CNRS - INP - UT3 - UT1 - UT2J
Institut de Recherche en Informatique de Toulouse

How can semantic metadata improve data sharing in Law?

Nathalie Aussenac-Gilles (CNRS) – aussenac@irit.fr

+ collaboraters from MELODI (J. Bouché-Pillon, C. Trojahn dos Santos (UT2J), M. Kamel (U Perpignan) , ADRIA (P. Zaraté (UT1C)) and LiLAC (Y. Chevalier (UT3)

Data and law

- Law produces more and more data in digital form
  - Legal texts, regulations, case reports …
  - Statistics and figures about cases and law decisions
  - …
- Lawyers and law decisions reuse more and more digital data
  - Videos, pictures or text from social media, web sites
  - Data produced by suspects or about suspects
  - Used as supports for investigations, proofs …
  - Perspectives for prevention of dangerous actions / events
- Law provides regulations and guidance about data accessibility, sharing, processing, …
  - RGPD, European law about the use of AI
  - AI act

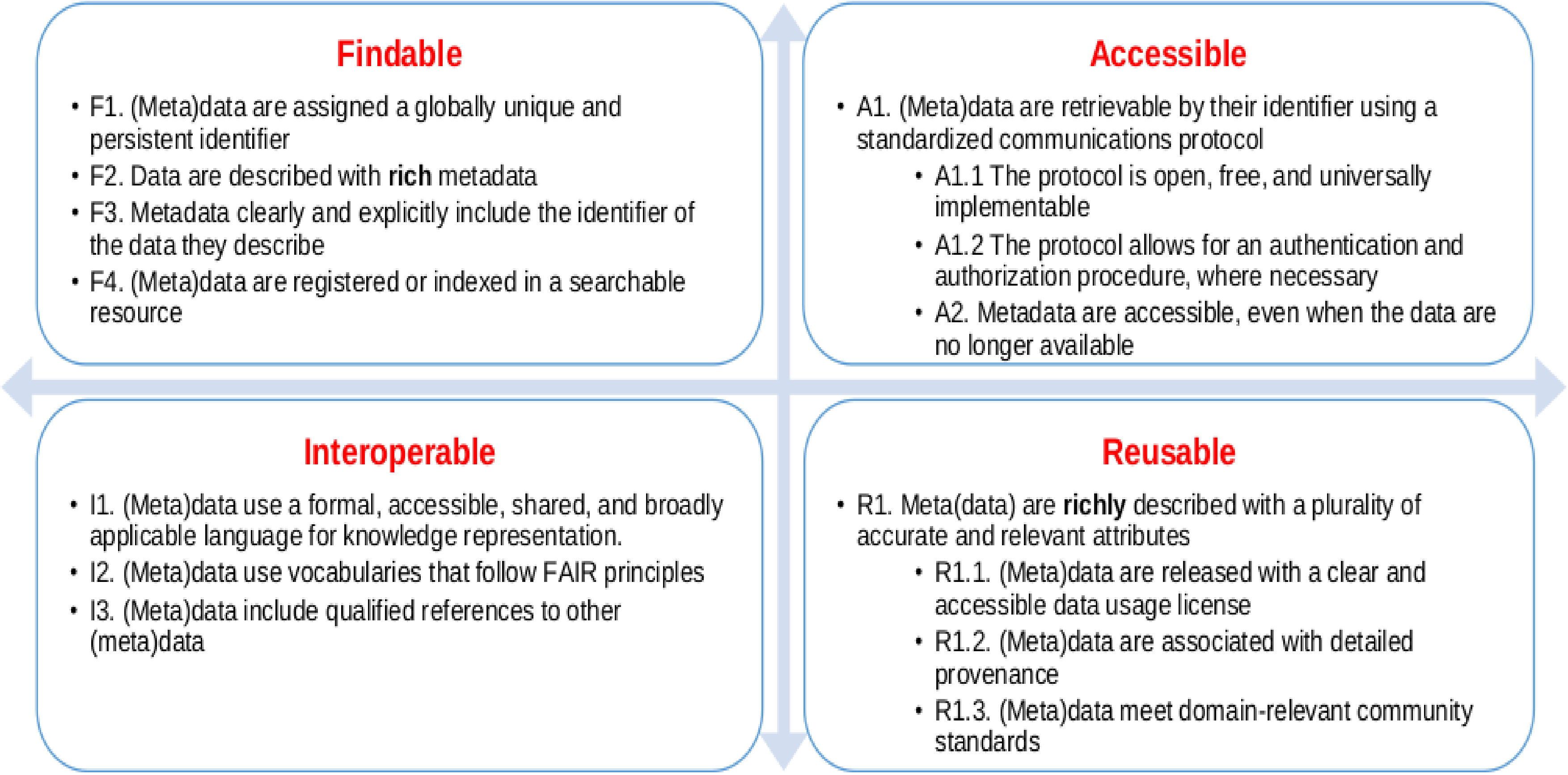
Issues in sharing datasets

- Finding the right dataset for the proper use
- Accessing to this dataset
- Being able to open, manage and operate the dataset
- Actually use the dataset

Issues in sharing datasets:
a first analysis: a technical point of view

- Finding the right dataset for the proper use
  - Add metadata, the richer the better
  - Standardize metadata vocabularies and their values
- Accessing to this dataset (how to download it?)
  - Standard protocols
- Being able to open, manage and operate the dataset
  - Use standard and open format to store the datasets
  - Use standard vocabularies to decribe and type the data
- Actually use the dataset
  - Inform about licences and access rights
  - Provide detailed descriptions of the data

The FAIR principles for data sharing



Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with **rich** metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorization procedure, where necessary
  - A2. Metadata are accessible, even when the data are no longer available

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

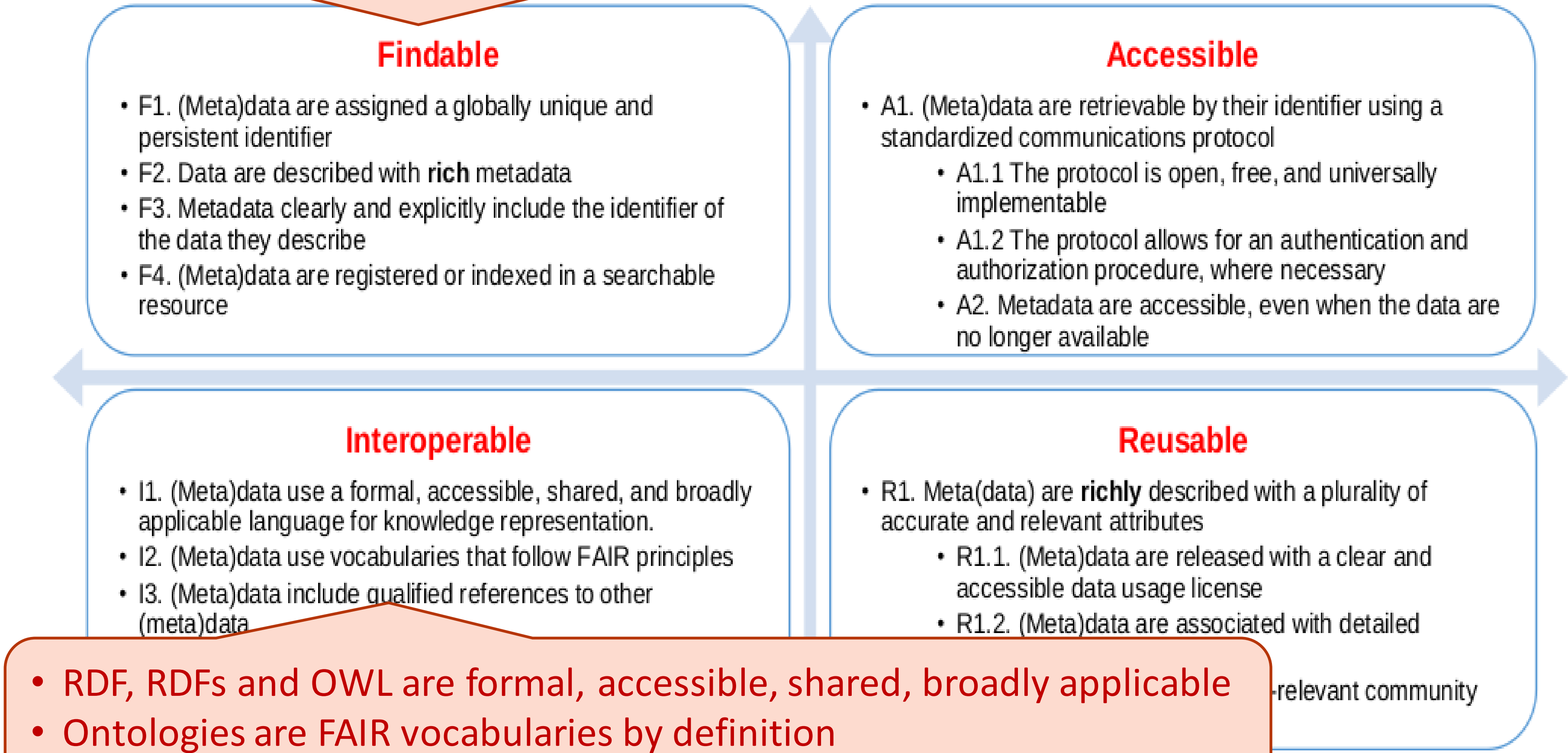
Reusable

- R1. Meta(data) are **richly** described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

Wilkinson, M., Dumontier, M., *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

How can ontologies and formal vocabularies help?

- Machine readable metadata with identifiers = semantic metadata
- RDF, RDFs and OWL can be used to represent metadata



Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with **rich** metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorization procedure, where necessary
  - A2. Metadata are accessible, even when the data are no longer available

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Reusable

- R1. Meta(data) are **richly** described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed

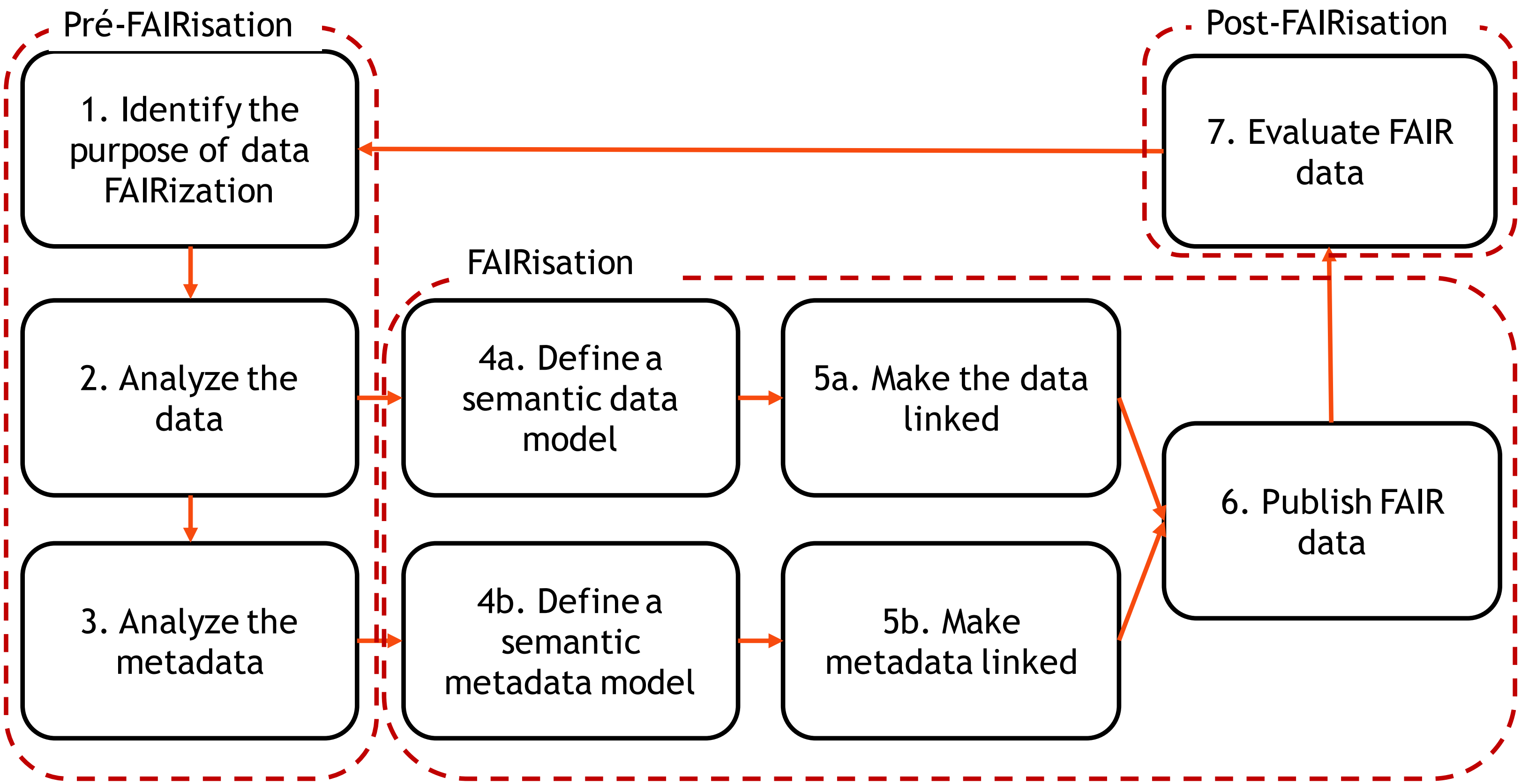
-relevant community

- RDF, RDFs and OWL are formal, accessible, shared, broadly applicable
- Ontologies are FAIR vocabularies by definition
- Linked Open Data refer to each other

How can ontologies and formal vocabularies help?

- Ontologies provide classes to define types of metadata and properties to describe them precisely
- Examples of widely reused ontologies for metadata
  - Dc-term: author, format, version, institution, licence
  - DCAT: catalogue metadata for datasets
  - Prov-o: provenance ontology
  - SSN: sensor data ontology
  - …
- Formal vocabularies and knowledge bases provide formal representations of entities that can used as metadata values
  - Geonames or OpenStreetMap-data for locations
  - Use identifiers for authors (i.e. ORCID for researchers)
  - Use wikidata for famous people, monuments, historical periods, events …
  - Use medical ontologies for names of diseases, anatomy, …
  - …

FAIRisation process (Jacobsen et al., 2020)



A success story: the BNF dataset

https://data.bnf.fr/semanticweb

- BNF = French National Library
- Each document is described with sematic metadata
- BNF-onto: a unique metadata schema for the entire collection https://data.bnf.fr/ontology/bnf-onto/
- This schema is made of various standard ontologies, vocabularies
  - Dc-term, RDA-registry, Foaf and SKOS
  - DBPedia, GeoNames, Ign, …
- It is compliant with international library norms
  - IFLA-LRM Library Reference model

National and international initiatives

- RDA working groups
  - INRAE as French leader for the French chapter
- Recherche.Data.gouv
  - French portal for research datasets
- EOSC supported projects
  - Adoption of the CKAN standard
  - FAIRimpact project
- FAIRsFAIR

Some limitations of the FAIR principles

- No reference to law… or proposed
  - Each data portal or catalogue can have its own schema
  - Web catalogues of datasets that collect metadata from dataset repositories or from other reference catalogues must translate original metadata into their own schema
- No information is asked about the dataset structure and the precise location of each data in the storage
- Very technical view … but semantics means more
  - Metadata are added by domain experts using domain concepts > which accessibility to non-experts ?
  - Metadata should be rich: definitions in natural language, labels in various languages, …

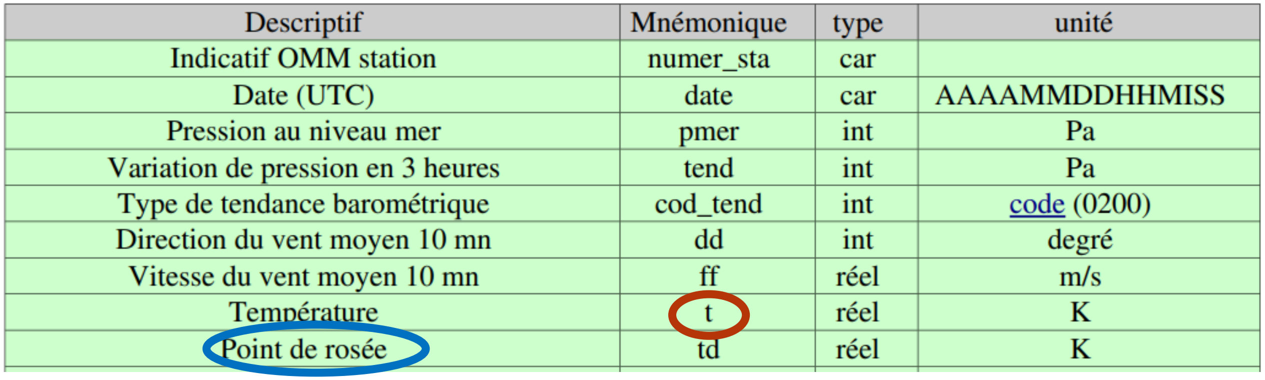
SEMANTICS 4 FAIR

- Goals
  - Reduce the gap between data users and data producers
  - Application to the METEO-France data portal
  - Make METEO-France datasets FAIR

- Methodology
  - Use semantic models
  - Reuse existing ontologies and vocabularies
  - Build a rich core ontology usable in any domain
  - Describe the dataset structure (starting with tabular data)
  - Make this ontology adaptable to knowledge domains
  - Make it easy to describe datasets with this ontology

Example: the SYNOP dataset

PDF file that explains the content of a table



Descriptif	Mnémonique	type	unité
Indicatif OMM station	numer_sta	car	
Date (UTC)	date	car	AAAAMMDDHHMISS
Pression au niveau mer	pmer	int	Pa
Variation de pression en 3 heures	tend	int	Pa
Type de tendance barométrique	cod_tend	int	code (0200)
Direction du vent moyen 10 mn	dd	int	degré
Vitesse du vent moyen 10 mn	ff	réel	m/s
Température	t	réel	K
Point de rosée	td	réel	K

numer_sta	date	pmer	ff	t	…
7005	20200201000000	100710	3.200000	285.450000	…
7015	20200201000000	100710	7.700000	284.950000	…
7020	20200201000000	100630	8.400000	284.150000	…
7027	20200201000000	100770	5.500000	285.650000	…
…	…	…	…	…	…

Extract of a table in the dataset

- Technical terms, acronyms, no definition
- No schema provided with the table
- The dataset is not self contained: the pdf file is « somewhere » on the web portal

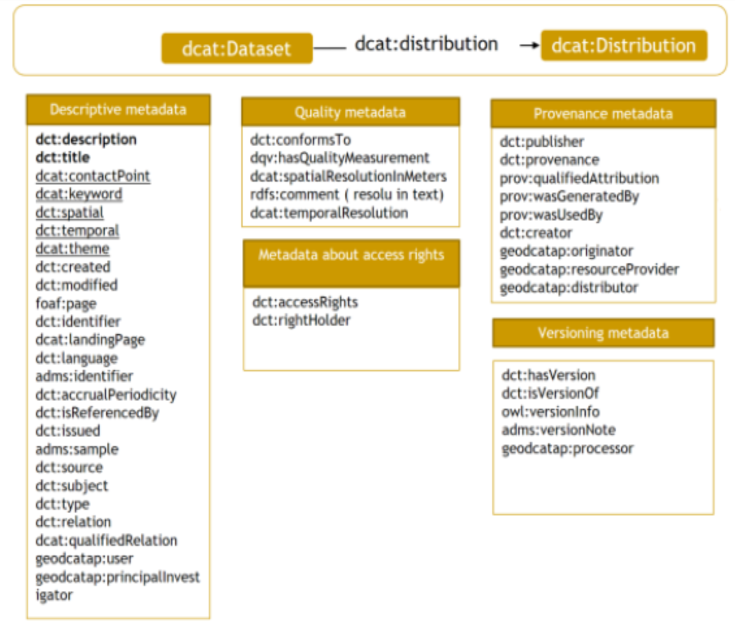
The Dataset Metadata Ontology (DMO-core) for tabular data

- A core model for representing both descriptive metadata and the internal structure of a dataset.
- Relies on existing FAIR vocabularies and ontologies and is itself compliant with the FAIR principles.
- dmo-core can be instanciated with domain-specific entities and definitions to provide domain understanding for data consumers

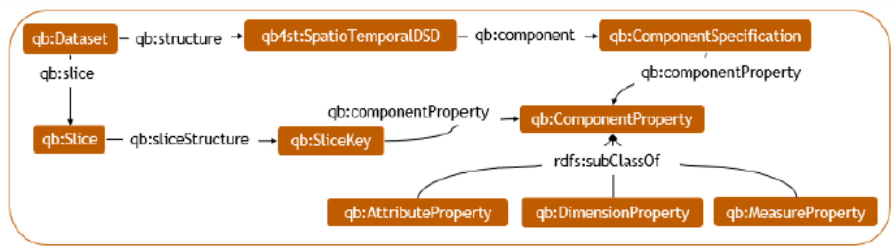
http://w3id.org/dmo

DMO-core: reused ontologies

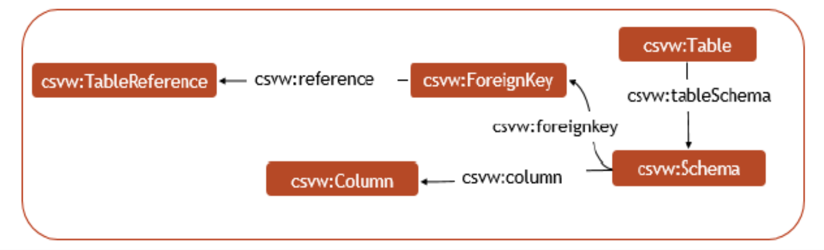
- DCAT: metadata for dataset documentation



- QB (RDF Data Cube): metadata to describe dataset structure



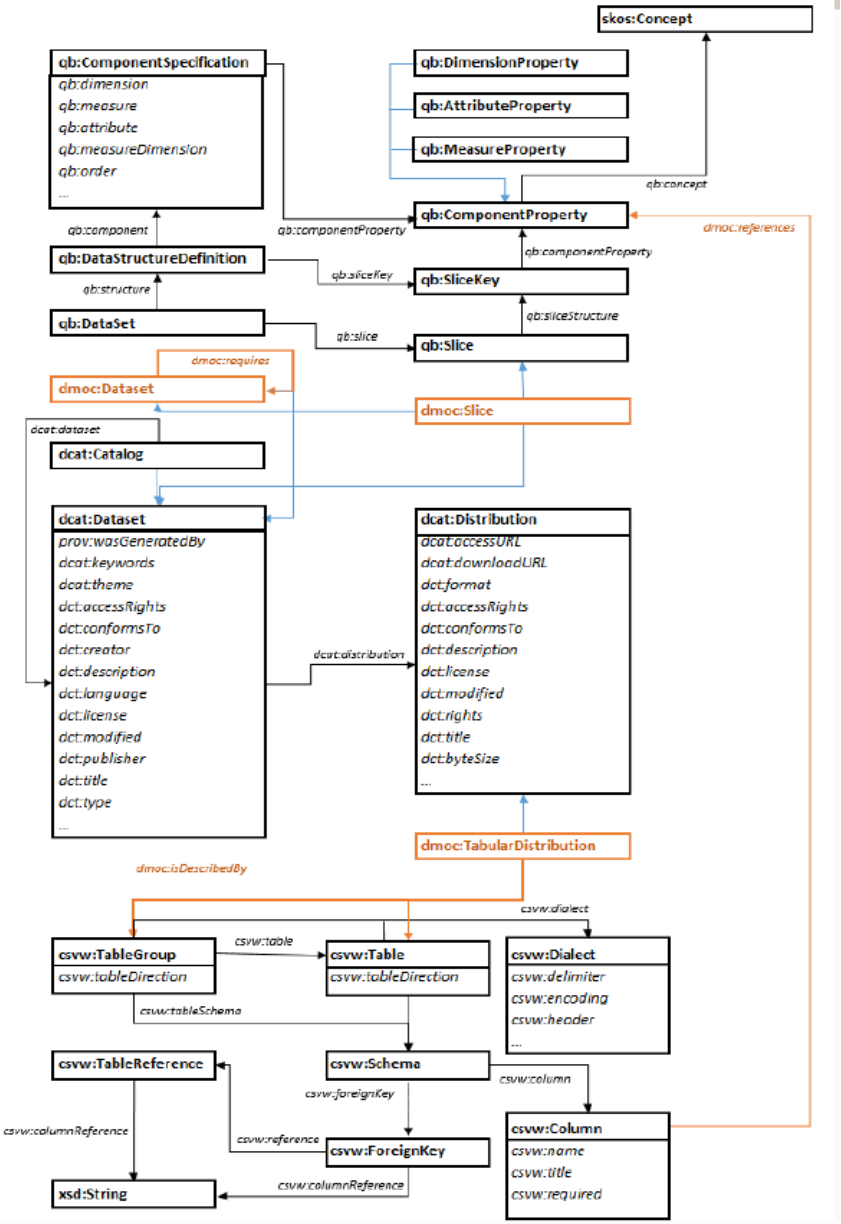
- CSVW: metadata for table description



The DMO-core ontology

Dmoc main concepts:

- dmoc:Dataset
  a qb:Dataset,
  a dcat:Dataset
- dmoc:Slice
  a qb:Slice,
  a dmoc:Dataset
- dmoc:TabularDistribution
  a dcat:Distribution



skos:Concept

qb:ComponentSpecification
qb:dimension
qb:measure
qb:attribute
qb:measureDimension
qb:order
...

qb:DimensionProperty

qb:AttributeProperty

qb:MeasureProperty

qb:concept

qb:ComponentProperty

dmoc:references

qb:component

qb:componentProperty

qb:DataStructureDefinition

qb:componentProperty

qb:sliceKey

qb:SliceKey

qb:structure

qb:sliceStructure

qb:DataSet

qb:slice

qb:Slice

dmoc:requires

dmoc:Dataset

dmoc:Slice

dcat:dataset

dcat:Catalog

dcat:Dataset
prov:wasGeneratedBy
dcat:keywords
dcat:theme
dct:accessRights
dct:conformsTo
dct:creator
dct:description
dct:language
dct:license
dct:modified
dct:publisher
dct:title
dct:type
...

dcat:distribution

dcat:Distribution
dcat:accessURL
dcat:downloadURL
dct:format
dct:accessRights
dct:conformsTo
dct:description
dct:license
dct:modified
dct:rights
dct:title
dct:byteSize
...

dmoc:TabularDistribution

dmoc:isDescribedBy

csvw:dialect

csvw:table

csvw:TableGroup
csvw:tableDirection

csvw:Table
csvw:tableDirection

csvw:Dialect
csvw:delimiter
csvw:encoding
csvw:header
...

csvw:tableSchema

csvw:TableReference

csvw:Schema

csvw:column

csvw:foreignKey

csvw:columnReference

csvw:reference

csvw:ForeignKey

csvw:Column
csvw:name
csvw:title
csvw:required

xsd:String

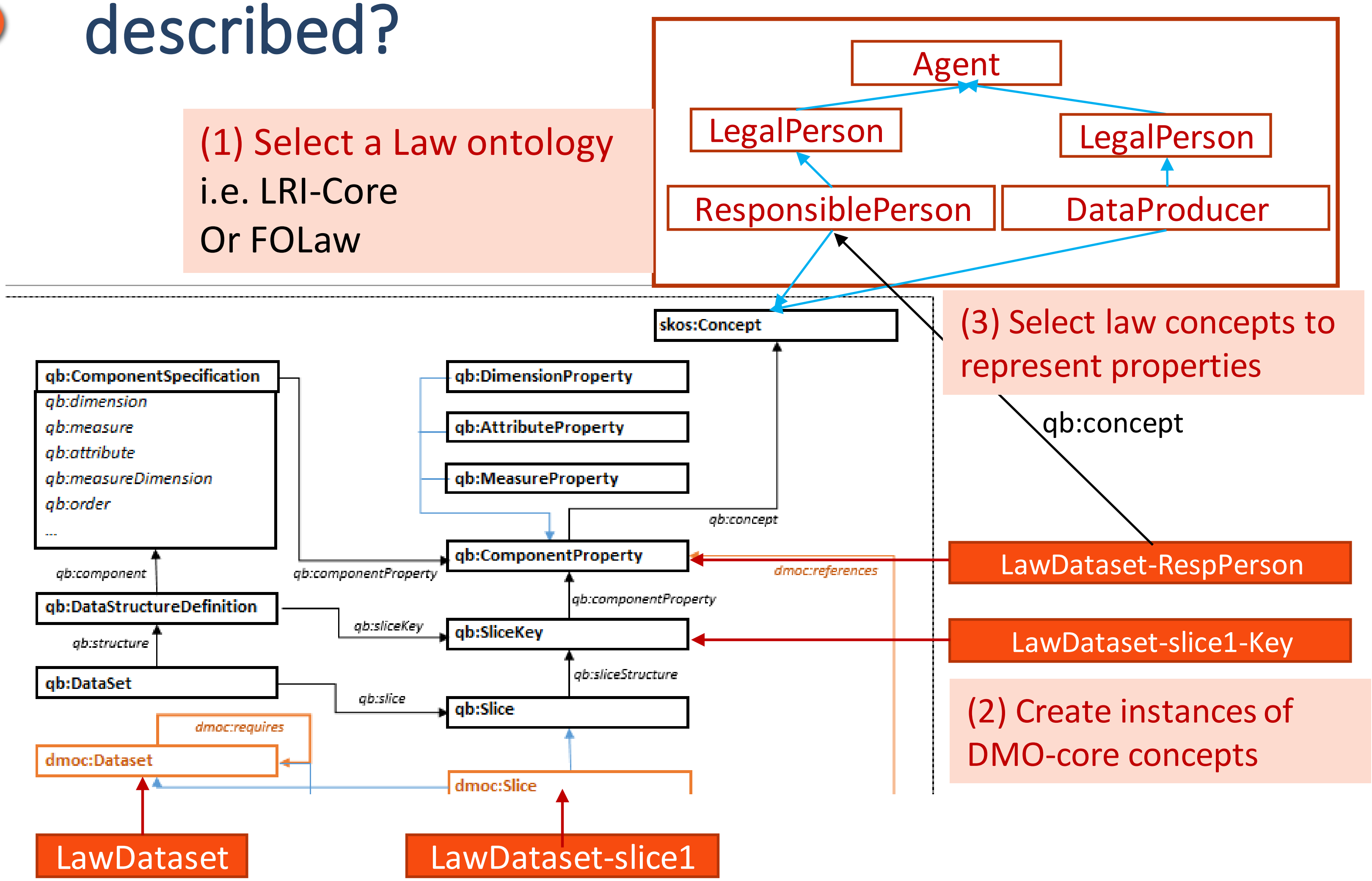
csvw:columnReference

→ subclassOf property

→ DMO-core properties

→ Domain ontology properties

How would a tabular dataset in Law be described?



(1) Select a Law ontology
i.e. LRI-Core
Or FOLaw

(3) Select law concepts to represent properties

(2) Create instances of DMO-core concepts

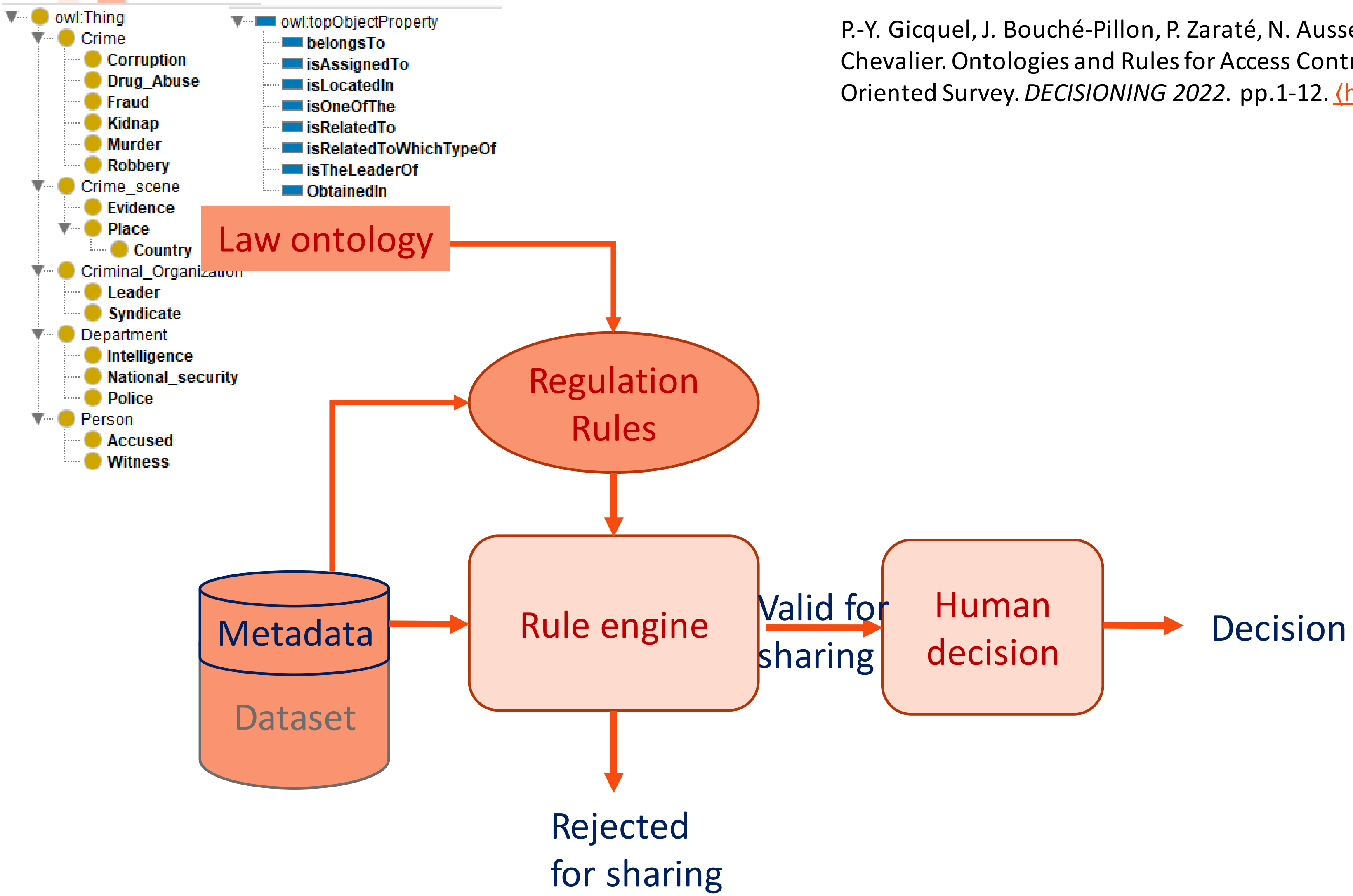
Engers, Tom & Boer, Alexander & Breuker, Joost & Valente, Andre & Winkels, Radboud. (2008). Ontologies in the Legal Domain. 10.1007/978-0-387-71611-4_13.

More limitations of the FAIR principles

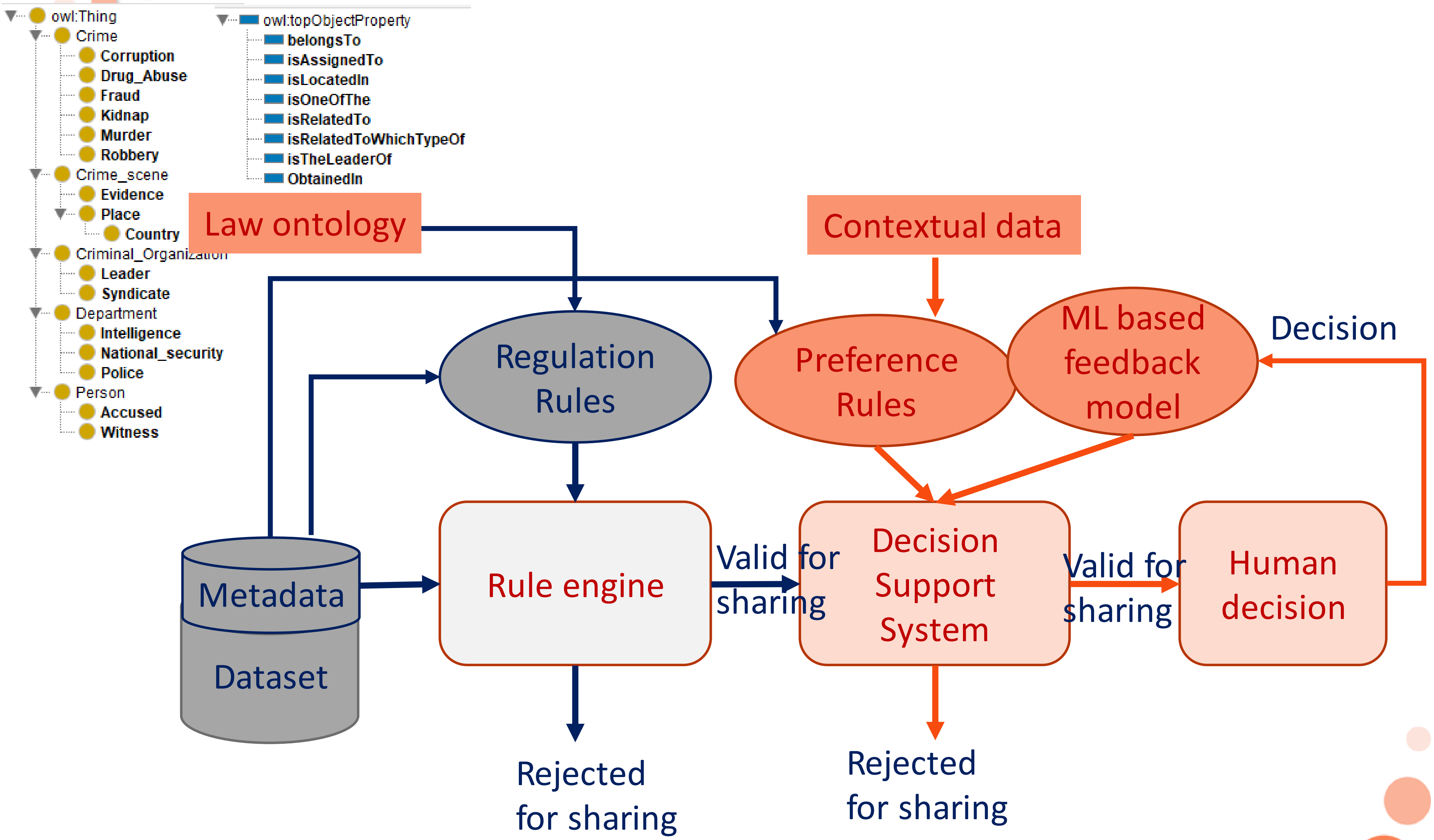
- Very technical view … but sharing raises more issues
- No reference to data sharing regulation / law
  - Regulation about personal data RGPD
  - Regulation about data analytics with AI
- No reference to sharing preferences, or particular sharing conditions according to the target user or use conditions
- Making formal law and preferences
  - Formal access control
  - Semantic rules using a law ontology
  - Representation of regulations and law
  - Representation of sharing preferences

Making formal sharing regulations: Ontology-Based Access Control

P.-Y. Gicquel, J. Bouché-Pillon, P. Zaraté, N. Aussenac-Gilles, Y. Chevalier. Ontologies and Rules for Access Control: a Feature Oriented Survey. *DECISIONING 2022*. pp.1-12. ⟨hal-03762626⟩



Making formal sharing regulations and preferences: an adaptive access control



- owl:Thing
  - Crime
    - Corruption
    - Drug_Abuse
    - Fraud
    - Kidnap
    - Murder
    - Robbery
  - Crime_scene
    - Evidence
    - Place
      - Country
  - Criminal_Organization
    - Leader
    - Syndicate
  - Department
    - Intelligence
    - National_security
    - Police
  - Person
    - Accused
    - Witness

- owl:topObjectProperty
  - belongsTo
  - isAssignedTo
  - isLocatedIn
  - isOneOfThe
  - isRelatedTo
  - isRelatedToWhichTypeOf
  - isTheLeaderOf
  - ObtainedIn

The road is open for new research

- Semantic metada can improve (law) dataset FAIRness
- Law ontologies, formal metadata and decision support systems can help to implement regulations and preferences about data sharing languages

- Investigation lines for the future
  - New types of metadata are required
  - Narrow collaboration between law and AI researchers
  - Build relevant law and context ontologies
  - Integrate case models or user feedback thanks to machine learning
  - Be allowed to access to past decisions