

L'Ingénierie des Connaissances: Web sémantique, web de données, données liées

ANF APSEM2018 : Apprentissage et sémantique

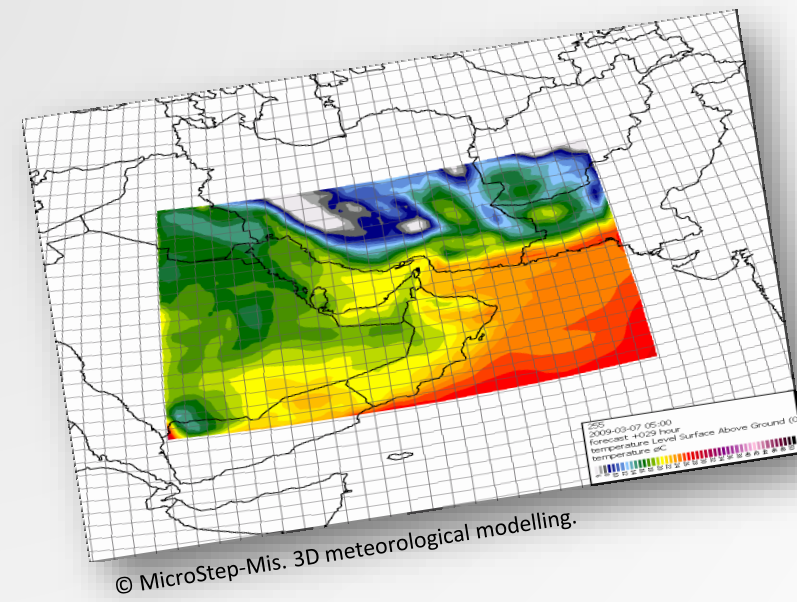
Toulouse, 12-15 novembre 2018

F. Michel

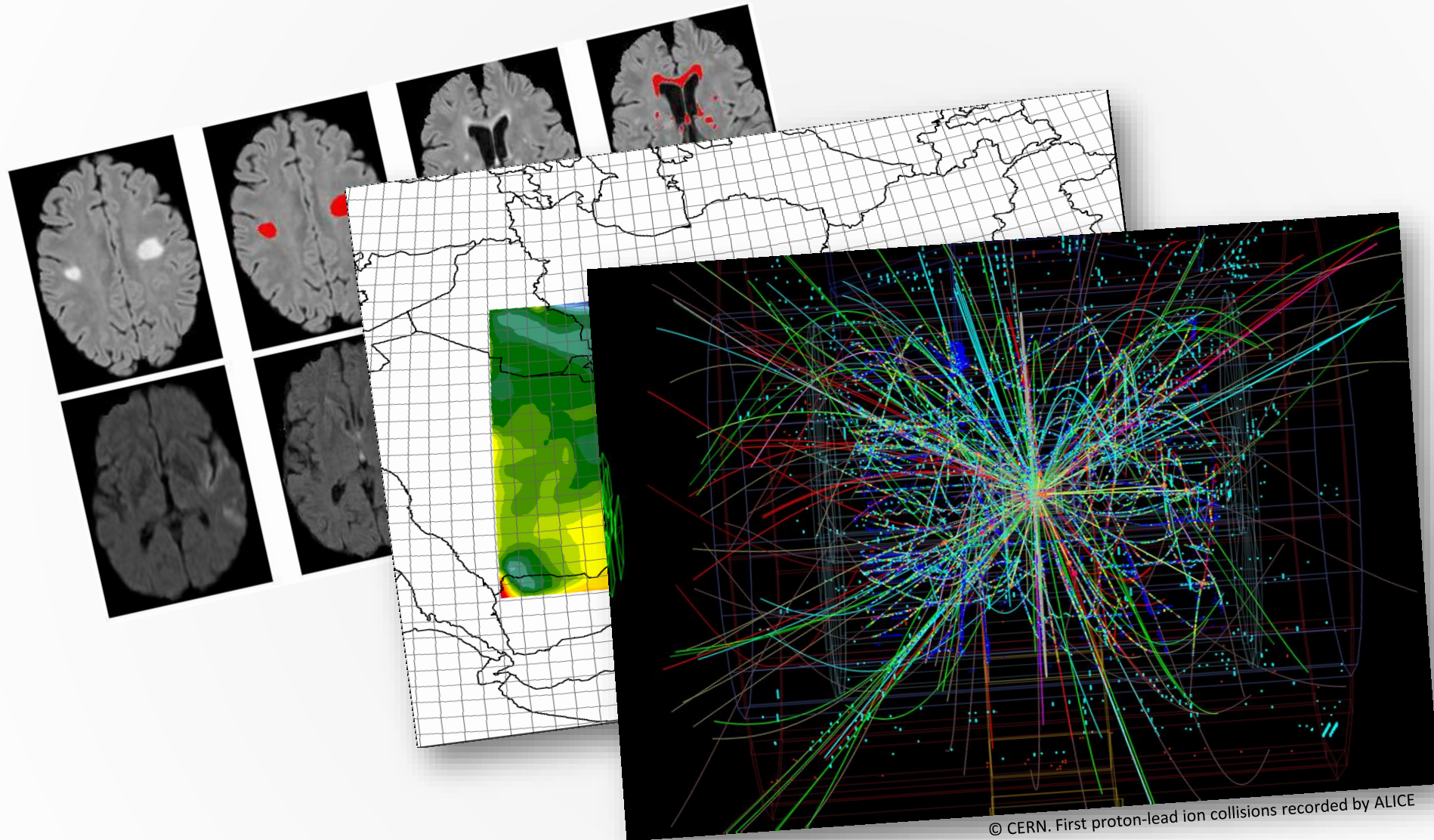
Université Côte d'Azur, CNRS, Inria, I3S, France



More data sources ⇒ More opportunities



To you, your data may mean this...



© CERN. First proton-lead ion collisions recorded by ALICE

Interoperability Challenges

Structural heterogeneity

⇒ Uniform representation format

Semantic heterogeneity

⇒ Controlled vocabularies, thesaurus, ontologies...

Common way to query the data



Agenda



The Semantic Web

Linked Data and the Web of Data

Publishing legacy data in RDF

The Semantic Web





Publish



Link

Vocabularies



*“The **Semantic Web** provides an environment where applications can **publish** and **link** data, define **vocabularies**, **query** data at web scale, and draw **inferences**.”* (adapted from W3C website)

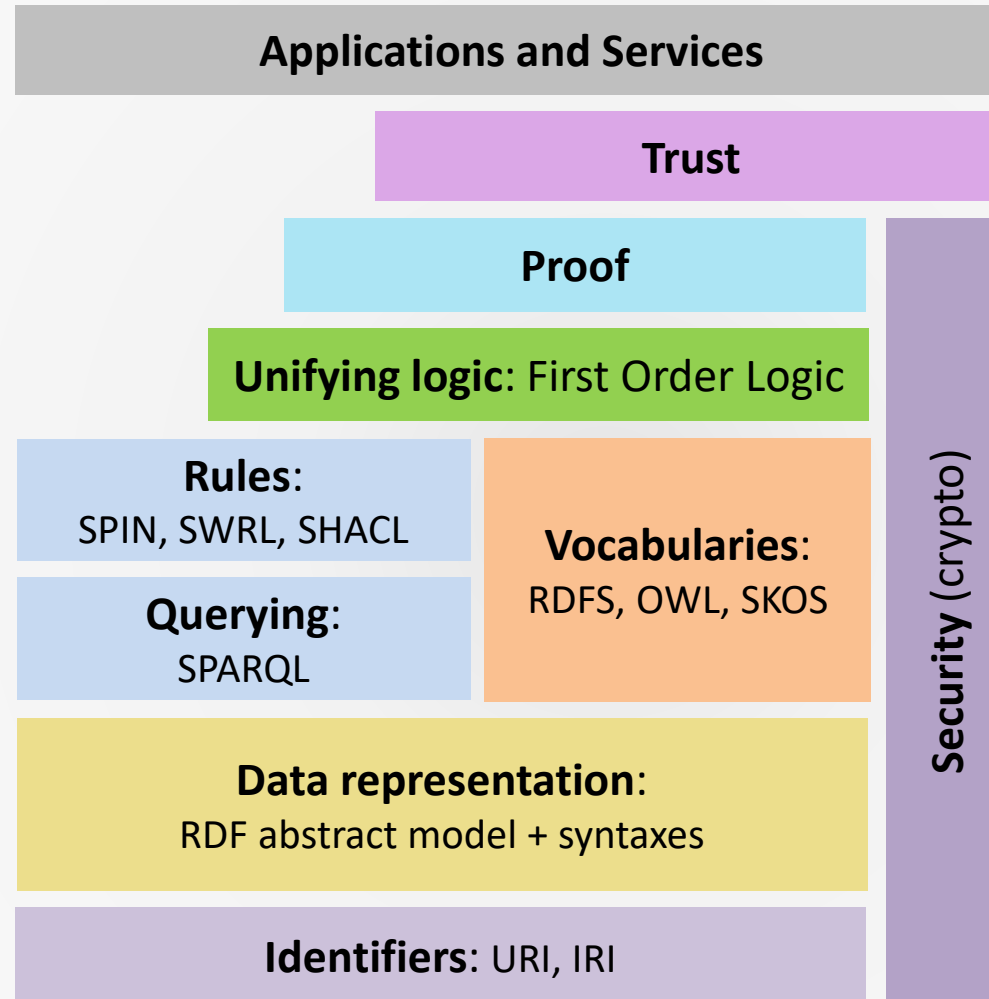


Querying

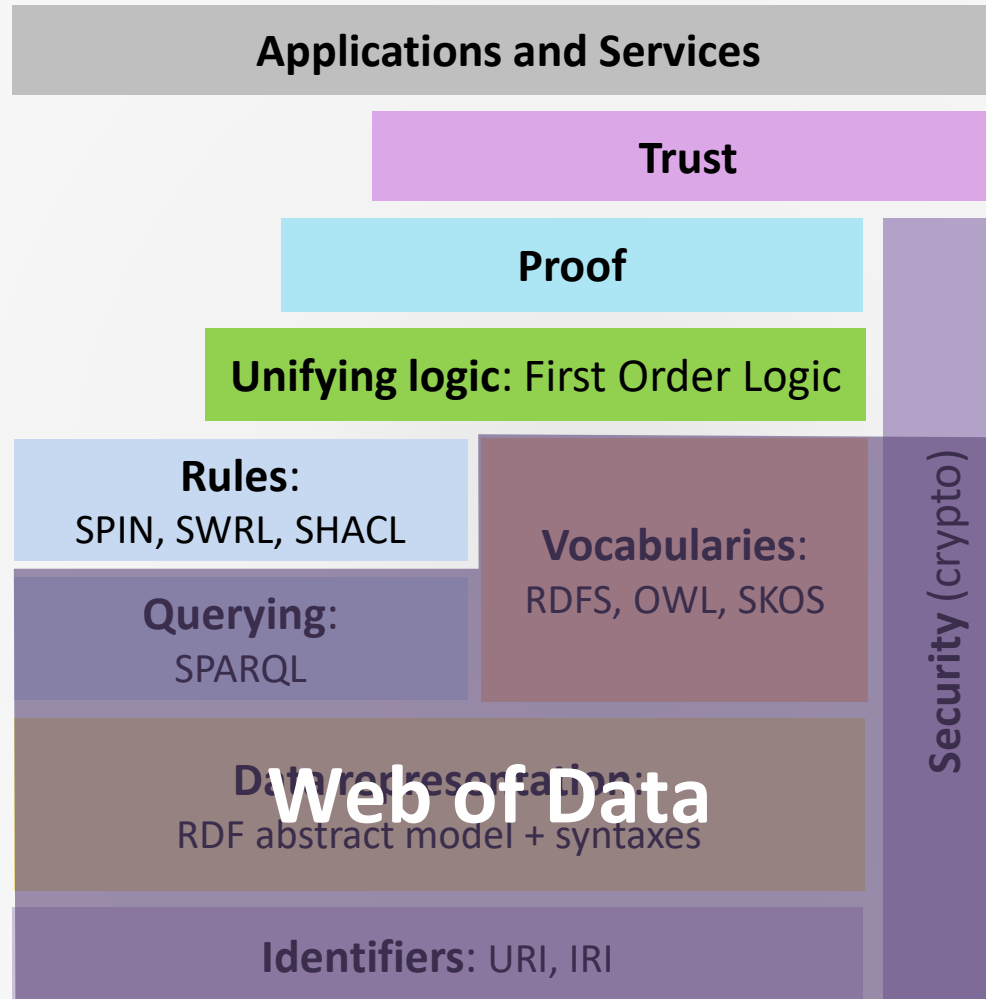


Inference

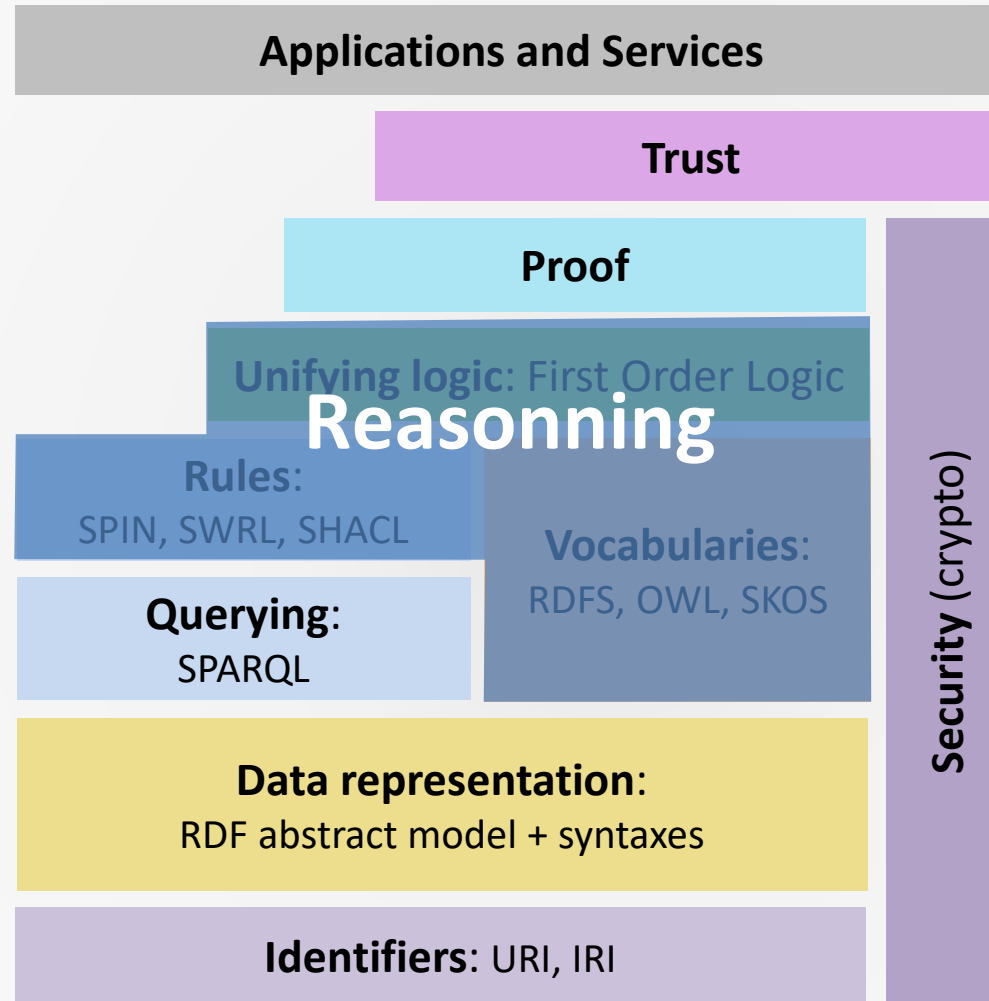
Standards of the Semantic Web



Standards of the Semantic Web



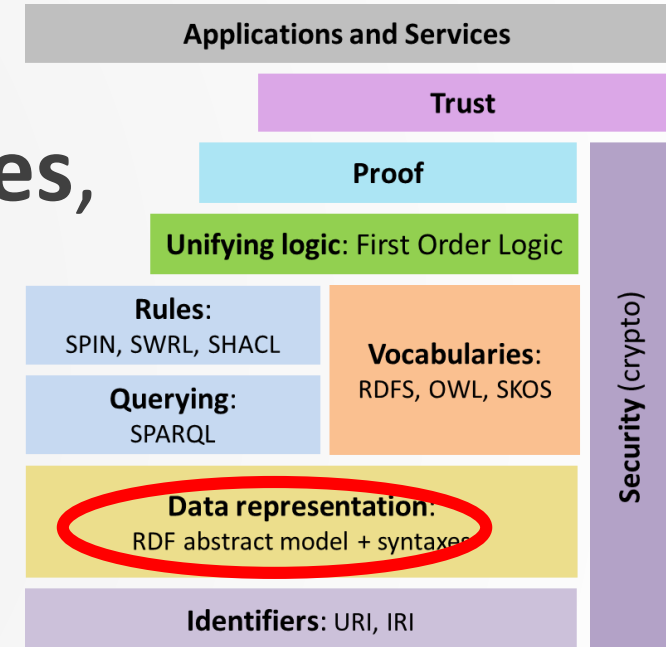
Standards of the Semantic Web



The Resource Description Framework

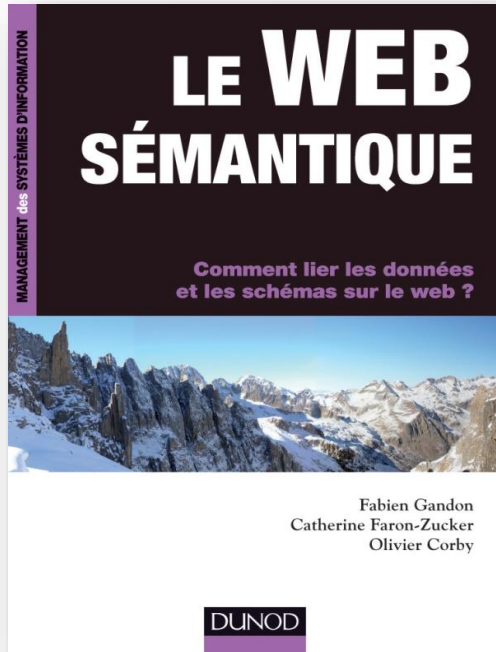


RDF is a conceptual model based on **triples**,
i.e. any fact consists of 3 components:
(**subject**, **predicate**, **object**)



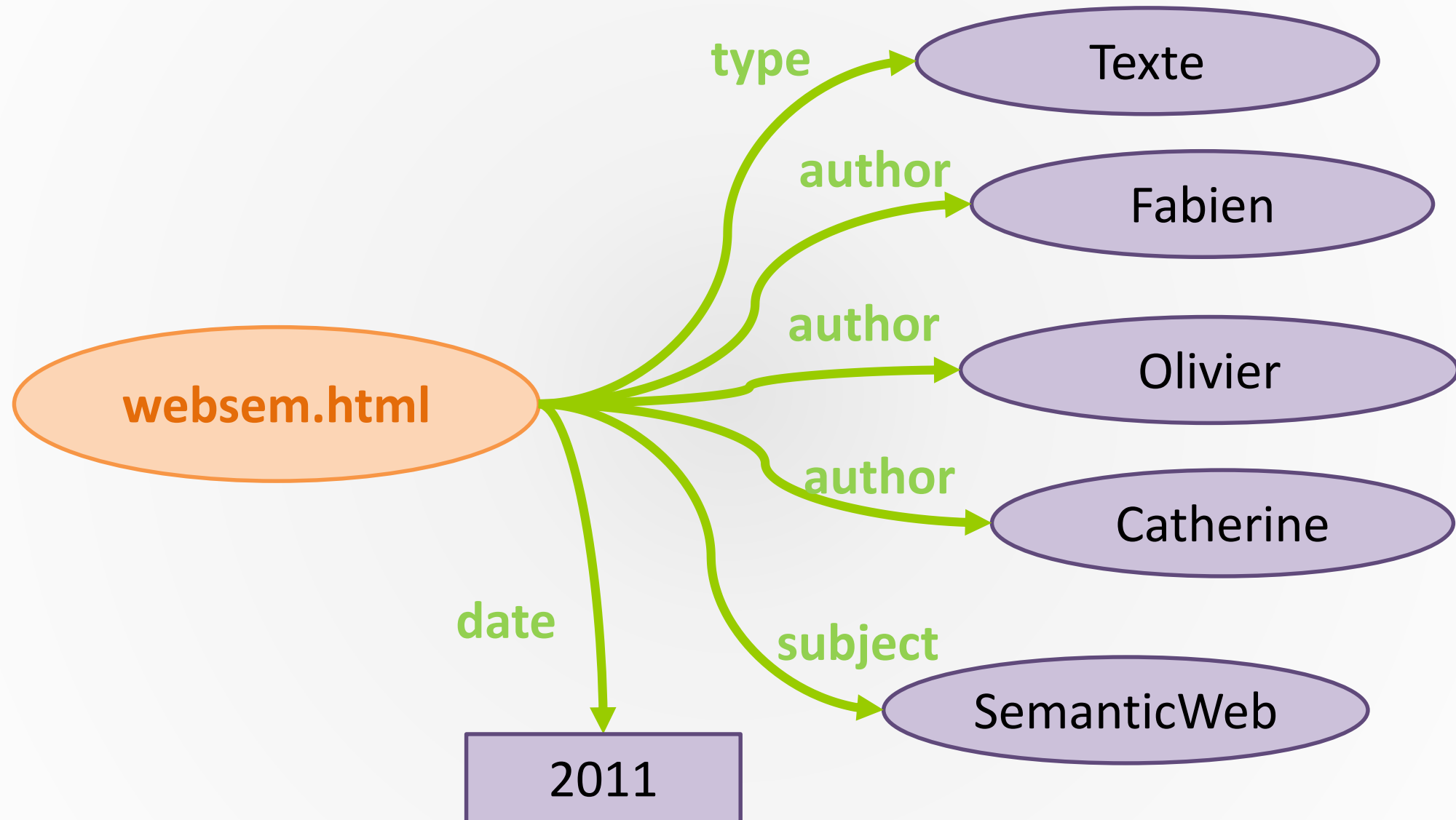
Source: C. Faron Zucker, O. Corby. Introduction au web de données et au web sémantique. Séminaire INRA Open Data Dec. 2014.

The Resource Description Framework

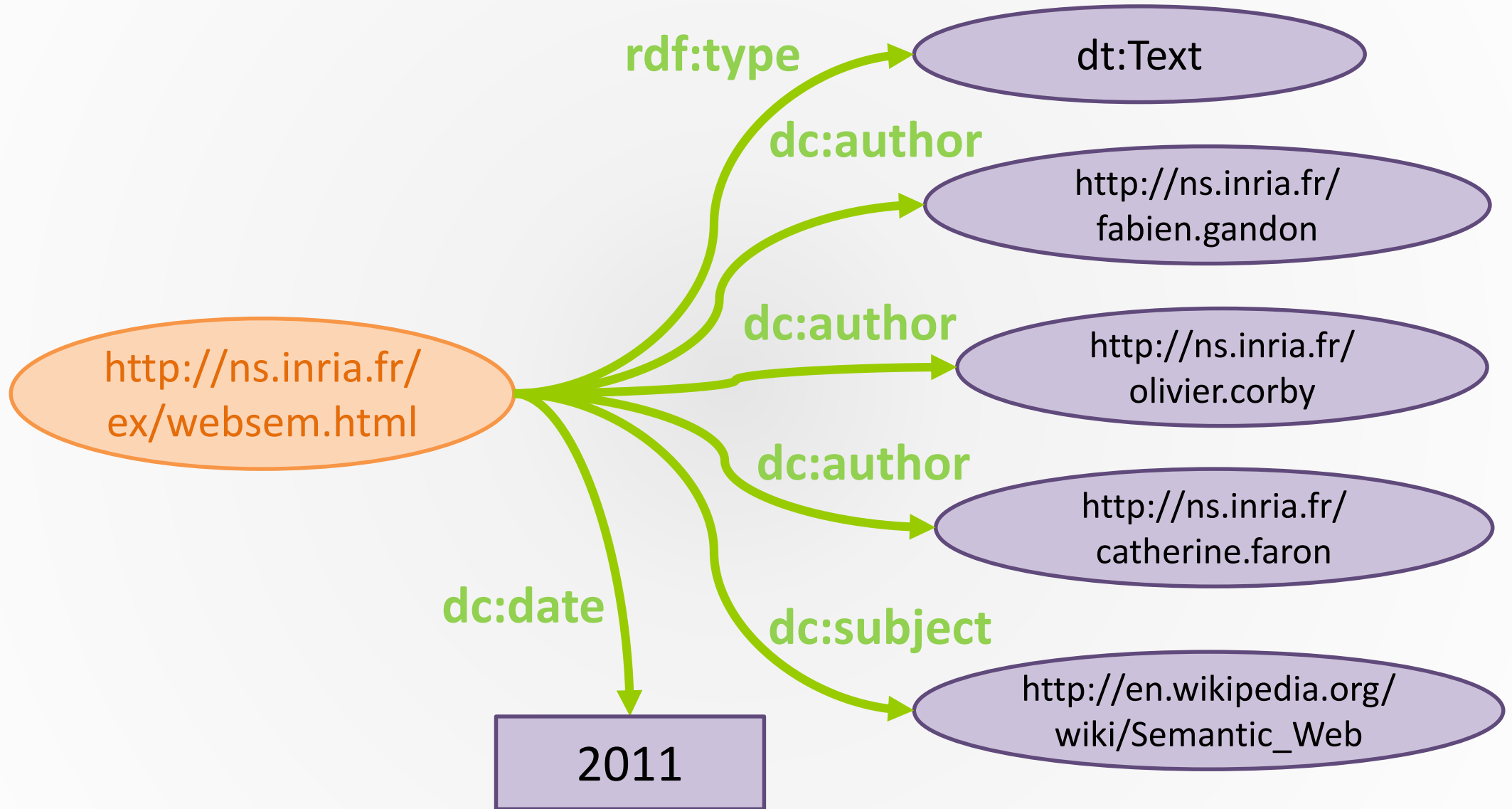


websem.html is a texte
websem.html has as author Fabien
websem.html has as author Olivier
websem.html has as author Catherine
websem.html has as subject Semantic Web
websem.html was written in 2011

The Resource Description Framework



The Resource Description Framework



RDF Syntaxes: N-Triples, Turtle, JSON-LD, Trig, RDF/XML...

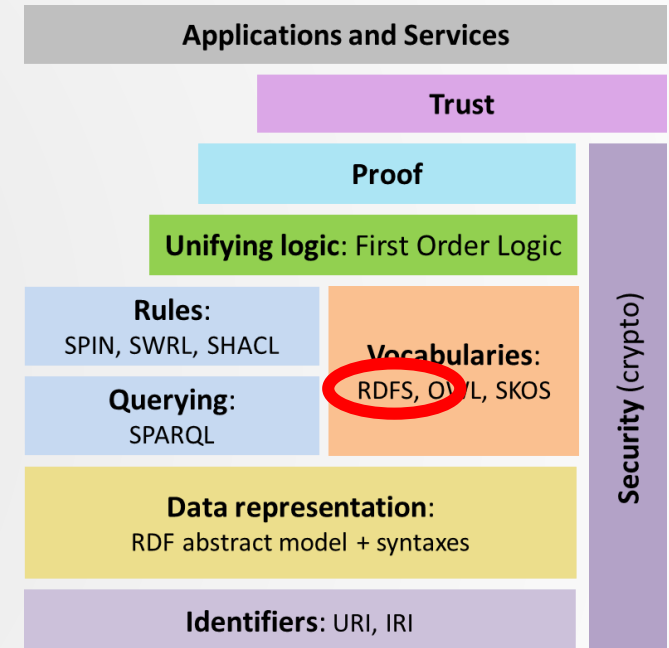
N-Triples

```
<http://inria.fr/ex/websem.html>  
  <http://purl.org/dc/elements/1.1/author>  
  <http://ns.inria.fr/catherine.faron> .  
  
<http://inria.fr/ex/websem.html>  
  <http://purl.org/dc/elements/1.1/theme> "Semantic Web" .
```

Turtle

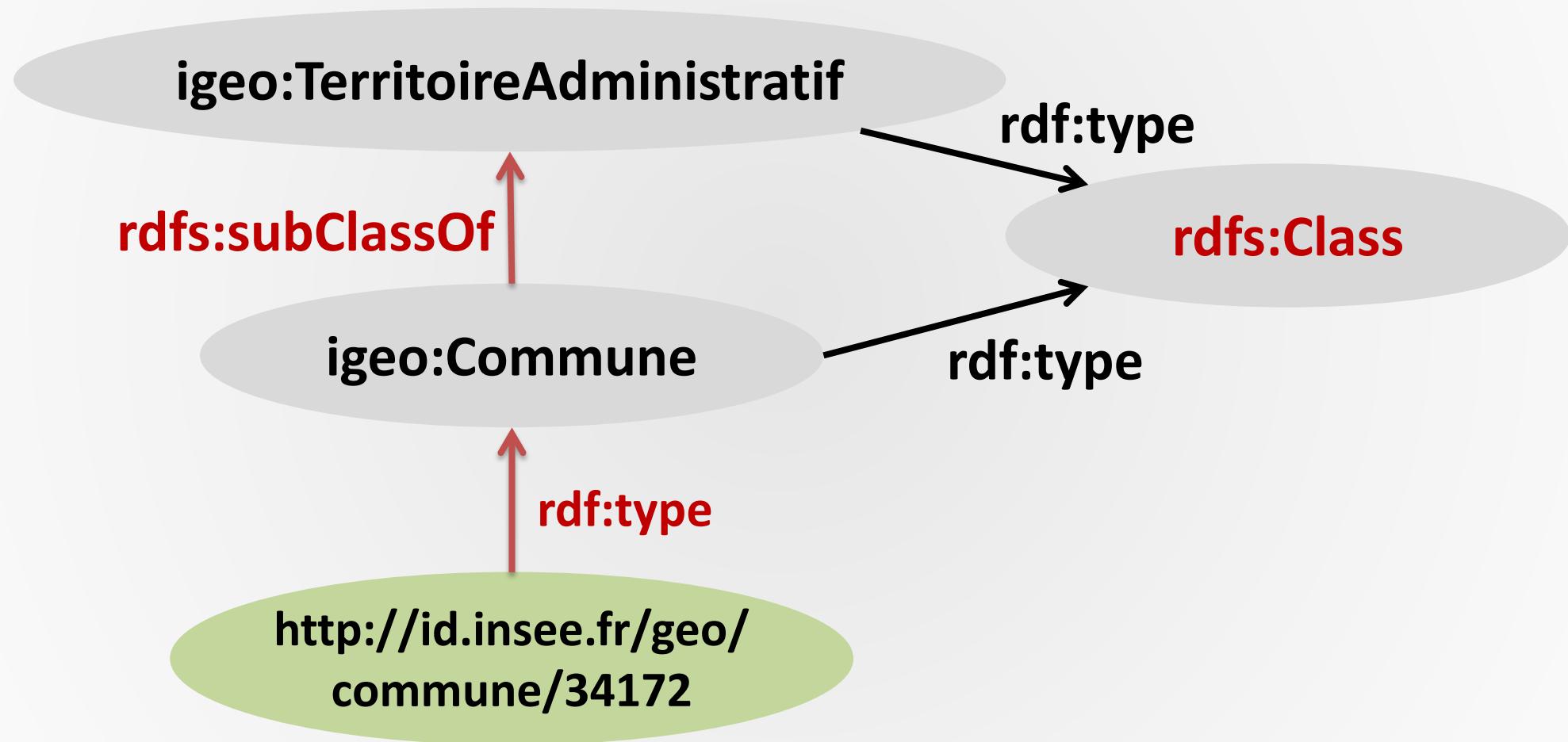
```
@prefix dc: <http://purl.org/dc/elements/1.1/> .  
<http://inria.fr/ex/websem.html>  
  dc:author <http://ns.inria.fr/catherine.faron> ;  
  dc:theme "Semantic Web" .
```

RDF Schemas define
classes of resources,
 their **properties**,
 and organize their **hierarchies**



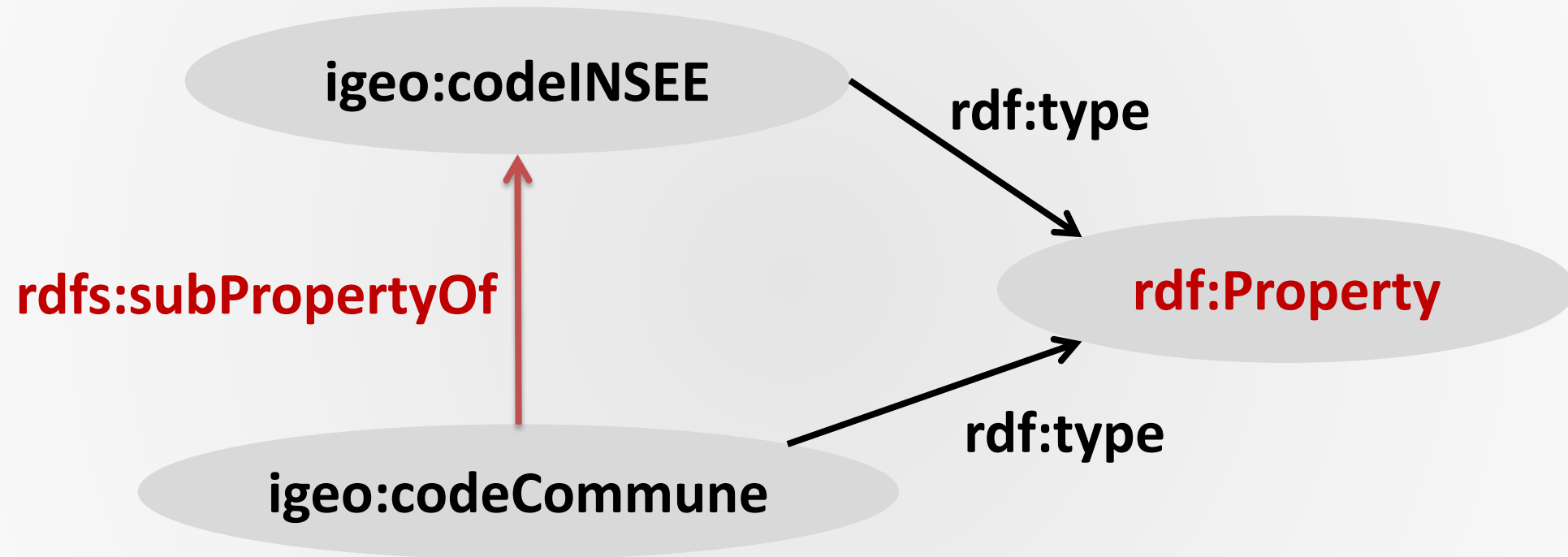
RDF Schema - Classes

```
@prefix igeo: <http://rdf.insee.fr/def/geo#> .
```



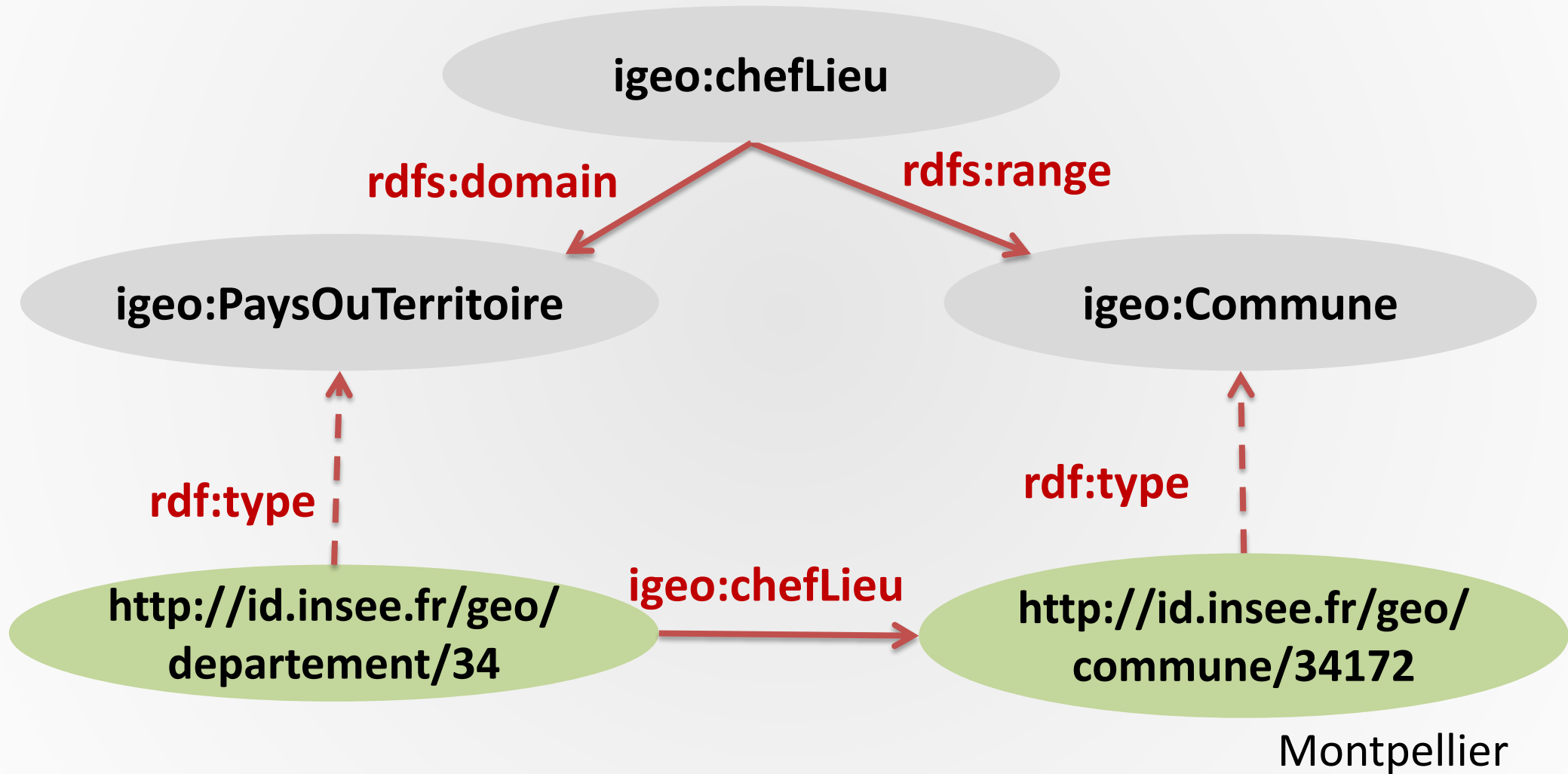
RDF Schema - Properties

```
@prefix igeo: <http://rdf.insee.fr/def/geo#> .
```



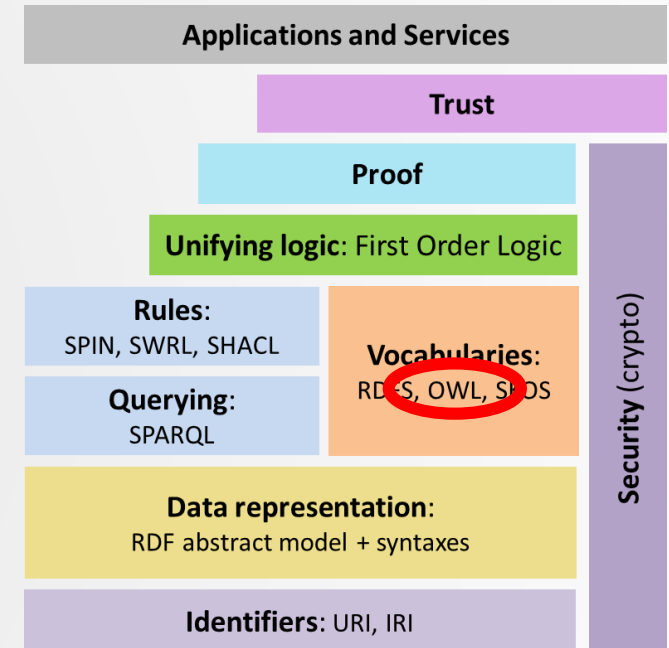
RDF Schema - Properties

```
@prefix igeo: <http://rdf.insee.fr/def/geo#> .
```


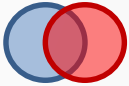
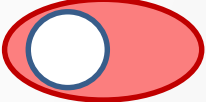




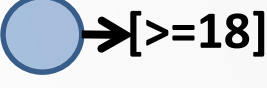






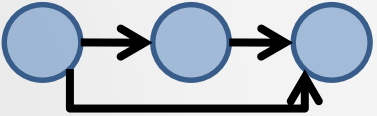



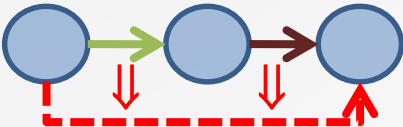
OWL

The Web Ontology Language

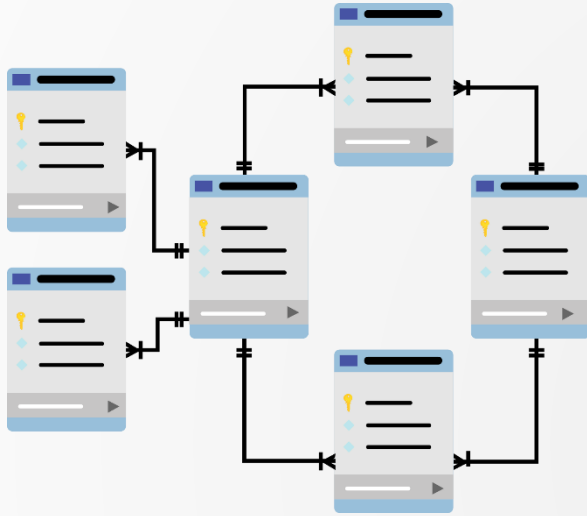


OWL in one slide...

-  def. by union
-  def. by intersection
-  def. by complement
-  def. by restriction
-  def. by cardinality
-  def. by equivalence
-  def. by enumeration
-  def. by value restrict.
-  class disjunction
- ...

-  (a)symmetric prop.
-  (irr)reflexive prop.
-  inverse prop.
-  transitive prop.
-  prop. disjunction
-  cardinality
-  indiv. prop. negation
-  chained prop.

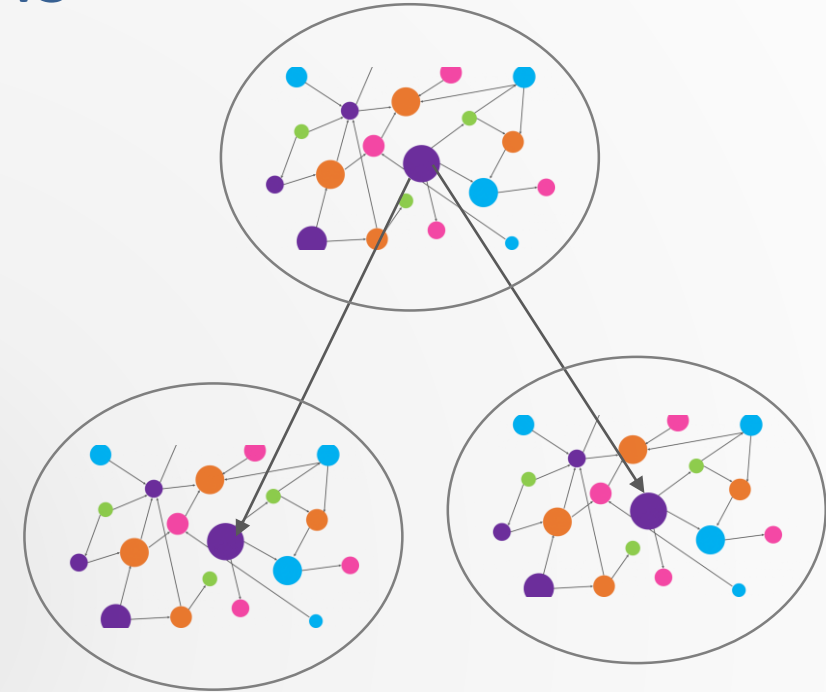
Closed vs. Open Worlds Assumptions



Closed World

Everything there is to know about a thing is stated in a single, closed DB.

- ⇒ Not asserted facts are false, i.e. only asserted facts are true.
- ⇒ A schema may define what can be stated (a schema may be violated).



Open World

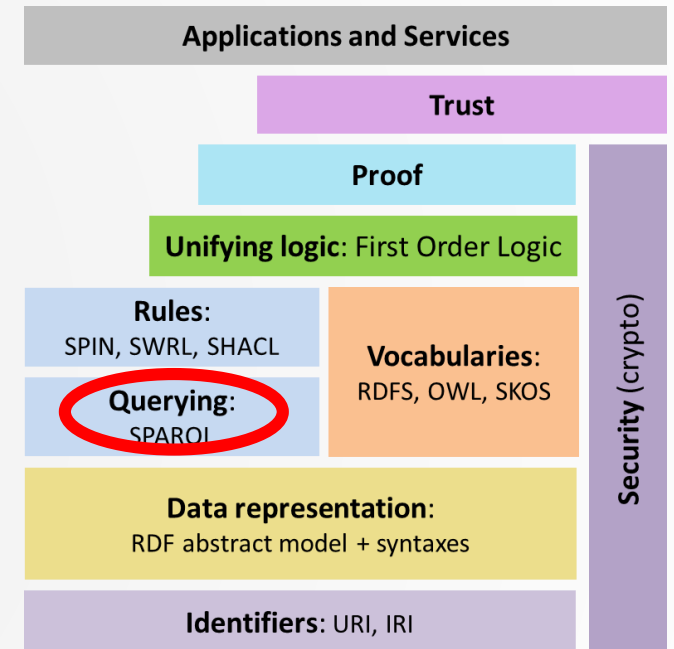
Knowledge is **distributed**.

Each RDF graph may state facts about a thing, irrespective of what others state.

- ⇒ Because a fact is not asserted does not mean it is false.
- ⇒ Every asserted fact is true (no schema)
- ⇒ But some facts may lead to inconsistencies

Querying RDF with SPARQL

SPARQL Protocol and RDF Query Language



SPARQL 1.1 Rec. 21 Mar. 2013

- Query Language (using the Turtle syntax)
 - CRUD operations
- Query results
 - Query Results Format XML, JSON, CSV/TCV
- Protocols
 - SPARQL Protocol
 - SPARQL Graph Store HTTP Protocol
- Entailment Regimes

SPARQL: triple patterns

Turtle syntax with “?” or “\$” to mark **variables**:

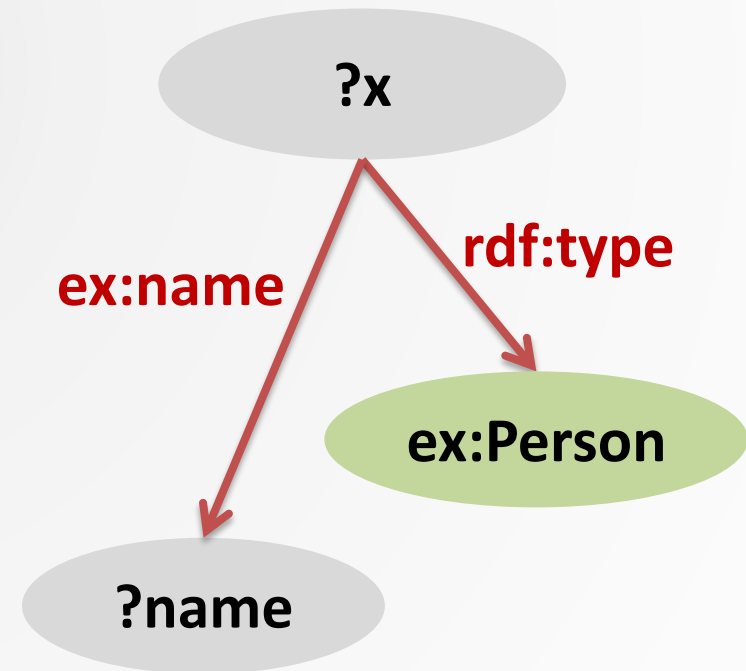
```
?x rdf:type ex:Person
```

Describe **patterns of triples** that we look for:

```
SELECT ?subject ?type  
WHERE { ?subject rdf:type ?type }
```

Default pattern: **conjunction** of triple patterns:

```
SELECT ?x WHERE  
{ ?x rdf:type ex:Person .  
  ?x ex:name ?name . }
```



SPARQL: namespace prefixes

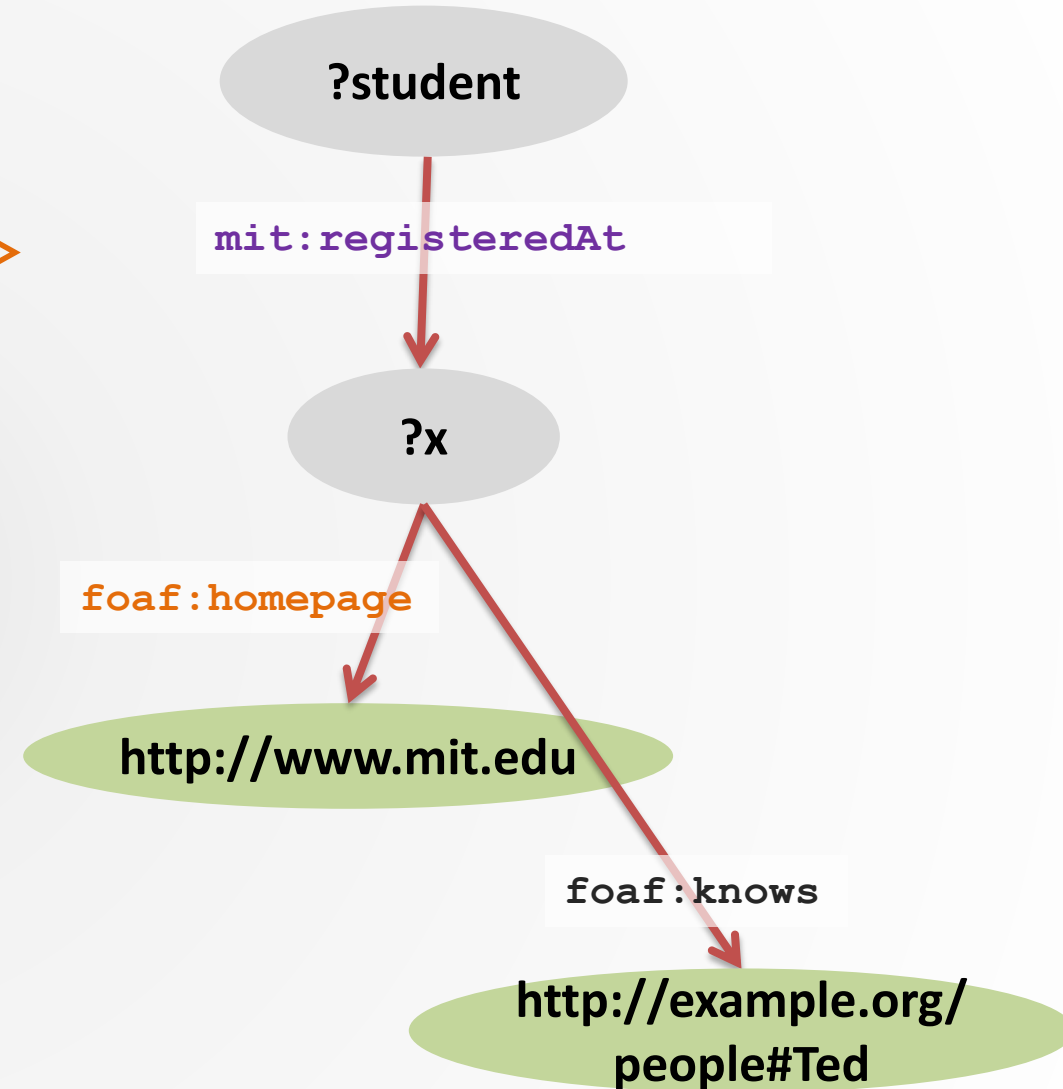
Declare prefixes of used vocabularies:

```
PREFIX mit: <http://www.mit.edu#>  
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

```
SELECT ?student  
WHERE {  
  ?student mit:registeredAt ?x .  
  ?x foaf:homepage <http://www.mit.edu> .  
}
```

Declare a base namespace for relative URIs:

```
BASE <http://example.org/people#>  
SELECT ?student  
WHERE { ?student foaf:knows <Ted> . }
```



SPARQL: language and typed literals

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?x ?f WHERE {
  ?x foaf:name "Steve"@en ; foaf:knows ?f .
}
```

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?x WHERE {
  ?x foaf:name "Steve"@en ;
     foaf:age "21"^^xsd:integer .
}
```

SPARQL: optional pattern

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?person ?name
WHERE {
  ?person foaf:homepage <http://fabien.info> .
  OPTIONAL { ?person foaf:name ?name . }
}
```

⇒ Variable **?name** is potentially **unbound**.

SPARQL alternative pattern

Merge the results of two graph patterns:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?person ?name
WHERE {
  ?person foaf:name ?name .
  { ?person foaf:homepage <http://fabien.info> . }
  UNION
  { ?person foaf:homepage <http://fabien.org> . }
}
```

SPARQL filters

```
PREFIX ex: <http://inria.fr/schema#>
SELECT ?person ?name
WHERE {
  ?person rdf:type ex:Person; ex:name ?name; ex:age ?age .
  FILTER (xsd:integer(?age) >= 18)
}
```

Other examples:

```
FILTER(?name IN ("fabien", "olivier", "catherine"))
```

```
FILTER(langMatches(lang(?name), "en"))
```

SPARQL additional features

- Solution modifiers:

ORDER BY, LIMIT, OFFSET, DISTINCT

- Aggregates

GROUP BY, HAVING

- Negation

NOT EXISTS, MINUS, NOT IN

WHERE { ?x a ex:Person MINUS { ?x foaf:knows ex:John } }

- Nested queries

- Named graphs

- Property paths

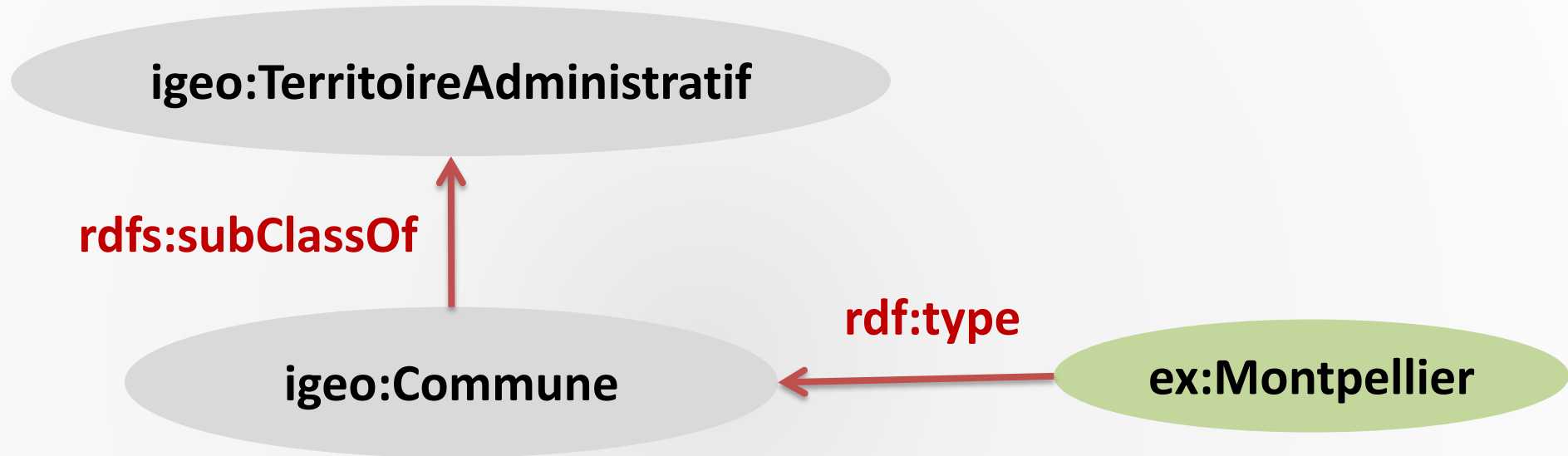
?x foaf:knows+ ?friend .

- ...

SPARQL JSON results

```
{
  "head": { "vars": [ "student" ] },
  "results": {
    "bindings": [
      { "student": {
        "type": "uri",
        "value": "http://www.mit.edu/data.rdf#joe" }
      },
      { "student": {
        "type": "uri",
        "value": "http://www.mit.edu/abcdef" }
      }
    ]
  }
}
```

SPARQL Entailments: infer knowledge

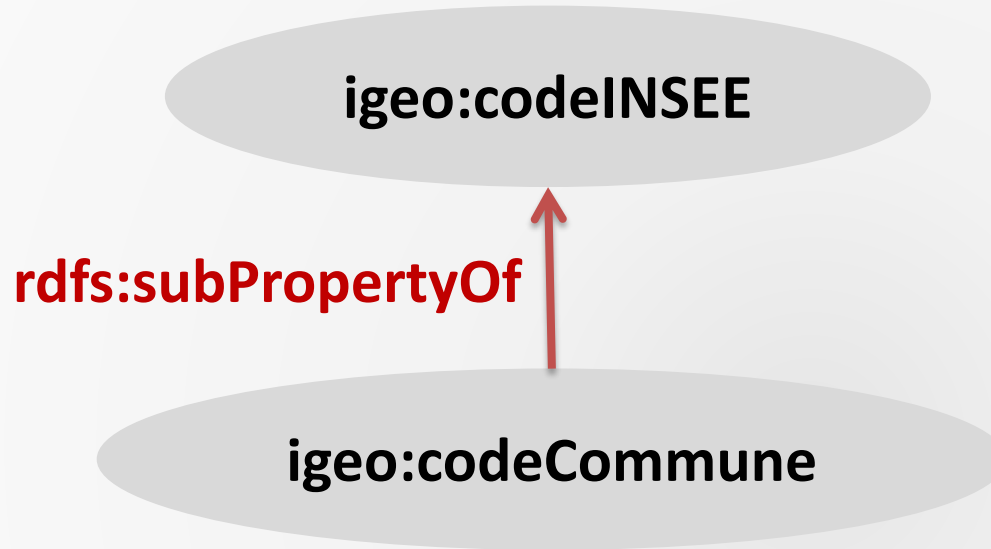


```
PREFIX igeo:<http://rdf.insee.fr/def/geo#>
```

```
SELECT ?x
```

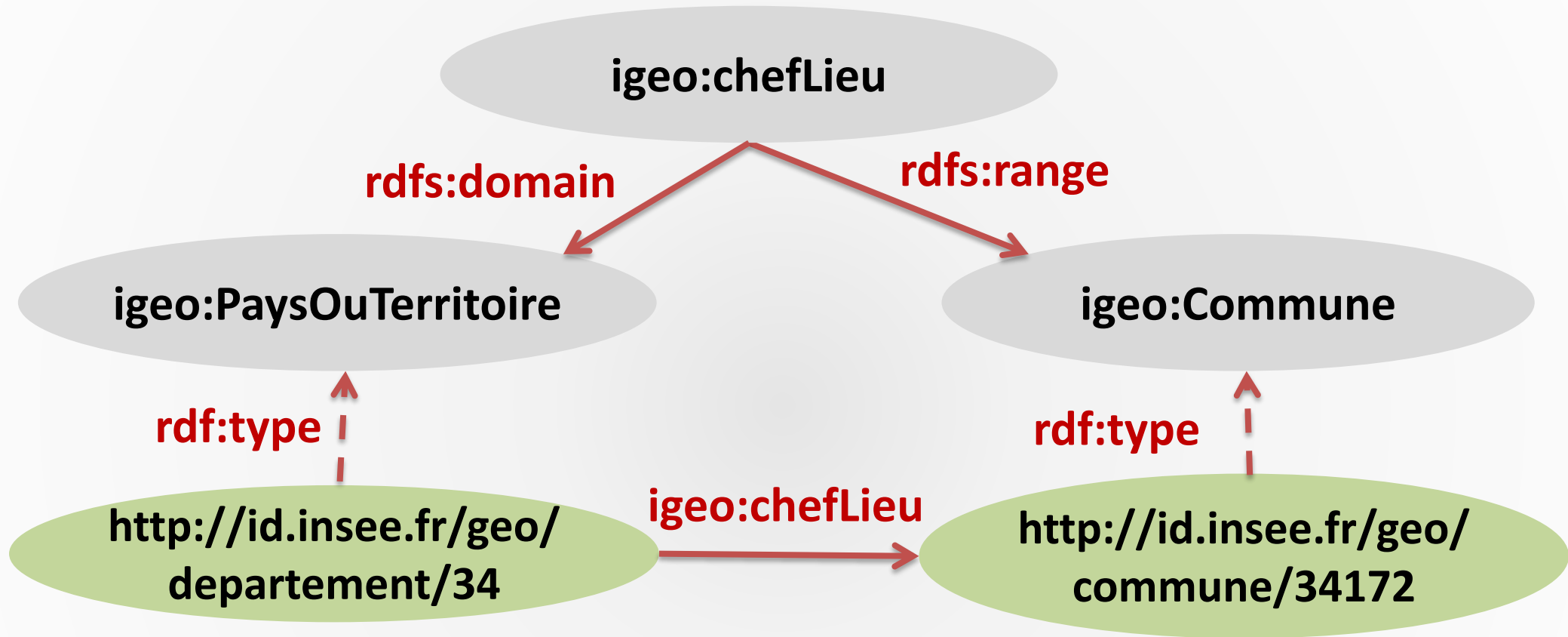
```
WHERE { ?x rdf:type igeo:TerritoireAdministratif }
```

SPARQL Entailments: infer knowledge



```
PREFIX igeo:<http://rdf.insee.fr/def/geo#>  
SELECT   ?x ?code  
WHERE { ?x igeo:codeINSEE ?code }
```


SPARQL Entailments: infer knowledge



SELECT ?x **WHERE** { ?x rdf:type igeo:Commune }

SELECT ?x **WHERE** { ?x rdf:type igeo:PaysOuTerritoire }

Agenda



The Semantic Web

Linked Data and the Web of Data

Publishing legacy data in RDF

The Web of Data

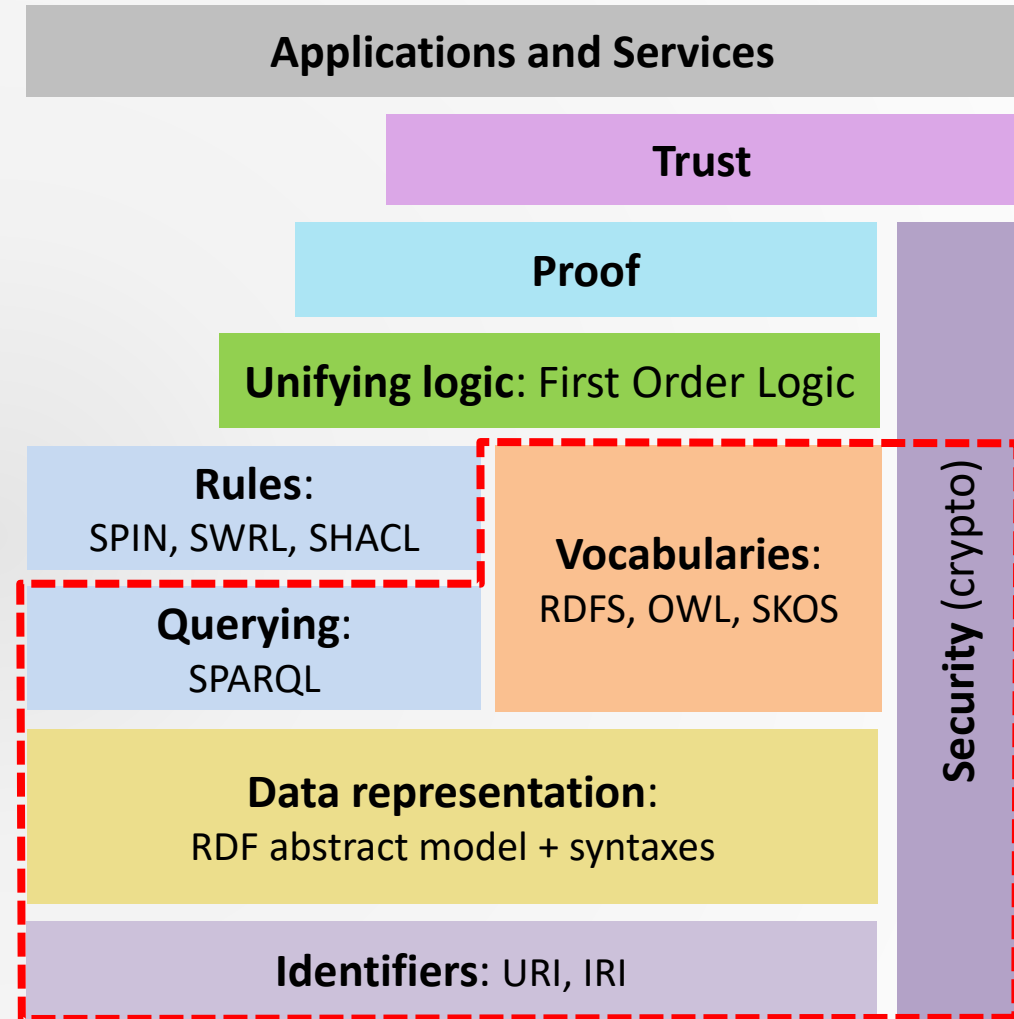
*aka. Data Web, Web 3.0,
Global Knowledge Graph...*



The Web of Data

First step in the deployment of the **Semantic Web**

Detractors would say the part of the Semantic Web that works...





Publish



Link

Vocabularies



Web of Data

*“The ~~Semantic Web~~ provides an environment where applications can **publish** and **link** data, define **vocabularies**, **query** data at web scale, ~~and draw inferences.~~”*



Querying



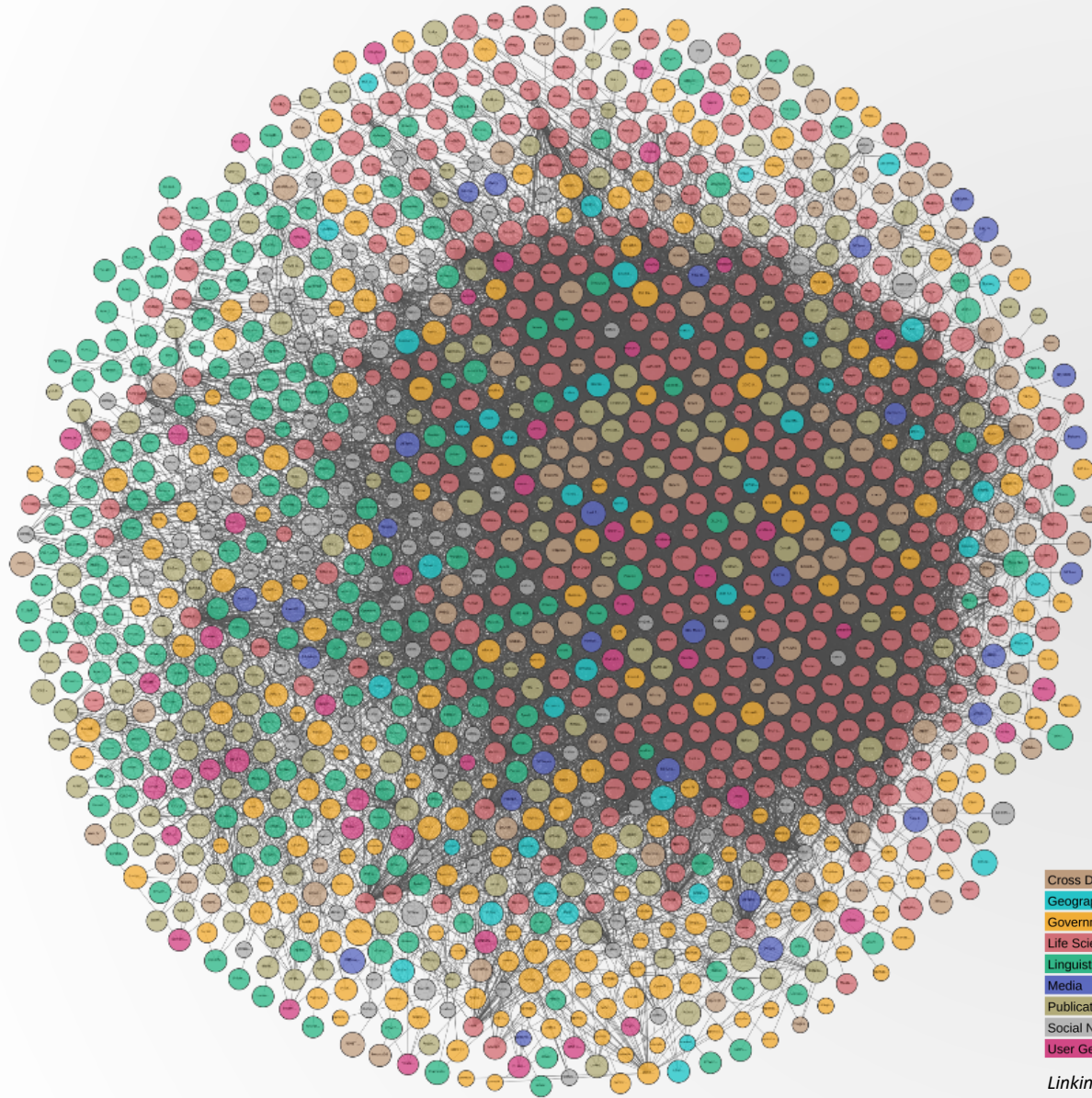
Inference

Linked Data principles



1. Use URIs to name things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)
4. Include **links to other URIs**, so they can **discover more things**

Linked Open Data Cloud: 1200+ linked datasets



- ✓ On the web, under open licenses
- ✓ Machine-readable (RDF)
- ✓ URIs to name things
- ✓ Common vocabularies
- ✓ Linked with each other
- ✓ Queryable

Iconic but partial view of the Web of Data
LOD Atlas: 25,000+ datasets



Linking Open Data cloud diagram, 2018. J.P. McCrae, A. Abele, P. Buitelaar, A. Jentzsch, V. Andryushechkin and R. Cyganiak.
<http://lod-cloud.net/>

Agenda



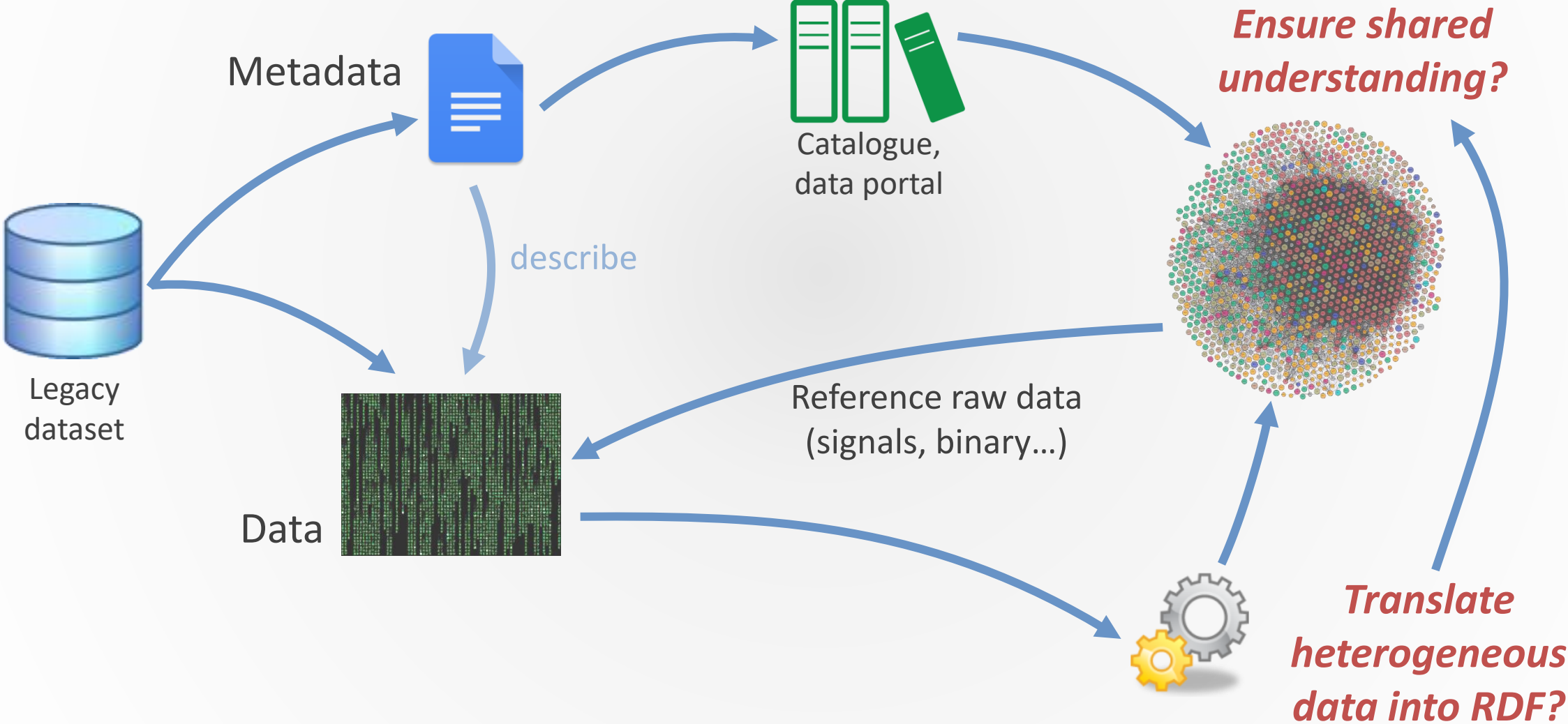
The Semantic Web

Linked Data and the Web of Data

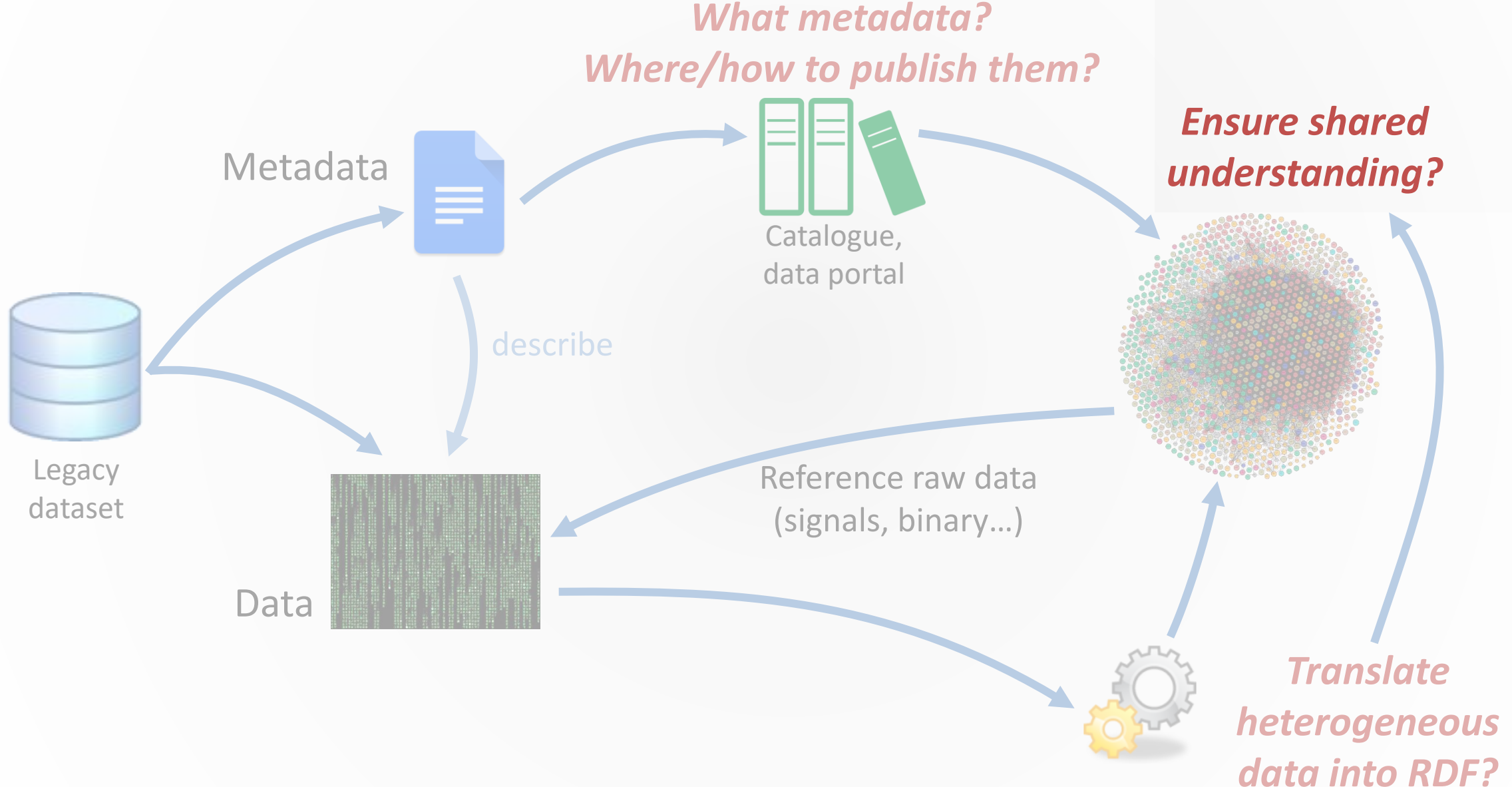
Publishing legacy data in RDF

Publishing legacy data in RDF raises tricky questions

*What metadata?
Where/how to publish them?*



Publishing legacy data in RDF raises tricky questions



Ensure shared understanding?

Need for common vocabularies with well defined semantics

- Controlled vocabulary, thesaurus, ontology
- How to define/model a vocabulary?
- Where to find existing vocabularies, how to reuse and/or them?

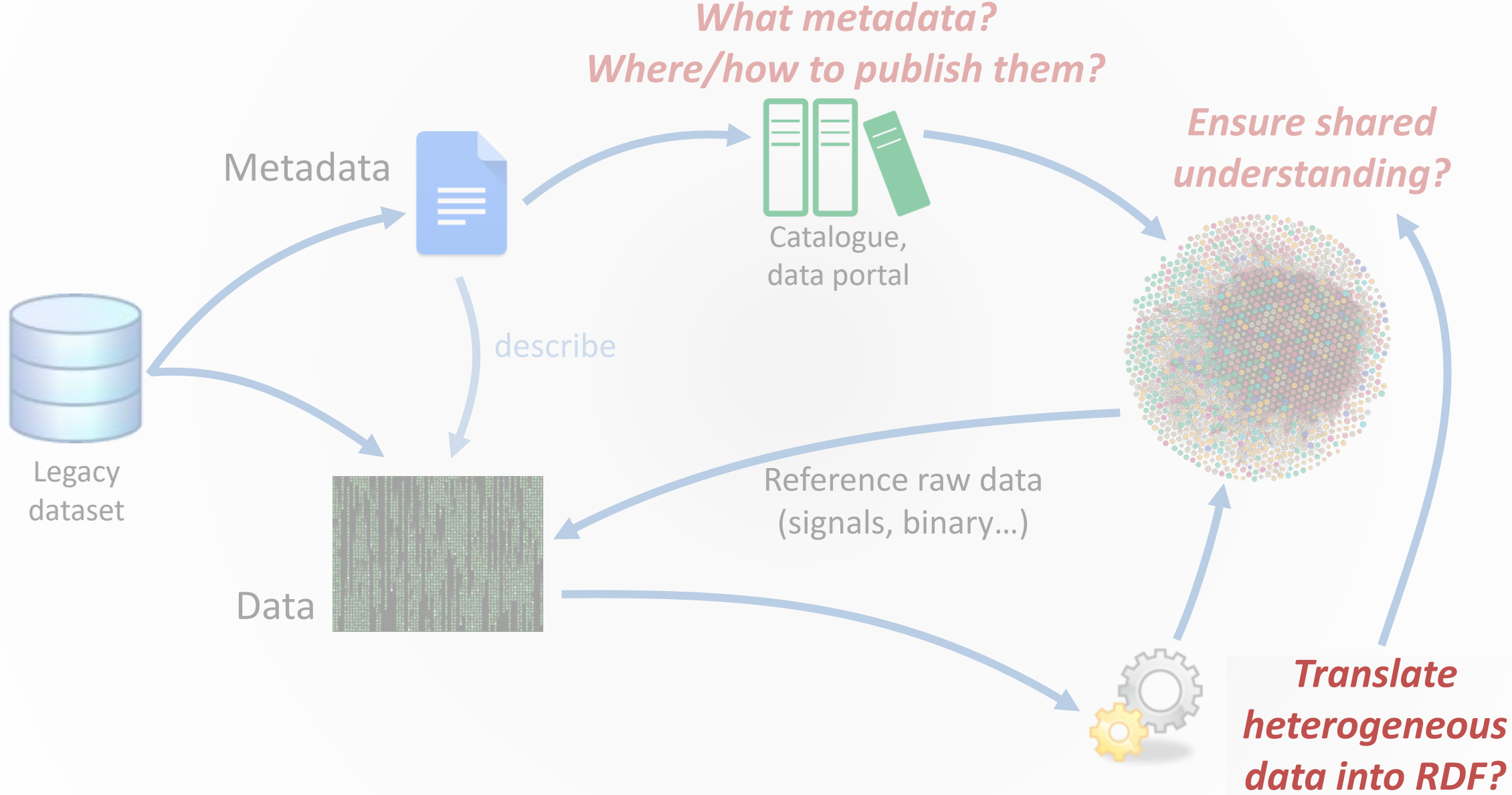
10h00-10h45 : Réutiliser/créer des vocabulaires contrôlés, des ontologies de domaine: LOV, BioPortal...

Nathalie Hernandez, Alban Gaignard

12h00-12h30 : Vocabulaire liés aux statistiques, formaliser les activités d'analyse.

Franck Cotton

Publishing legacy data in RDF raises tricky questions



Many methods for many types of data sources

HTML

RDFa, Microformats

XML

AstroGrid-D, SPARQL2XQuery, XSPARQL

CSV/TSV/Spreadsheets

XLWrap, Linked CSV, CSVW, RML

JSON

TARQL, JSON-LD, RML

Relational Databases

D2RQ, R₂O, Ultrawrap, Triplify, SM
R2RML: Morph-RDB, ontop, Virtuoso

NoSQL

xR2RML (MongoDB), ontop (MongoDB),
[Mugnier et al, 2016] (key-value stores)

Web APIs

SPARQL Micro-services, Linked REST APIs

Multiple formats

RML, TARQL, Apache Any23, DataLift,
SPARQL-Generate

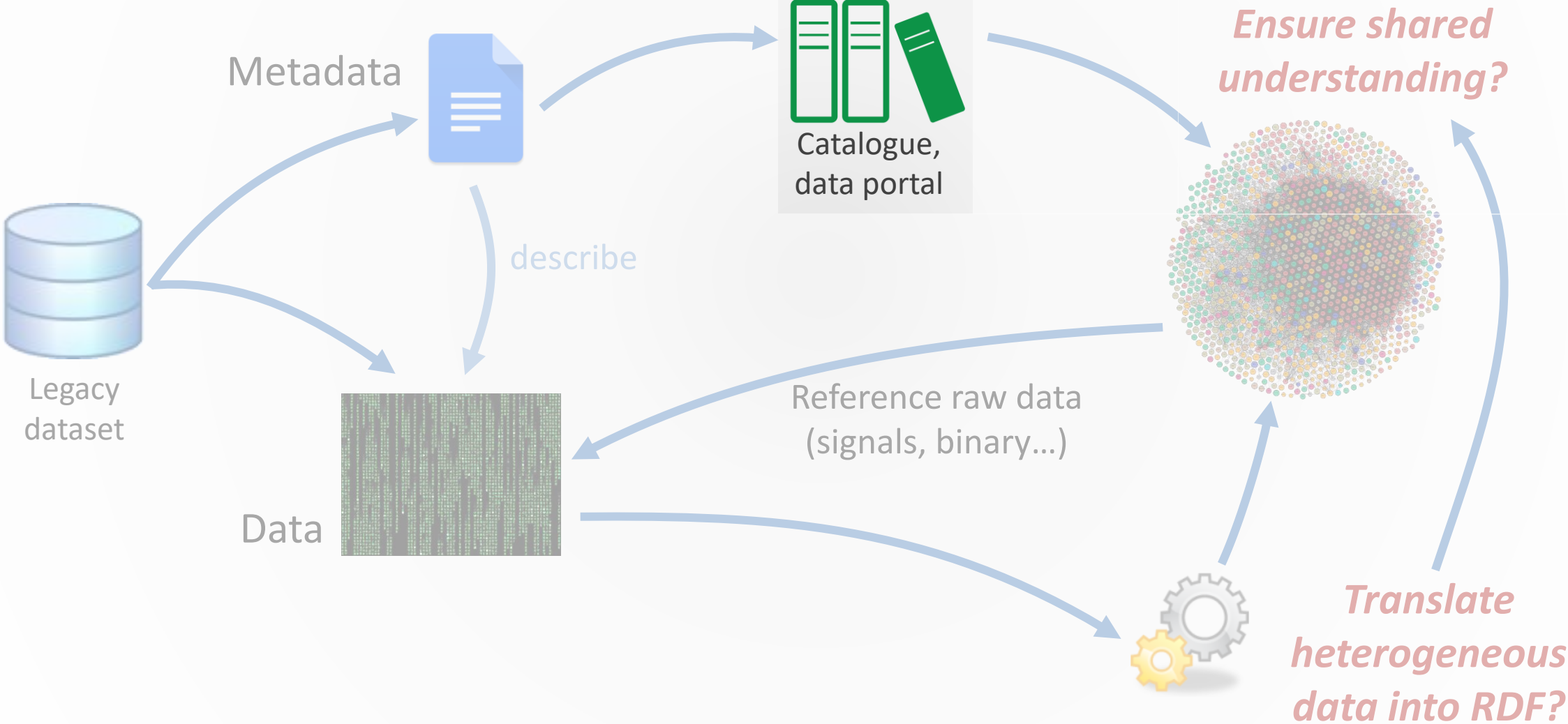
Interpret data as RDF, annotate data with RDF

11h15-11h40 : Comment annoter sémantiquement des données existantes (Web Annotation, CSV on the Web, JSON-LD...).

Nathalie Hernandez

Publishing legacy data in RDF raises tricky questions

What metadata?
Where/how to publish them?



Vocabularies to describe datasets and dataset catalogues

- Metadata vocabularies
Schema.org, DCAT, VOID, HCLS
- Data portals and catalogues
CKAN, data.gov.*, Google Dataset Search

11h40-12h00 : Décrire et Publier des jeux de données sur le web: vocabulaires, catalogues et portails.

Franck Michel

Thank you!