Queryable Provenance Metadata For GDPR Compliance

Harshvardhan J. Pandit, Declan O’Sullivan, Dave Lewis
ADAPT Centre - Trinity College Dublin - Dublin, Ireland
https://openscience.adaptcentre.ie/
pandith@tcd.ie | @coolharsh55
Overview

1. GDPR (what, why, who, where, when guidance by regulatory authorities)
2. GDPR Readiness Checklist by Ireland’s DPC
3. Semantification of queries
4. Implementation & Demonstration
5. Related Work

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
General Data Protection Regulation

- Enforced from 25th May 2018
- Fines: 4% global turnover or 20 million whichever is higher
- Obligations and rights based on use of consent and legal basis
- Necessary documentation
- Impact Assessments
- Data Privacy Officer
- Rights for Data Subjects
- Distinction between Controllers and Processors
- Sharing with Named Third Parties
- Privacy Seals
GDPR Readiness Checklist

ON YOUR MARK
GET READY
ANY TIME NOW
REALLY SOON

http://gdprandyou.ie/

GDPR Readiness Checklist

PREPARING YOUR ORGANISATION FOR THE GENERAL DATA PROTECTION REGULATION
YOUR READINESS CHECKLIST

DATA PROTECTION COMMISSIONER

25TH MAY 2018
GET AWARE AND GET PREPARED

Structure & Layout
1. 13 pages
2. 63 questions
3. 9 sections

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
<table>
<thead>
<tr>
<th>Category of Questions</th>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Comments/Remedial Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal data</td>
<td>Consent based data processing (Articles 7, 8 and 9 and further guidance available on GDPRandYou.ie)</td>
<td><strong>BINARY</strong></td>
<td>✓</td>
<td>more information</td>
</tr>
<tr>
<td></td>
<td>Have you reviewed your organisation’s mechanisms for collecting consent to ensure that it is freely given, specific, informed and that it is a clear indication that an individual has chosen to agree to the processing of their data by way of statement or a clear affirmative action?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>If personal data that you currently hold on the basis of consent does not meet the required standard under the GDPR, have you re-sought the individual’s consent to ensure compliance with the GDPR?</td>
<td>✓</td>
<td>✓</td>
<td>specifics</td>
</tr>
<tr>
<td></td>
<td>Are procedures in place to demonstrate that an individual has consented to their data being processed?</td>
<td>✓</td>
<td>✓</td>
<td>details</td>
</tr>
<tr>
<td></td>
<td>Are procedures in place to allow an individual to withdraw their consent to the processing of their personal data?</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Children’s personal data (Article 8)</td>
<td>Where online services are provided to a child, are procedures in place to verify age and get consent of a parent/ legal guardian, where required?</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Analysis

Consider each question a ‘query’

Three categories of queries:

1. Demonstrative
   - Demonstrate a process or an activity
   - Real

2. Evaluative
   - Evaluate a criteria
   - Ideal

3. Assistive
   - Cannot demonstrate or evaluate
   - Therefore retrieve all relevant information that can assist in demonstrating or evaluation

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
**Queryable Provenance Metadata for GDPR Compliance** at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55

---

## Analysis - notes

- **Article in GDPR**
- **Data involved**

**https://w3id.org/GDPRep/checklist-demo/notes**

---

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>Title</th>
<th>Comment</th>
<th>GDPR</th>
<th>To implement</th>
<th>Data</th>
<th>Data Comment</th>
<th>Model Comment</th>
<th>Instance Comment</th>
<th>Automated</th>
<th>Authorite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>General</td>
<td>General</td>
<td>Fibre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>General</td>
<td>General</td>
<td>Article</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>General</td>
<td>General</td>
<td>eq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>General</td>
<td>General</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- "Fibered NOT IMPLEMENTED can this be assessed on the model of the system or it requires instances?"
- "The CAN be ran on instances for data subject-specific queries. This is not the original query mean"
SPARQL queries

- Nearly arbitrary number
- 33 SPARQL queries

- Ontologies
  - GDPRov
  - GDPRtEXT

prefixes

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX gdprov: <http://purl.org/adaptcentre/openscience/ontologies/gdprov#>
PREFIX gdprtext: <http://purl.org/adaptcentre/openscience/ontologies/GDPRtEXT#>
PREFIX p-plan: <http://purl.org/net/p-plan#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX this: <http://example.com/ontology/shoppingapp#>
```

G5. legal basis for processing

```
SELECT DISTINCT ?process ?legal where {
  ?data a ?data_type .
  ?data_type rdfs:subClassOf gdprov:PersonalData .
  ?step_type rdfs:subClassOf gdprov:DataStep .
  ?step gdprov:usesData ?data .
  OPTIONAL (?step gdprov:hasLegalBasis ?legal ) .
  OPTIONAL (?process gdprov:hasLegalBasis ?legal ) .
} ORDER BY ?process
```
Implementation

- proof-of-concept demonstration
- example use-case: online shopping service
- GDPRov & GDPRtEXT ontologies
- Protege (environment) → FACT++ (reasoner)

proof of concept that compliance questions CAN be expressed as SPARQL queries

0. define instances
1. add metadata
2. run reasoner

- easy to use
- checks common human errors
- visual tool
- integrates with reasoners
- can execute SPARQL

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
Demonstration

- online demo for querying of ‘readiness checklist’ information
- aims
  - convert static document to interactive/automated environment
  - use semantic web to create a graph of information
- same layout and format as original document
- queries SPARQL endpoint on page load (browser)

https://w3id.org/GDPRep/checklist-demo
Elements of personal data included within each data category

List each type of personal data included within each category of personal data e.g. name, address, banking details, purchasing history, online browsing history, video and images.

G2. Types of Personal Data

```
SELECT DISTINCT ?data ?type
WHERE {
  ?data a ?type.
  ?type rdfs:subClassOf <gprov:PersonalData> .
  FILTER(regex(str(?data), "http://example.com/ontology/shoppingapp")) .
} ORDER BY ?data ?type.
```

<table>
<thead>
<tr>
<th>data</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>this:AnonymisedUserProfile</td>
<td><a href="">gprov:AnonymisedData</a></td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:CustomerInfo</td>
</tr>
<tr>
<td>this:CustomerBankAC</td>
<td><a href="">gprov:SensitiveData</a></td>
</tr>
<tr>
<td>this:CustomerCardDetails</td>
<td><a href="">gprov:SensitiveData</a></td>
</tr>
<tr>
<td>this:CustomerContactNo</td>
<td>this:CustomerInfo</td>
</tr>
<tr>
<td>this:CustomerEmail</td>
<td>this:CustomerInfo</td>
</tr>
<tr>
<td>this:CustomerName</td>
<td>this:CustomerInfo</td>
</tr>
</tbody>
</table>

Showing 1 to 7 of 7 entries
Invalid/Non-existant/Empty/Null Queries

Not all questions from the GDPR Readiness Checklist could be interpreted into SPARQL queries.

Retrospective Consent

If personal data that you currently hold on the basis of consent does not meet the required standard under the GDPR, have you re-sought the individual’s consent to ensure compliance with the GDPR?

Does not contain provenance metadata OR Is currently not implemented

\[
\text{total questions} = 63 \\
\text{SPARQL queries} = 33 \\
\text{not-implemented} = 63 - 33 = 30
\]

This happened because not all queries were quantitative IR questions.
Queries that provide information cannot be evaluated.

A1. personal data purposes

```
SELECT DISTINCT ?data ?process WHERE {
  ?StepType rdfs:subClassOf gdprov:DataStep.
  ?step a ?StepType.
  ?DataType rdfs:subClassOf gdprov:PersonalData.
  ?data a ?DataType.
  ?action ?data.
  ?step gdprov:isPartOfProcess ?process
}
ORDER BY ?data ?process
```

SPARQL query to retrieve relevant information.

<table>
<thead>
<tr>
<th>data</th>
<th>process</th>
</tr>
</thead>
<tbody>
<tr>
<td>this:AnonymisedUserProfile</td>
<td>this:RemoveUserAccountProcess</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:AdGenProcess</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:HandleRightDataPortability</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:HandleSAR</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:NewUserSignUpProcess</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:OrderProcess</td>
</tr>
<tr>
<td>this:CustomerAddress</td>
<td>this:RemoveUserAccountProcess</td>
</tr>
</tbody>
</table>

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

http://openscience.adaptcentre.ie/  |  pandith@tcd.ie  |  @coolharsh55
Assumptions and Limitations

1. Does not assess compliance
   
   - not in scope
   - not the aim of the work
   - requires more data

2. Depends on consent
   
   - as in, the modelling of consent and how to interpret it
   - pre-GDPR notion of consent was a confusion carnival

3. Interpretation of results is not clear
   
   - results are just tabular data, what to do with it?
   - SPARQL ASK queries
   - why haven’t you done them?

Doing them now!

Check out my POSTER

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
Proof of Concept!

My client couldn’t have killed anyone with this arrow, and I can prove it!

I’d like to examine your proof, Zeno. You may approach the bench.

But never reach it!

Started out as an evaluation of how to express these compliance related questions as queries over semantic metadata

Success!!!
Technologies

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
Keep your friends close, your ‘peers’ closer...

SPECIAL PROJECT
- semantic web based compliance framework
- OWL reasoning to evaluate compliance
- web-based dashboard

PrOnto Ontology
- OWL modelling of compliance related concepts and terms
- describe data (metadata) with relation to compliance

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://opencience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
Data Privacy Vocabularies and Controls Community Group

The mission of the W3C Data Privacy Vocabularies and Controls CG (DPVC) is to develop a taxonomy of privacy terms, which include in particular terms from European General Data Protection Regulation (GDPR), such as taxonomy of personal data as well as a classification of purposes (i.e., purposes for data collection), disclosures, consent, and processing of personal data.

The Community Group officially started on 25th of May 2018, the official GDPR coming into force, as a result of the W3C Workshop on Data Privacy Vocabularies in Vienna earlier this year.

Data Privacy Vocabularies and Controls Community Group

https://www.w3.org/community/dpvcg/

Data Privacy Vocabularies and Controls Community Group

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
Future Work

1. check GDPR compliance using SHACL

2. build a ‘knowledge graph’ with compliance related information

3. create a ‘unit testing’ approach towards compliance

“Queryable Provenance Metadata for GDPR Compliance” at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55
"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit

http://openscience.adaptcentre.ie/ | pandith@tcd.ie | @coolharsh55