



Engaging Content
Engaging People

Queryable Provenance Metadata For GDPR Compliance

→ that's me!

Harshvardhan J. Pandit, Declan O'Sullivan, Dave Lewis

ADAPT Centre - Trinity College Dublin - Dublin, Ireland

<https://openscience.adaptcentre.ie/>
pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)

↓
email

↓
twitter

→ all our GDPR
related work
CHECK
IT OUT!

- 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](https://twitter.com/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

www.adaptcentre.ie



Effort by Ireland's
Data Protection Commissioner



[Home](#) [Individuals](#) [Organisations](#) [Resources](#) [Media](#) [Contact Us](#)

<http://gdprandyou.ie/>
website with guidelines
& resources

GDPR

Coming May 25th 2018

2018

The General Data Protection Regulation significantly changes data protection law in Europe, strengthening the rights of individuals and increasing the obligations on organisations. Get aware, and get prepared.



“GDPR Readiness Checklist”

PREPARING YOUR ORGANISATION FOR THE GENERAL DATA PROTECTION REGULATION

YOUR READINESS CHECKLIST



DATA PROTECTION COMMISSIONER

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](https://twitter.com/coolharsh55)



GDPR Readiness Checklist (pg.10)

www.adaptcentre.ie

category of questions

Personal data

	Question	Yes BINARY	No BINARY	Comments/ Remedial Action
Consent based data processing (Articles 7, 8 and 9 and further guidance available on GDPRandYou.ie)	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?	✓		more information
sub	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?		✓	specific
	Are procedures in place to demonstrate that an individual has consented to their data being processed?		✓	details
	Are procedures in place to allow an individual to withdraw their consent to the processing of their personal data?	✓		
Children's personal data (Article 8)	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?	✓		

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)



→ our approach

Consider each question a 'query'

Three categories of queries:

1. Demonstrative → demonstrate a process or an activity
 2. Evaluative → evaluate a criteria
 3. Assistive → cannot demonstrate or evaluate, therefore retrieve all relevant information that can assist in demonstrating or evaluation
- } ideal
- real

Analysis - notes

identified

<https://w3id.org/GDPRRep/checklist-demo/notes>

Article in GDPR

data involved

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1														
2	ID	Category	Title	Comment	GDPR	Type	To Implement?	Data	Data Comment	Model based?	Model Comment	Instance based?	Instance Comment?	Automate
3	G1	General	Categories of personal data and data subjects	List the categories of data subjects and personal data collected and retained e.g. current employee data, retired employee data; customer data (sales information); marketing database; CCTV footage.	demonstr	Y	personal data, data subjects	subclasses that have other subclasses can be considered as categories in this case		Y	this only needs information about the classes, not the instances	N	instances are difficult to aggregate into categories, and would need some abstract information to efficiently do so	Y
4	G2	General	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.	demonstr	Y	personal data	subclasses that do not have other subclasses can be considered by types within categories		Y	this only needs information about the classes, not the instances	N	instances are difficult to aggregate into categories, and would need some abstract information to efficiently do so	Y
5	G3	General	Source of the personal data	List the source(s) of the personal data e.g. collected directly from individuals; from third parties (if third party identify the data controller as this information will be necessary to meet obligations under Article 14).	type demonstr	Y	personal data, steps that collect data, entities that provide data	can this be assessed on the model of the system or it requires instances?		Y	there could be fixed models where data is collected directly from data subjects or some data provider which can be shown through the abstract model	Y	instances can show who the actual data providers are, if they can change with time. Ideally, the change should be reflected in the model	Y
6	G4	General	Purposes for which personal data is processed	Within each category of personal data list the purposes for the data is collected and retained e.g. marketing, service enhancement, research, product development, systems integrity, HR matters, advertising.	demonstr	Y	results of G1, processes acting on data	get all plans that contain steps that act on the data, then aggregate them based on categories		Y	run this over the model only as it enquires about the state of the system and not about a particular instance	N	this CAN be run on instances for data subject specific queries, but this is not what the original query meant	Y
7	G5	General	Legal basis for each processing purpose (non-special categories of personal data)	For each purpose that personal data is processed, list the legal basis on which it is based e.g. consent, contract, legal obligation (Article 6).	demonstr	Y	results of G4, processes acting on data	get legal basis in steps within plans from G4		Y	legal basis does not change in instances, so query this on models	N	this CAN be run on instances for data subject specific queries, but this is not what the original query meant	Y
8	G6	General	Special categories of personal data	If special categories of personal data are collected and retained, set out details of the nature of the data e.g. health, genetic, biometric data.	demonstr	Y	special category personal data	subclasses under special category of personal data		Y	as with normal categories of data, this query only needs information about category, not specifics	N	instances are difficult to aggregate into categories, and would need some abstract information to efficiently do so	Y
9	G7	General	Legal basis for processing special categories of personal data	List the legal basis on which special categories of personal data are collected and retained e.g. explicit consent, legislative basis (Article 9).	demonstr	Y	results of G6, steps that collect data, steps that store data	get all steps that collect or store special categories of data, then retrieve their legal basis		Y	same as G5, this is information about the abstract model	N	this CAN be run on instances for data subject specific queries, but this is not what the original query meant	Y
10	G8	General	Retention period	For each category of personal data, list the period for which the data will be retained e.g. one month? one year? As a general rule data must be retained for no longer than is necessary for the purpose for which it was collected in the first place.		N	results of G1, steps that store data	this is interpretative based on how retention time is calculated; ideally, this will be a part of the consent or policy that feeds into the provenance graph						
11	G9	General	Action required to be GDPR compliant?	Identify actions that are required to ensure all personal data processing operations are GDPR compliant e.g. this may include deleting data where there is no further purpose for retention.		N		this is very vague and does not depend on classes nor direct links to provenance graphs. A list of processes or plans can be linked to show 'actions' but these would still need to be combined with some form of documentation						
12	P1	PersonalData	Validity of Consent	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?	7,8,9	assistive	Y	consent, steps that acquire consent	This cannot be directly evaluated because of conditions such as freely given, specific, etc. which are qualitative. But, the information about how the consent was collected can be presented to make an informed decision.	Y	identify steps that collect consent along with static data content such as privacy policy and T&C that are used along with the form mechanism used to collect consent.	N	This is assuming that the instances follow the abstract model. So the mechanism that they used to collect consent is the same as that referenced in the abstract model. Therefore, this is already considered to be evaluated under the abstract model. However, this CAN be used to retrieve and evaluate the consent mechanism for a particular data subject.	Y

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

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



SPARQL queries

www.adaptcentre.ie

- (1) nearly arbitrary number
- (2) based on questions in checklist

available here

<https://w3id.org/GDPRRep/checklist-demo/sparql-queries>

● 33 SPARQL queries

● Ontologies

- GDPRProv

- GDPRtEXT
 - GDPR as linked data resource
 - Vocabulary of terms/concepts

prefixes

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX gdprov: <http://purl.org/adaptcentre/opensource/ontologies/gdprov#>
PREFIX gdprtext: <http://purl.org/adaptcentre/opensource/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#>
```

↑ non-existent

G5. legal basis for processing

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

Implementation

www.adaptcentre.ie

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

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



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

- ① define instances
- ② add metadata
- ③ run reasoner



Demonstration

www.adaptcentre.ie

<https://w3id.org/GDPRRep/checklist-demo>

- 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)

↓
executes SPARQL
on page refresh



Query G2: Personal Data in Data Category

www.adaptcentre.ie

category → 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.

→ *category description*

Query ID → G2. Types of Personal Data

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

SPARQL

	data	type
1	this:AnonymisedUserProfile	gdprov:AnonymisedData
2	this:CustomerAddress	this:CustomerInfo
3	this:CustomerBankAC	gdprov:SensitiveData
4	this:CustomerCardDetails	gdprov:SensitiveData
5	this:CustomerContactNo	this:CustomerInfo
6	this:CustomerEmail	this:CustomerInfo
7	this:CustomerName	this:CustomerInfo

Showing 1 to 7 of 7 entries

Results

data category

type of data



Invalid/Non-existent/Empty/Null Queries

www.adaptcentre.ie

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?

→ can be checked with additional data

Does not contain provenance metadata OR Is currently not implemented

$$\begin{array}{lcl} \text{total questions} & = & 63 \\ \text{SPARQL queries} & = & 33 \\ \text{not-implemented} & = & \text{DO-THE-MATH} \\ & & 63 - 33 = 30 \end{array}$$

that's a lot!
↓
this happened because not
all queries were
quantitative IR questions



Queries that provide information

Purpose Limitation

Is personal data only used for the purposes for which it was originally collected? → 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 .
  ?step ?action ?data .
  ?step gdprov:isPartOfProcess ?process
} ORDER BY ?data ?process
```

} SPARQL query to retrieve relevant information

RAW RESPONSE TABLE PIVOT TABLE GOOGLE CHART Search: Show 50 entries

data	process
1 this:AnonymisedUserProfile	is used in
2 this:CustomerAddress	this:RemoveUserAccountProcess
3 this:CustomerAddress	this:AdGenProcess
4 this:CustomerAddress	this:HandleRightDataPortability
5 this:CustomerAddress	this:HandleSAR
6 this:CustomerAddress	this>NewUserSignUpProcess
7 this:CustomerAddress	this:OrderProcess
	this:DemolishUserAccountProcess

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

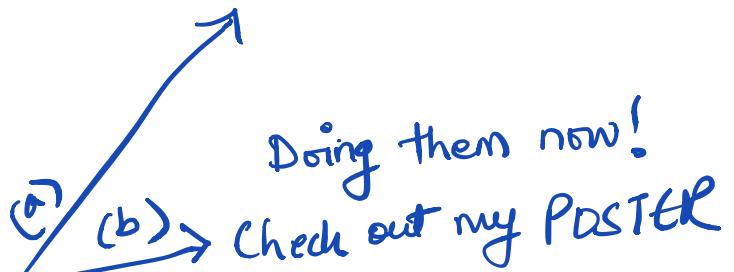
<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)



Assumptions and Limitations

www.adaptcentre.ie

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 ↓
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?
 - pre-GDPR notion of consent was a confusion carnival



"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)



Proof of Concept!

www.adaptcentre.ie



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

SUCCESS!!!

Technologies

www.adaptcentre.ie



protégé



Virtuoso Open Source



SPARQL



- colour highlight
- query & display

YAS GUI



browser
(privacy conscious choice)



wishlist



"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

Presented by: Harshvardhan J. Pandit

<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)

Keep your friends close, your ‘peers’ closer...

www.adaptcentre.ie

SPECIAL Privacy Dashboard

Group by:

- Purpose
- Data
- Processing
- Recipient
- Storage

8/31/2018 11:59:25 PM 9/1/2018 12:45:43 AM 9/1/2018 2:28:00 AM

Data processed on Fri Aug 31 2018
23:09:52 GMT+0200 (CEST)

AusPurpose
Browsing

Data processed on Fri Aug 31 2018
23:09:49 GMT+0200 (CEST)

Admin
Browsing
Charity
AusPurpose
Arts

Data processed on Fri Aug 31 2018
23:09:38 GMT+0200 (CEST)

Admin

Logo: TU Berlin

Name: Technical University of Berlin

Address: Ernst-Reuter-Platz 7 10587 Berlin

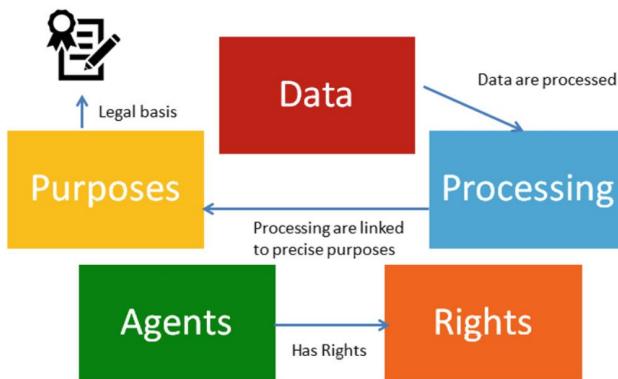
Email address: privacy@tu-berlin.de

Privacy policy: Privacy policy

Review consent: Review consent

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



COMMUNITY & BUSINESS GROUPS

[Home](#) / Data Privacy Vocabularies...

DATA PRIVACY VOCABULARIES AND CONTROLS

COMMUNITY GROUP

The mission of the W3C Data Privacy Vocabularies and Controls CG (DPVCG) develop a taxonomy of privacy terms, which include in particular terms from European General Data Protection Regulation (GDPR), such as a taxonomy of It is the goal of the CG to harmonize related efforts and bring together stakeholders that already have brought forward proposals to dev data as well as a classification of purposes (i.e., purposes for data collection), logs about personal data processing, enable data portability for data subjects, etc. The exact scope of use cases related to making pers of disclosures, consent, and processing such personal data.

The Community Group shall officially start 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.

It is the goal of the CG to harmonize related efforts and bring together stakeholders that already have brought forward proposals to develop respective vocabularies for semantic interoperability and interchange of transparency logs about personal data processing, enable data portability for data subjects, etc. The exact scope of use cases related to making personal data processing interoperable by respective standards in order to ease proof of compliance with the GDPR and related privacy protection regulations will be the first deliverable of the CG.

More concretely, the following steps and deliverables are planned so far.

Data Privacy Vocabularies and Controls Community Group

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

Use-Cases [\[edit\]](#)

- SPECIAL/Proximus use case - personalized touristic recommendations
- SPECIAL/DT use case - mobile network quality measurements
- SPECIAL/TR use case - 'Know Your Customer' (finance, anti-money-laundering)
- DECODE/DEC01 use case - Online voting system with privacy
- DECODE/DEC02 use case - Rental Register
- DECODE/DEC03 use case - Sharing sensor data

Data Privacy Vocabularies and Controls Community Group [\[edit\]](#)

The mission of the W3C Data Privacy Vocabularies and Controls CG (DPVCG) is to develop a taxonomy of privacy terms, which include in particular terms from the GDPR coming into force, as a result of the W3C Workshop on Data Privacy Vocabularies in Vienna earlier this year. The Community Group officially started on 25th of May 2018, the official data of the GDPR coming into force, as a result of the W3C Workshop on Data Privacy Vocabularies in Vienna earlier this year. It is the goal of the CG to harmonize related efforts and bring together stakeholders that already have brought forward proposals to develop respective vocabularies for semantic interoperability and interchange of transparency logs about personal data processing, enable data portability for data subjects, etc. The exact scope of use cases related to making personal data processing interoperable by respective standards in order to ease proof of compliance with the GDPR and related privacy protection regulations will be the first deliverable of the CG.

More concretely, the following steps and deliverables are planned so far:

1. Use cases and requirements: in a first step we will collect and align common requirements from industry and also from other stakeholders. The outcome shall be a prioritized list of requirements for what needs to be covered by shared vocabularies to enable interoperability.
2. Alignment of vocabularies and identification of overlaps: in a second document, we will collect existing vocabularies and standardize them to cover the requirements prioritized in step one.
3. Glossary of GDPR terms: a third deliverable will be an understandable glossary of common terms from the GDPR and how they should be used.
4. Vocabularies based on the heterogeneity or homogeneity of the agreed upon use cases and requirements, we will define a single vocabulary for purposes/processing, disclosure/consent, anonymisation, and transparency logs.

Timeline [\[edit\]](#)

For the moment, we plan the following milestones:

1. 24 May 2018: Presentation of this initial charter draft to initial stakeholders
2. 25 May 2018: Launch of the CG by registration as a proposed W3C Community Group
3. 26-30 May 2018 until 30 June 2018: dissemination of invitations to participate in the CG & feedback collection on the present charter
4. We have started 2-weekly Telephone Conferences on 23 July, see below.
5. 29-31 August 2018: 1st Face-to-face meeting co-located at MyData2018 in Helsinki, Finland, agreement on first steps and regular meetings
6. 12-14 November 2018: 2nd Face-to-face meeting co-located with the European Big Data Value Forum 2018 in Vienna, Austria. The topics will be:
 - ODRL
 - P3P

"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018

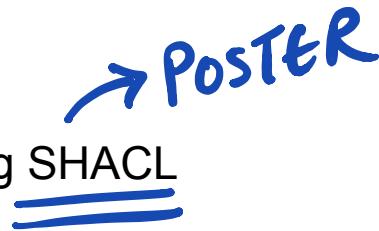
Presented by: Harshvardhan J. Pandit

<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)



Future Work

www.adaptcentre.ie



1. check GDPR compliance using SHACL
2. build a 'knowledge graph' with compliance related information

vision paper to be presented at ISWC workshop CK6
3. create a 'unit testing' approach towards compliance

test one thing,
but test it well



"Queryable Provenance Metadata for GDPR Compliance" at SEMANTiCS 2018
Presented by: Harshvardhan J. Pandit
<http://openscience.adaptcentre.ie/> | pandith@tcd.ie | [@coolharsh55](https://twitter.com/coolharsh55)



I'm a SPARQL endpoint, Query me!

www.adaptcentre.ie

! How extensible
is this approach?
very!

Q what do you
think about
using non
semantic web
tools → NO THANKS

! CAN WE
COLLABORATE
YES ☺

Q can GDPR compliance
be checked in this
automated manner
SOME OF IT

! did a
lawyer
check this?
(I DON'T LIKE YOU!)

How do you
define the
metadata for
the queries

JUST MODEL THE SYSTEM AS RDF

! Is the
demo
* really *
online?

Ja !

!!!!!! maybe you
should use
this approach....
OKAY....

