Le Web des données ouvertes (Linked Open Data)

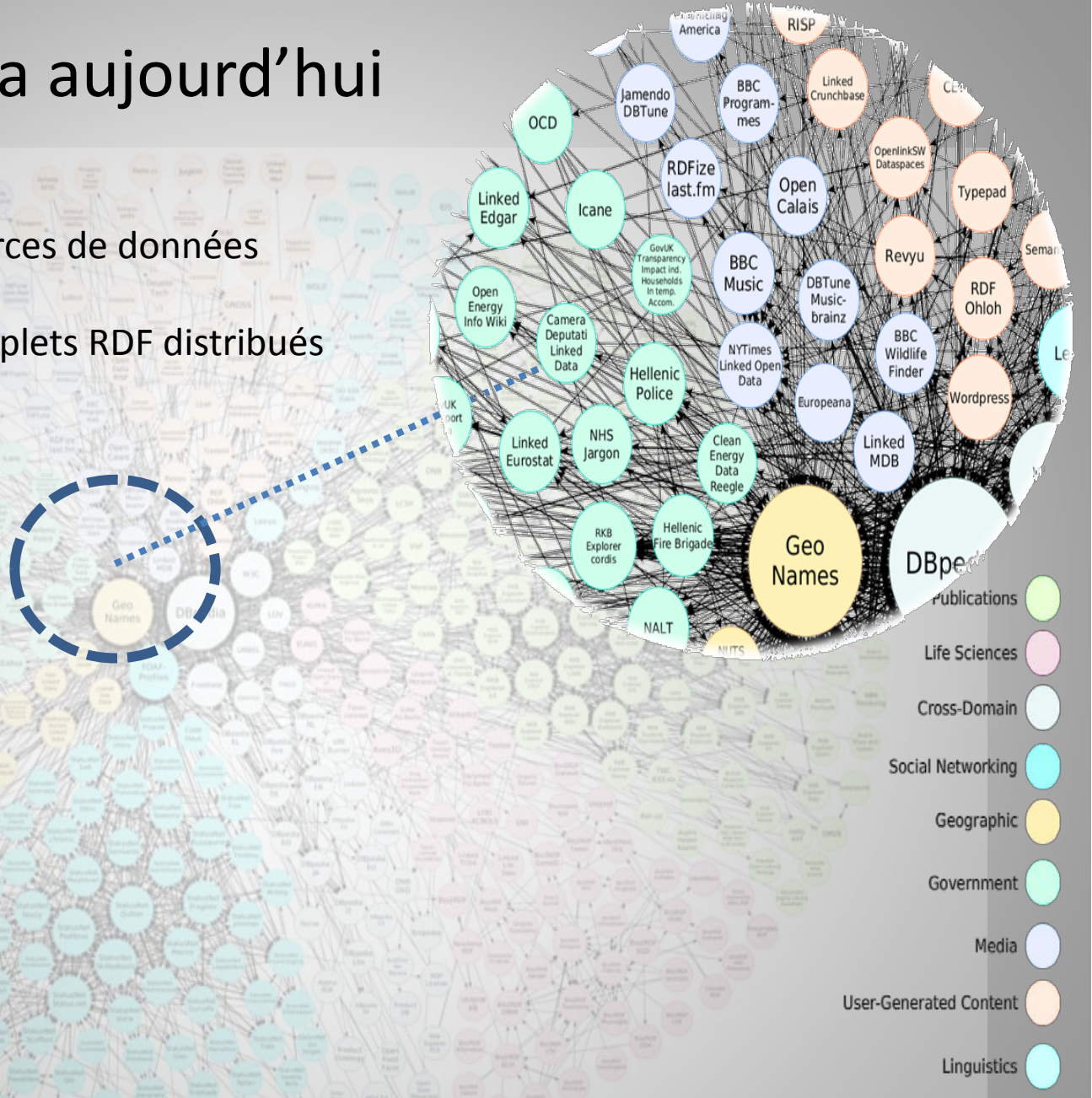
Marie-Christine Rousset

LIG

Université de Grenoble (UJF) et Institut Universitaire de France

Linked Open Data aujourd'hui

-Plusieurs milliers de sources de données accessibles sur le web
- Plusieurs milliards de triplets RDF distribués
- Domaines variés



Les principes sous-jacents

"Les données liées sont un ensemble de principes de conception pour le partage de données lisibles par machine sur le Web pour une utilisation par les administrations publiques, les entreprises et les citoyens."

EC ISA Case Study: How Linked Data is transforming eGovernment

Les **quatre principes de conception** des données liées (par Tim Berners Lee):

1. Utiliser des identificateur de ressources uniformes(URI) pour les noms des choses.
2. Utiliser des URIs HTTPde sorte que les gens puissent consulter ces adresses.
3. Quand quelqu'un consulte une URI, fournir des informations utiles, en utilisant les standards (RDF *, SPARQL).
4. Inclure des liens vers d'autres URIs afin qu'ils puissent découvrir plus de choses.

Voir aussi:
http://www.youtube.com/watch?v=4x_xzT5eF5Q
http://www.w3.org/DesignIssues/LinkedData.html
http://www.youtube.com/watch?v=uju4wT9uBIA

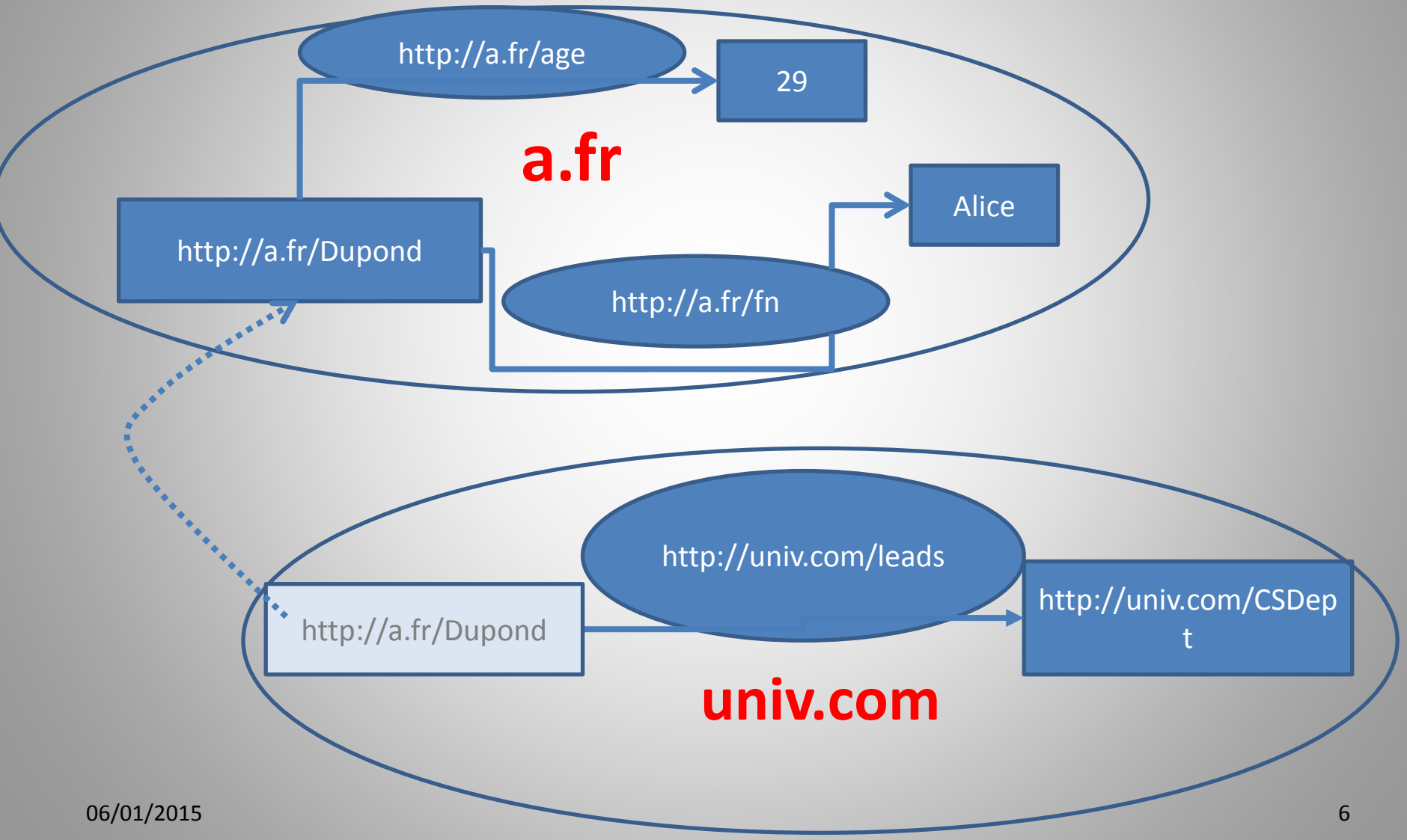
Les standards à la base du Linked Open Data

- URIs et espaces de noms (namespaces)
  - Pour dénoter et nommer de façon non ambigue des entités
    - URI: Uniform Resource Identifier
    - Namespace: pour lever les ambiguités sur des termes qui pourraient être homonymes sinon
      - Matérialisé par un préfixe qui est une URL
      - Pas d'homonyme au sein d'un même espace de noms
- RDF (Resource Description Framework)
  - Pour déclarer des faits connus sur ces entités sous la forme de triplets <sujet, relation/propriété, objet/valeur>
- RDFS (RDF Schema) et OWL
  - Pour structurer les entités par rapport à une hiérarchie de classes et donner une sémantique aux relations utilisées
- SPARQL
  - Pour poser des requêtes par des points d'accès via un service web
    - http://fr.dbpedia.org/sparql
    - http://rdf.insee.fr/sparql
    - https://gate.d5.mpi-inf.mpg.de/webyagospotlx/WebInterface

URIs et espaces de noms

- http://a.fr/Dupond est un URI
- http://a.fr est un espace de noms (dans lequel il n'y a qu'un Dupond

Linked data: un ensemble distribué de triplestores et de namespaces accessibles par des addresses sur le web



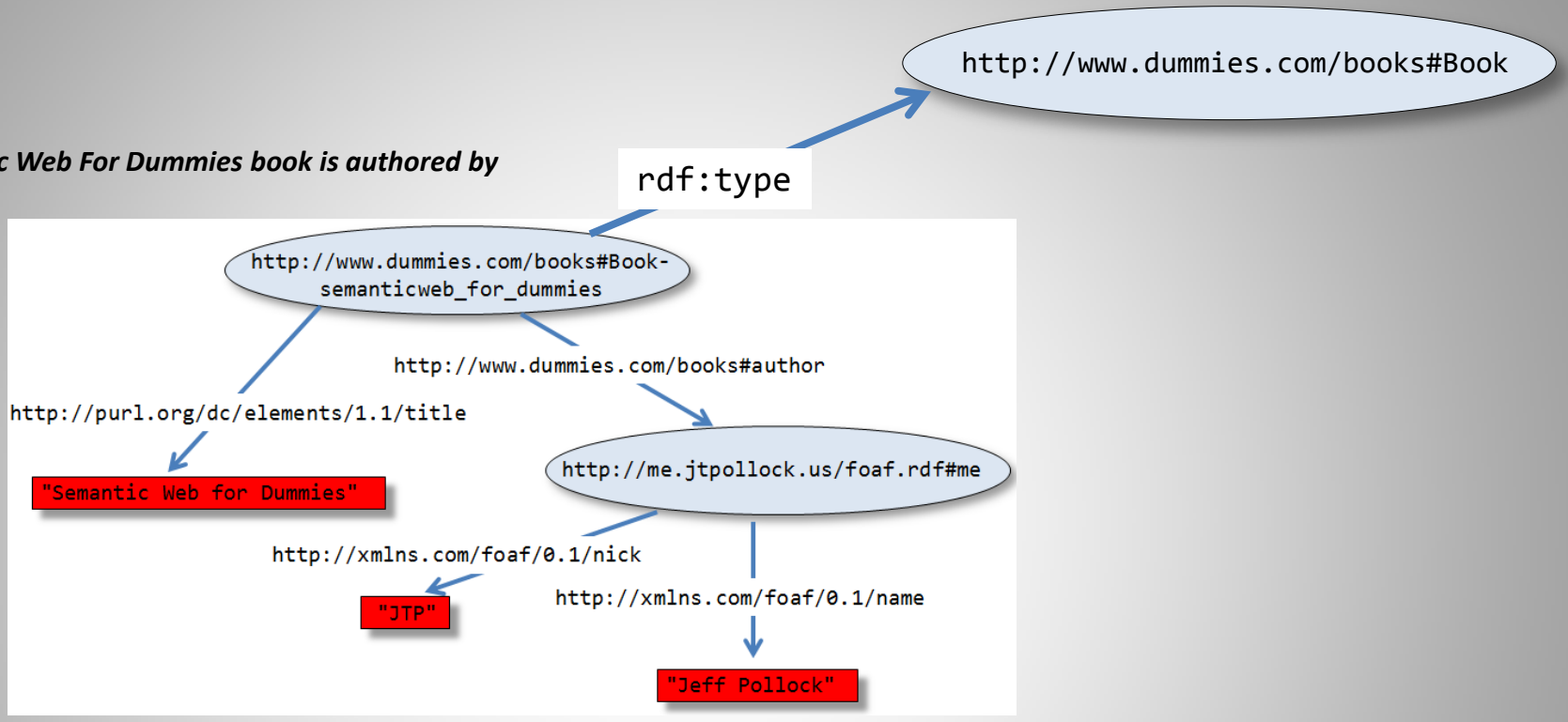
Certains namespaces offrent des vocabulaires standardisés: **une bonne pratique est de les réutiliser**

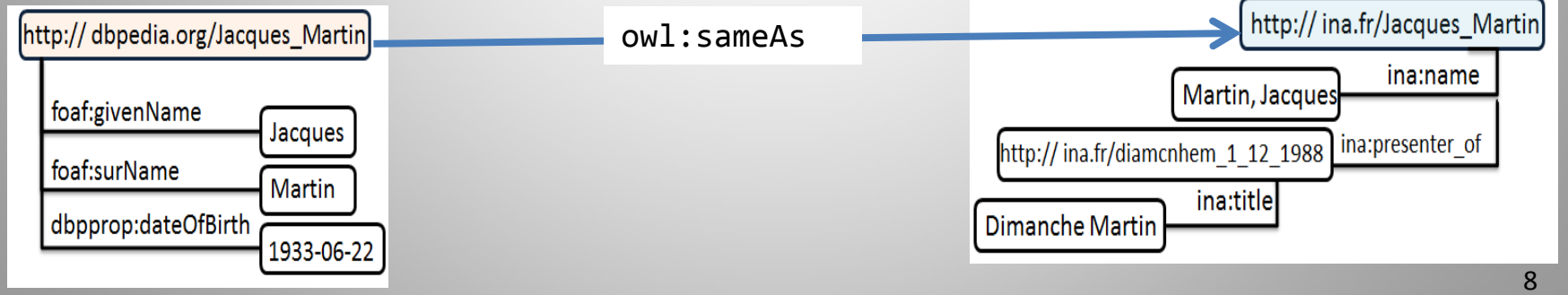
- Dublin Core: vocabulaire prédéfini pour décrire des documents
  - <http://purl.org/dc/terms/>
  - préfixe **dc:** (dc:author, dc:title, dc:publisher, …)
- Schema.org namespace: vocabulaire pour le e-commerce
  - http://schema.org/
  - Préfixe **schema:** (schema:organization, …)
- FOAF namespace: vocabulaire pour les réseaux sociaux
  - http://xmlns.com/foaf/0.1/
  - préfixe **foaf:** (foaf: name, foaf: title, foaf: knows, …)

- RDF: vocabulaire dédié pour définir des faits RDF
  - http://www.w3.org/1999/02/22-rdf-syntax-ns/
  - préfixe **rdf:** (rdf:type, rdf:resource, rdf:Propery, rdf:datatype, rdf:about, rdf:description, rdf:Bag, ..)
- RDFS: vocabulaire pour décrire le schéma de RDF
  - http://www.w3.org/2000/01/rdf-schema/
  - préfixe **rdfs:** (rdfs:subClassOf, rdfs:range, rdfs:domain, rdfs:subProperty, …)
- Espace RDF de l'INSEE:
  - http://rdf.insee.fr/
  - préfixes **igeo:** et **idemo:** (idemo:PopulationLegale, igeo: Region, igeo:Commune, …)

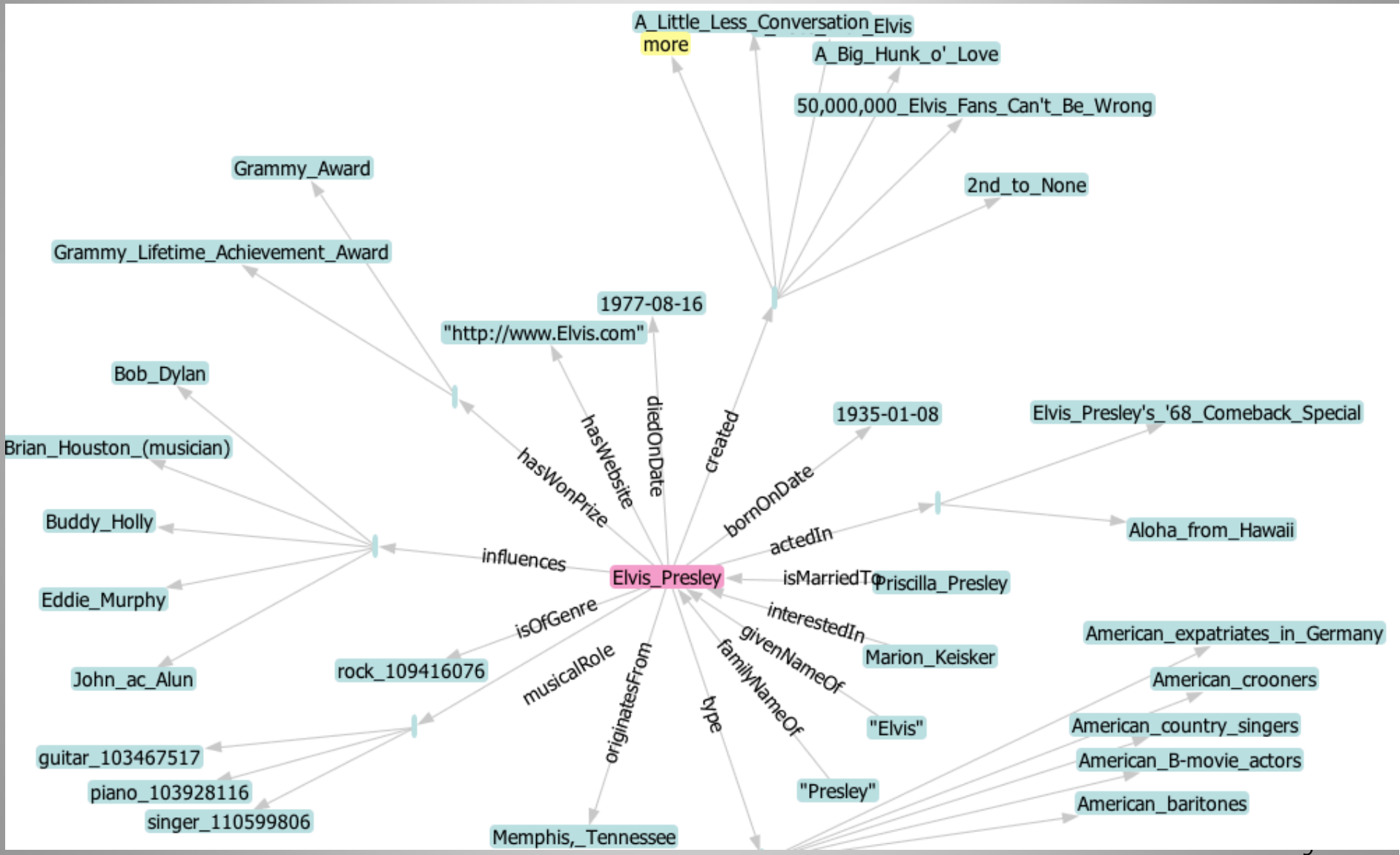
Le modèle de (meta)-données RDF

The Semantic Web For Dummies book is authored by Jeff Pollock



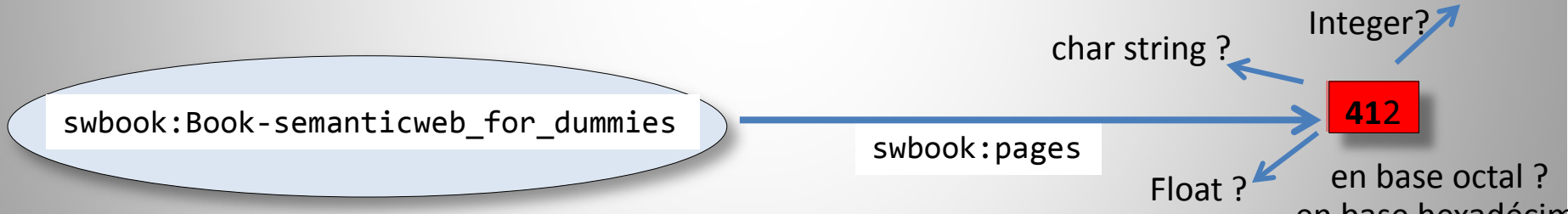


Exemple d’un graphe RDF dans Yago



Lien entre RDF et XML

- Typage des littéraux : XML Schema
  - W3C recommandations :
    - http://www.w3.org/TR/xmlschema-2/
    - http://www.w3.org/TR/rdf-mt/
  - hiérarchie de types de données prédéfinis
    - des types primitifs (string, float, decimal, etc.)
    - des types dérivés (integer,long, date, etc .)




```
"literal"^^type_URI
```

```
@prefix swbook: <http://www.dummies.com/books#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
swbook:Book-semanticweb_for_dummies swbook:pages
        "412"^^xsd:integer.
```

XML/RDF : une sérialisation de RDF

- Sérialisations de RDF
  - fournissent une manière de convertir le modèle abstrait en un format concret (un fichier ou un autre flux de données)
  - plusieurs formats de sérialisation
    - XML/RDF (format d'échange normatif/standard pour la sérialisation)
    - N-Triples, Turtle (Terse RDF Triple Language), N3 (Notation3)

une déclaration XML (déclare que c'est un document XML)

espaces de noms XML

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:books="http://www.dummies.com/books#">
    <rdf:Description
       rdf:about="http://www.dummies.com/books#Book-semanticweb_for_dummies">
       <books:author
            rdf:resource="http://me.jtpollock.us/foaf.rdf#me" />
    </rdf:Description>
</rdf:RDF>
```

l'élément **rdf:Description** pour des déclarations sur des ressources

Focus sur 3 sources

- DPPedia: http://fr.dbpedia.org
  - Universités de Leipzig et de Berlin + OpenLink Software
  - version RDF des fiches wikipedia sur des entités (personnes, lieux, etc…): 4 millions d'entités, 470 millions de faits
    - Exploration et extraction automatique d'entités et de relations de pages web (Wikipedia)
  - version entièrement francophone depuis 2012
- Yago: http://www.mpi-inf.mpg.de/yago-naga/yago/
  - Univ. Saarbrucken, Max Plank Institute
  - 120 millions de faits sur 10 millions d'entités
- FreeBase: http://www.freebase.com/
  - création collaborative d'une base de données du Web,
    - Initiée par la société Metaweb, rachetée par Google en 2010
  - 2 millions de faits sur 40.000 entités

SPARQL (le SQL de RDF)

Données

```
<book1> <title> "SPARQL Tutorial" .
```

Requête

```
SELECT ?title
WHERE
{
  <book1> <title> ?title .
}
```

Résultat

```
title
-----
SPARQL Tutorial
```

SPARQL par l'exemple

- Que sait-on sur Paris dans fr.dbpedia ?

```
SELECT ?p ?o
WHERE { <http://fr.dbpedia.org/resource/Paris> ?p ?o . }
LIMIT 100

SELECT ?s ?p
WHERE { ?s ?p <http://fr.dbpedia.org/resource/Paris> . }
LIMIT 100
```

- Quelles sont les communes d'Ile de France ?

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?commune
WHERE {
?commune dbpedia-owl:region <http://fr.dbpedia.org/resource/Île-de-France> .
?commune rdf:type dbpedia-owl:PopulatedPlace }
```

Exemples (suite)

- Quelles sont les communes d'Ile de France de plus de 100.000 habitants et leurs maires?

```
SELECT ?commune ?n
WHERE {
?commune <http://dbpedia.org/ontology/region>
   <http://fr.dbpedia.org/resource/Île-de-France> .
?commune rdf:type dbpedia-owl:PopulatedPlace .
?commune dbpedia-owl:populationTotal ?population .
?commune prop-fr:maire ?n
FILTER (?population > 100000)
}
```

Exemples (suite)

- Quelles sont les entités dont le nom contient « Martin » ?

```
select distinct  ?z ?y
 where
{
?z <http://xmlns.com/foaf/0.1/name> ?y.
?z rdf:type <http://xmlns.com/foaf/0.1/Person>.
FILTER (regex(?y, "Martin", "i"))
 }
LIMIT 100
```

Comment construire un triplestore RDF pour une application/domaine particulier(e)s ?

- Application non connectée à Linked Open Data
  - Utilisation d'un environnement d'édition et de visualisation de données RDF
    - Topbraid composer, Protégé, Jena, …
- Application connectée à Linked Open Data:
  - Extraction pour réutilisation de triplets ou d'URIs de sources existantes
    - exploration / interrogation de Linked Data via les points d'entrée SPARQL
  - Enrichissement en ajoutant de nouveaux triplets à l'aide d'outils précédents

Différentes façons de déclarer des liens dans Linked Open Data

- owl:equivalentClass
- owl:sameAs
- rdfs:seeAlso
- skos:closeMatch
- skos:exactMatch
- skos:related
- foaf:homepage
- foaf:topic
- foaf:based_near
- foaf:maker/foaf:made
- foaf:page
- foaf:primaryTopic

- Example:

http://dbpedia.org/resource/Canberra
owl:sameAs
http://rdf.freebase.com/rdf/en.canberra

• 3.6 millions de liens sameAs entre Freebase et DBpedia 3.8 en Juin 2012 (2.4 millions au départ, en 2008)

• 45 millions de faits sameAs déclarés dans DBPedia (avec d'autres sources de Linked Data).

Les défis du Linked Open Data

- Recherche et interrogation de sources de données dont on ne connait pas le vocabulaire
- Ajouter de la sémantique et du raisonnement
  - Déclarer des ontologies RDFS ou OWL en notation RDF
- Déclaration de classes et de relations de spécialisation entre classes
  - <Staff, rdf:type , rdfs:Class>
  - <AcademicStaff, rdfs:subClassOf, Staff>
- Déclaration d'appartenance d'instances à une classe
  - <Dupond, rdf:type, AcademicStaff>
- Déclaration de propriétés , de relations de spécialisation entre propriétés, de typage des arguments des propriétés
    - <RegisteredTo, rdfs:domain, Student>
    - <RegisteredTo, rdfs:range, Course>
  - Exploiter ces ontologies comme des règles logiques pour saturer les triplets ou récrire les requêtes

Sémantique logique des déclarations RDFS

si <?i , rdf:type, ?c>, <?c, rdfs:subClassOf, ?d> **alors** <?i, rdf:type, ?d>

si <?b , rdfs:subClassOf, ?c>, <?c, rdfs:subClassOf, ?d> **alors** <?b, rdfs:subClassOf, ?d>

si <?p , rdfs:subPropertyOf, ?q>, <?q, rdfs:subPropertyOf, ?r>

alors <?p, rdfs:subPropertyOf, ?r>

si <?p, rdfs:domain, ?c> , <?i , ?p, ?j> **alors** <<?i, rdf:type, ?c>

si <?p, rdfs:range, ?c> , <?i , ?p, ?j> **alors** <<?j, rdf:type, ?c>

si <?p, rdf:domain, ?c> **alors** <?p, rdf:type, rdf:Property>

si <?p, rdf:range, ?c> **alors** <?p, rdf:type, rdf:Property>

si <?i, ?p, ?j> **alors** <?p, rdf:type, rdf:Property>

si <?i, rdf:type, ?c> **alors** <?c, rdf:type, rdfs:Class>

Les défis du Linked Open Data

- Recherche et interrogation de sources de données dont on ne connait pas le vocabulaire
- Ajouter de la sémantique et du raisonnement
  - Déclarer des ontologies RDFS ou OWL en notation RDF

OWL étend l'expressivité de RDFS

- déclaration de disjonction entre classes
  <Student, owl:disjointWith, Course>
- déclaration de propriétés fonctionnelles
  <HasName, rdf:type, owl:FunctionalProperty>
  <Leads, rdf:type , owl:InverseFunctionalProperty>
- déclaration de propriétés inverses
  <child, owl:inverseOf, parent>
- déclaration de propriétés symétriques ou transitives
  <near, rdf:type, owl:SymmetricProperty>
  <friend, rdf:type, owl:TransitiveProperty>

Les défis du Linked Open Data

- Recherche et interrogation de sources de données dont on ne connait pas le vocabulaire
- Ajouter de la sémantique et du raisonnement
  - Déclarer des ontologies RDFS ou OWL en notation RDF

Contraintes OWL complexes exprimables en RDF par plusieurs triplets à l'aide de mots-clefs réservés du namespace de OWL, et de nœuds vides

owl:Restriction
owl:onProperty
owl:someValuesFrom
owl:allValuesFrom
owl:minCardinality
owl: maxCardinality
owl:union

Illustration par l'exemple

- Seuls des professeurs ou des directeurs de recherche peuvent enseigner à des étudiants de master
  - Définir la classe anonyme des objets qui « enseignent à au moins un étudiant de master »

    <_a , rdfs:subClassOf, owl:Restriction>

    <_a, owl:onProperty, TeachesTo>

    <_a, owl:someValuesFrom, MasterStudent>
  - Définir la classe anonyme de l'union des deux classes Professeur et Directeur de Recherche

    <_b, owl:unionOf, (Professor, DirecteurRecherche)>
  - Spécifier que les objets qui « enseignent à au moins un étudiant de master » appartiennent à l'union des deux classes Professeur et Directeur de Recherche

    <_a, rdfs:subClassOf, _b>

Sémantique logique de OWL

- Les règles logiques capturent certaines contraintes exprimées en OWL:

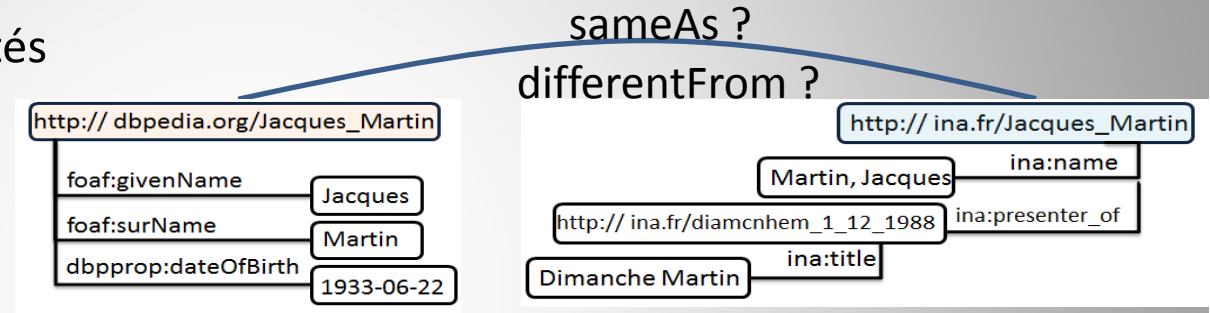
  si <?p, owl:inverse, ?q> , <?i, ?p, ?j> **alors** <?j, ?q, ?i>

  si <?p, rdf:type, owl:TranstiveProperty> , <?i, ?p, ?j> , <?j, ?p, ?l> **alors** <?i, ?p, ?l>

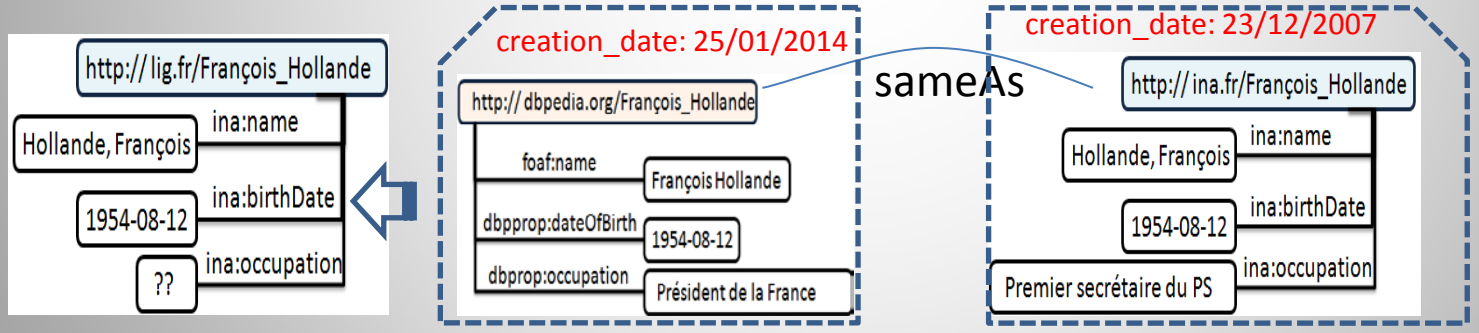
- Mais pas toutes:
  - Règles existentielles ou avec négation
  - Pouvoir d'expression des logiques de description
  - Nécessite des moteurs d'inférences adaptés (tel Pellet) dont certains sont disponibles dans les éditeurs d'ontologies existants

Les défis du Linked Open Data

- Recherche et interrogation de sources de données dont on ne connait pas le vocabulaire
- Ajouter de la sémantique et du raisonnement
- Découverte automatique de liens entre entités de différentes sources
  - Par agrégation de similarités
  - Par inférence



- Fusion de données



- Qualité des données
  - Exploitation de la provenance
  - Modélisation et exploitation de la confiance dans les (sources de) données
  - crowdsourcing