LIÈGE université
Sciences Appliquées

Just-In-Time Integration and Generation of Datasets

Christophe Debruyne

Montefiore Institute @ University of Liège

2022-11-30 @ KG4DI FWO Scientific Research Network Kick-Off

Generating GDPR-compliant datasets

Note: compliant with respect to the informed consent obtained by an organization.

Problem: datasets are created and used for a specific purpose and data processing is the subject of various internal and external regulations (e.g., informed consent).

- Too much focus on post-hoc compliance.
- Ensuring compliance with respect to informed consent is challenging.
- Can we ensure compliance at earlier stages?

Unlike other initiatives (which we will discuss later), we focus on consent information that has been stored, not the context (e.g., via a form) nor the process.

Generating GDPR-compliant datasets

Note: compliant with respect to the informed consent obtained by an organization.

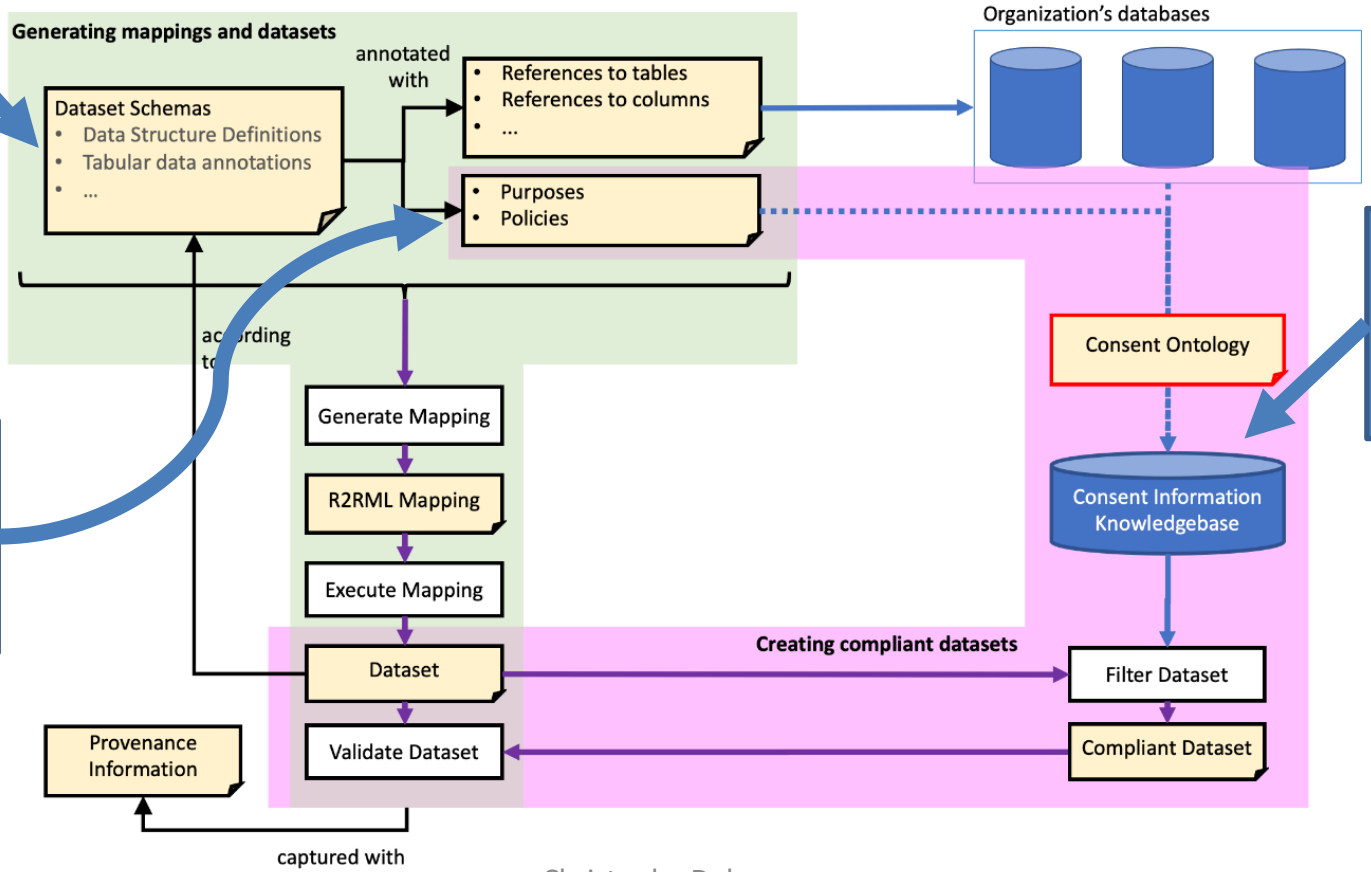
Approach

- Context-model: a knowledge graph of terms and conditions, schemas for datasets, data processing purposes, and informed consent.
- Just-in-time dataset compilation:
  - What dataset schema?
  - Where to get data?
  - Who gave their consent?
  - Compile the dataset.
- Context-aware data integration: a sequence of SPARQL graph queries rendering the process fully transparent, traceable, and declarative.

A knowledge graph of T&C, data processing purposes, and informed consent.
Building of "intelligent" agents for generating and integrating data.

A representation of the data that we will process and references to databases.

RDF Data Cube Vocabulary.



Relating schemas to data processing purposes

A representation of the purposes and policies, and a KB based on the consent gathered

Annotating the dataset schemas

```
# PREFIXES OMMITED FOR BREVITY
@base <http://www.example.org/>
dct:identifier a rdf:Property, qb:DimensionProperty ;
  rr:template "http://data.example.com/user/{id}" ;
  rdfs:label "user id"@en ;
  rdfs:subPropertyOf sdmx-dimension:refPeriod ;
  rdfs:range owl:Thing .
 foaf:mbox a rdf:Property, qb:MeasureProperty ;
 rr:template "mailto:{email}" ;
  rdfs:label "email address"@en ;
  rdfs:subPropertyOf sdmx-measure:obsValue ;
  rdfs:range owl:Thing .
<#dsd-le> a qb:DataStructureDefinition;
  rr:tableName "user";
  ont:forPurpose <http://data.example.com/purpose/8> ;
  ont:forPolicy <http://data.example.com/policy/10> ;
  qb:component [ qb:dimension dct:identifier ];
  qb:component [ qb:measure foaf:mbox ] .
```

A simple data structure definition (DSD).

Yellow → Annotations for the generation of an R2RML mapping

Cyan → Linking the DSD to a purpose of a policy

Engaging with the KG

```
DESCRIBE ?consent WHERE {
  ?consent ont:forInclusion ?inclusion .
  { # GET LATEST INCLUSION OF PURPOSE FOR A POLICY
    SELECT ?inclusion WHERE {
      ?inclusion ont:ofPurpose <.../purpose> .
      ?inclusion ont:ofPolicy <.../policy> .
      <.../policy> dcterms:created ?dt . }
    ORDER BY DESC(?dt) LIMIT 1 }
  ?consent ont:givenBy ?user .
  ?consent ont:registeredOn ?datetime .
  # GET LATEST CONSENT INFORMATION FOR EACH USER
  FILTER NOT EXISTS {
    [ ont:forInclusion ?inclusion ;
      ont:givenBy ?user ;
      ont:registeredOn ?datetime2 ]
    FILTER(?datetime2 > ?datetime)
  }
}
```

Retrieving the latest consent information for a specific purpose of the latest version of a policy.

DESCRIBE query returns a graph which we will use to manipulate the dataset.

Implementation details

- Both consent information and endpoint are behind the service, one just needs an annotated DSD. Governance platforms can be adopted to guide one to identifiers for a policy and purpose.
- Intermediate graphs that are generated allow for a posteriori analysis.
- Both governance problems are outside the scope of this study and considered future work.

Related Work

- Both SPIRIT (Westphal et al. 2018) and SPECIAL (Kirrane et al. 2018) studies have similar concepts (data subject, purpose, …), but aim to analyze compliance a posteriori (both) or a priori (SPECIAL). Our goal was to generate compliant datasets “just in time”
- Pandit, O’Sullivan and Lewis (2018) proposed an ontology for the operational representation of informed consent and allows one to analyze these representations w.r.t. annotated logs and questionnaires.
- Fatema et al (2017) proposed a model for representing the informed consent an organization has obtained, but there is no explicit notion of policies or support for revisions.

Given the overlap in terminology, it is clear there is an opportunity in aligning the vocabularies.

Summary and future

- Integrating knowledge and data
- Formalizing aspects of GDPR
- Datasets of demand in a declarative and transparent manner

- This work also validated aspects of the Data Privacy Vocabulary (DPV).

Sources

- Christophe Debruyne, Dave Lewis, Declan O'Sullivan: Generating Executable Mappings from RDF Data Cube Data Structure Definitions. OTM Conferences (2) 2018: 333-350

- Christophe Debruyne, Harshvardhan J. Pandit, Dave Lewis, Declan O'Sullivan: "Just-in-time" generation of datasets by considering structured representations of given consent for GDPR compliance. Knowl. Inf. Syst. 62(9): 3615-3640 (2020)