 28th International Data Protection and Privacy Commissioners Conference, the Global Privacy Standard (GPS) [19] was accepted as a unified set of principles that reflects appropriate variants of the FIPPs. The GPS principles harmonise various sets of the FIPPs into universal privacy principles. We adopt the GPS principles to place limitations on the stages of the lifecycle and associated activities.

Figure 1 illustrates the main stages of the abstract personal data lifecycle model along with relevant lifecycle roles. We describe each stage in terms of associated activities and their dependencies on other stages and relevant GPS principles. Each stage has a set of metadata as outputs to describe associated data about the manner in which these activities may be conducted.

0) **Conceptual modelling:** This is a preliminary stage: it is a prerequisite to any personal data lifecycle.
Activities: This stage involves a set of activities to construct conceptual models that represent key and relevant concepts, associated properties, relationships and constraints that restrict the semantics of the concepts and conceptual relationships according to the purposes for which they are created and intended users. There are two distinct activities: conceptual modelling and data modelling. The former is to represent privacy-related concepts in relation to the context of data protection, whereas the latter is to represent the minimum amount of required data in relation to the context in which it is processed.
The conceptual model, which can be represented through UML diagrams [20], provides vocabulary terms that can be used to facilitate communication with non-expert stakeholders.
Dependency: In Figure 1 conceptual modelling is represented by a dotted line to emphasise that it is not a core part of the personal data lifecycle. The outputs of this stage are two models to be used by multiple stakeholders. The first is a conceptual model that represents privacy-related concepts, associated properties and relationships as shared knowledge for a specific domain. The second is a data model that represents context-related objects, associated properties and semantic relationships as shared knowledge for a specific application.
Principles: In order to carry out the essential activities of the conceptual modelling, appropriate techniques need to be used to refine privacy-related concepts and their meanings, including formal and informal text analysis techniques. The most appropriate technique needs to be adopted to ensure that the conceptual model is based upon a widely-used set of terms. To achieve this goal, universal or standard privacy principles that represent the commonly-used concepts need to be considered as a source of knowledge. As such, the GPS principles serve as the foundation for conceptual modelling and associated activities.

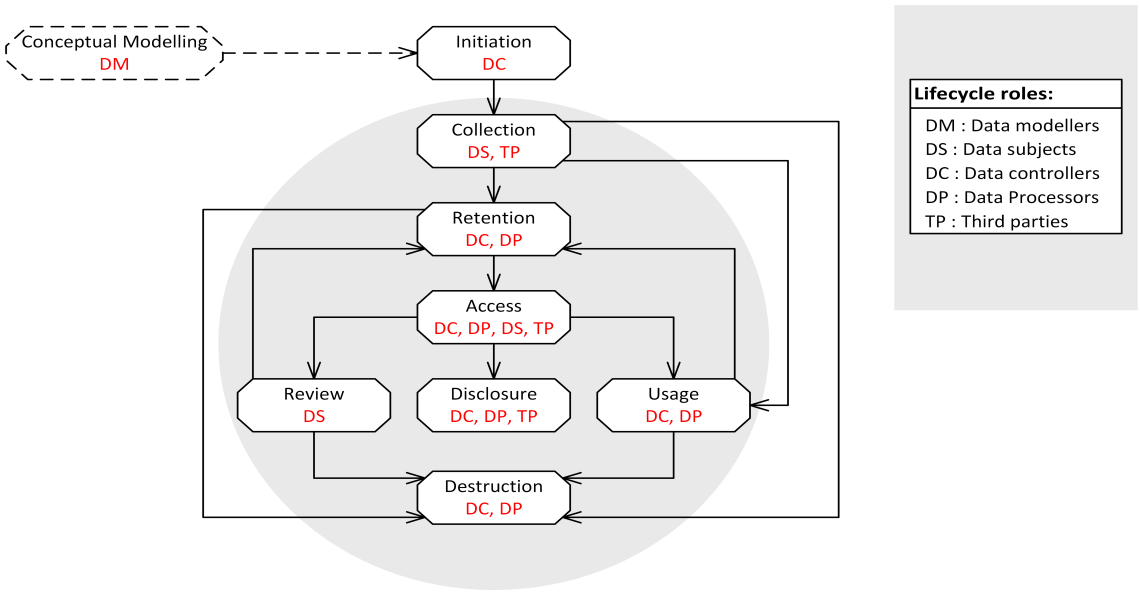

Fig. 1. The Abstract Personal Data Lifecycle (APDL) Model.

1) **Initiation:** This stage precedes the collection of personal data and represents the first stage.
   Activities: This stage involves a set of activities to specify a 'processing plan', in relation to the context. The processing plan includes: the elements of personal data to be collected; the sources of this data; the methods of collection; the purposes — along with their lawfulness, fairness and proportionality — for, and the manner in which, personal data is collected, retained, used and disclosed; the choices available to data subjects and the consent to be obtained; involved actors, associated roles and responsibilities; relevant privacy regulations or standard principles; and domain-specific constraints that govern personal data processing. The elements of personal data need to be adequate, relevant and not excessive in relation to the purposes for which personal data is collected. This implies that the minimum amount of personal data needs to be clearly specified in this stage. In addition, it is important to explicitly explain whether the purposes for which personal data is collected involve the intention to collect, derive or acquire additional personal data items, whether from existing items or external sources.
   Dependency: The input is from the conceptual modelling stage, as well as relevant legal and standard frameworks. The output is a processing plan that determines the purpose for, and the manner in which, personal data is processed.
   Principles: The GPS principles serve as the foundation for initiation and associated activities.
2) **Collection:** It follows the initiation stage and precedes the retention stage.
   Activities: This stage involves a set of activities that represents the creation of personal data values, whether these are directly collected from data subjects, or have been acquired from external sources. In both cases, it is important that these values have not existed in the lifecycle before the collection stage.
   The values of the specified items of personal data are collected through various methods of collection. To ensure that these values are collected by fair and lawful means, the methods of collection need to be limited to those identified and reviewed in the processing plan and for which explicit or implicit consent is obtained. Thus the most aspect in this stage is the personal data values and associated sources, rather than the methods of collection themselves. The methods of collecting personal data are considered as significant in the context of data protection; therefore, data that describes these methods can be created as metadata.
   Dependency: The input is from the initiation stage. The output is a set of personal data values that correspond to the specified personal data elements.
   Principles: Collection Limitation.
3) **Retention:** It follows the collection stage and precedes the access stage.
   Activities: This stage involves a set of activities that represents the act of continued storage of personal data in repositories or digital storage media. Retention involves three distinct activities: primary storage, archiving and backup.
   Dependency: The input is from the collection stage. The output is a set of personal data values that correspond to the specified personal data elements.
   Principles: Use, Retention, and Disclosure Limitation.
4) **Access:** It follows the retention stage. In particular, it follows the activities of primary storage and occasionally follows the activities of archiving for regulatory compli-

ance purposes.

Activities: This stage involves a set of activities that represents the act of specifying and retrieving personal data stored in repositories or digital storage media: personal data is made accessible for use by involved actors, whether they are internal users, external users or data subjects. Involved actors gain access to the stored personal data and are able to retrieve this data to perform specific actions according to their roles and responsibilities as specified in the processing plan. Data retrieval is not restricted to specific mechanisms, such as using query languages; rather, it can be accomplished by using interfaces or any mechanisms that allow the stored data to be searched, retrieved and appropriately displayed. Retrieval mechanisms are considered as significant in the context of data protection; therefore, data that describes these mechanisms can be created as metadata.

Dependency: The input is from the retention stage. The output is a set of personal data values that correspond to the specified personal data elements.

Principles: Access is considered as the preceding stage of the usage, disclosure and review stages of the stored personal data. Thus, the following principles need to be applied in this stage to govern associated activities: Use, Retention, Disclosure Limitation, and Access.

5) **Review:** It follows the access stage.
   Activities: This stage involves a set of activities that represents the act of providing data subjects with control over their personal data. There are two distinct activities to provide the control: refinement and storage. Refinement refers to the activities of the use of the previously accessed and retrieved personal data by data subjects to, for example, make sure that their personal data is accurate. The most important point in this stage is providing data subjects with access to exercise control over personal data, rather than merely the means by which data subjects can review, update and correct this data. The methods of collecting personal data are considered as significant in the context of data protection; therefore, data that describes these methods can be created as metadata.
   Dependency: The input is from the retention and access stages. The output is a set of reviewed, updated or corrected personal data values that correspond to the specified personal data elements.
   Principles: Access and Accuracy.
6) **Disclosure:** It follows the access stage.
   Activities: This stage involves a set of activities that represents the act of preparing the previously accessed and retrieved personal data for external use. It involves the act of disseminating the prepared data to be used by external actors, such as third parties, to perform further actions, including manipulating and combining several data items from various sources. These actions may include further processing for historical, statistical or scientific purposes. Most importantly, the disseminated personal data items need to be used only for the specified purposes in the processing plan and with the consent and knowledge of data subjects. The used means of disseminating personal data are considered as significant in the context of privacy and data processing; therefore, data that describes these means can be created as metadata.
   Dependency: The input is from the retention and access stages. The output is a set of personal data values that correspond to the specified personal data elements.
   Principles: Use, Retention, and Disclosure Limitation.
7) **Usage:** It follows the access stage.
   Activities: This stage involves a set of activities that represents the use and manipulation of personal data There are two distinct activities: refinement and storage. Refinement includes deriving new personal data items by mining or combining several data values from internal or external sources. Storage refers to the activities of re-storing the manipulated personal data in the primary storage media for operational purposes.
   Dependency: The input is from the retention and access stages. The output is a set of personal data values that correspond to the specified personal data elements.
   Principles: Use, Retention, and Disclosure Limitation.
8) **Destruction:** It follows the initiation, collection, retention and access stages, and is the final stage.
   Activities: This stage involves a set of activities that represents the act of removing personal data items from repositories or digital storage media in accordance with relevant retention policies. These activities include: completely and permanently erasing personal data or destroying digital storage media; removing or redacting specific items of personal data that can serve as identifiers or quasi-identifiers; and disposing of original, archived and backup copies of personal data in accordance with relevant destruction policies. This indicates that the most important point in this stage is the permanent destruction, disposal, erasure or redaction of personal data, rather than merely the methods of storage. The methods of destroying personal data are considered as significant in the context of data protection; therefore, data that describes these methods can be created as metadata.
   Dependency: The input is from the retention stage. The output is a set of destroyed personal data items that correspond to the specified personal data elements.
   Principles: Use, Retention, and Disclosure Limitation.

Table I summarises the personal data lifecycle stages, associated activities, dependency in terms of inputs and outputs, and relevant GPS principles.

In summary, the APDL model represents the stages through which personal data moves during its lifecycle along with associated activities and involved actors. These activities need to be performed in an ordered manner to indicate how and when to move from one stage to another. As per the ADLM, the granularity levels can be classified into two levels as *coarse*, which means that all personal data items are processed in each cycle, or *fine*, which means that only a number are

TABLE I
THE STAGES, ASSOCIATED ACTIVITIES, DEPENDENCY AND RELEVANT GPS PRINCIPLES OF THE APDL MODEL.

Stage	Activities	Input	Output	GPS Principles
Conceptual Modelling	Specification Conceptualisation Representation	Domain knowledge	Conceptual model Data model	All GPS principles
Initiation	Processing plan specification	Conceptual model Legal frameworks and standards	Specified processing plan	All GPS principles
Collection	Personal data creation	Specified processing plan A set of personal data	A set of personal data and metadata	Collection Limitation
Retention	Primary storage Archiving Backup	Specified processing plan A set of personal data and metadata	A set of personal data and metadata	Use, Retention, and Disclosure Limitation
Access	Personal data specification Personal data retrieval	Specified processing plan A set of personal data	A set of personal data and metadata	Use, Retention, and Disclosure Limitation Access
Usage	Refinement Storage	Specified processing plan A set of personal data	A set of personal data and metadata	Use, Retention, and Disclosure Limitation
Disclosure	Personal data preparation Personal data dissemination	Specified processing plan A set of personal data	A set of personal data and metadata	Use, Retention, and Disclosure Limitation
Review	Refinement Storage	Specified processing plan A set of personal data	A set of personal data and metadata	Access Accuracy
Destruction	Erasure of personal data Destruction of storage media Redaction of data identifiers or quasi-identifiers Disposal of original, archived and backup copies	Specified processing plan A set of personal data	A set of personal data and metadata Conformity certificate	Use, Retention, and Disclosure Limitation

processed in each cycle. The level of granularity helps support the application of data minimisation as a foundational step for privacy engineering. This can be achieved by restricting the processing of personal data to the minimum amount necessary according to the purpose of each processing activity.

B. Lifecycle roles

We now analyse the roles that may be played by actors in the ADLM and specialise these roles for the purposes of the APDL model. Each actor may have one or many roles and each role will be typically associated with one or many stages.

1) *Data modellers.* Data modellers are involved in the conceptual modelling stage and play the role of establishing the context in which personal data is processed. Actors who play this role are responsible for defining a conceptual model of the domain of interest.
2) *Data subjects.* Data subjects are involved in the collection stage of the lifecycle with the capability of providing their personal data. Such actors can actively participate in the creation of the personal data values. Data subjects may be involved in the access and review stages of the lifecycle with the capability of accessing and updating their personal data. Actors with this ability can access, review, update or correct their personal data to ensure that the retained personal data is accurate.
3) *Data controllers.* Data controllers are actors who specify the purposes for, and the manner in which, personal data is to be collected and processed. They are involved in the planning, retention, access, usage, disclosure and destruction stages with administrative capabilities. Such actors are responsible for handling personal data items without changing its format or meaning. If the data controller is a data processor, administrators are responsible for archiving, making backup copies, disclosing and destroying personal data items. The administrative capabilities may also include other activities, such as those related to compliance monitoring and audit trails. Data controllers are also involved in the access and usage stage of the lifecycle with different levels of user capabilities. Such actors use and manipulate the retained personal data items according to the purposes for which this data is collected. They perform data processing activities, including update or modification, consultation or other actions as per the processing plan.
4) *Data processors.* Data processors are actors who process the collected personal data on behalf of the data controller. They are involved in the retention, access and usage stages and process personal data items without changing its format or meaning. Such actors are responsible for archiving, making backup copies and destroying personal data items according to the data controller instructions. The role of data processors may also include other responsibilities, such as those related to operations and performance monitoring.
5) *Third parties.* Third parties are actors other than data subjects, data controllers or data processors. They may be involved in the collection stage with data-providing capabilities, i.e. they may be external sources other than data subjects. Such actors actively participate in the creation of the personal data values in the lifecycle model. In addition, third parties may be involved in the disclosure

stage of the lifecycle with data-receiving capabilities. Such actors receive and use the disclosed personal data items only for the purposes specified in the processing plan and with the consent or knowledge of data subjects.

V. An example

We now consider again the European Citizens' Initiative (ECI) with the aim of instantiating a personal data lifecycle.

1) *Lifecycle stages.*

(a) *Conceptual modelling.* Conceptual modelling for the purpose of representing privacy-related concepts, their properties, relationships and constraints is a prerequisite to any personal data lifecycle. As such, we consider conceptual modelling only for the purpose of representing the minimum amount of required data for the ePetition system that implements the ECI.
The specification of the required data is driven by the specification of purposes for which personal data is to be processed. In this case, the main purpose of collecting and processing signatories' personal data is to verify and certify the valid number of the submitted statements of support. In order for the purpose to be fulfilled, a minimum amount required of data needs to be appropriately specified. Initially, we can draw a partial data model diagram that represents: Signatory, Organiser, Petition, Address, NationalAuthority, DataProtectionAuthority and EuropeanCommission classes, as illustrated in Figure 2. The relationships of the classes can of course be directly modelled by associations in the UML.

(b) *Initiation.* The personal data lifecycle that underlies the ePetition system in the context of participatory democracy is not unique: a data processing plan must precede any collection and processing of personal data. In accordance with the EU Data Protection Directive [15] and the Regulation (EU) No. 211/2011 on the Citizens' Initiative [14], organisers are required to notify the Data Protection Authority in the EU Member State where the personal data will be processed before the collection of statements of support. Such a notification requires a complete processing plan that may serve as the basis of developing a privacy notice. The processing plan needs to outline: the elements of personal data to be collected along with its sources; the purposes for, and the manner in which, this data is processed; the collection methods; the choices available to data subjects and the consent to be obtained; the involved actors and their assigned roles and responsibilities; relevant regulations and standards; and the domain-specific constraints.

(c) *Collection.* Once the registration of an initiative has been confirmed, the relevant Data Protection Authority has been notified and the online collection system has been certified by competent national authorities, the organisers may use an online collection system to collect the specified personal data from at least one million EU citizens who act as signatories. The specified data is collected with a specific time limit — the collection period is no longer than twelve months starting from the date of registration of an initiative [14]. Most importantly, the collected personal data values must not exist in the lifecycle before the collection to prevent duplicate statements of support. In order for organisers to collect adequate, relevant and not excessive personal data, the collection system must generate statements of support in an appropriate form.

(d) *Retention.* During the collection period, the statements of support that have been submitted by signatories are required to be persistently stored in a primary storage media for operation purposes. One might also assume the existence of copies of the original personal data for operational recovery purposes.
Once the collection period is finished and the personal data is sent for verification and certification, competent national authorities have three months to certify the number of valid statements of support. During this period, the retained data is no longer needed for regular use by the organisers and can be archived as historical data for compliance purposes. Having submitted the received certificates, organisers have one month to destroy the retained personal data and any copies thereof or 18 months from the date of the registration of the initiative, whichever is the earlier [14]. Signatories' personal data or any copies thereof may be retained beyond the specified retention time for the purpose of legal or administrative proceedings relating to an initiative. This requires retaining statements of support and any copies thereof for one week after the date of conclusion [14].

(e) *Access.* During the collection period, organisers need to monitor the collection of statements of support that have been submitted. Once the collection of the statements of support have been de-activated at the end of the collection period, organisers need to export signatories' personal data from statements of support and display the current signatures distribution, which are classified according to the Member State of signatories or the date of submission. These activities require specifying and retrieving the retained statements of support. In particular, signatories' personal data needs to be made accessible for use by involved actors, in this case, internal users who acting as organisers.

(f) *Review.* Data subjects are unable to access their personal data once they have submitted their statements of support. In particular, the ePetition system that implements the ECI does not provide signatories with full control over their personal data, i.e. review, update or correct, to make sure that their personal data is accurate, complete and up-to-date.

(g) *Disclosure.* Statements of support are used only for verification and certification; they cannot be disclosed to any other parties.

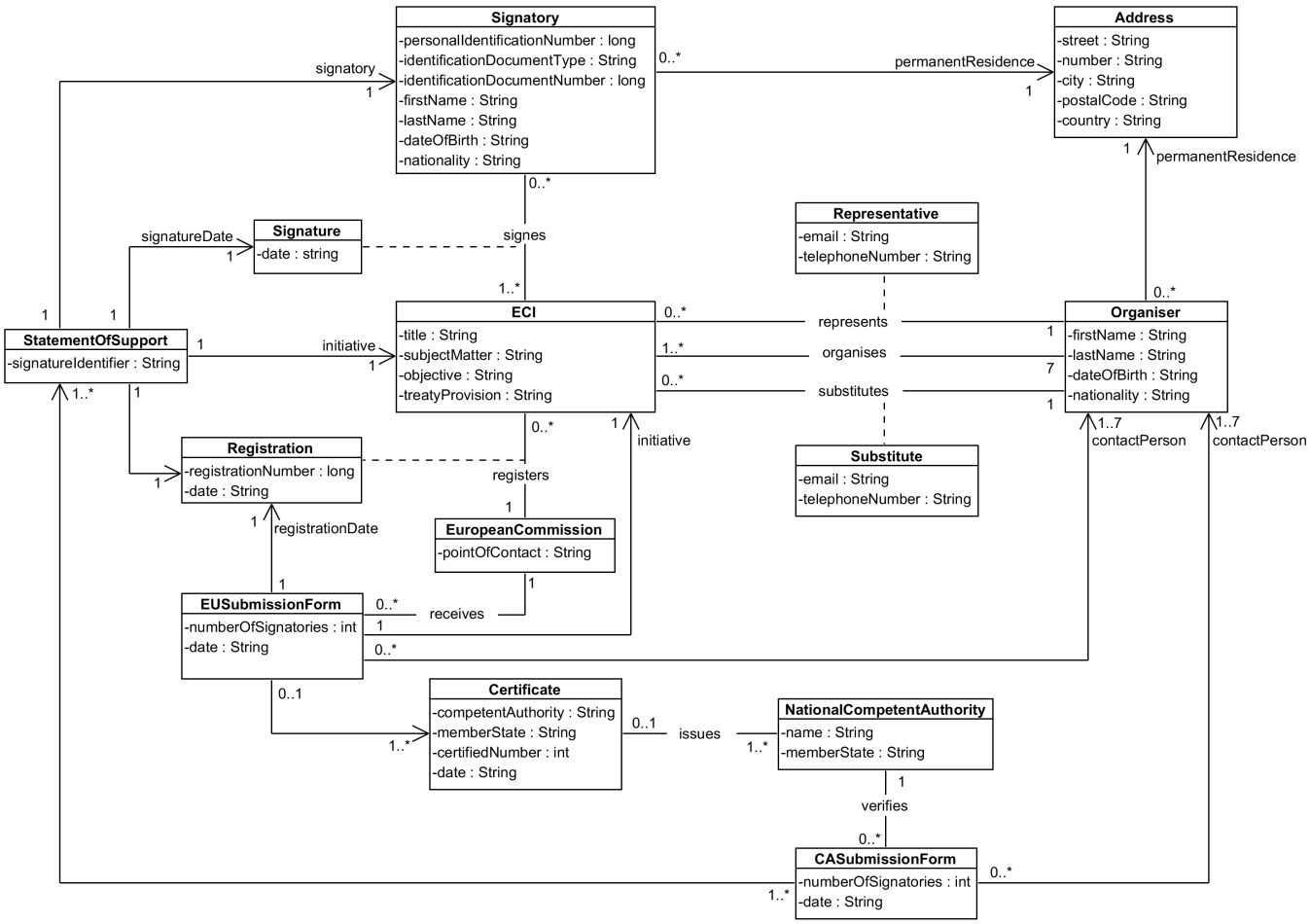

Fig. 2. The data model diagram for the European Citizens Initiative (ECI).

(h) *Usage.* Signatories' personal data is collected and processed for verifying and certifying the number of valid statements of support. In this case, organisers use and manipulate this data to fulfil the specified purpose. These include monitoring, deleting, exporting, preparing and sending statements of support to relevant competent authorities. These activities include refinements; however, they do not include any storage activities that involve re-storing the manipulated personal data in the primary storage. The use of signatories' personal data is accomplished by relevant competent authorities as they conduct the verification process and produce certificates for valid statements of support.

(i) *Destruction.* Removing statements of support is the final stage of the lifecycle. Signatories' personal data that has been collected and stored as statements of support are required by law to be destroyed after a specific time limit, as explained in the retention stage. Statements of support need to be completely and permanently erased, and digital storage media needs to be destroyed. Original, archived or backup copies of the retained statements of support need to be disposed in accordance with relevant retention and destruction policies.

2) *Lifecycle roles.*

The *data modeller* role may be assigned to capable actors who are able to define an appropriate conceptual model for the context of participatory democracy and, in particular, for the ePetition system.

Citizens or permanent residents of the EU Member States act in their capacities as *data subjects* who are able to provide their personal data. They actively participate in the creation of personal data with the aim of supporting an initiative. However, data subjects are not able to access and review their personal data once they have submitted statements of support. Data subjects are mainly involved in the collection stage. Organisers and competent national authorities act in their capacities as *data controllers*. Organisers are responsible for specifying the purpose for, and the manner in which, the required personal data is to be collected and processed. They are responsible for collecting, monitoring, preparing and sending personal data to competent national authorities. The competent national authorities are responsible for verifying and

certifying the number of valid statements of support for an ECI.

There are three possible options for *data controllers*. First, data controllers may act in their capacity as data controllers and processors at the same time if they are capable of operating the online collection system. Second, the European Commission may act in its capacity as a hosting service provider by providing the OCS. The third case is a third party that acts in its capacity as a hosting service provider. In all cases, data processors are responsible for handling personal data without changing its format or meaning. They are responsible for archiving, making backup copies and destroying this data according to the data controllers' instructions.

There is no *third party* involved in the collection or disclosure stages of the personal data lifecycle.

VI. Discussion

We have introduced the APDL as an abstract model to represent the main stages of the personal data lifecycle along with associated activities and involved actors at a high level of abstraction. Each stage is an abstraction of a set of logically related data processing activities. This classification is based on the relevant GPS principles, the nature of processing activities, and the role type of involved actors and their assigned responsibilities.

The APDL model is expressed in terms of cycles that reflect the nature of data processing. The activities associated with the initiation stage, for example, can be carried out in a repetitive manner, i.e. in terms of iterations. This manner allows more flexibility in evolution of the processing plan, e.g. the inclusion, modification and removal of items at any time, such as processing additional data items, acquiring data items from different sources or processing data for secondary purposes. In particular, it allows a new iteration to start at any time with initial planning. Obviously, this manner is adequate for a domain in which the processing plan is undergoing continuous evolution to adapt to rapid changes in system requirements. This gives the APDL model the possibility to be applied to various domains, including dynamic and interconnected scenarios where data is collected from different sources with different formats.

In addition, the lifecycle roles give the APDL model the possibility to classify data processing activities according to the involved actors and their assigned roles and responsibilities. For each stage of the lifecycle, data processing activities can be assigned to lifecycle roles. A lifecycle role combines activities with respect to who is responsible for them. As such, the data lifecycle is a way of describing data processing, with the possibility of expressing how processing activities are performed, when they take place, i.e. lifecycle stages, and where they take place, i.e. lifecycle roles. This supports the applicability of the model when there are more than one domain, as well as when data is collected and processed collaboratively by multiple stakeholders by determining who is responsible for which lifecycle stage and what is their level of authority with respect to the decisions and activities performed.

Additionally, we should note that we limit our model to those terms that are necessary to define the fundamental concepts of the personal data lifecycle. Those might be further refined and extended by developing a conceptual model that represents all relevant concepts, associated properties and relationships. For example, the lifecycle may be characterised by properties that help support its application in various domains, such as the type of the lifecycle, the homogeneity of data and the centrality of the system underlying the lifecycle. Such properties give the model the possibility to represent and document data — processing activities in a meaningful manner.

Most importantly, the APDL model is considered as a first step towards privacy analysis by providing a common language that is understood by multiple stakeholders. It needs to be complemented by an appropriate meta-model that is represented in a widely-used modelling language to increase its usefulness in the context of software engineering. In doing so, engineers will benefit from the model in terms of conducting risk analysis and making design decisions by selecting from existing strategies, patterns and technologies.

VII. Conclusions

The integration of privacy into the early stages of the design process is increasingly important — PIAs and PbD are now mandated by, for example, the EU's GDPR. PIAs can be used by stakeholders to identify high level privacy requirements, which, in turn, need to be translated into technical system requirements. As such, a PIA needs to be complemented by a sufficiently robust model that represents data processing activities in a way that is amenable to risk analysis.

We have presented the APDL model as an abstract model for personal data lifecycles — where the data lifecycle is defined by its stages, associated activities and involved actors. The APDL model distinguishes between the types of operations that can be performed on personal data. For each operation, it outlines various distinct activities in relation to the GPS principles with the aim of governing the behaviour of these operations. The separation is important for several reasons: it helps support the manageability and traceability of the flow of personal data during its lifecycle; it is necessary for ensuring and demonstrating compliance with legal frameworks and standards; it reflects the extent to which the flow of personal data is appropriate in terms of involved actors and their assigned roles and responsibilities; and it facilitates the identification of data processing activities that may lead to privacy violations or harms.

Our next task is to define a conceptual model that describes the problem and its solution in terms of the domain vocabulary as a prerequisite to any data lifecycle. Based on this, we intend to define a profile that allows the APDL model to be represented in the Unified Modeling Language (UML). We also plan to use additional case studies with the aim of further validating the applicability of the APDL model.

REFERENCES

[1] S. S. Shapiro, "Privacy by design: Moving from art to Practice," *Communications of the ACM*, vol. 53, no. 6, pp. 27–29, 2010.

[2] M. Kost, J. C. Freytag, F. Kargl, and A. Kung, "Privacy Verification using Ontologies," in *Proceedings of the Sixth International Conference on Availability, Reliability and Security (AReS 2011)*. IEEE, 2011, pp. 627–632.

[3] S. Gürses, C. Troncoso, and C. Diaz, "Engineering privacy by design," *Computers, Privacy & Data Protection*, vol. 14, 2011.

[4] S. Spiekermann, "The Challenges of Privacy by Design," *Communications of the ACM*, vol. 55, no. 7, pp. 38–40, 2012.

[5] A. Cavoukian, *Privacy by Design . . . Take the Challenge*. Office of the Information and Privacy Commissioner of Ontario, 2009.

[6] Cavoukian, A., "Privacy by Design: The 7 Foundational Principles Implementation and Mapping of Fair Information Practices," https://www.ipc.on.ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=953, 2010.

[7] D. Wright, "The State of the Art in Privacy Impact Assessment," *Computer Law & Security Review*, vol. 28, no. 1, pp. 54–61, 2012.

[8] A. Cavoukian, S. Shapiro, and R. J. Cronk, "Privacy Engineering: Proactively Embedding Privacy by Design," https://www.privacybydesign.ca/content/uploads/2014/01/pbd-priv-engineering.pdf, 2014.

[9] T. Antignac, R. Scandariato, and G. Schneider, "A Privacy-Aware Conceptual Model for Handling Personal Data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, 2016, pp. 942–957.

[10] K. Möller, "Lifecycle Models of Data-centric Systems and Domains: The Abstract Data Lifecycle Model," *Semantic Web*, vol. 4, no. 1, pp. 67–88, 2013.

[11] T. Dillon, E. Chang, M. Hadzic, and P. Wongthongtham, "Differentiating Conceptual Modelling from Data Modelling, Knowledge Modelling and Ontology Modelling and a Notation for Ontology Modelling," in *Proceedings of the fifth Asia-Pacific conference on Conceptual Modelling (APCCM '08) - Volume 79*. ACM, 2008, pp. 7–17.

[12] European Commission, "The European Citizens' Initiative," http://ec.europa.eu/citizens-initiative/public/welcome, 2012.

[13] European Commission, "Commission Implementing Regulation (EU) No 1179/2011 of 17 November 2011 laying down technical specifications for online collection systems pursuant to Regulation (EU) No 211/2011 of the European Parliament and of the Council on the citizens' initiative," http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:301:0003:0009:EN:PDF, 2011.

[14] European Commission, "Regulation (EU) No 211/2011 of the European Parliament and of the Council on the citizens' initiative," http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02011R0211-20131008&from=EN, 2011.

[15] The European Union: Official Journal of the European Communities, "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=EN, 1995.

[16] United States Department of Health, Education and Welfare: Secretary's Advisory Committee on Automated Personal Data Systems, *Records, Computers and the Rights of Citizens: Report*. [Cambridge? Mass.]: [MIT Press], 1973.

[17] Organisation for Economic Co-operation and Development (OECD), "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data," http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm, 2013.

[18] American Institute of Certified Public Accountants and Canadian Institute of Chartered Accountants (AICPA/CICA), "Generally Accepted Privacy Principles," http://www.aicpa.org/InterestAreas/InformationTechnology/Resources/Privacy/GenerallyAcceptedPrivacyPrinciples/DownloadableDocuments/GAPP_PRAC_%200909.pdf, 2009.

[19] A. Cavoukian, "Creation of a Global Privacy Standard," https://www.ipc.on.ca/images/Resources/gps.pdf, 2006.

[20] Object Management Group (OMG), "Unified Modeling Language (UML)," http://www.omg.org/spec/UML/, 2015.