



Sophia'Conf 2014

Bases de données RDF –
versatilité, puissance et scalabilité

Agenda

Introduction

- Vision du Web sémantique par Atos

Bases de données RDF

- Modèle W3C vs Vision entreprise
- Triples stores
- SPARQL – un langage et un protocole
- Triples stores – points forts, points faibles

Retours d'expérience

- Mise en œuvre #1 : BnF SPAR
- Mise en œuvre #2 : SIAF Thesaurus W
- Mise en œuvre #3 : Datalift
- Mise en œuvre #4 : Agence Européenne de l'Environnement

Vision du Web sémantique par Atos

Premières mises en œuvre dès 2009

Technologie stratégique

- **Publication de données**
→ Linked Open Data & référentiels d'entreprise
- **Croisement de données hétérogènes**

Participation aux projets de recherche

- Socle technique & intégration
- Pérennisation & réutilisation
→ approche **open-source**

Approche Big Data

- **Projet Waves** : croisement de flux de données issues de réseaux de capteurs
- **Projet TriSHaPeD** : Triple Store Haute Performance Distribuée

Modèle W3C vs Vision entreprise

W3C

- ▶ 1 objet = 1 URI (URL) = 1 document
- ▶ Documents sur le web, se référant les uns les autres
- ▶ RDF : aucune contrainte de contenu

Vision entreprise

- ▶ Entrepôts de données centralisés
- ▶ Vocabulaires contrôlés (schéma SQL, XML...)
- ▶ Langage de requêtage
- ▶ Accès externes limités (réseaux fermés, DMZ, proxies...)
- ▶ Contrôle d'accès aux données

Documents vs Graphe local

RDF/XML – Données chez AtoS

```

▶ <?xml version="1.0"?>
▶ <rdf:RDF
  ▶ xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  ▶ xmlns:adc="http://atos.net/metadata#"
  ▶ xmlns:hp="http://hp.com/techdata#">
  ▶ <Description about="http://atos.net/adc/superdome453">
    ▶ <adc:localisation>SalleRenoir</adc:localisation>
    ▶ <adc:type>http://hp.com/techdata/HPIS9000
      </adc:type>
  ▶ </Description>
  ▶ <Description about="http://atos.net/adc/SalleRenoir">
    ▶ <adc:temperature>16° C</adc:temperature>
    ▶ <adc:climatisation>on</adc:climatisation>
  ▶ </Description>
▶ </rdf:RDF>

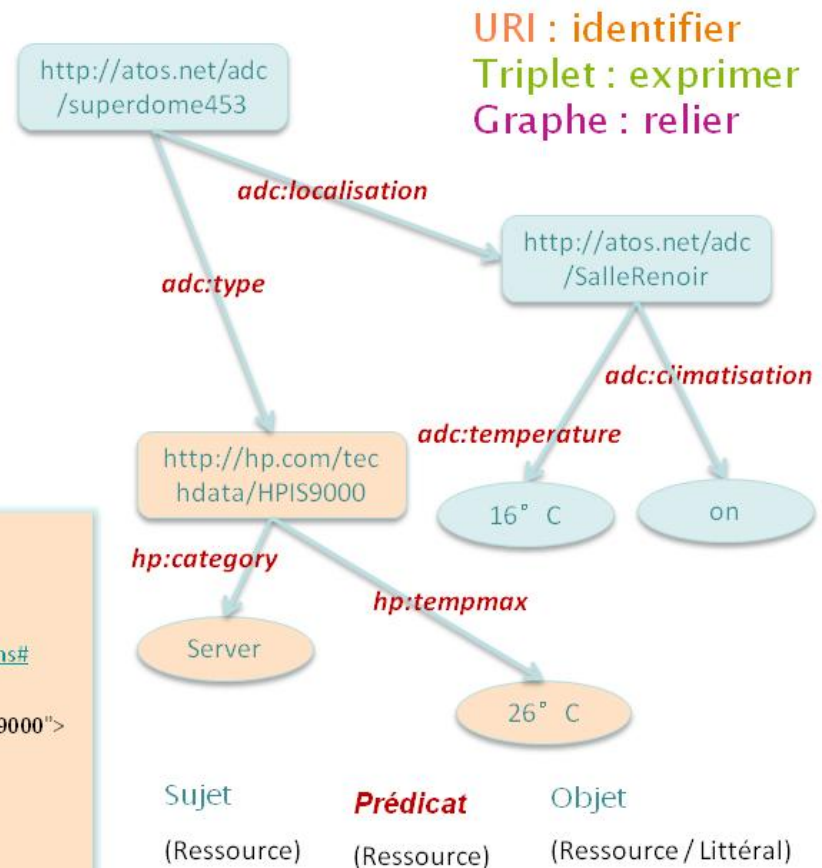
```

RDF/XML – Données chez HP

```

▶ <?xml version="1.0"?>
▶ <rdf:RDF
  ▶ xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  ▶ xmlns:hp="http://hp.com/techdata#">
  ▶ <Description about="http://hp.com/techdata/HPIS9000">
    ▶ <hp:category>server</hp:category>
    ▶ <hp:tempmax>26° C</hp:tempmax>
  ▶ </Description>
▶ </rdf:RDF>

```



Triple stores, les bases de données RDF

- ▶ **Stockage optimisé**
 - Index (jusqu'à 15 !) & compression (1 URI = 1 entier)
- ▶ **Natif** (Sesame, OWLIM, Virtuoso, AllegroGraph, BigData, Jena (Fuseki / TDB)...))
- ▶ **Ou relationnel** (Sesame, Oracle, Jena (SQL DB)...))

- ▶ **Scalables**
 - Plusieurs milliards de triples
- ▶ **Schema-less**
- ▶ **Support de l'inférence** (RDFS / OWL)
- ▶ **Partitionnement des données**
 - Graphes nommés

- ▶ **Langage de requêtage & manipulation évolué & standard : SPARQL**

SPARQL – Un langage et un protocole

« SPARQL Protocol And RDF Query Language »

► Défini par le W3C

- SPARQL 1.0 (2008) : lecture seule
- SPARQL 1.1 (2013) : modification des données

► Points d'accès (endpoints) normalisés

- HTTP (REST)
- SOAP
- Fédération
- Graph Access Protocol

► Extensions

- GeoSPARQL

The screenshot shows a web-based SPARQL query editor. At the top, it says 'Editeur de requête SPARQL'. Below that, there are tabs for 'Données publiées', 'Format de la réponse', and several output formats: HTML, RDF/XML, N3/Turtle, NTriples, TriG, TriX, and CSV. The 'Requête' field contains a SPARQL query:

```
SELECT ?auteur ?serie ?resume WHERE
{
  ?s1 glenat:serie ?serie ;
    glenat:Auteur ?auteur .
  ?s2 rdfs:label ?nom ;
    dbp:abstract ?resume .
  FILTER(STR(?serie) = STR(?nom))
  FILTER(LANGMATCHES(LANG(?nom), "fr"))
  FILTER(LANGMATCHES(LANG(?resume), "fr"))
}
```

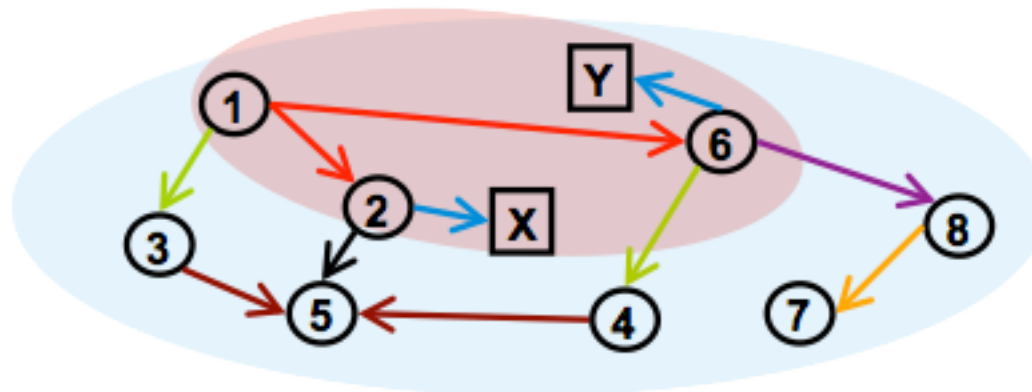
Below the query, there are buttons for 'Requêtes prédéfinies' (Toutes les entrées, Sujets, Types, Contextes, Exemple 1) and a 'Nombre maximum de résultats' field set to 5000. An 'Exécuter la requête' button is on the right. The results are displayed in a table with columns 's', 'p', and 'o'.

s	p	o
http://localhost:9091/paris/kiosques/1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://localhost:9091/paris/kiosques#kiosques-ouverts-a-
http://localhost:9091/paris/kiosques/1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2006/vcard/ns#VCard
http://localhost:9091/paris/kiosques/1	http://localhost:9091/paris/kiosques#adresse	"PCE DU CHATELET"
http://localhost:9091/paris/kiosques/1	http://localhost:9091/paris/kiosques#arrdt	"75001"
http://localhost:9091/paris/kiosques/1	http://localhost:9091/paris/kiosques#ouv	"08:30:00.0000^^"
http://localhost:9091/paris/kiosques/1	http://localhost:9091/paris/kiosques#ferm	"20:00:00.0000^^"
http://localhost:9091/paris/kiosques/1	http://localhost:9091/paris/kiosques#ouv-dim	true
http://localhost:9091/paris/kiosques/1	http://www.w3.org/2006/vcard/ns#adr	_:node17e90b9uox1
http://localhost:9091/paris/kiosques/1	http://www.w3.org/2006/vcard/ns#fn	"Kiosque PCE DU CHATELET"
http://localhost:9091/paris/kiosques/1	http://www.w3.org/2006/vcard/ns#geo	_:node17e90b9uox1

At the bottom, there is a pagination bar showing 'Page 1 of 1' and a total of 200 results.

SPARQL – Requêtage sur graphes

Soit le graphe suivant enregistré dans un triple store :



SPARQL permet d'extraire un sous-ensemble de ce graphe par expression de contraintes sous la forme d'équations

Exemple :

Je cherche les ressources liées à 1 par prédicat « rouge » et les chaînes de caractères liée à ces ressources par le prédicat « bleu » :

1. Les ressources liées à 1 par le prédicat « rouge » : `<1> <rouge> ?ressource`
2. Les chaînes liées à ces ressources par le prédicat « bleu » : `?ressource <bleu> ?string`

Triple stores – Points forts, points faibles



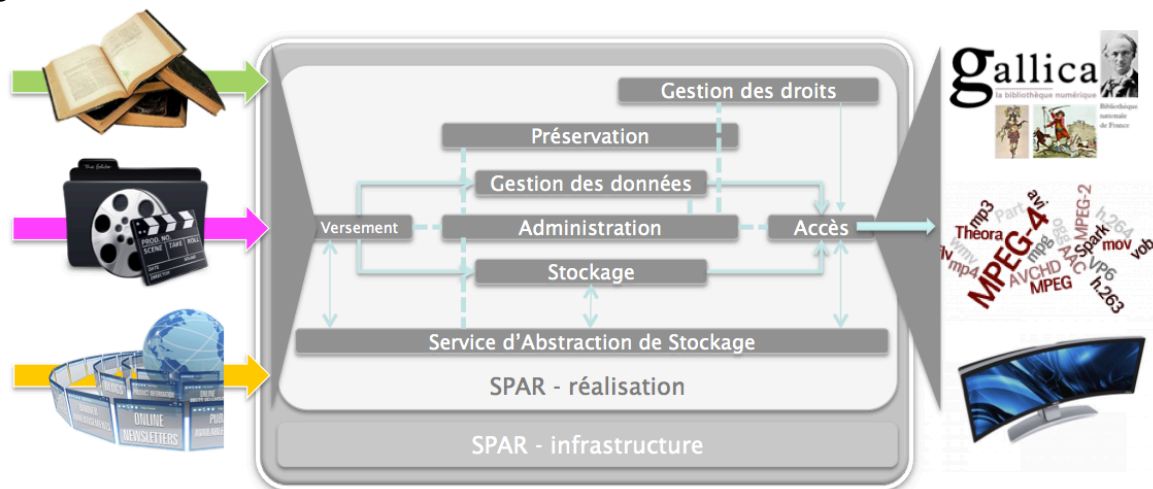
- Pas de schéma prédéfini
- Découverte de la donnée
- Interopérabilité
(langage & protocole)

- Temps de réponse
- Occupation mémoire
- Pas de partitionnement naturel
des données (distribution)
- Pas de limite à la complexité des
requêtes
→ Bases répliquées en cluster
actif/actif (2 nœuds)
- Inférence : à double tranchant

Mise en œuvre #1 : BnF SPAR

Systeme de Préservation et d'Archivage Réparti

- ▶ **Objectif : préservation du patrimoine numérique sur 50 ans**
 - Modèle OAIS (*Open Archival Information System*)
 - Stockage des données sur bandes magnétiques (robots), bi-site
- ▶ **Indexation des métadonnées techniques en RDF**
 - 3 bases (complet, sélection, référence)
 - Reconstructibles à tous instant (disque ou bandes)
 - Plusieurs milliards de triples
 - 1 millions de graphes
 - 250 To de documents (x2)



Mise en œuvre #2 : SIAF Thesaurus W

Service Interministériel des Archives de France

<http://data.culture.fr/thesaurus/page/vocabulaires>

► Triple store + frontal de publication

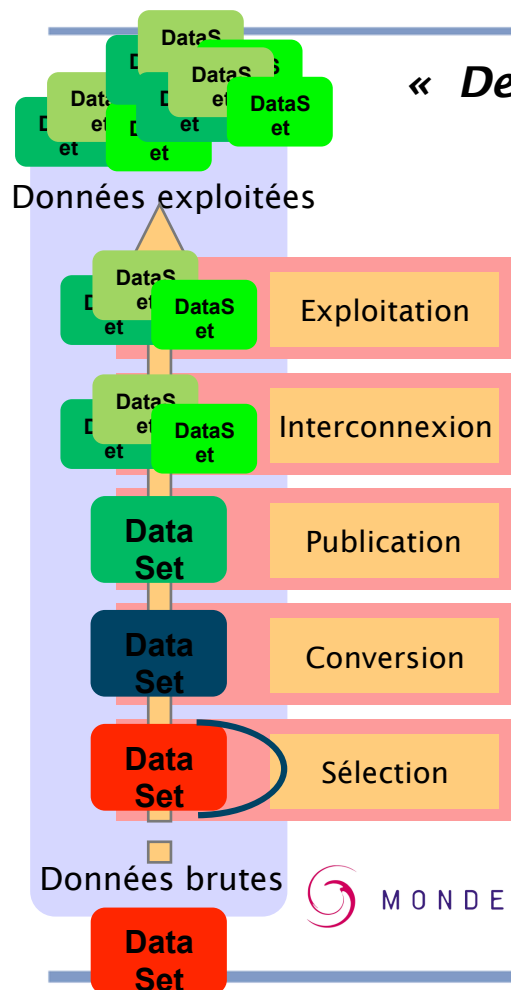
- 1 thesaurus SKOS = 1 graphe
- Publication par suppression et création de graphes

► Frontal

- URI non déréférençables (ARK), préfixées de l'URL du site
- Négociation de contenu HTML ou RDF (XML, Turtle, N3) par redirection
- GET URL → SPARQL DESCRIBE
- 1 modèle de page HTML par type SKOS

The screenshot shows the website 'LES VOCABULAIRES DU MINISTÈRE DE LA CULTURE ET DE LA COMMUNICATION'. It features a search bar with the text 'Chercher le concept' and a 'Chercher' button. Below the search bar, there is a list of available thesauruses with blue links: 'Thésaurus-matières pour l'indexation des archives locales', 'Liste d'autorité « Typologie documentaire » pour l'indexation des archives locales', 'Thésaurus de la désignation des oeuvres architecturales et des espaces aménagés', 'Liste d'autorité « Contexte historique » pour l'indexation des archives locales', and 'Liste d'autorité « Actions » pour l'indexation des archives locales'. At the bottom, there are links for 'En savoir plus', 'SPARQL endpoint', and 'Site du producteur'.

Mise en œuvre #3 : Datalift



« De la donnée brute à la donnée sémantique interconnectée »

- ▶ **Plate-forme open source**
- ▶ **Construction**
 - Transformation CVS, SQL, XML, ShapeFile... → RDF
 - Mapping vers des vocabulaires catalogués (<http://lov.okfn.org/>)
 - Interconnexion (SILK)
- ▶ **Publication**
 - SPARQL endpoints
 - Négociation de contenu
 - Contrôle d'accès (S4AC)
- ▶ **Propulse** : <http://data.insee.fr> & <http://data.ign.fr>

MONDECA

EURECOM
Sophia Antipolis

INSEE

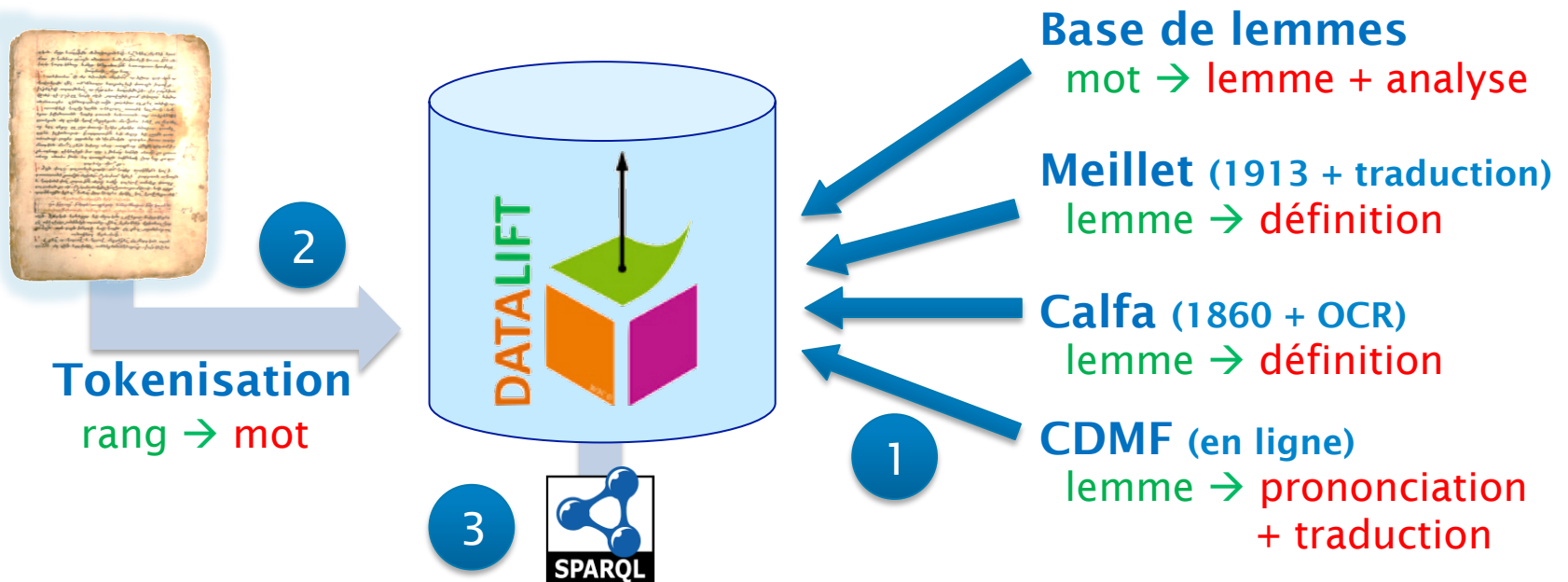
Inria
INVENTEURS DU MONDE NUMÉRIQUE

LIRMM

IGN
INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET FORESTIÈRE

fing

Mise en œuvre #3 : Assistance à la traduction de l'arménien hellénistique



76 լինին	M: (§ 117 a), je suis, εἶμι.			
76 լինին	C: ես, է, էի, էիր, էր, էր, իցեմ, եղէ v. auxil. être ; սիրեցից զսա մինչեւ եւս, je l'aimerai tant que je vivrai ; յուսել -, je suis en train de manger ; զի			
77 այսպէս	adv. ainsi, comme ceci adv.			
77 այսպէս	M: (adv. § 25, 37 C b et § 162), ainsi, de cette façon (այս + պէս).			
77 այսպէս	C: adv. ainsi, comme cela, de même, de cette façon, manière, sorte ; որպէս ... -, comme... ainsi, de même que.			
78 եւ	conj. et conj.			
78 եւ	M: (§ 164), aussi, et, même.			
78 եւ	C: conj. et ; même ; ոչ մին եւ ոչ միւսն, ni l'un ni l'autre ; - այլն ad v. et caetera ; - ալի conj. or, donc ; եւ, եւ, et, et ; - զի, - քանզի conj. puisque.			
79 այլադանորն				
80 մարմին	noun.nom.acc.loc.sg. corps nm.			
80 մարմին	M: մարմնայ corps, chair, σῶμα.			
80 մարմին	C: մեց sm. corps ; chair ; corps, cadavre ; homme, personne ; corps, ensemble ; corps, consistance ; - առնուլ, prendre chair, s'incarner.			

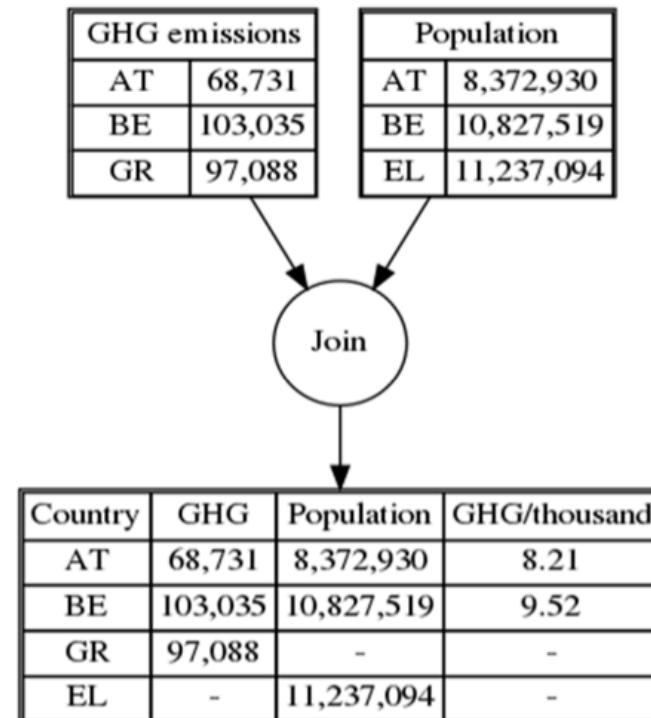
Mise en œuvre #4 : Agence Européenne de l'Environnement

Avant

Formats propriétaires

locality	Field2	Field3	Field4
[tsieb010] - BIP pro Kopf in KK3			
BIP pro Kopf in Kaufkraftstanda			
geo\time	1995		1996
EU (27 Länder)	100		100
EU (25 Länder)	105		105
EU (15 Länder)	116		116
Euroraum (16 Länder)	114		114
Euroraum (15 Länder)	116		115
Belgien	129		126
Bulgarien	32		28
Tschechische Republik	73		75
Dänemark	132		133
Deutschland	129		127
Estland	36	(b)	38
Irland	103		108
Griechenland	84		84
Spanien	92		92
Frankreich	116		115

Référentiels hétérogènes



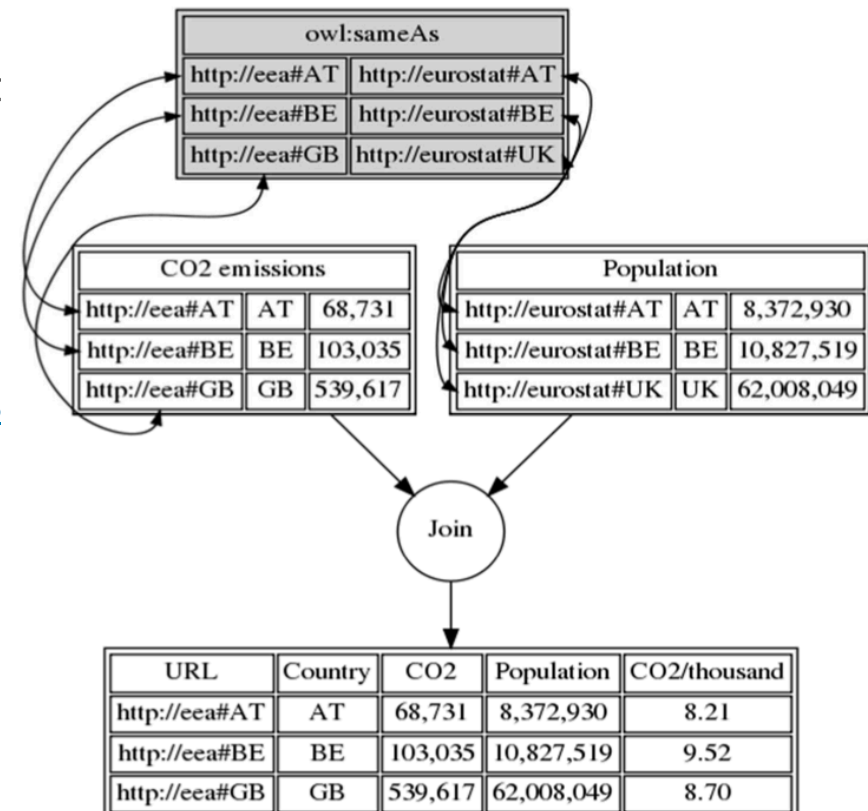
Mise en œuvre #4 : Agence Européenne de l'Environnement

Migration

- ▶ Données états membres en RDF uniquement
 - Vocabulaires et référentiels libres
- ▶ Ontologie de mise en correspondance des référentiels

<http://eurostat.europa.eu/countries#UK> =
<http://eea.europa.eu/countries.rdf#GB>
- ▶ Triple store supportant l'inférence
owl:sameAs
- ▶ Interrogation SPARQL

Après



Merci

For more information, please contact:

Laurent BIHANIC
laurent.bihanic@atos.net

Atos France
River Ouest
80, quai Voltaire
95877 Bezons Cedex

atos.net

Atos, the Atos logo, Atos Consulting, Atos Worldline, Atos Sphere, Atos Cloud and Atos WorldGrid are registered trademarks of Atos SA. July 2011

© 2011 Atos Consulting. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

Your business technologists. **Powering progress**

