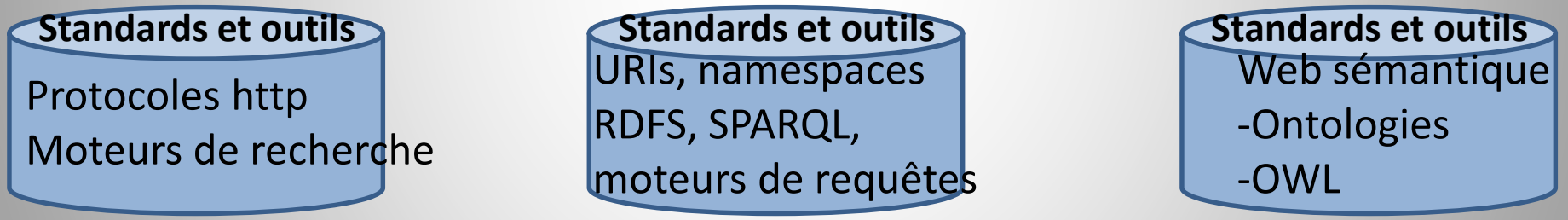
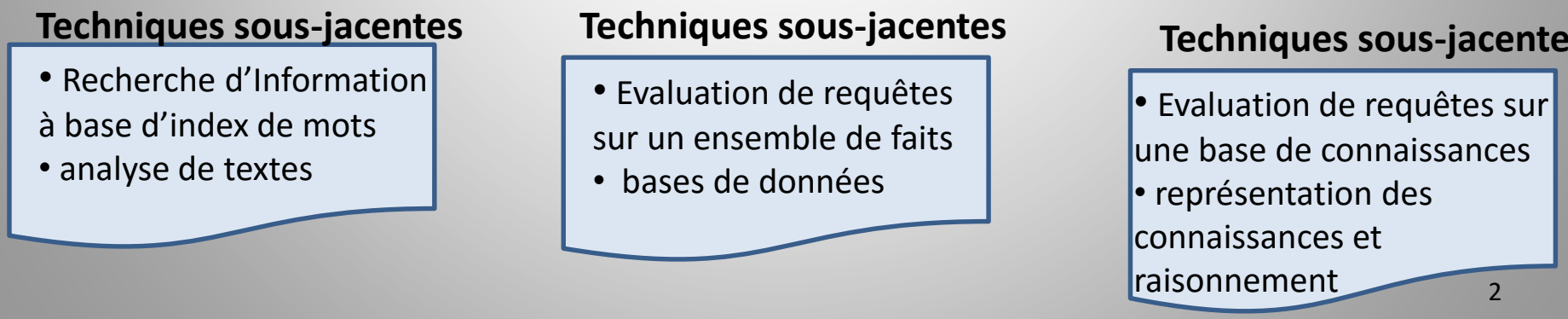
Laboratoire d'Informatique de Grenoble

Web sémantique: défis et enjeux

Marie-Christine Rousset
Laboratoire d'Informatique de Grenoble
Université Grenoble Alpes et Institut Universitaire de France

Problématique: l'évolution du web





Différents paradigmes: illustration

scientists born in Europe who received a Nobel Prize ?

Mots-clés : « scientists, born in Europe , Nobel Prize »

requête SPARQL

select **x** where <y, hasName , **x**>
<y, hasWonPrize, NobelPrize>
<y, type, Scientist>, <y, bornIn, Europe>

Web textuel → Web des données → Web des connaissances

Un ensemble ordonné de pages web susceptibles de contenir des éléments de réponses de la requête

The Royal Swedish Academy of ***Sciences*** awards the ***Nobel Prize*** in Physics, the **...** listen (help·info)) was ***born*** on 21 October 1833 in Stockholm, Sweden, into a **.....** The heavy focus on ***European*** and Swedish authors has been a subject of **...**

RDFS store (e.g. YAGO)

<http://.../marie-curie, hasName, « Marie Curie »>
<http://.../marie-curie, type, Scientist>
<http://.../marie-curie, hasWonPrize, NobelPrize>
<http://.../marie-curie, bornIn, Europe>
<http://.../einstein, hasName, « Albert Einstein »>
<http://.../einstein, type, Physicist>
<http://.../einstein, hasWonPrize, NobelPrize>
<http://.../einstein, birthPlace, Ulm>
<Physicist, subClassOf, Scientist>
<Ulm, locatedIn, Germany>,
<Germany, partOf, Europe>

Différents paradigmes: illustration

scientists born in Europe who received a Nobel Prize ?

Mots-clés : « scientists, born in Europe , Nobel Prize »

requête SPARQL

select **x** where <y, hasName , **x**>
<y, hasWonPrize, NobelPrize>
<y, type, Scientist>, <y, bornIn, Europe>

Web textuel → Web des données → Web des connaissances

Un ensemble ordonné de pages web susceptibles de contenir des éléments de réponses de la requête

The Royal Swedish Academy of ***Sciences*** awards the ***Nobel Prize*** in Physics, the **...** listen (help·info)) was ***born*** on 21 October 1833 in Stockholm, Sweden, into a **.....** The heavy focus on ***European*** and Swedish authors has been a subject of **...**

+ Extraction d'entités nommées

«Alfred Nobel », « Albert Einstein », « Albert Camus », « Marie Curie »

RDFS store (e.g. YAGO)

<http://.../marie-curie, hasName, « Marie Curie »>
<http://.../marie-curie, type, Scientist>
<http://.../marie-curie, hasWonPrize, NobelPrize>
<http://.../marie-curie, bornIn, Europe>
<http://.../einstein, hasName, « Albert Einstein »>
<http://.../einstein, type, Physicist>
<http://.../einstein, hasWonPrize, NobelPrize>
<http://.../einstein, birthPlace, Ulm>
<Physicist, subClassOf, Scientist>
<Ulm, locatedIn, Germany>,
<Germany, partOf, Europe>

Différents paradigmes: illustration

scientists born in Europe who received a Nobel Prize ?

Mots-clés : « scientists, born in Europe , Nobel Prize »

requête SPARQL

select **x** where <y, hasName , **x**>
<y, hasWonPrize, NobelPrize>
<y, type, Scientist>, <y, bornIn, Europe>

Web textuel → Web des données → Web des connaissances

Un ensemble ordonné de pages web susceptibles de contenir des éléments de réponses de la requête

The Royal Swedish Academy of ***Sciences*** awards the ***Nobel Prize*** in Physics, the **...** listen (help·info)) was ***born*** on 21 October 1833 in Stockholm, Sweden, into a **.....** The heavy focus on ***European*** and Swedish authors has been a subject of **...**

Réponses fausses

+ Extraction d'entités nommées

«Alfred Nobel », « Albert Einstein », « Albert Camus », « Marie Curie »

RDFS store (e.g. YAGO)

<http://.../marie-curie, hasName, « Marie Curie »>
<http://.../marie-curie, type, Scientist>
<http://.../marie-curie, hasWonPrize, NobelPrize>
<http://.../marie-curie, bornIn, Europe>
<http://.../einstein, hasName, « Albert Einstein »>
<http://.../einstein, type, Physicist>
<http://.../einstein, hasWonPrize, NobelPrize>
<http://.../einstein, birthPlace, Ulm>
<Physicist, subClassOf, Scientist>
<Ulm, locatedIn, Germany>,
<Germany, partOf, Europe>

Différents paradigmes: illustration

scientists born in Europe who received a Nobel Prize ?

{y → http://.../marie-curie,
x → « Marie Curie »}

Mots-clés : « scientists, born in Europe , Nobel Prize »

requête SPARQL

select **x** where <y, hasName , **x**>
<y, hasWonPrize, NobelPrize>
<y, type, Scientist>, <y, bornIn, Europe>

Web textuel → Web des données → Web des connaissances

Un ensemble ordonné de pages web susceptibles de contenir des éléments de réponses de la requête

The Royal Swedish Academy of ***Sciences*** awards the ***Nobel Prize*** in Physics, the **...** listen (help·info)) was ***born*** on 21 October 1833 in Stockholm, Sweden, into a **.....** The heavy focus on ***European*** and Swedish authors has been a subject of **...**

Réponses fausses

+ Extraction d'entités nommées

«Alfred Nobel », « Albert Einstein », « Albert Camus », « Marie Curie »

SPARQL Evaluation

RDFS store (e.g. YAGO)

<http://.../marie-curie, hasName, « Marie Curie »>
<http://.../marie-curie, type, Scientist>
<http://.../marie-curie, hasWonPrize, NobelPrize>
<http://.../marie-curie, bornIn, Europe>
<http://.../einstein, hasName, « Albert Einstein »>
<http://.../einstein, type, Physicist>
<http://.../einstein, hasWonPrize, NobelPrize>
<http://.../einstein, birthPlace, Ulm>
<Physicist, subClassOf, Scientist>
<Ulm, locatedIn, Germany>,
<Germany, partOf, Europe>

« Marie Curie »

Différents paradigmes: illustration

scientists born in Europe who received a Nobel Prize ?

{y → http://.../einstein,
x → « Albert Einstein »}

Mots-clés : « scientists, born in Europe , Nobel Prize »

requête SPARQL

select **x** where <y, hasName , **x**>
<y, hasWonPrize, NobelPrize>
<y, type, Scientist>, <y, bornIn, Europe>

Web textuel → Web des données → Web des connaissances

Un ensemble ordonné de pages web susceptibles de contenir des éléments de réponses de la requête

The Royal Swedish Academy of ***Sciences*** awards the ***Nobel Prize*** in Physics, the **...** listen (help·info)) was ***born*** on 21 October 1833 in Stockholm, Sweden, into a **.....** The heavy focus on ***European*** and Swedish authors has been a subject of **...**

Réponses fausses

+ Extraction d'entités nommées

«Alfred Nobel », « Albert Einstein », « Albert Camus », « Marie Curie »

RDFS store (e.g. YAGO)

<http://.../marie-curie, hasName, « Marie Curie »>
<http://.../marie-curie, type, Scientist>
<http://.../marie-curie, hasWonPrize, NobelPrize>
<http://.../marie-curie, bornIn, Europe>
<http://.../einstein, hasName, « Albert Einstein »>
<http://.../einstein, type, Physicist>
<http://.../einstein, hasWonPrize, NobelPrize>
<http://.../einstein, birthPlace, Ulm>
<Physicist, subClassOf, Scientist>
<Ulm, locatedIn, Germany>,
<Germany, partOf, Europe>

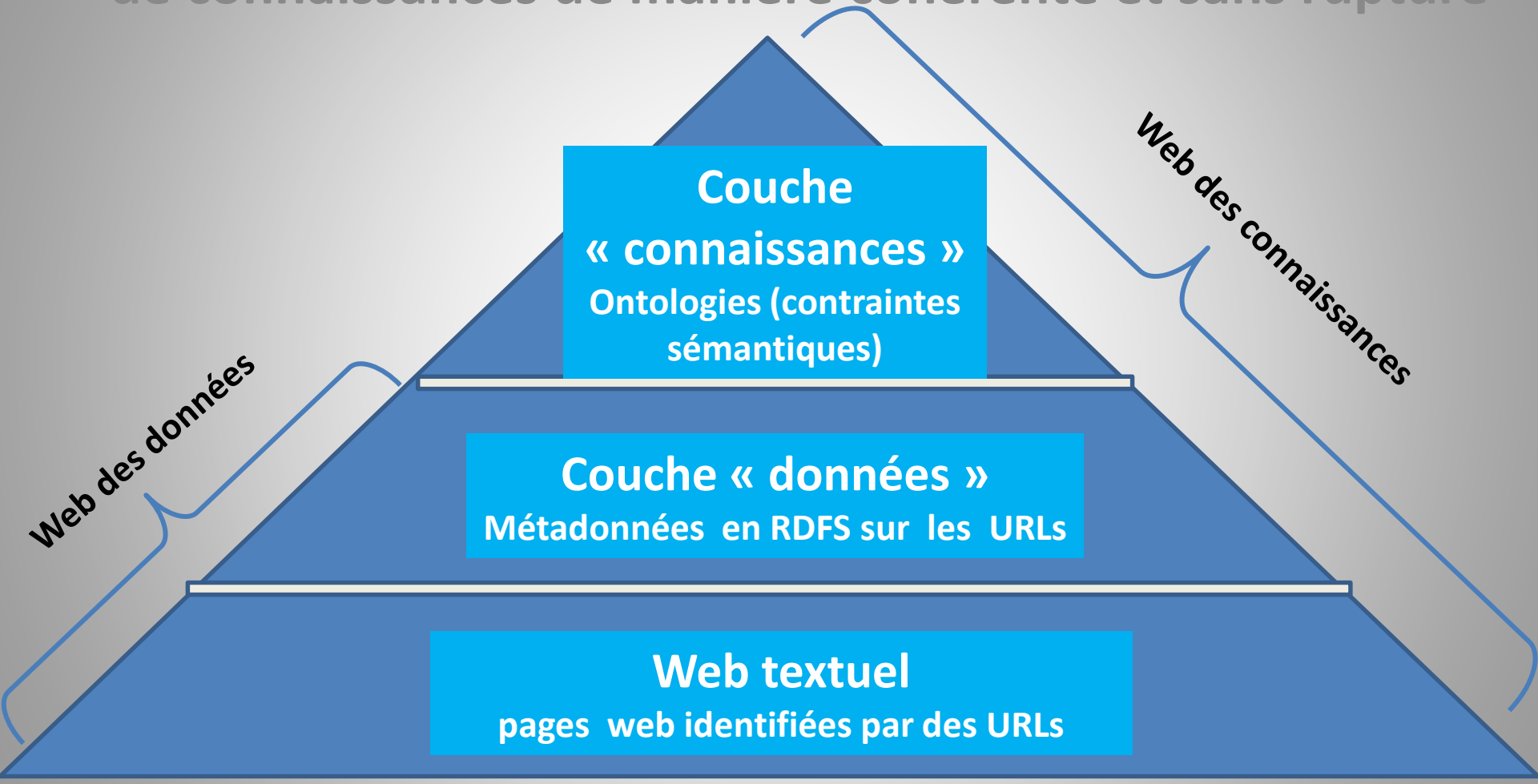
SPARQL Evaluation + inférence

« Marie Curie »

« Marie Curie », « Albert Einstein »

Une vision globale et cohérente

Le web est en train d'évoluer d'un web textuel à un web de connaissances de manière cohérente et sans rupture



Open Data (données ouvertes)

- Tendance à l'échelle mondiale de mettre à disposition des citoyens de plus en plus de données numériques produites par des services publics
  - par souci de transparence
  - en vue de réutilisation (croisement avec d'autres données, extraction d'informations)
- Principes de publication des données ouvertes :
  - doivent être publiées sous une **forme structurée** dans un format non propriétaire le plus « standard » possible pour faciliter l'**interopérabilité** entre machines et traitements.
  - selon une méthodologie et une licence garantissant son libre accès, sans restriction technique, juridique ou financière

Les standards à la base du Web sémantique

- http, URLs et espaces de noms (namespaces)
  - Pour dénoter et nommer de façon non ambigue des entités
    - URLs: Uniform Resource Locator
    - 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://rdf.insee.fr/sparql

Illustration: interrogation de DBpedia.fr avec SPARQL

- DBpedia:
  - 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 : http://fr.dbpedia.org
- http://fr.dbpedia.org/sparql
  - Quelles sont les communes d'Ile de France de plus de 100.000 habitants et leurs maires?

```
SELECT ?commune ?maire
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 ?maire
FILTER (?population > 100000) }
```

Illustration: interrogation de DBpedia.fr avec SPARQL

- DBpedia:
  - 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 : http://fr.dbpedia.org
- http://fr.dbpedia.org/sparql
  - Quelles sont les communes d'Ile de France de plus de 100.000 habitants et leurs maires?

commune	maire
http://fr.dbpedia.org/resource/Paris	"Anne Hidalgo"@fr
http://fr.dbpedia.org/resource/Boulogne-Billancourt	"Pierre-Christophe Baguet"@fr
http://fr.dbpedia.org/resource/Montreuil_(Seine-Saint-Denis)	"Patrice Bessac"@fr
http://fr.dbpedia.org/resource/Saint-Denis_(Seine-Saint-Denis)	"Didier Paillard"@fr
http://fr.dbpedia.org/resource/Val-de-Marne	http://fr.dbpedia.org/resource/Laurent_Cathala
http://fr.dbpedia.org/resource/Argenteuil_(Val-d'Oise)	"Georges Mothron"@fr

Enjeux

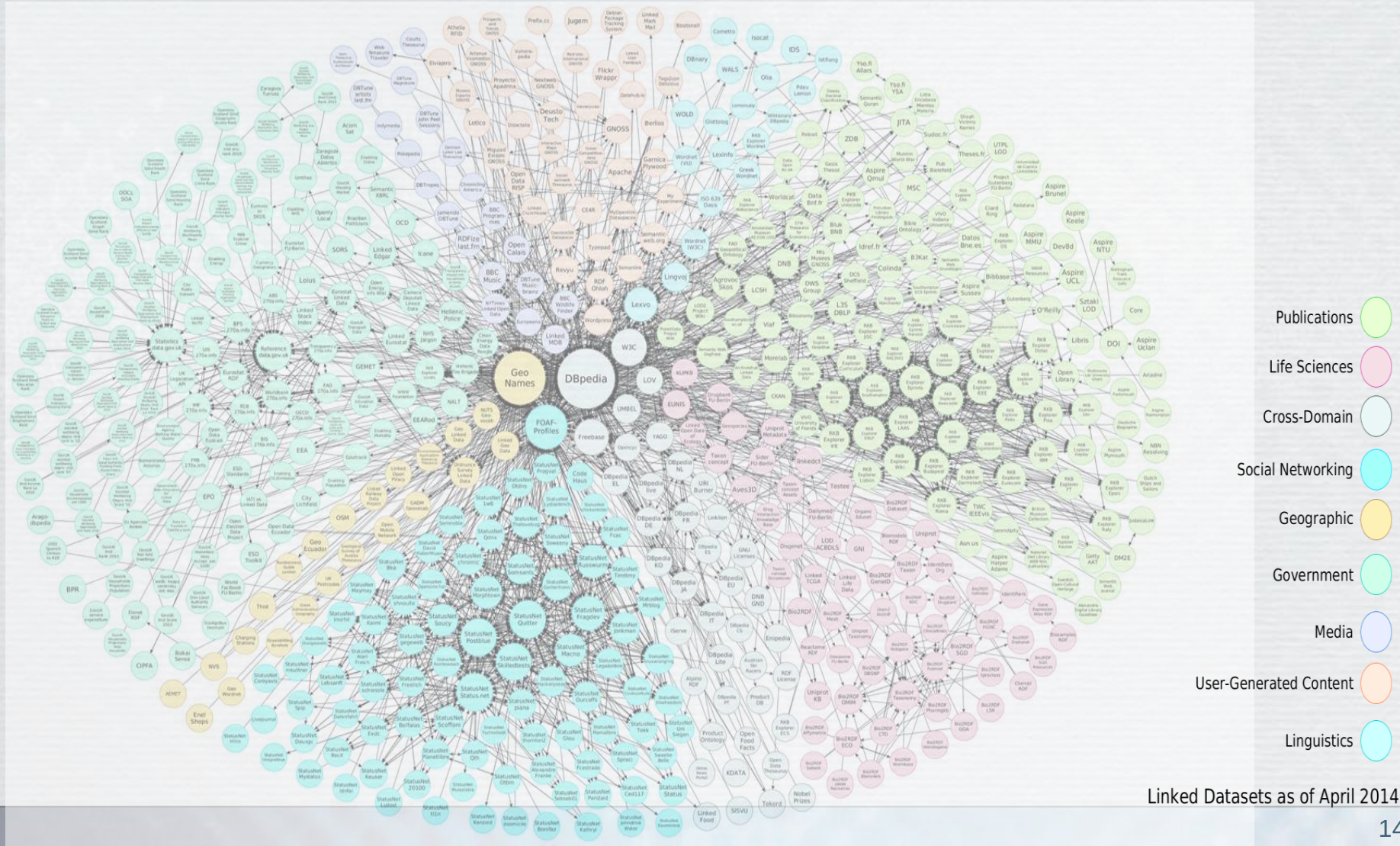
- La création de valeur ajoutée, de nouveaux usages, la découverte de nouvelles corrélations, passent par le croisement de données provenant de différentes sources hétérogènes et souvent séparées en **silos de données**

⇒ **Linked Data :** initiative du W3C et de Tim Berners-Lee pour promouvoir sa vision du **web sémantique**

  - où les données, distribuées à l'échelle du web, sont liées et peuvent être interrogées de façon automatique (par des humains mais aussi des applications informatiques) **quels que soient leur lieu de stockage et sans avoir à les dupliquer.**

Linked Open Data aujourd'hui

des milliers de sources de données RDF addressables sur le Web, des milliards de triplets



Différentes façons de déclarer des liens dans LOD

- 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

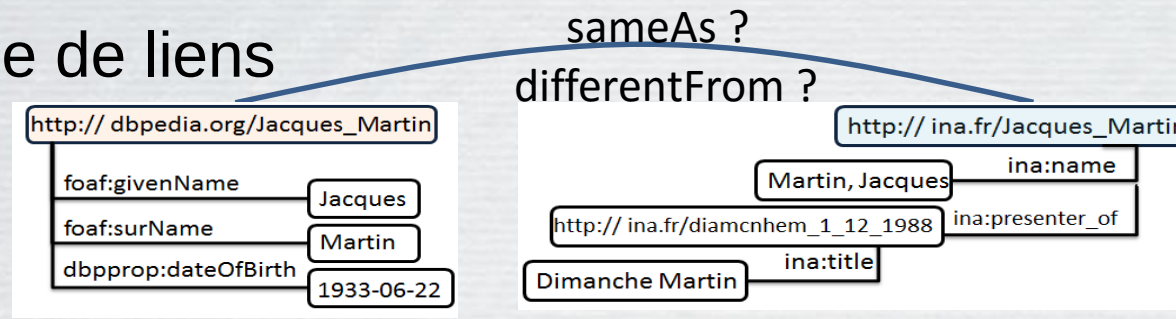
45 millions de faits sameAs déclarés dans DBpedia (avec d'autres sources de Linked Data).

Exemples d'applications développées à l'aide du Linked Data

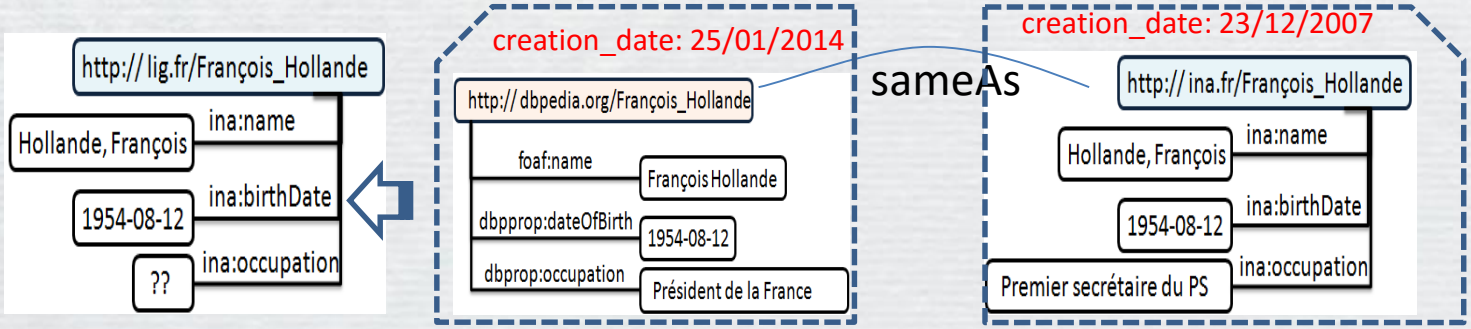
- BBC Olympics Data Service
  - pour créer à la volée du contenu intéractif sur les différents évènements en intégrant en temps réel un contenu riche (présent en format RDF dans la plateforme de la BBC) associé aux sportifs ou personnalités participant à tel ou tel évènement.
- EPA Linked Data
  - pour visualiser sur une carte et fournir rapidement des informations à la demande sur des installations susceptibles d'émettre des substances chimiques dans une certaine zone géographique, et croiser ces informations avec des rapports ou des études gouvernementales et faire des statistiques d'évolution dans le temps, etc …
  - nécessite de croiser des données de Environmental Protection Agency , de geonames, et de dat.gov

Défis du Linked Open Data

- Découverte automatique de liens
  - Par agrégation de similarités
  - Par inférence



- Fusion de données



- Qualité des données