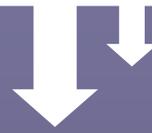


**Take your content further**  
Extend visibility || Increase relevance || Drive revenue



# Publication de données liées et réutilisation de vocabulaires

semweb.pro



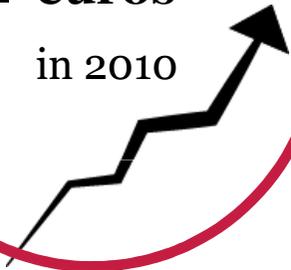
# PLAN

- 1. QUI SOMMES-NOUS ? MONDECA**
- 2. PROBLEMATIQUE : LA PUBLICATION DE DONNEES**
- 3. LA REUTILISATION DES VOCABULAIRES AVEC LE LOV**
- 4. OUTILS DE SEMANTISATION DES DONNEES**
- 5. ALIGNEMENT DE VOCABULAIRES**

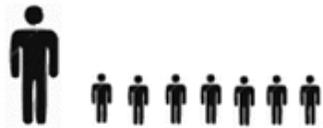
SOCIETE

# Vue d'ensemble

**2** millions  
euros  
in 2010



**18** people

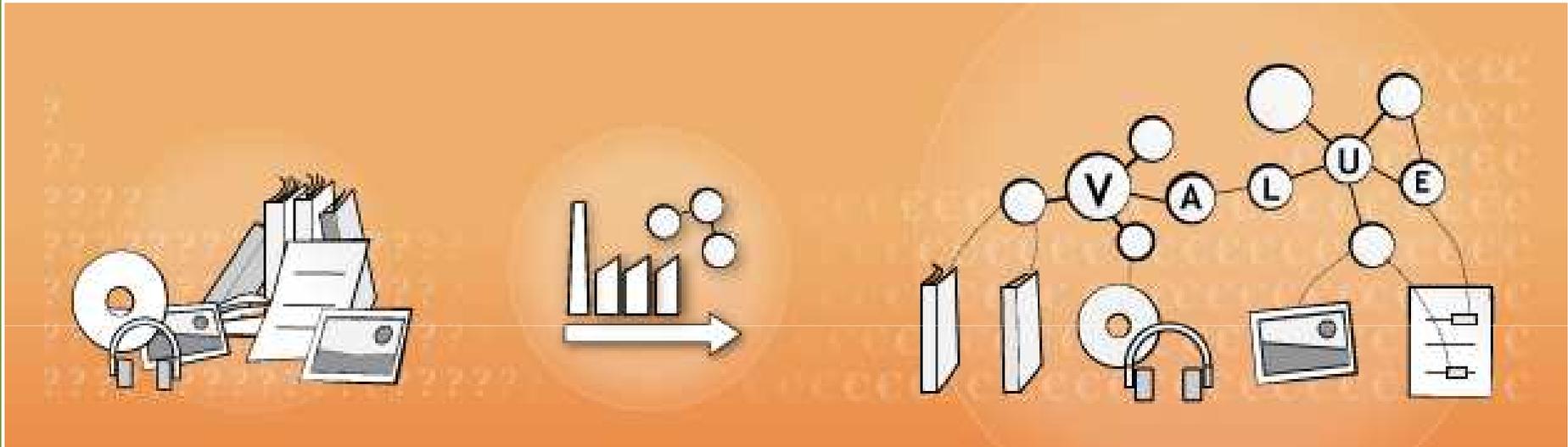


Gartner, 2011

**Taxonomy & ontology  
management**

SOCIETE

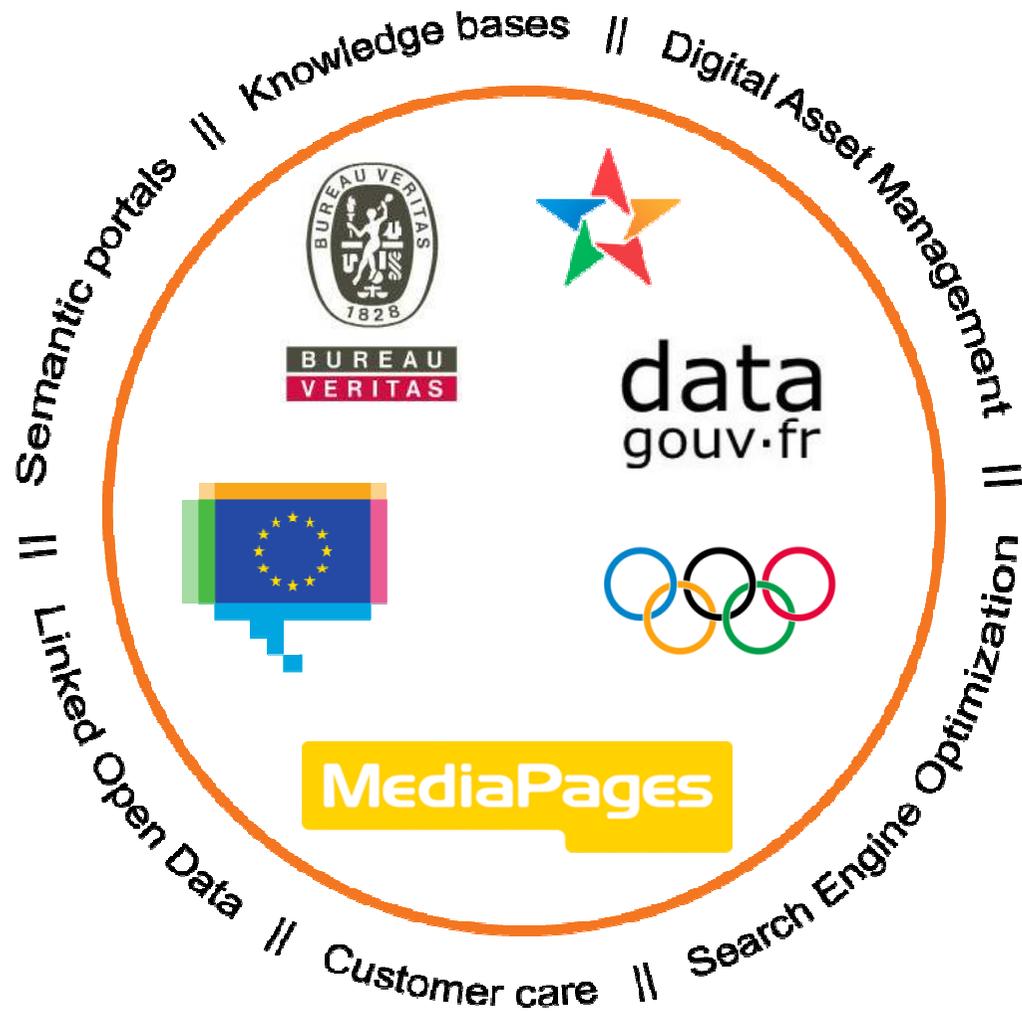
## Smart Content : la vision de Mondeca



- Valoriser toutes les informations et contenus
- Révéler les éléments clefs
- Agréger les contenus et intégrer de l'information complémentaire
- Mettre en réseau les contenus pour offrir richesse et performance

SOCIETE

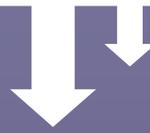
>60 références client



# Le projet Datalift



- Partners : INRIA EXMO, INRIA Edelweiss, ATOS, IGN, INSEE, EURECOM, FING, INSEE
- Objective: to develop a platform to publish and interlink datasets on the Web of Data
- Summary: Datalift will both publish datasets coming from a network of partners and data providers and propose a set of tools for easing the datasets publication process. The project will provide tools allowing to facilitate each step of the publication process:
  - selecting ontologies for publishing data
  - converting data to the appropriate format (RDF using the selected ontology)
  - publishing the linked data
  - interlinking data with other data sources
- Mondeca's role: data conversion tools, import interfaces, open data publishing, mapping wit LOD



# PLAN

1. QUI SOMMES-NOUS ? MONDECA
2. PROBLEMATIQUE : LA PUBLICATION DE DONNEES
3. LA REUTILISATION DES VOCABULAIRES AVEC LE LOV
4. OUTILS DE SEMANTISATION DES DONNEES
5. ALIGNEMENT DE VOCABULAIRES

## Je dois publier des données...

- Quelles sont les questions à me poser ?
- Comment modéliser mon jeu de données ?
- Comment mettre au bon format mes données pour les publier ?
- Exemple support : résultats du premier tour de l'élection présidentielle publiés par data.gouv.fr
  - Data.gouv.fr : chercher « élections », et prendre les données par départements
  - Modélisation du jeu de données
  - Transformation du jeu de données
  - S'appliquerait aussi à la question de la réutilisation des données ouvertes !

TITRE PREMIERE PARTIE

# Data is King



# Data

*Variante : « Content is King, but Data is God »*

## Publier ses données – pourquoi ?

- SEO : [schema.org](http://schema.org)
- Pour favoriser l'émergence d'un écosystème autour des données : feedback, curateurs, utilisateurs...
  - Pour permettre leur réutilisation et attirer de nouveaux clients
  - Pour faciliter leur mise en relation avec d'autres données et augmenter leurs chances d'être trouvée
  - Pour permettre à d'autres données de les référencer et favoriser leur accès
- Pour faciliter la réutilisation des données par plusieurs applications, en interne, dans l'entreprise

## Publier ses données sur le web (de données)

1. Quelle Licence ?
2. Quelle Modélisation ? (et quels vocabulaires réutiliser ?)
3. Quels Identifiants ?
4. Quels Liens avec d'autres données ?
5. Quel Format ?
6. Quel Mécanisme de publication ?
7. Quelle Evolution dans le temps ?

# Les identifiants, cas de l'INSEE

- Article dans « Documentaliste, Sciences de l'information », décembre 2011, dossier sur le Web sémantique
- INSEE : publication de nomenclatures officielles
  - Attribution de « codes » aux entités
  - Activités, produits, services, etc.
  - Code Officiel Géographique (COG)
    - Découpage administratif et statistique du territoire
- Code d'une commune
  - 05065 : commune de Guillestre
  - Valable uniquement dans un contexte où l'on sait que c'est la valeur d'un code commune
- Pour la publication des données
  - Génération d'URI à partir du code
  - <http://data.insee.fr/geo/Commune/05065>
  - Génération des données facilitée
  - Réutilisation des données facilitée, pour des applications qui s'appuient déjà sur les codes

## Le format, cas de la SNOMED

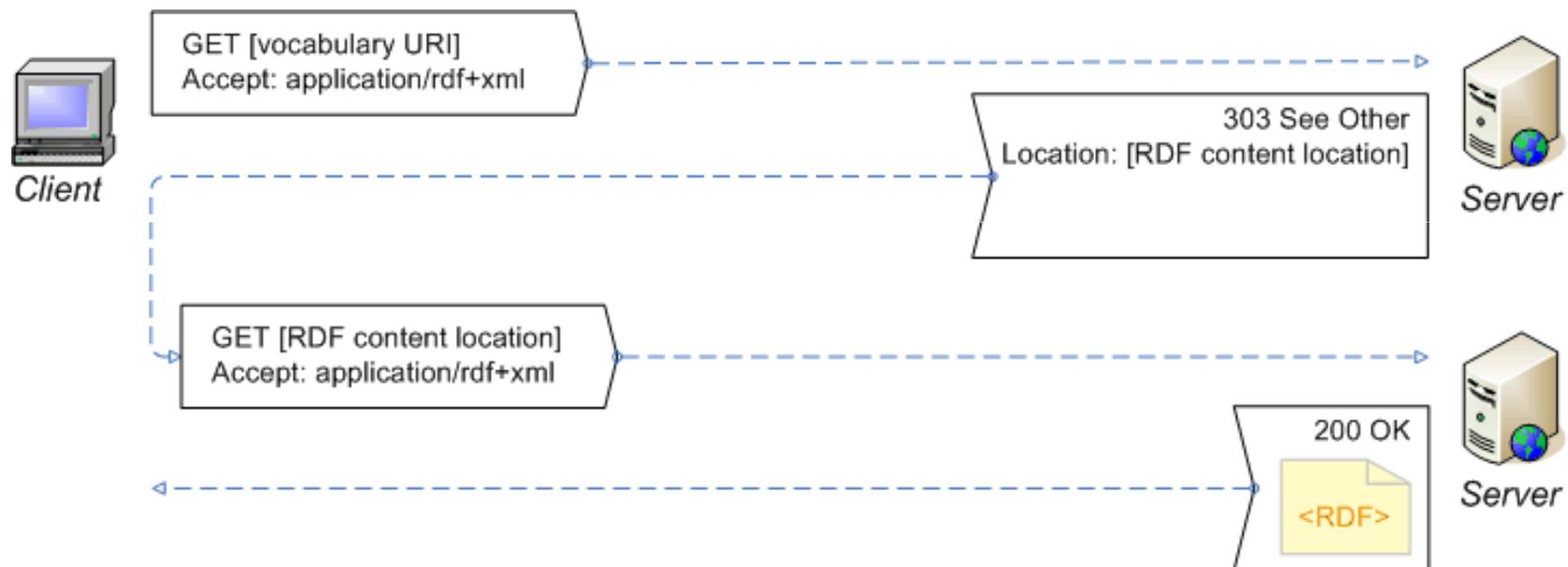
	A	B	C	D	E
1	LIGN	TERMCODE	FMOC	FCLASS	FNOMEN
2	2	D0-00000			Chapitre 0 Maladies de la peau et des tissus sous-cutanés
3	3	D0-00000		-	Section 0-0 Maladies de la peau et des tissus sous-cutanés: termes généraux, types histologiques et infections
4	4	D0-00000		0	0-00 Maladies de la peau et des tissus sous-cutanés: termes généraux et types histologiques
5	5	D0-00000		00	0-000 Maladies de la peau et des tissus sous-cutanés: termes généraux
6	6	D0-00000		01	maladie de la peau et du tissu sous-cutané
7	7	D0-00004		01	maladie de la peau

- Tableau Excel à sémantique ambiguë
  - 1 ligne par terme
  - Mais plusieurs fois le même « TERMCODE »...
  - Sans explication dans la documentation
- Il faut désambigüiser en fonction de l'ordre d'apparition dans le tableau...
- Un format de publication sémantique en RDF aurait levé toute ambiguïté sur les identifiants des concepts

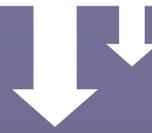
# La publication : le cas de service-public.fr

- Répertoire de l'administration française
  - Géré par la DILA (Direction de l'Information Légale et Administrative)
- Problématique d'identifiants : quel identifiant utiliser pour identifier les services ?
- Problématique de publication :
  - Quelles données publier ?
  - Sous quels formats ?
    - RDF pour les données brutes
    - HTML pour l'internaute
    - XML pour des services partenaires
  - Une bonne solution serait : mécanisme de négociation de contenu pour que chaque type d'utilisateur accède au format approprié

# La négociation de contenu



- <http://validator.linkeddata.org>



# PLAN

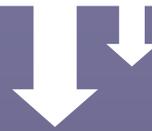
- 1. QUI SOMMES-NOUS ? MONDECA**
- 2. PROBLEMATIQUE : LA PUBLICATION DE DONNEES**
- 3. LA REUTILISATION DES VOCABULAIRES AVEC LE LOV**
- 4. OUTILS DE SEMANTISATION DES DONNEES**
- 5. ALIGNEMENT DE VOCABULAIRES**



## Réutilisation des vocabulaires et LOV

---

*Où l'on bascule vers un autre jeu de transparents !*



# PLAN

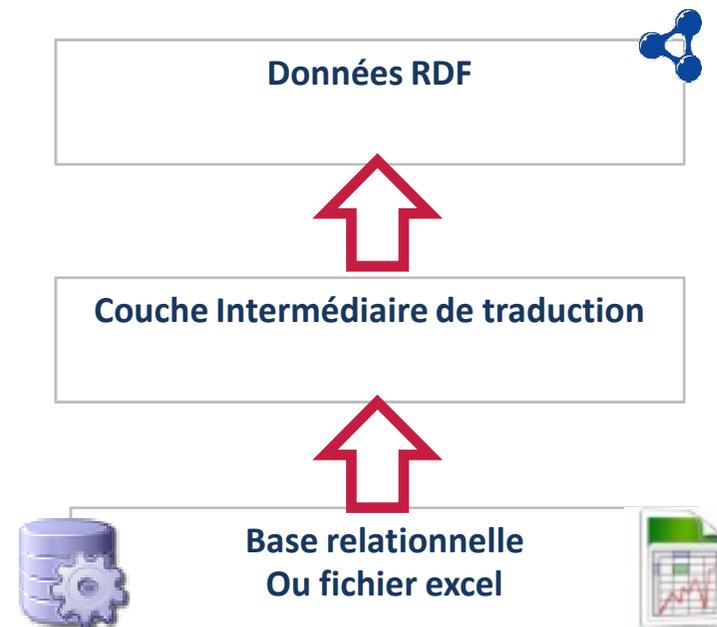
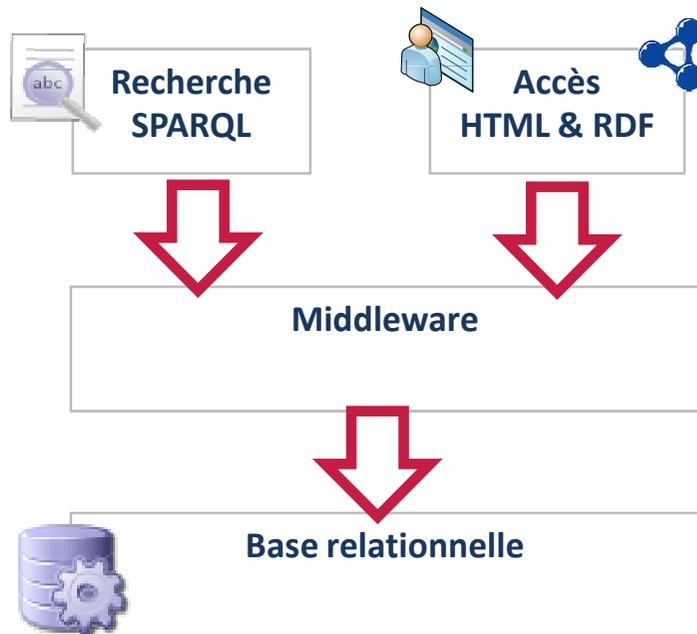
- 1. QUI SOMMES-NOUS ? MONDECA**
- 2. PROBLEMATIQUE : LA PUBLICATION DE DONNEES**
- 3. LA REUTILISATION DES VOCABULAIRES AVEC LE LOV**
- 4. OUTILS DE SEMANTISATION DES DONNEES**
- 5. ALIGNEMENT DE VOCABULAIRES**

## Formater ses données pour le web (de données)

- Sources de données actuelles :
  - Bases de données
  - Fichiers Excel ou CSV
- Problématique : comment passer de données dans des bases fermées à des formats sémantiques ?

## Vers des données dans des formats sémantiques

- Accéder aux données via un middleware
- Traduire les données dans des formats « sémantiques »



## D2RQ: Accès à une base relationnelle en SPARQL

- <http://d2rq.org> : système d'accès à une base relationnelle avec les technologies du web sémantique
  - approche « middleware » : accéder à une base relationnelle via SPARQL
  - Ou approche « transformation » : transformer une base relationnelle en RDF
- Open-Source
- D2RQ Mapping Language
- Intérêt : *transformer le contenu de bases relationnelles en RDF*

# D2RQ Mapping Language

```
# D2RQ Namespace @prefix
d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
# Namespace of the ontology
@prefix : <http://annotation.semanticweb.org/iswc/iswc.daml#> .

# Namespace of the mapping file; does not appear in mapped data
@prefix map: <file:///Users/d2r/example.ttl#> .

# Other namespaces
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

map:Database1 a d2rq:Database;
d2rq:jdbcDSN "jdbc:mysql://localhost/iswc";
d2rq:jdbcDriver "com.mysql.jdbc.Driver";
d2rq:username "user";
d2rq:password "password";
.
```

# D2RQ Mapping Language (continued)

```
# -----  
# CREATE TABLE Conferences (ConfID int, Name text, Location text);  
  
map:Conference a d2rq:ClassMap;  
d2rq:dataStorage map:Database1;  
d2rq:class :Conference;  
d2rq:uriPattern "http://conferences.org/comp/confno@@Conferences.ConfID@";  
.  
  
map:eventTitle a d2rq:PropertyBridge;  
d2rq:belongsToClassMap map:Conference;  
d2rq:property :eventTitle;  
d2rq:column "Conferences.Name";  
d2rq:datatype xsd:string;  
.  
  
map:location a d2rq:PropertyBridge;  
d2rq:belongsToClassMap map:Conference;  
d2rq:property :location;  
d2rq:column "Conferences.Location";  
d2rq:datatype xsd:string;  
.
```

## DSPL : visualiser ses données

- DataSet Publishing Language
  - <https://developers.google.com/public-data>
- Format supporté par Google
- Ajout de métadonnées sur un fichier CSV permettant de le visualiser dans le portail <http://google.com/publicadata>
- Pas en phase avec les standards du web sémantique (RDF, OWL)
- Orienté visualisation
- Nécessite que les données soient bien structurées avant publication
- Intérêt : *avoir rapidement des visualisations intéressantes sur les données*



## DSPL : exemple

---

*Où l'on voit l'exemple en direct !*

## Google Refine with RDF extension

- <http://code.google.com/p/google-refine/>
- Un outil permettant de faire des manipulations sur des jeux de données
  - ~ Excel orienté data
- Extension RDF proposée par le laboratoire DERI
  - <http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/>
  - Permet de « réconcilier » les valeurs avec celles d'un endpoint SPARQL (exemple : DBPedia)
- Intérêt : *faire des manipulations manuelles sur un jeu de données*

# Google Refine with RDF extension

**RDF Schema Alignment**

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

**Base URI:** <http://data.fingal.ie/councilor/> [edit](#)

**RDF Skeleton** [RDF Preview](#)

Available Prefixes: `dc` `rdfs` `foaf` `rdf` `void` [+add prefix](#) [manage prefixes](#)

<b>name</b> URI ✕ foaf:Person ✕ Councilor <a href="#">add rdf:type</a>	<input type="checkbox"/> ✕ > foaf:name →	<input type="checkbox"/> <b>name</b> cell	
	<input type="checkbox"/> ✕ > :party →	<input type="checkbox"/> <b>party</b> URI ✕ Party <a href="#">add rdf:type</a>	<input type="checkbox"/> ✕ > rdfs:label → <input type="checkbox"/> <b>party</b> cell <a href="#">add property</a>
	<input type="checkbox"/> ✕ > foaf:mbox →	<input type="checkbox"/> <b>email</b> cell	
	<input type="checkbox"/> ✕ > foaf:phone →	<input type="checkbox"/> <b>tel</b> cell	
	<input type="checkbox"/> ✕ > foaf:depiction →	<input type="checkbox"/> <b>image</b> URI <a href="#">add rdf:type</a>	<input type="checkbox"/> ...
	<input type="checkbox"/> ✕ > :councilorOf →	<input type="checkbox"/> <b>area</b> URI ✕ CouncilorDistrict <a href="#">add rdf:type</a>	<input type="checkbox"/> ✕ > rdfs:label → <input type="checkbox"/> <b>area</b> cell <a href="#">add property</a>
	<input type="checkbox"/> ✕ > :currentCouncilor →	<input type="checkbox"/> <b>current member</b> cell	

[Add another root node](#)

OK Cancel

## XLWrap : de excel vers RDF

- <http://xlwrap.sourceforge.net/> : Permet de rendre accessible les données d'un fichier excel ou csv en SPARQL
  - Approche « middleware » : donne accès aux données d'un fichier excel ou csv via SPARQL
  - Approche « transformation » : transforme le contenu d'un fichier excel ou csv en RDF : mais nécessite d'écrire du code
- Fichier de paramétrage pour expliciter comment transformer les lignes et les colonnes en RDF
- Permet de traiter des fichiers excel compliqués
  - Pas simplement « 1 ligne = 1 entité, 1 colonne = 1 propriété »
- Intérêt : *transformer le contenu de fichiers Excel en RDF*

# XLWrap : exemple simple

	A	B	C
1	First name	Second name	E-Mail address
2	Tom	Houston	th@ex.com
3	Tim	Presley	jp@ex.com
4	...	...	...

```
{ [] a xl:Mapping ;
  xl:template [
    xl:fileName "file:employees.xls" ;
    xl:templateGraph :Persons ;
    xl:transform [ a xl:RowShift ]
  ] .
}

:Persons {
  [xl:uri "'http://example.org/' & URLENCODE(A2 & B2)"^^xl:Expr] a foaf:Person;
  foaf:name "A2 & ' ' & B2"^^xl:Expr ;
  foaf:mbox_sha1sum "SHA(C2)"^^xl:Expr ;
}
```

# XLWrap : exemple avancé – hiérarchie

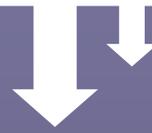
	A	B	C	D	E
1	Corporate	EMEA	Company Germany	Sales Germany	
2	Corporate	EMEA	Company Germany	Marketing Germany	
3	Corporate	Americas	Company USA	Sales USA	
4	Corporate	Americas	Company USA	Marketing USA	
5	Corporate	Shared Services	HR		
6	Corporate	Shared Services	IT		



## **XLWrap : exemple**

---

*Où l'on voit l'exemple en direct !*

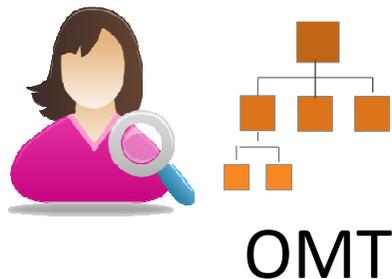


# PLAN

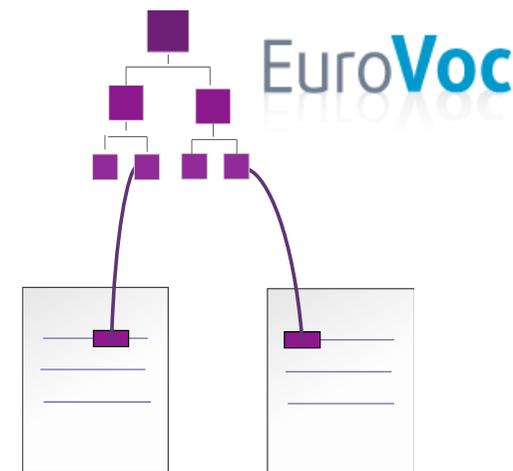
- 1. QUI SOMMES-NOUS ? MONDECA**
- 2. PROBLEMATIQUE : LA PUBLICATION DE DONNEES**
- 3. LA REUTILISATION DES VOCABULAIRES AVEC LE LOV**
- 4. OUTILS DE SEMANTISATION DES DONNEES**
- 5. ALIGNEMENT DE VOCABULAIRES**

# Alignement de vocabulaires

- Pourquoi ?
  - Si le contenu est annoté sur un vocabulaire A, et que l'utilisateur cherche avec un vocabulaire B ?
  - Permet d'interroger des corpus annoté sur un vocabulaire différent de celui de la recherche
- Exemples
  - « adult training » <exact match> « adult education »
  - « advertiser » <related match> « advertising »



Alignement de  
vocabulaires





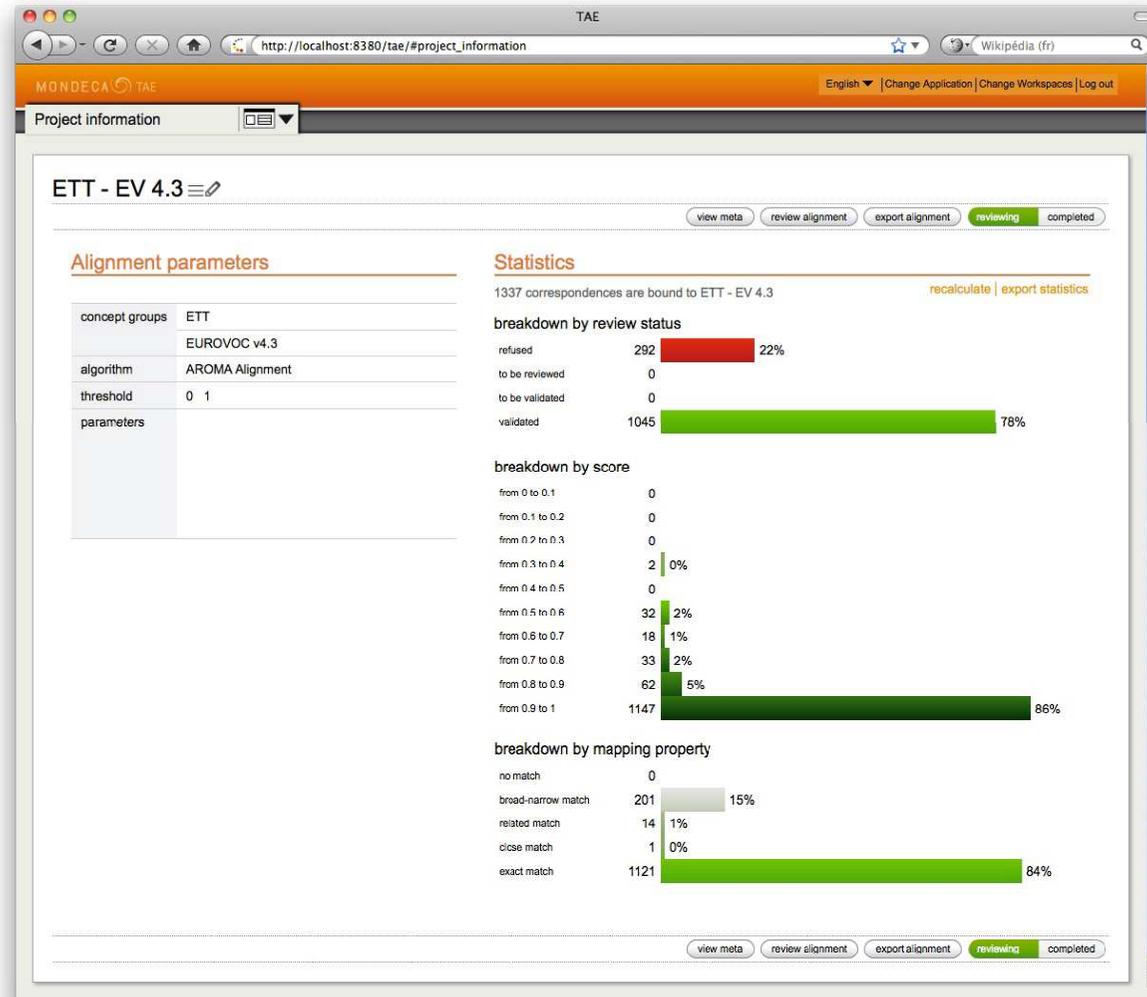
## Utilisation des alignements dans un moteur de recherche

- Plutôt au moment de l'*indexation*
- On traduit les annotations des documents d'origine en utilisant l'alignement
  - Du vocabulaire A vers le vocabulaire B
- On enrichit l'index avec les concepts du vocabulaire B
  - L'index contient donc l'annotation sur le vocabulaire A et sur le vocabulaire B
- On peut ensuite rechercher sur le corpus en utilisant les concepts du vocabulaire B

The screenshot shows the TAE web application interface. At the top, there is a browser window with the URL `http://localhost:8380/tae/#new_project`. Below the browser, the application header includes the MONDECA logo and navigation links for "English", "Change Application", "Change Workspaces", and "Log out". The main content area is titled "My Project" and contains three panels:

- Thesaurus#1**: ETT > Basic VET vocabulary > Education. Level 3 terms include Education, Information and communication, Learning, Sciences and technology, and Training.
- Thesaurus#2**: EUROVOC v4.3 > 32 EDUCATION AND C. Level 3 terms include 3206 education, 3211 teaching, 3216 organisation of teaching, 3221 documentation, 3226 communications, 3231 information and information processing, and 3236 information technology and data processing.
- Alignment process**: Features an "AROMA Alignment" dropdown, a "Threshold" slider (0 to 1), and input fields for "Parameter File" and "Alignment File".

Three arrows point to the interface: a red arrow to the "My Project" title, an orange arrow to the "Thesaurus#1" panel, and a green arrow to the "Alignment process" panel.



## ALIGNMENT REVIEW

Alignment review

ETT - EV 4.3

MAPPING ATTRIBUTES T#1 T#2

Score

Status

With concept

Mapping type

Result list

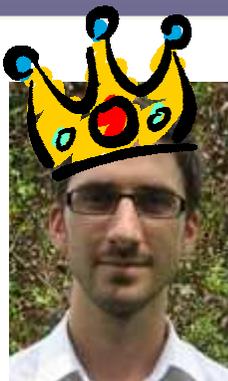
1336 mappings

add to tasks export

Actions: validate - pending - rejected - suppress

	Concept T1	Mapping type	Concept T2	Score	Status	Comment
1	A level	exact match	balance sheet	0.7	refused	
2	Abruzzi	exact match	Abruzzi	1.0	validated	
3	absenteeism	exact match	absenteeism	1.0	validated	
4	access to education	exact match	access to education	1.0	validated	
5	access to employment	exact match	job access	2.0	validated	
6	access to employment	narrow match	market access	1.0	refused	
7	access to information	exact match	access to information	1.0	validated	
8	accountant	exact match	accountant	1.0	validated	
9	accreditation of education and training providers	exact match	EACI	1.0	refused	
10	action research	broad match	research programme	1.0	refused	
11	action research	broad match	research	2.0	validated	
12	addiction	exact match	drug addiction	1.0	validated	
13	administration	exact match	administrative structures	1.0	validated	
14	administrative structure	exact match	administrative organisation	1.0	validated	
15	adult	exact match	adult	1.0	validated	
16	adult training	exact match	adult education	1.0	validated	
17	advanced technology	exact match	new technology	1.0	validated	
18	advertiser	related match	advertising	0.8	validated	
19	Afghanisan	exact match	Afghanistan	0.9	validated	
20	Africa	exact match	Africa	0.9	validated	

add to tasks export



**Pierre-Yves  
Vandebussche**  
Researcher – LOVer



3 Cité Nollez  
75018 Paris, France  
+33 1 44 92 35 00

[pierre-yves.vandebussche@mondeca.com](mailto:pierre-yves.vandebussche@mondeca.com)  
[www.mondeca.com](http://www.mondeca.com)

MERCI !

**Prochain rendez-vous avec  
les données ouvertes !**

**Jeudi 10 mai 14h à La Cantine  
Mondeca – ATOS – eMakina**  
« Données ouvertes, mode  
d'emploi »



**Thomas Francart**  
CTO



3 Cité Nollez  
75018 Paris, France  
+33 1 44 92 35 04

[thomas.francart@mondeca.com](mailto:thomas.francart@mondeca.com)  
[www.mondeca.com](http://www.mondeca.com)

## Data is king – other logo

