



**HAL**  
open science

# Integrating Heterogeneous Data Sources in the Web of Data

Franck Michel

► **To cite this version:**

Franck Michel. Integrating Heterogeneous Data Sources in the Web of Data. École thématique. Journées du développement logiciel (JDEV 2017), Marseille, France. 2017. cel-01585312

**HAL Id: cel-01585312**

**<https://hal.science/cel-01585312v1>**

Submitted on 11 Sep 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Integrating Heterogeneous Data Sources in the Web of Data

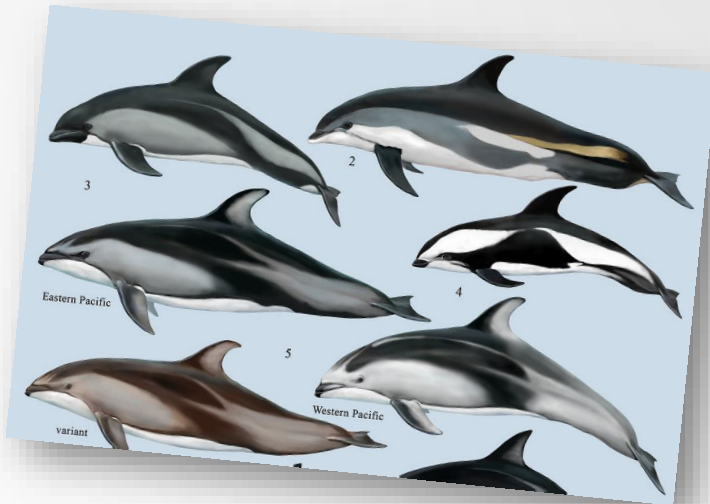
JDEV<sup>2017</sup>

Franck Michel

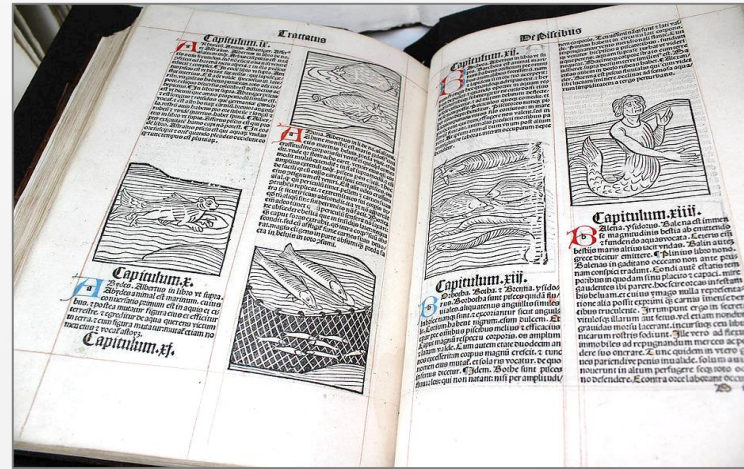
UNIVERSITÉ  
CÔTE D'AZUR



# More data sources $\Rightarrow$ More opportunities



# Example: study history of zoological knowledge



First Natural History Encycloedia, 1485.



Conservation biology\*



Archaeological excavation

# Example: study history of zoological knowledge

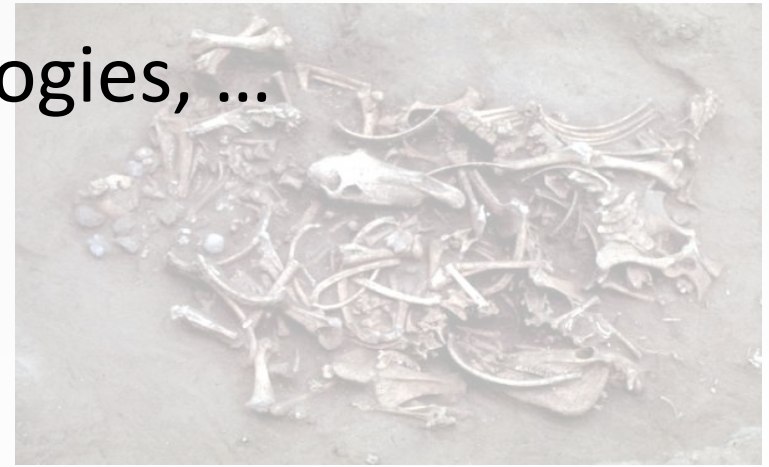


Knowledge formalizations

Controlled vocabularies,  
taxonomies  
domain ontologies, ...

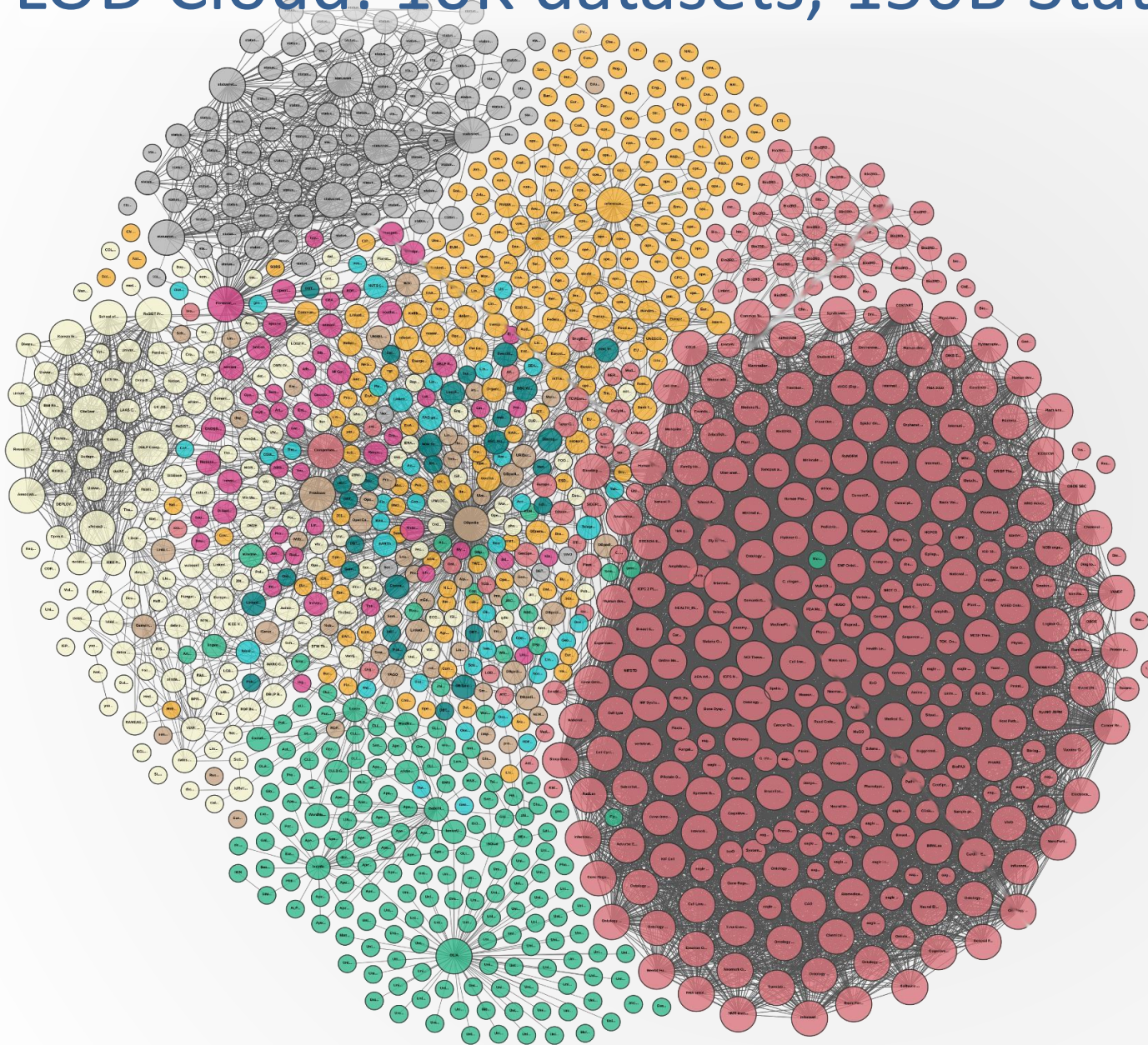


Conservation biology\*



Archaeological excavation

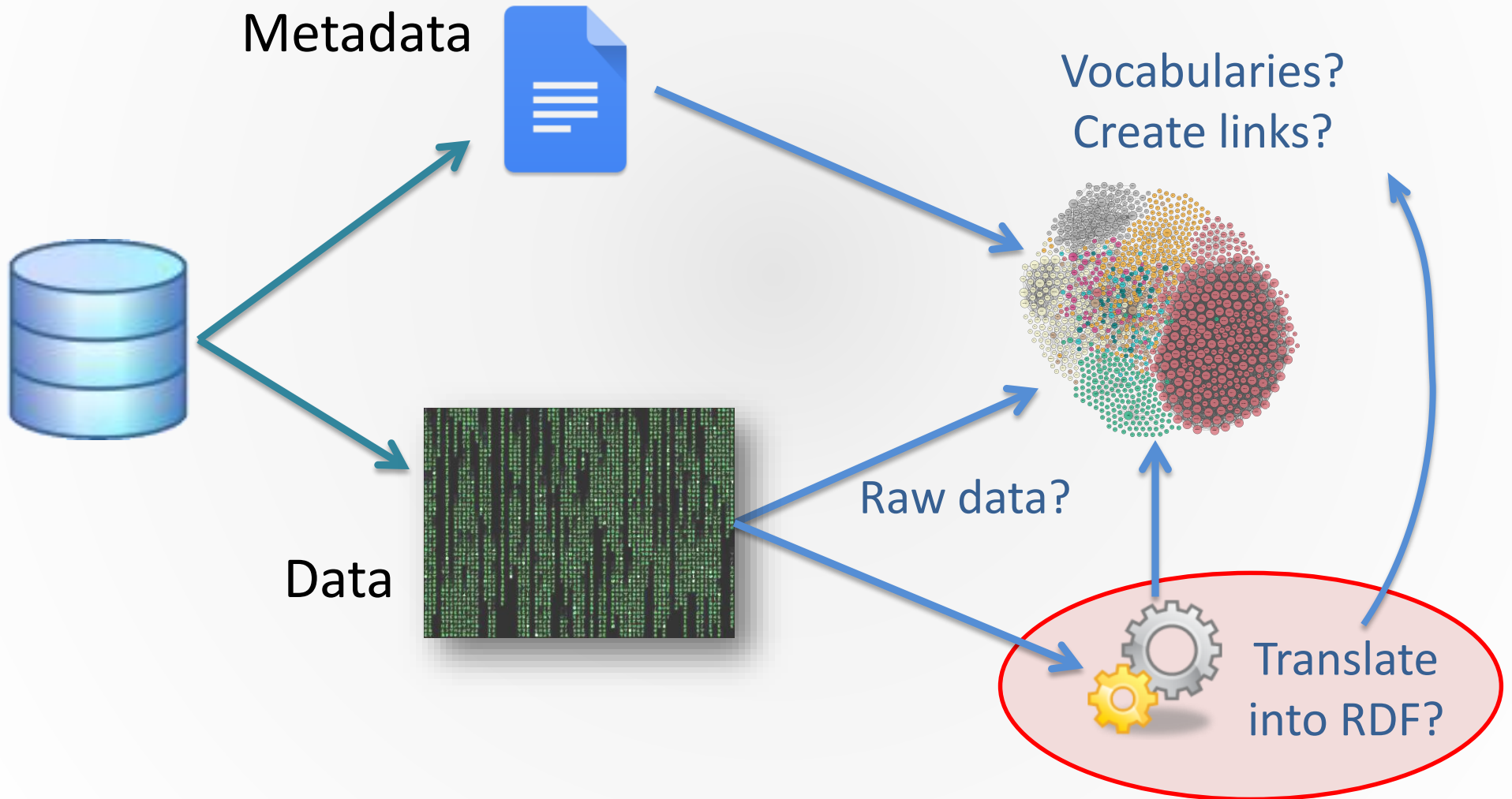
# LOD Cloud: 10K datasets, 150B Statements



- ✓ On the Web
- ✓ In RDF
- ✓ Under open licences
- ✓ Interlinked



# Publishing legacy data in RDF raises tricky questions



# Agenda



Describe the translation of heterogeneous data into RDF

Choose vocabularies to represent RDF data

Access the RDF data produced

The key importance of metadata



# Agenda



**Describe the translation of heterogeneous data into RDF**

Choose vocabularies to represent RDF data

Access the RDF data produced

The key importance of metadata

# Data Sources Have Heterogeneous Data Models

## Documents



## Relational DB

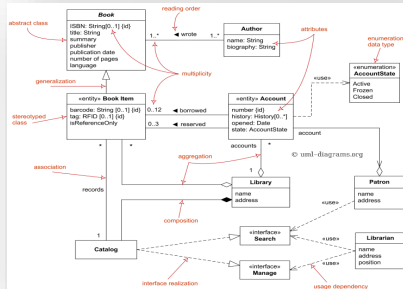
| ID | NAME |  |
|----|------|--|
|    |      |  |
|    |      |  |
|    |      |  |

## Native XML DBs

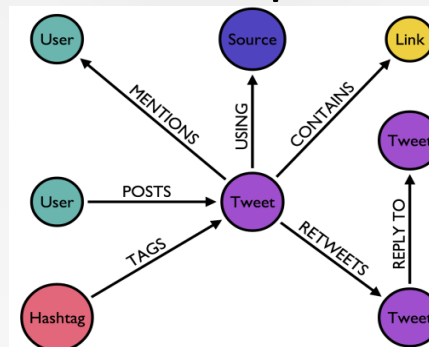
```

<Books>
  <Book ISBN="0553212419">
    <title>Sherlock Holmes: Complete Novels...</title>
    <author>Sir Arthur Conan Doyle</author>
  </Book>
  <Book ISBN="0743273567">
    <title>The Great Gatsby</title>
    <author>F. Scott Fitzgerald</author>
  </Book>
  <Book ISBN="0684826976">
    <title>Undaunted Courage</title>
    <author>Stephen E. Ambrose</author>
  </Book>
  <Book ISBN="0743203178">
    <title>Nothing Like It In the World</title>
    <author>Stephen E. Ambrose</author>
  </Book>
</Books>
    
```

## Object-Oriented



## Graph



# Agenda



## Describe the translation of heterogeneous data into RDF

- HTML data with RDFa

- CSV data

- Relational data

- NoSQL data

Choose vocabularies to represent RDF data

Access the RDF data produced

The key importance of metadata

# Agenda



## Describe the translation of heterogeneous data into RDF

HTML data with RDFa

CSV data

Relational data

NoSQL data

Choose vocabularies to represent RDF data

Access the RDF data produced

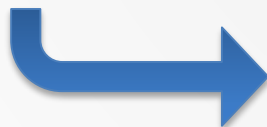
The key importance of metadata



<http://devlog.cnrs.fr/>



```
<body vocab="http://schema.org/">
  <div resource="/jdev2017" typeof="Event">
    <h2 property="title">JDEV 2017</h2>
    <p>Date: <span property="startDate">2017-07-04</span></p>
    ...
    <p>T2 - Ingénierie et web des données.
      <a property="url" href="http://devlog.cnrs.fr/jdev2017/t2">More...</a>
    </p>
  </div>
</body>
```



```
prefix sch: <http://schema.org/>
<http://devlog.cnrs.fr/jdev2017>
  rdf:type sch:Event ;
  sch:title "JDEV 2017";
  sch:startDate "2015-10-20" ;
  sch:url <http://devlog.cnrs.fr/jdev2017/t2> .
```

# Agenda



## Describe the translation of heterogeneous data into RDF

HTML data with RDFa

CSV data

Relational data

NoSQL data

Choose vocabularies to represent RDF data

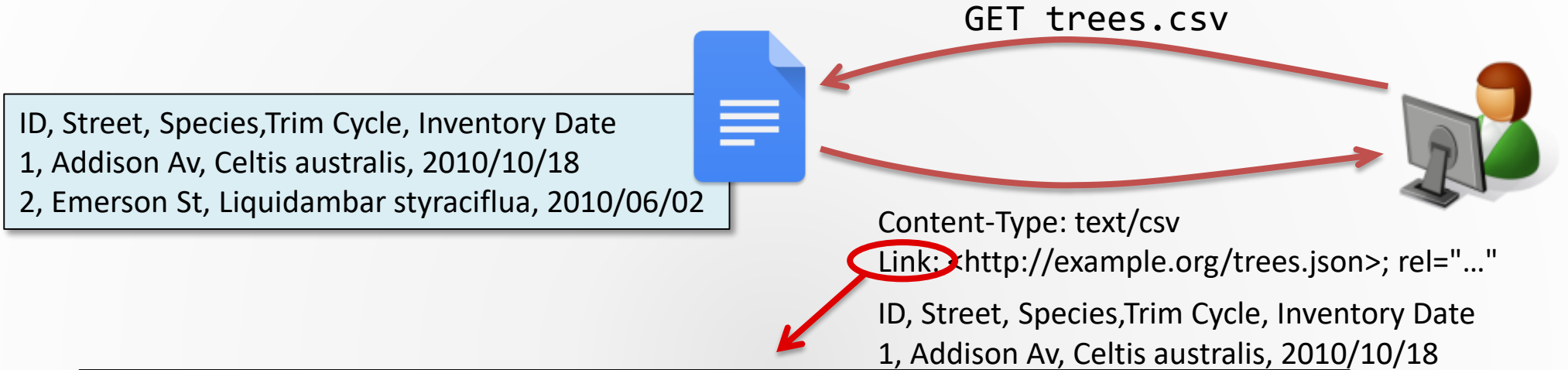
Access the RDF data produced

The key importance of metadata

# CSVW: CSV on the Web



<https://www.w3.org/2013/csvw/>



```
{ "@context": ["http://www.w3.org/ns/csvw", {"@language": "en"}],  
  "url": "trees.csv", "dc:title": "Trees",  
  "dc:license": { "@id": "http://.../..cc-by/" },  
  "dc:modified": { "@value": "2010-12-31", "@type": "xsd:date" },  
  "tableSchema": {  
    "columns": [  
      { "name": "ID", "titles": ["ID", "Identifier"], "datatype": "string", "required": true },  
      { "name": "Street", "titles": "On Street", "dc:description": "...", "datatype": "string" }, ...  
    ],  
    "primaryKey": "ID", "aboutUrl": "#id-{ID}" } }
```

# Agenda



## Describe the translation of heterogeneous data into RDF

HTML data with RDFa

CSV data

**Relational data**

NoSQL data

Choose vocabularies to represent RDF data

Access the RDF data produced

The key importance of metadata



# Direct Mapping of a RDB to RDF

| Table: PEOPLE |           |                    |
|---------------|-----------|--------------------|
| ID            | FNAME     | WROTE (FK BOOK/ID) |
| 7             | Catherine | 22                 |
| 8             | Olivier   | 22                 |
| ...           | ...       | ...                |

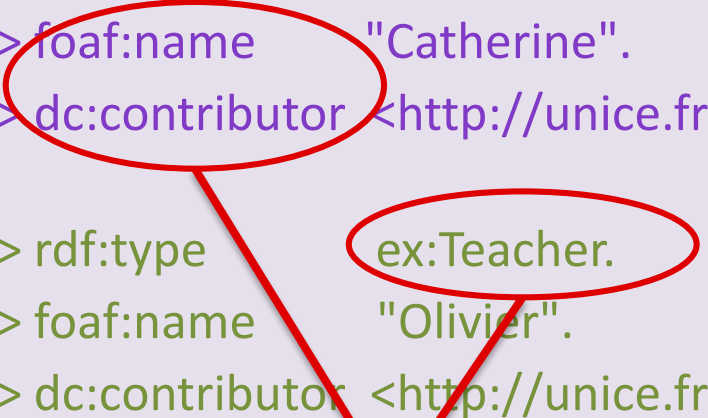
```
<PEOPLE/ID=7> rdf:type          <PEOPLE> .
<PEOPLE/ID=7> <PEOPLE#FNAME> "Catherine" .
<PEOPLE/ID=7> <PEOPLE#WROTE> <BOOK/ID=22> .

<PEOPLE/ID=8> rdf:type          <People> .
<PEOPLE/ID=8> <PEOPLE#FNAME> "Olivier" .
<PEOPLE/ID=8> <PEOPLE#WROTE> <BOOK/ID=22> .
```

# Custom Mapping of a RDB to RDF

| Table: PEOPLE |           |                    |
|---------------|-----------|--------------------|
| ID            | FNAME     | WROTE (FK BOOK/ID) |
| 7             | Catherine | 22                 |
| 8             | Olivier   | 22                 |
| ...           | ...       | ...                |

```
<http://unice.fr/staff/7> rdf:type          ex:Teacher.  
<http://unice.fr/staff/7> foaf:name      "Catherine".  
<http://unice.fr/staff/7> dc:contributor <http://unice.fr/book/22>.  
  
<http://unice.fr/staff/8> rdf:type          ex:Teacher.  
<http://unice.fr/staff/8> foaf:name      "Olivier".  
<http://unice.fr/staff/8> dc:contributor <http://unice.fr/book/22>.
```



**Existing vocabularies**

```
<#MapPeople>
  rr:logicalTable [ rr:tableName "PEOPLE" ];
  rr:subjectMap [
    rr:template "http://unice.fr/staff/{ID}";
    rr:class ex:Teacher;
  ];
  rr:predicateObjectMap [
    rr:predicate foaf:name;
    rr:objectMap [ rr:column "FNAME" ];
  ].
```



```
<http://unice.fr/staff/7> rdf:type ex:Teacher.
<http://unice.fr/staff/7> foaf:name "Catherine".
<http://unice.fr/staff/8> rdf:type ex:Teacher.
<http://unice.fr/staff/8> foaf:name "Olivier".
```

# Agenda



## Describe the translation of heterogeneous data into RDF

HTML data with RDFa

CSV data

Relational data

**NoSQL data**

Choose vocabularies to represent RDF data

Access the RDF data produced

The key importance of metadata

# Mapping of a NoSQL DBs to RDF with xR2RML

```
{ "id": 106,  
  "firstname": "John",  
  "emails": [ "john@foo.com", "john@example.org" ]  
}
```



```
<#MapMbox> xR2RML  
  xrr:logicalSource [ xrr:query "db.people.find({'emails':{$ne: null}})" ];  
  rr:subjectMap [ rr:template "http://example.org/member/{$.id}" ];  
  rr:predicateObjectMap [  
    rr:predicate foaf:mbox;  
    rr:objectMap [ xrr:reference "$.emails.*"; rr:termType rr:Literal ]  
  ].
```

```
<http://example.org/member/106> foaf:mbox "john@foo.com".  
<http://example.org/member/106> foaf:mbox "john@example.org".
```

# 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, R<sub>2</sub>O, Ultrawrap, Triplify, SM  
R2RML: Morph-RDB, ontop, Virtuoso

## NoSQL

xR2RML (MongoDB), ontop (MongoDB),  
[Mugnier et al, 2016]

## Multiple formats

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

# Agenda



Describe the translation of heterogeneous data into RDF

**Choose vocabularies to represent RDF data**

Access the RDF data produced

The key importance of metadata

# Direct mapping: create my own vocabulary

Can be derived from an existing schema

May seem easier: “*I do whatever I want*”

But no added semantics, need to link my vocabulary with other ones



# Custom Mapping: reuse existing vocabularies

Large variety of existing vocabularies

But may be difficult to find the appropriate one

Partial coverage of the domain

Granularity: too high (cumbersome), too low (useless)

Different points of view

Frequently, a mixed approach is used

# Agenda



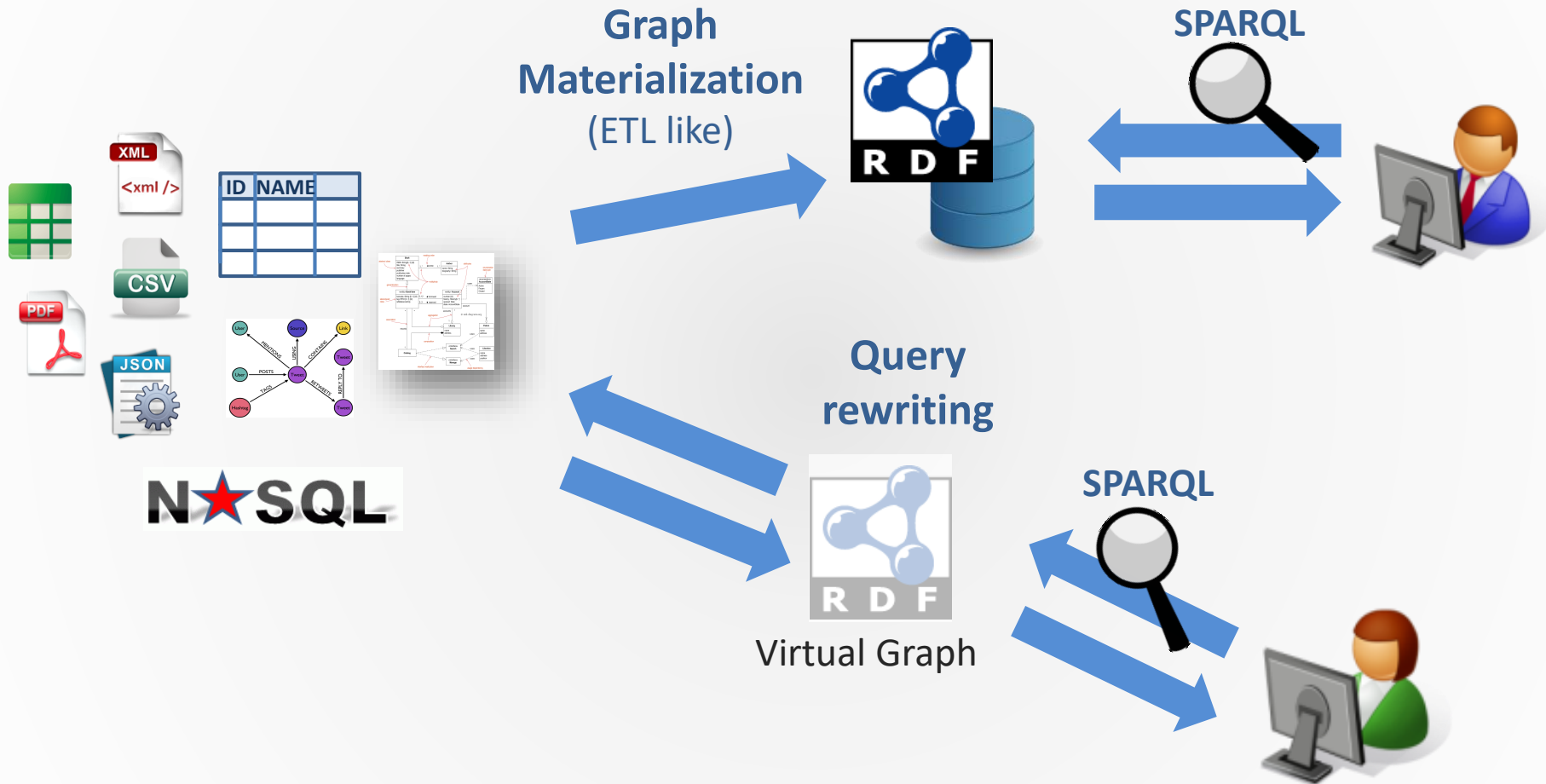
Describe the translation of heterogeneous data into RDF

Choose vocabularies to represent RDF data

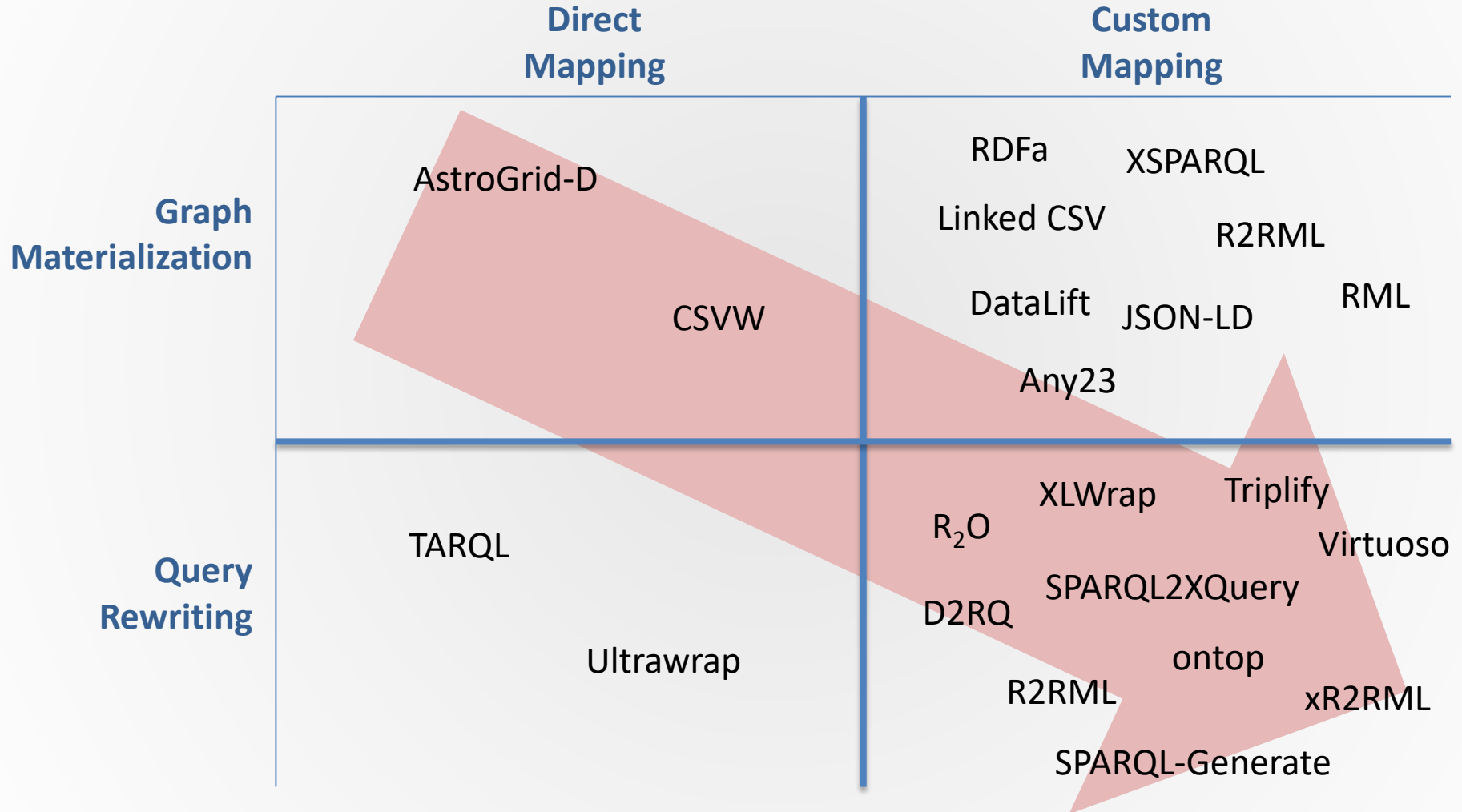
**Access the RDF data produced**

The key importance of metadata

# Two approaches to translate existing data sources in RDF



# A large variety of approaches



# Agenda



Describe the translation of heterogeneous data into RDF

Choose vocabularies to represent RDF data

Access the RDF data produced

**The key importance of metadata**

Making datasets **discoverable** and **reusable**

requires


**high-quality, comprehensive** metadata

 Recherche

Thématiques



CONTRIBUEZ !

 / Jeux de données 1 à 20 sur 27346

Trier par

---



## Population

Ce jeu de données permet d'accéder aux résultats des recensements de la population, à des séries chronologiques de la Banque de Données Macro-économiques de l'Insee sur le thème de la population et à d'autres données...

 01/1901 à 07/2015  Mensuelle  France  Commune française  23  32



## Indicateur Avancé Sanitaire IAS® - SYNDROME GRIPPAL

L'objectif de l'Indicateur Avancé Sanitaire (IAS®) "Syndrome Grippal" est de contribuer à la surveillance des syndromes grippaux en France en apportant des informations complémentaires à celles du réseau Sentinelles

# Metadata are the key to enable dataset reuse

|                |   |
|----------------|---|
| Context        | Identification, authors, dates, license, version, reference articles  |
| Access         | Format, structure, location (dwld), query method  |
| Meaning        | What do the data represent? What concepts, entities, semantics?   |
| Interpretation | Units (cm or inches, left/right)...   |
| Provenance     | Acquired with which equipment, parameters, protocols?<br>Derived from which dataset? With which processing?<br>Dataset-level or entity-level provenance |
| Statistics     | Number of triples per property of class, links to other datasets...   |
| ...            |   |



# Several works normalize metadata descriptions and how to use metadata

- CSVW: CSV on the Web
- DCAT: Data Catalog Vocabulary  
DCAT extensions, application profiles  
W3C Dataset Exchange Working Group
- VoID: Vocabulary of Interlinked Datasets
- HCLS: Health Care & Life Sciences Dataset Profile
- ...



Thank  
you!