A UML Profile for Privacy-Aware Data Lifecycle Models

Majed Alshammari and Andrew Simpson

Department of Computer Science
University of Oxford

September 15, 2017 - Oslo, Norway

Outline

The plurality and contextuality of privacy

- Privacy is a **multi-faceted** concept that has legal, social, political and technical aspects.
- It is **subjective** in nature and and culturally variable: it is influenced by a variety of factors, including societal demands — which evolve over time — and technological developments.

Privacy by Design

- Privacy by Design (PbD) has emerged as a proactive approach for embedding privacy into **the early stages of the design** of information technology, business practices, physical designs and networked infrastructures.
- It is an **engineering** and **strategic management approach** that aims to
  - meet legal obligations,
  - mitigate potential risks,
  - achieve accountability and
  - enhance user trust.

Privacy by Design

The appropriate implementation of the PbD approach requires:

- engineering innovation,
- clear commitment at managerial level,
- multiple stakeholders' participation, and
- social controls — legislation, code of conducts, and public response.

Motivation

- Typically, legal frameworks and standards are given at a high level of abstraction without relying on rigorous models that explicitly specify privacy-related concepts and associated properties.
- The principles of PbD are given at a high level of abstraction, which, in turn, leads to challenges with regards to translating these principles into engineering activities and artefacts.
- There is a lack of systematic methodologies that address the complexity and variability of privacy by identifying and addressing potential privacy risks, and ensuring and demonstrating privacy compliance.

Motivation

These challenges lead to a disconnect between policy-makers' intentions and software engineering reality with respect to:

- conceptualisations of privacy,
- its related concepts, and
- the ways in which systems can be developed to comply with legal frameworks and standards and to meet data subjects' expectations.

The first step of engineering PbD

- To **realise** PbD and translate its principles into engineering activities, a systematic methodology for Privacy Impact Assessments (PIAs) is required.
- A **PIA** is defined as a process that identifies and mitigates the impact of an initiative on privacy with multiple stakeholders' participation.
  - It needs to be complemented by an appropriate privacy risk model.
  - Such a risk model needs to be complemented by a sufficiently robust model that captures the main factors that have an impact on privacy risks along with their meanings, properties and relationships.

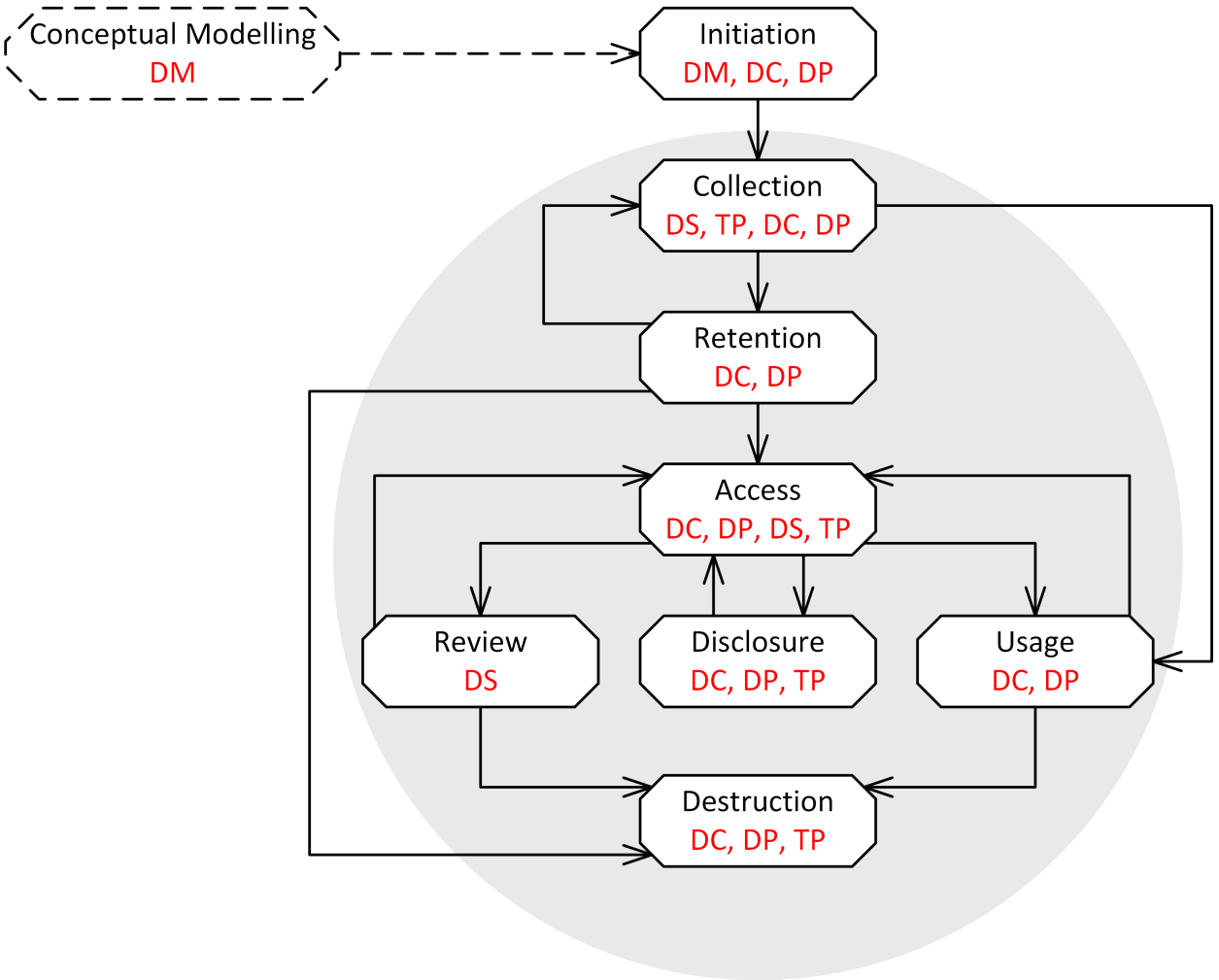
An overview of the APDL model

- The APDL model was developed to serve as an abstract model for personal data lifecycles — where a data lifecycle is defined by a set of **stages** through which personal data moves during its lifetime, associated **activities**, and involved **actors**.
- It aims to specify and represent the minimum amount of personal data, along with possible data-processing activities, necessary for the specified purpose.
- It distinguishes between the main types of operations that can be performed on personal data during its lifetime.
- For each operation, it outlines distinct activities that can be conducted in an ordered and planned manner.

An overview of the APDL model

- It categorises these activities into the following lifecycle stages: initiation, collection, retention, access, review, usage, disclosure and destruction.
- Each lifecycle stage is restricted by a set of the GPS principles to govern the behaviour of associated activities.
- Each stage involves a set of activities that may be performed by one or more types of role: data modellers, data subjects, data controllers, data processors and third parties.
- Each role specifies a set of related activities that are expected to be performed together by different actors according to their capabilities and responsibilities.

The APDL model



Lifecycle roles:

DM : Data modellers
DS : Data subjects
DC : Data controllers
DP : Data Processors
TP : Third parties

Modelling principles

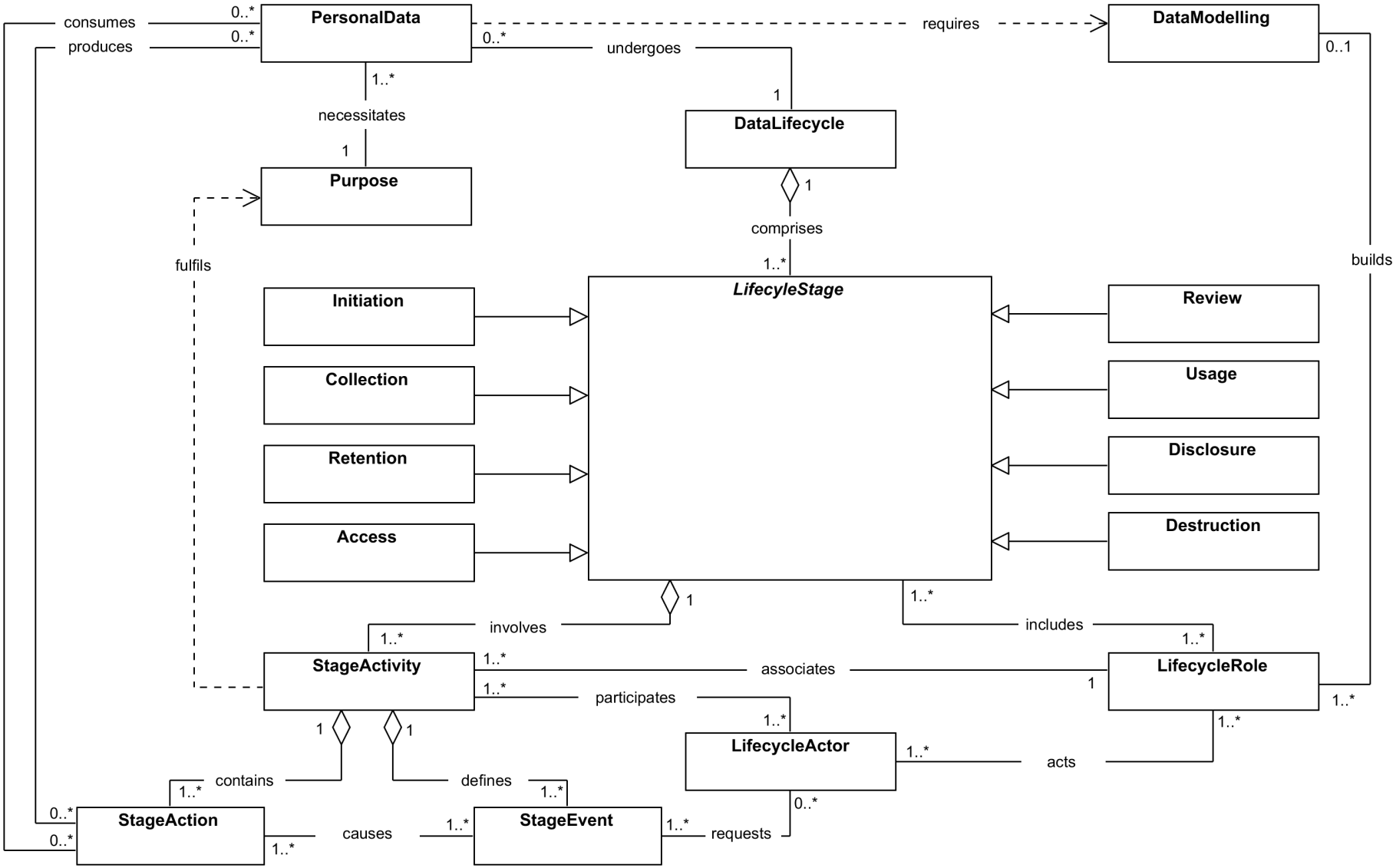
In order to develop an appropriate conceptual model, essential principles for the core parts of the model, which will be used as the basis for our UML profile, need to be specified.

- The **purpose** of modelling is to describe precisely the key privacy-related concepts, associated properties and relationships in the context of data protection.
- The **scope** of the modelling is identified by a list of key concepts proposed by the APDL model.
- Informal text analysis is used as an appropriate **technique** for deriving useful and potentially usable concepts, associated properties and relationships.
- Concept classification is used as a **conceptualisation approach** to analyse and classify relevant terms into concepts and processing activities that can be represented in a fine-grained manner as actions.

A Conceptual model for the APDL

- The conceptual model is intended to be used as a **common language** for privacy engineering to consider protection, manageability and traceability of personal data. Such a language is provided with the ability to express stakeholders' expectations and concerns.
- It involves concepts that are currently **not supported** by existing modelling languages, such as purposes, activities, actions, events, roles and actors.
- It is used as a **preliminary acquisition step** for requirements analysis to represent data-processing activities in a contextual manner.

The meta-model of the APDL profile



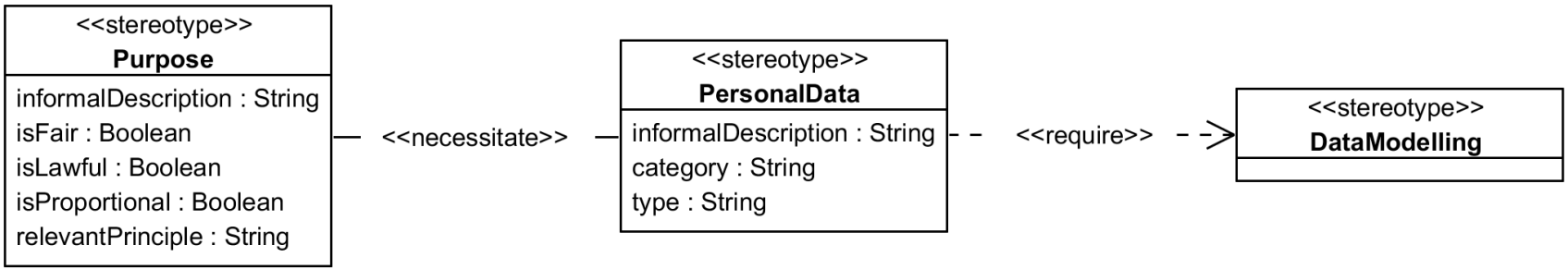
A UML profile for the APDL model

- A **UML Profile** describes lightweight extension mechanism to the UML by defining **stereotypes, tagged values,** and **constraints**.
- The **stereotypes** of the APDL profile are defined to extend existing metaclasses with the aim of using privacy-related terminology whether in place of, or in addition to, the terminology used for the extended metaclasses.
- The abstract syntax of the APDL is specified by extending three elements of the UML metamodel — the metaclass *Class*, the metaclass *Association* and the metaclass *Dependency* — with additional properties and constraints.

Purpose, personal data and data modelling

- **Purpose** represents goals and reasons for which personal data is collected and processed.
- **PersonalData** represents the minimum necessary amount of data that is sufficiently related to an identified or identifiable individual in support of the specified purpose.
- **DataModelling** represents the relevant objects, associated properties, relationships and constraints for the purpose of specifying the required data.

Purpose, PersonalData and DataModelling



Lifecycle and associated stages

- **DataLifecycle** represents the main characteristics of the personal data lifecycle in terms of the openness of the processed data and the centrality of its underlying system.
- **LifecycleStage** represents the concept of a generic lifecycle stage that models all possible stages through which personal data moves during its lifecycle in more repetitive and circular flows.

Lifecycle and associated stages

- **Initiation** represents a complete processing plan that can be referred to before and during the processing of personal data.
- **Collection** represents the act of recording, capturing or collecting personal data values, whether these are directly collected from data subjects, or have been acquired from external sources.
- **Retention** represents the act of organising, structuring or storing personal data values in repositories or digital storage media for operational, compliance or operational recovery purposes.

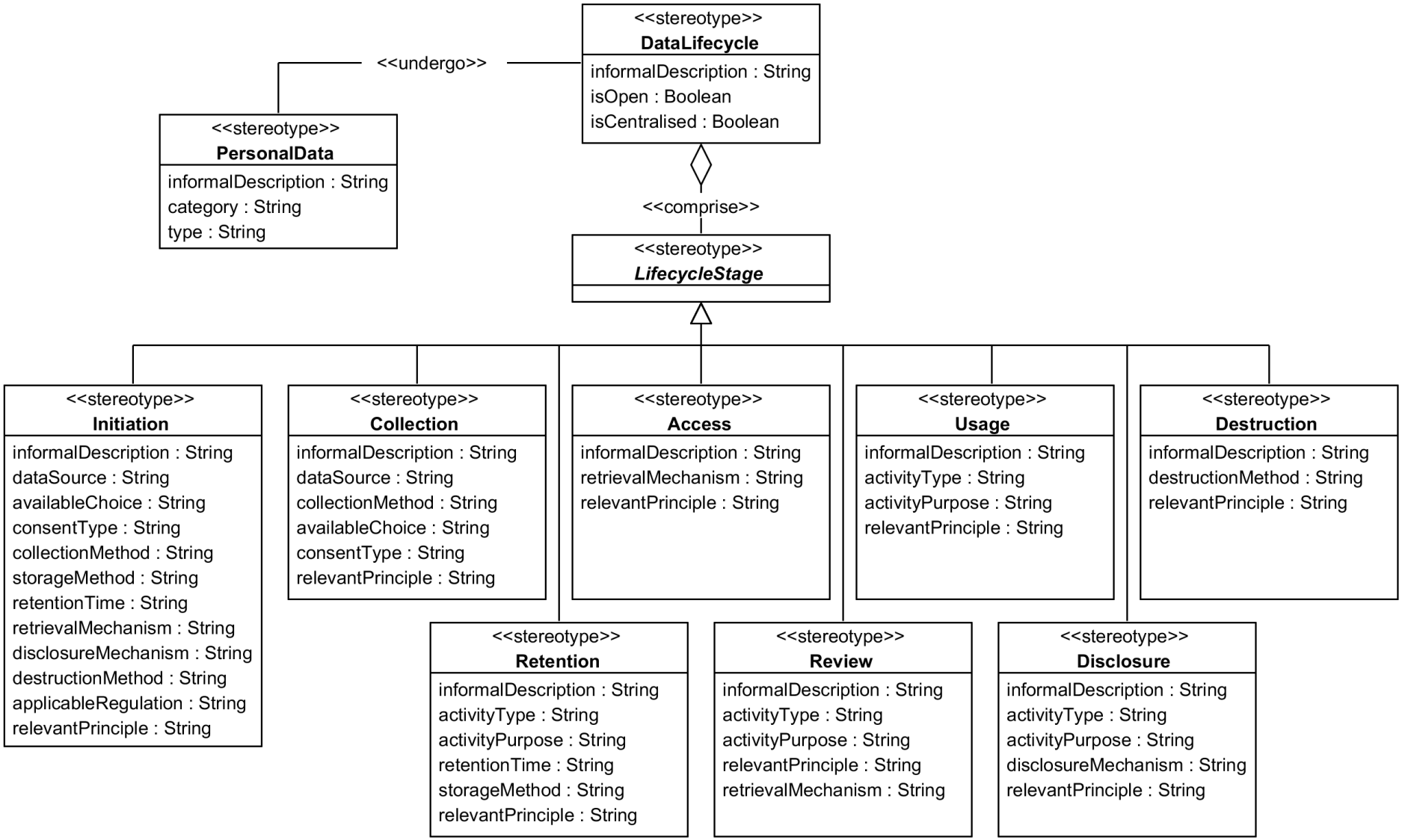
Lifecycle and associated stages

- **Access** represents the act of specifying, and retrieving or consulting personal data values that are stored in repositories or digital storage media.
- **Review** represents the act of implementing the access right and rectifying personal data values by data subjects to ensure that their data is accurate, complete and up-to-date.
- **Disclosure** represents the act of disseminating, making available or transmitting personal data for external use by third parties.

Lifecycle and associated stages

- **Usage** represents the act of classifying, altering, adapting, refining, aligning, analysing, integrating or using personal data items.
- **Destruction** represents the act of erasing, destroying, redacting or disposing of personal data.

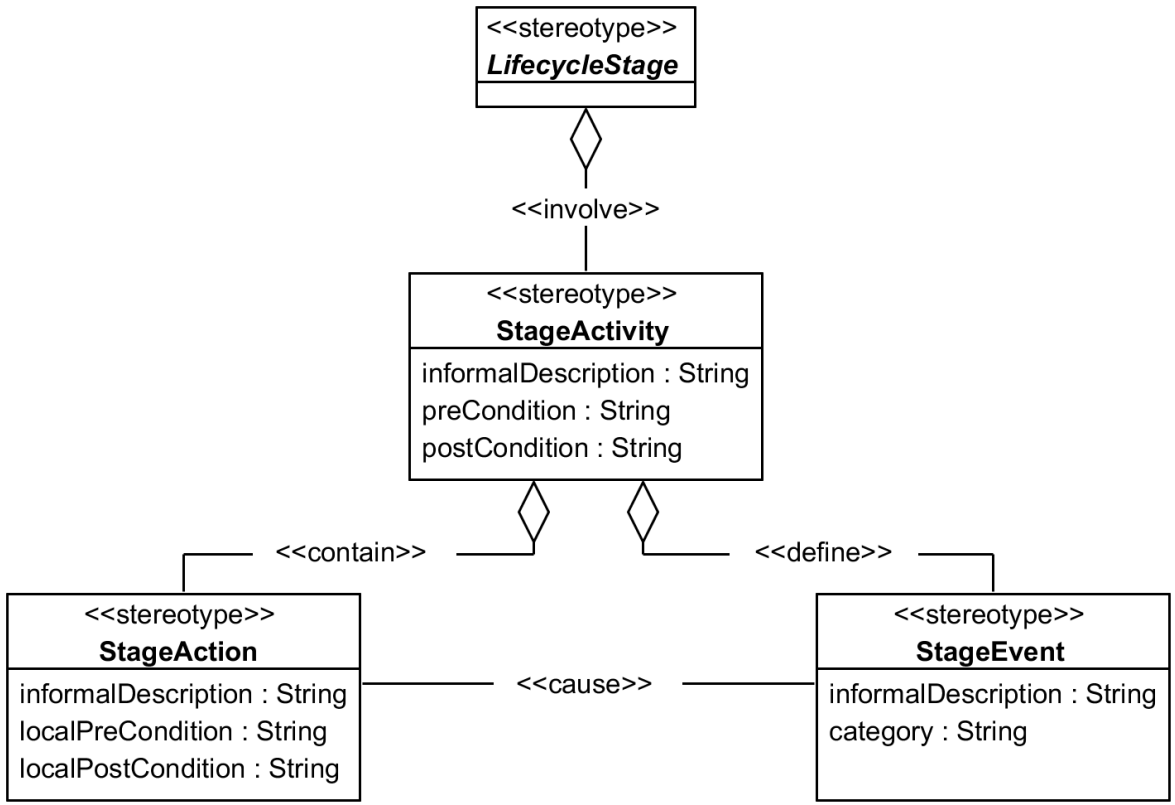
DataLifecycle, LifecycleStage and its specialisations



Stage activities, events and actions

- **StageActivity** represents data-processing activities that constitute the operations performing on personal data in each stage of the lifecycle.
- **StageEvent** represents occurrences that may happen at specific points in time which have consequences on personal data.
- **StageAction** represents single execution steps within an activity. Actions are the fundamental units that describe personal data processing activities in a fine-grained manner.

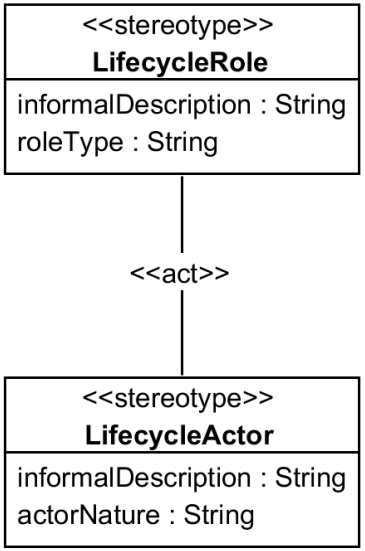
StageActivity, StageEvent and StageAction



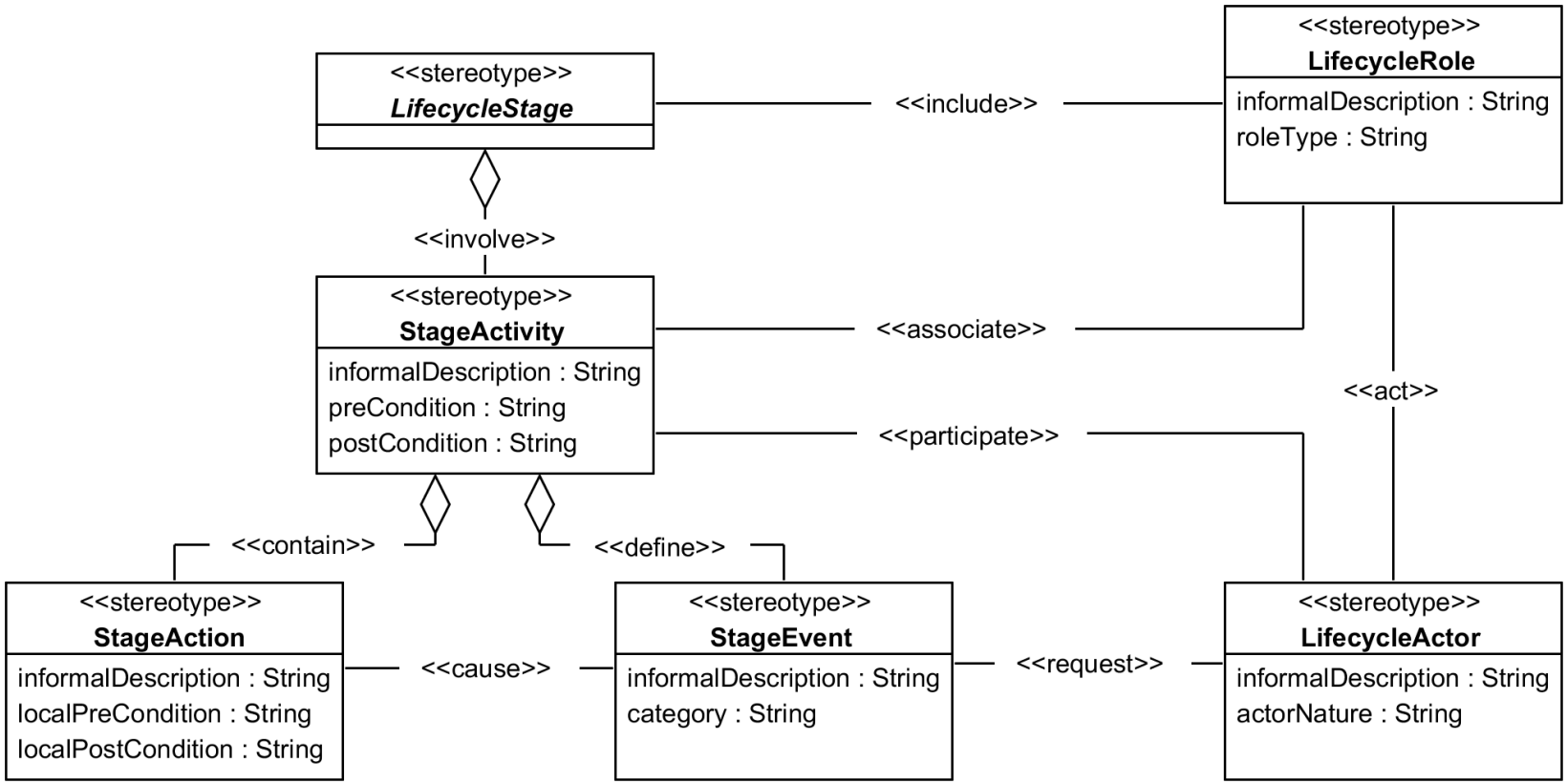
Lifecycle roles and actors

- **LifecycleRole** represents the way in which a concerned actor participates in a set of related activities of the personal data lifecycle.
- Each role represents a set of responsibilities that are logically related to each other, either by their objectives or the actors that may play the role.
- **LifecycleActor** represents an external or internal entity that is capable of, and responsible for, performing the activities of the role to which it is assigned

LifecycleRole and LifecycleActor



LifecycleStage, LifecycleRole and LifecycleActor



A way forward

- The conceptual model is limited to those terms that are necessary to define the fundamental concepts of the personal data lifecycle. The scope of modelling might be further extended by refining or defining all relevant concepts, associated meanings, properties and relationships.
- The UML profile needs to be further validated by using additional case studies from different domains.

Current work

An Analytical Approach for Analysing Potential Privacy Risks

- A **risk model** that represents the main factors that have an impact on privacy risks along with their meanings, properties and relationships.
- A **risk analysis approach** that identifies, analyses and assesses privacy risks in a comprehensive, contextual and non-reductive manner.
- It is intended to **adopt** the APDL model to capture all relevant and useful information for risk analysis and compliance checking.

Conclusion

- The UML profile for the APDL model provides foundations for analysing functional requirements and assessing potential privacy risks in a contextual and comprehensive manner.
- It represents privacy-related concepts using the standard extension mechanisms of the UML meta-model.
  - **Stereotypes** and **tagged values** are used to represent key aspects of privacy principles as requirements and assumptions.
  - **Constraints** provide criteria for the evaluation of these aspects to determine whether the representation of data-processing activities fulfils these requirements.

Conclusion

- The lifecycle stages and roles can be used to classify processing activities into multiple partitions, according to the nature of these activities, the capabilities and responsibilities of involved actors, the organisational units performing these activities or the geographical location at which these activities are performed.
- The APDL meta-model is a way of describing data-processing at a fined-grained level, with the possibility of expressing **how** activities are performed, **what** are their effects in terms of changes of states, **when** they take place in terms of lifecycle stages, and **where** they take place in terms of lifecycle roles.

Thank you !

Questions?