La 4ème édition des journées
« Big Data Mining and Visualization »

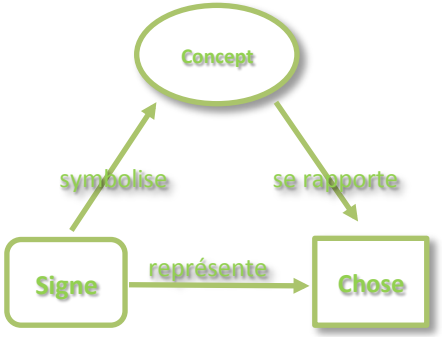
Approche **sémantique** pour la **recommandation de documents** textuels dans un contexte **Big Data** appliquée à la veille économique.

Christophe CRUZ| 19 juin 2015 | Lyon

Laboratoire LE2I – UMR CNRS 6306 – Université de Bourgogne

Axes de recherche

Modélisation et représentation des connaissances (MRC)



Web sémantique

Système centré utilisateur (SCU)

Système de recommandation

Big Data Value (BDV)

Véracité et valeur

Positionnement

Approche Numérique

- Taille, Dimension, poids
- Couleur, Type de matériaux
- Valeur financière

Approche Symbolique

- Espaces et usages
- Organisation architecturale
- Confort, dessert et facilités d'accès

Evaluation quantitative
entre éléments spécifiques

Evaluation qualitative
Sémantique sur un environnement

↓

SEMIOTIQUE

Positionnement

Point de vue sémiotique :

Sémiotique: étudie le processus de signification c'est-à-dire la production, la codification et la communication de signes.

Syntaxe: un ensemble de signes

Sémantique: relation entre les signes et des entités du monde réel

Pragmatique: **quels signes sont utilisés dans quel but ?**

Social : **qui utilise quels signes ?**

Plan

1. Le contexte industriel
2. Le transfert de technologie
3. La quête de la sémantique
4. La problématique Big Data
5. La solution First Eco Pro'fil
6. Conclusion

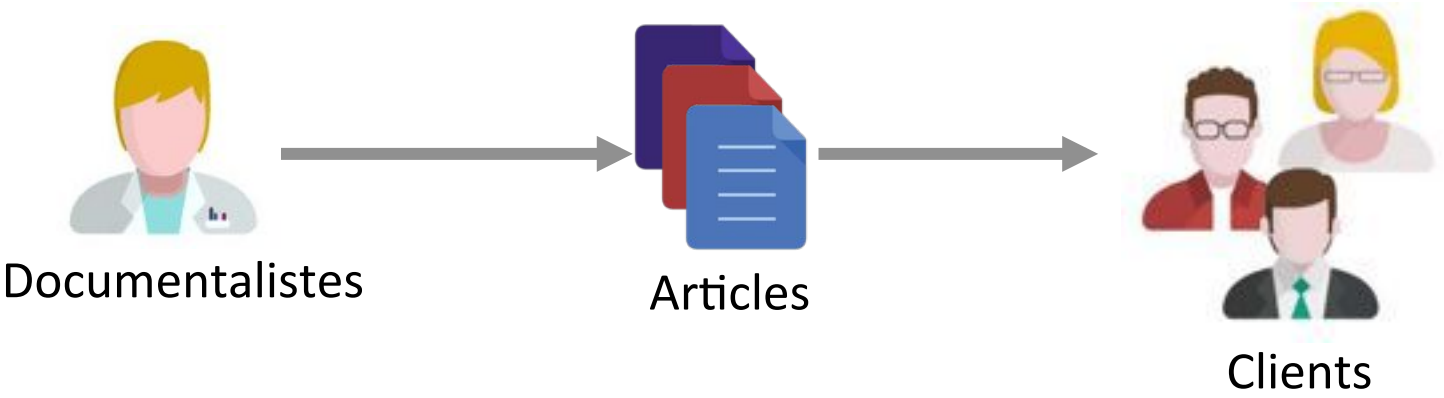
Le contexte industriel …

La solution First Pro'Fil

FIRSTECO
VOS DÉCISIONS PRENNENT UN TEMPS D'AVANCE

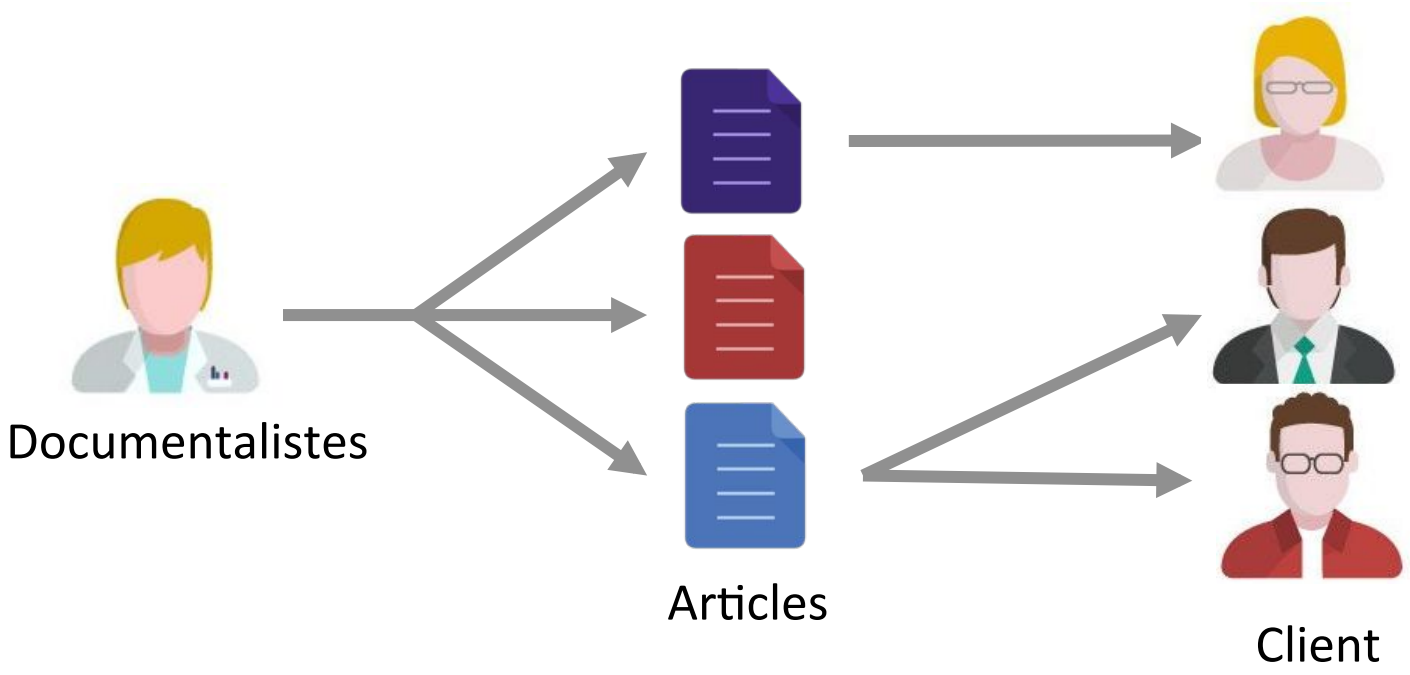
Contexte

FIRSTECO
VOS DÉCISIONS PRENNENT UN TEMPS D'AVANCE



Contexte

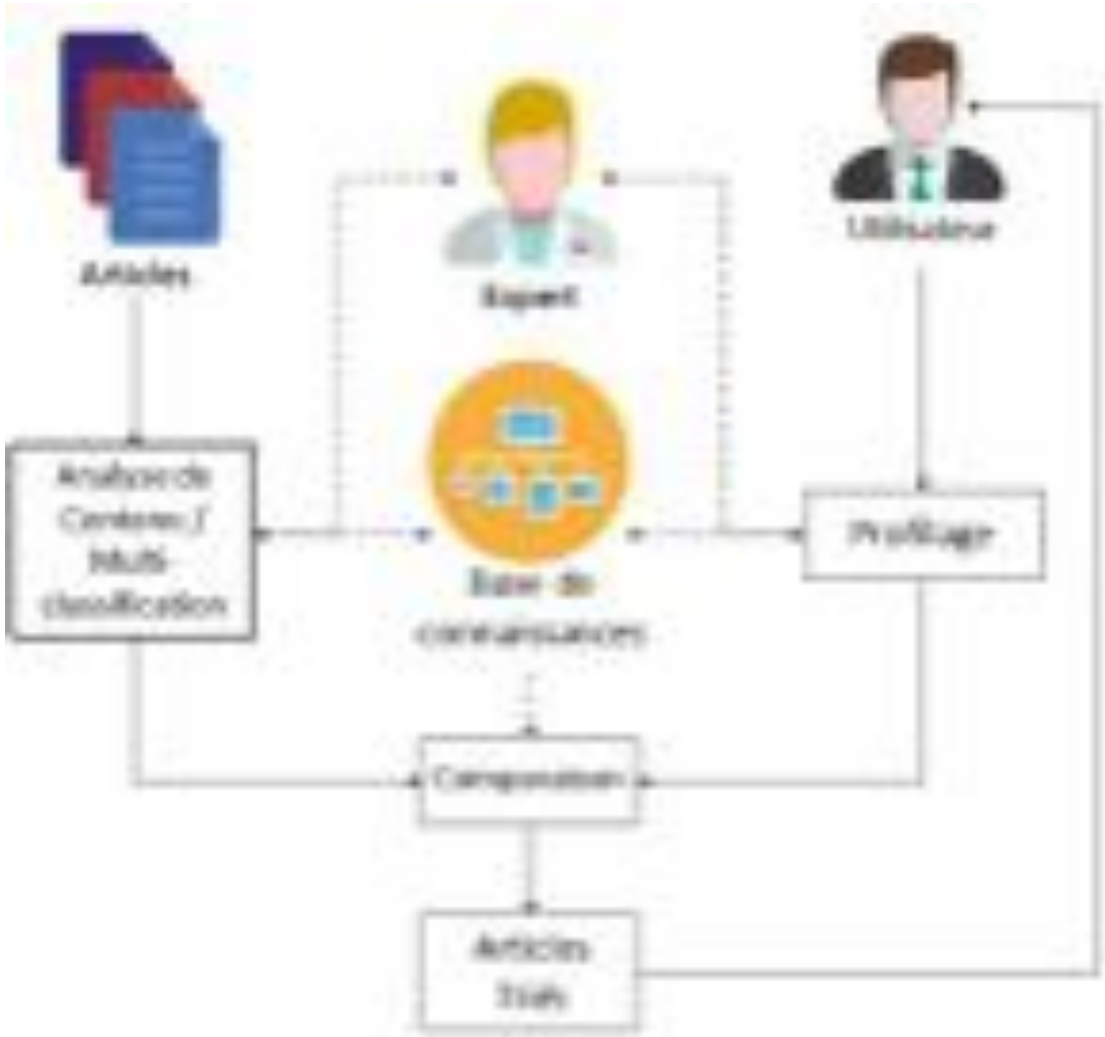
FIRSTECOPRO'FIL
BOOSTEZ VOTRE BUSINESS AVEC LA VEILLE PERSONNALISÉE



Contexte

L'UNIQUE PORTAIL DE VEILLE ÉCONOMIQUE ET COMMERCIALE À VOTRE IMAGE

Contexte scientifique



Transfert de technologie …

Actualis SARL

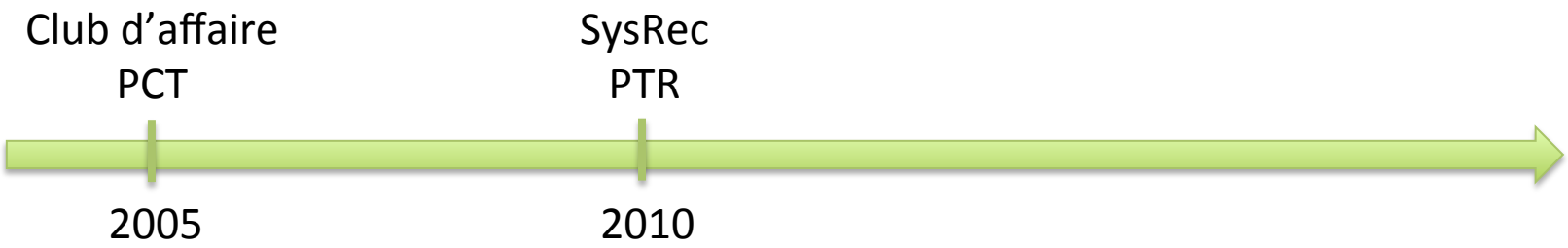
Historique

Club d'affaire
PCT

2005

BOURGOGNE
INNOVATION

Historique

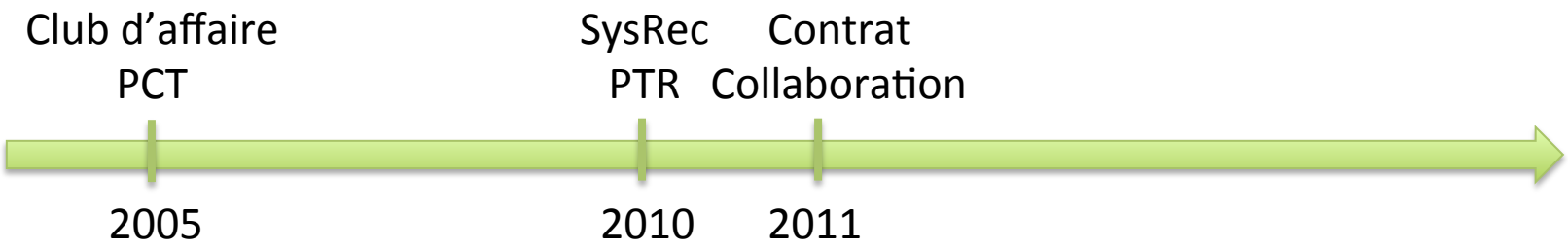


BOURGOGNE
INNOVATION

Historique

Rupture de politique - R&D

- Équipe de développement
- Changement du « process » métier
- Montée en compétence
- Homogénéisation des compétences et savoirs

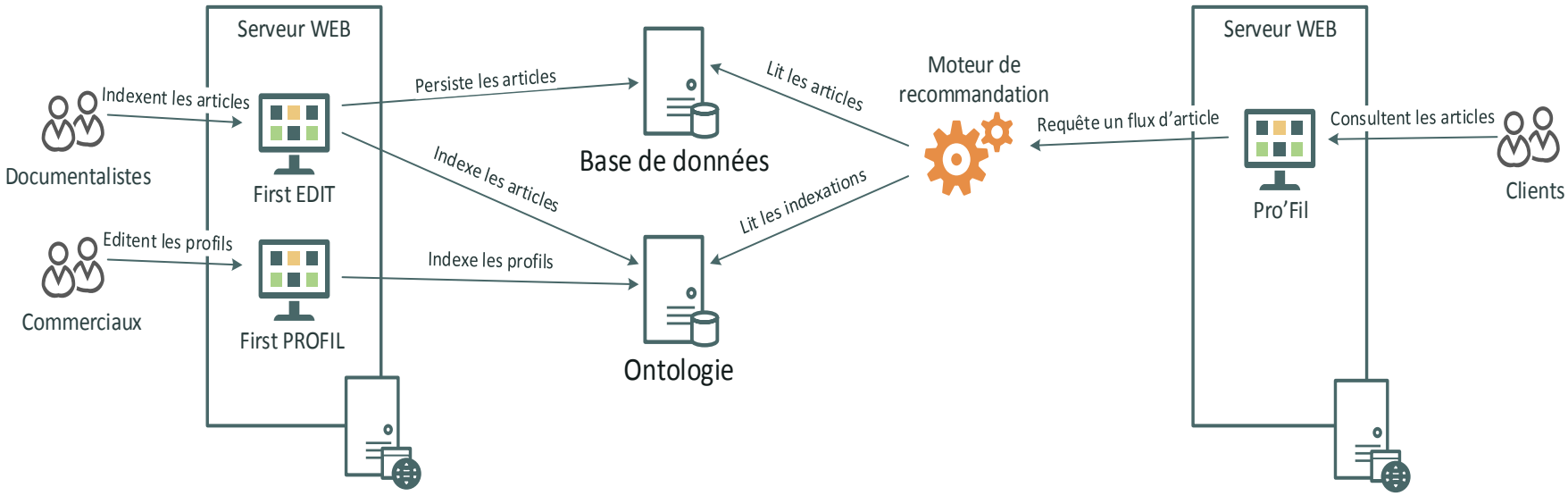


anrt
association nationale
recherche technologie

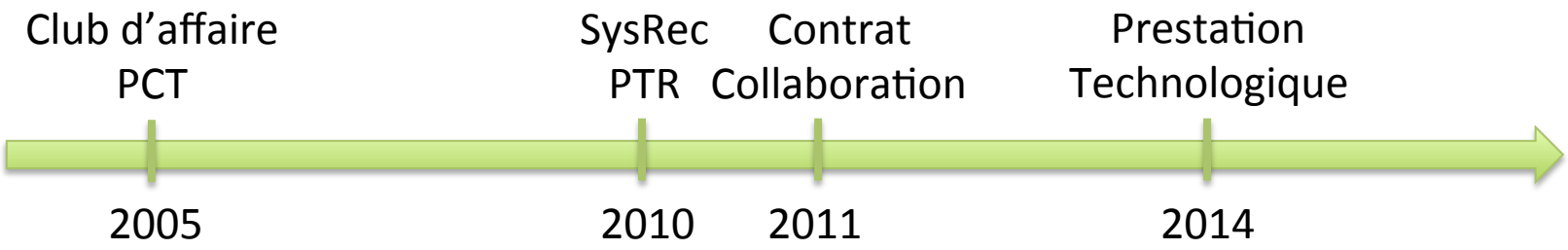
Welience
Innover, c'est notre métier

Historique

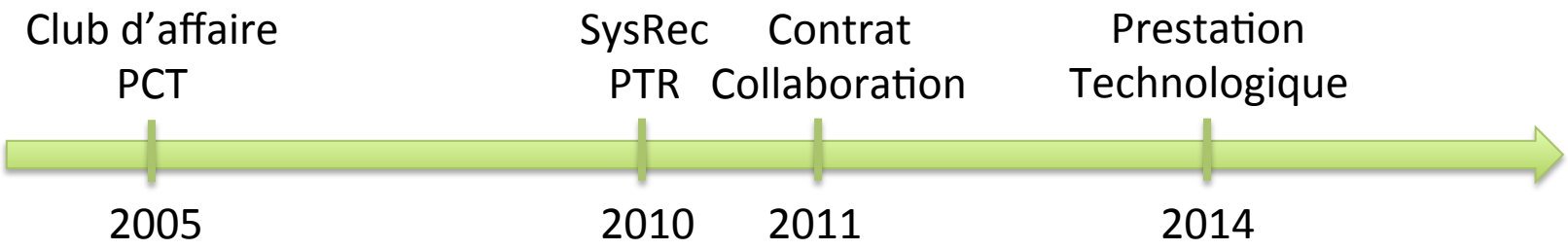
Architecture complète



Historique


Club d'affaire
PCT
SysRec
PTR
Contrat
Collaboration
Prestation
Technologique
2005
2010
2011
2014

Historique



SATT GRAND EST
Welience

Bourgogne
Conseil régional

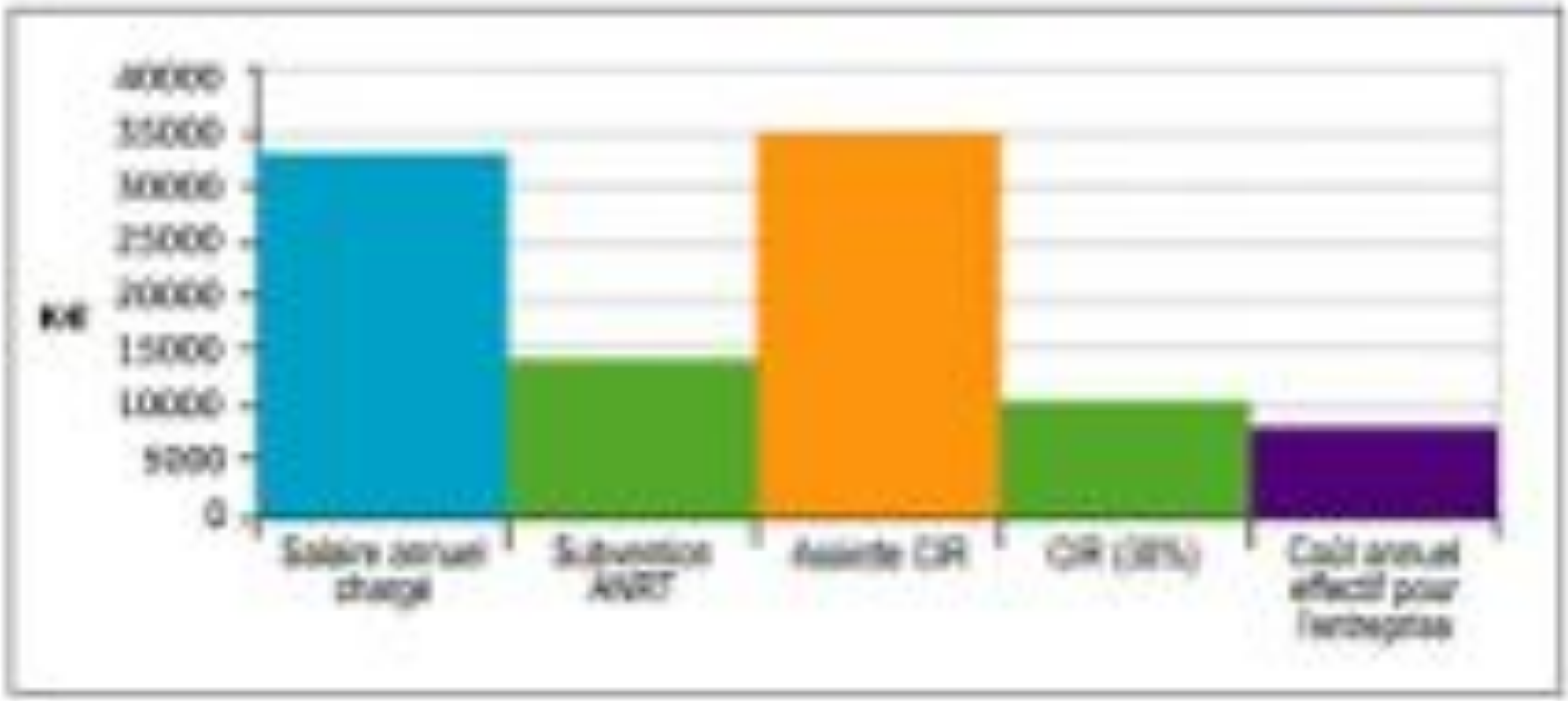
Transfert de technologie …

Parlons argent !

Le doctorant

Illustration à l'aide d'un exemple simplifié

Un exemple permet de concrétiser l'impact pour une entreprise de l'embauche d'un doctorant dans le cadre d'une CIFre.
Pour un salaire chargé du doctorant de 32 878 €, le coût effectif pour l'entreprise est de [illegible] €.



Coûts	Année 1	Année 2	Année 3	Informations
Doctorant	33 000,00	33 000,00	33 000,00	
CRI	-24 700,00	-24 700,00	-24 700,00	dont 14k sub ANRT + CRI

http://www.impots.gouv.fr/portal/deploiement/p1/fichedescriptive_6914/fichedescriptive_6914.pdf
http://www.agence-nationale-recherche.fr/missions-et-organisation/credit-d-impot-recherche/cir-en-detail/

Coûts	Année 1	Année 2	Année 3	Informations
Doctorant	33 000,00	33 000,00	33 000,00	
CRI	-24 700,00	-24 700,00	-24 700,00	dont 14k sub ANRT + CRI
Contrat de collab.	42 000,00	42 000,00	42 000,00	126 000 pour labo
Taux 2,0	84 000,00	84 000,00	84 000,00	doublement somme éligible
CRI	-25 200,00	-25 200,00	-25 200,00	30%/an

http://www.impots.gouv.fr/portal/deploiement/p1/fichedescriptive_6914/fichedescriptive_6914.pdf
http://www.agence-nationale-recherche.fr/missions-et-organisation/credit-d-impot-recherche/cir-en-detail/

Coûts	Année 1	Année 2	Année 3	Informations
Doctorant	33 000,00	33 000,00	33 000,00	
CRI	-24 700,00	-24 700,00	-24 700,00	dont 14k sub ANRT + CRI
Contrat de collab.	42 000,00	42 000,00	42 000,00	126 000 pour labo
Taux 2,0	84 000,00	84 000,00	84 000,00	doublement somme éligible
CRI	-25 200,00	-25 200,00	-25 200,00	30%/an
Dépenses (2 ing.)	200 000,00	200 000,00	200 000,00	incompressible
Taux 1,5	300 000,00	300 000,00	300 000,00	éligible
CRI	-90 000,00	-90 000,00	-90 000,00	30%/an

http://www.impots.gouv.fr/portal/deploiement/p1/fichedescriptive_6914/fichedescriptive_6914.pdf
http://www.agence-nationale-recherche.fr/missions-et-organisation/credit-d-impot-recherche/cir-en-detail/

Coûts	Année 1	Année 2	Année 3	Informations
Doctorant	33 000,00	33 000,00	33 000,00	
CRI	-24 700,00	-24 700,00	-24 700,00	dont 14k sub ANRT + CRI
Contrat de collab.	42 000,00	42 000,00	42 000,00	126 000 pour labo
Taux 2,0	84 000,00	84 000,00	84 000,00	doublement somme éligible
CRI	-25 200,00	-25 200,00	-25 200,00	30%/an
Dépenses (2 ing.)	200 000,00	200 000,00	200 000,00	incompressible
Taux 1,5	300 000,00	300 000,00	300 000,00	éligible
CRI	-90 000,00	-90 000,00	-90 000,00	30%/an
Coût total	275 000,00	275 000,00	275 000,00	825 000,00
CRI total	-139 900,00	-139 900,00	-139 900,00	-419 700,00
Coût total - CRI	135 100,00	135 100,00	135 100,00	0,50
Coût total Rech.	50 300,00	50 300,00	50 300,00	

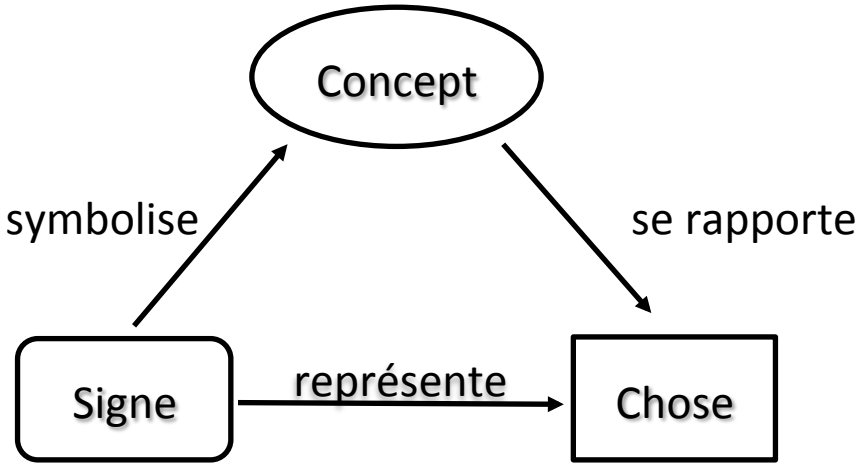
http://www.impots.gouv.fr/portal/deploiement/p1/fichedescriptive_6914/fichedescriptive_6914.pdf
http://www.agence-nationale-recherche.fr/missions-et-organisation/credit-d-impot-recherche/cir-en-detail/

La quête de la sémantique …

Quête de la sémantique

Complexe

Simple



Ogden et Richards (1923)

Modèles abstraits

Quête de la sémantique

Raisonnement

Quête de la sémantique

Du Web-documentaire …

au Web des données

Echange et partage

Les systèmes axiomatiques

(1906, 1978)

Kurt Gödel

"Systèmes logiques"

"Systèmes de décisions"

Alan Turing

(1912, 1954)

Décidabilité $\exists Algo \rightarrow V \cup F$

Calculabilité $\exists Algo$ pour $f(x)$ en temps fini

Les logiques descriptives

Phase 0 (1965-1980)

Pre-DL - introduction réseaux sémantiques et frames.

Critiques liées au manque de **sémantique formelle**.

KL-One: premier système de logique de description

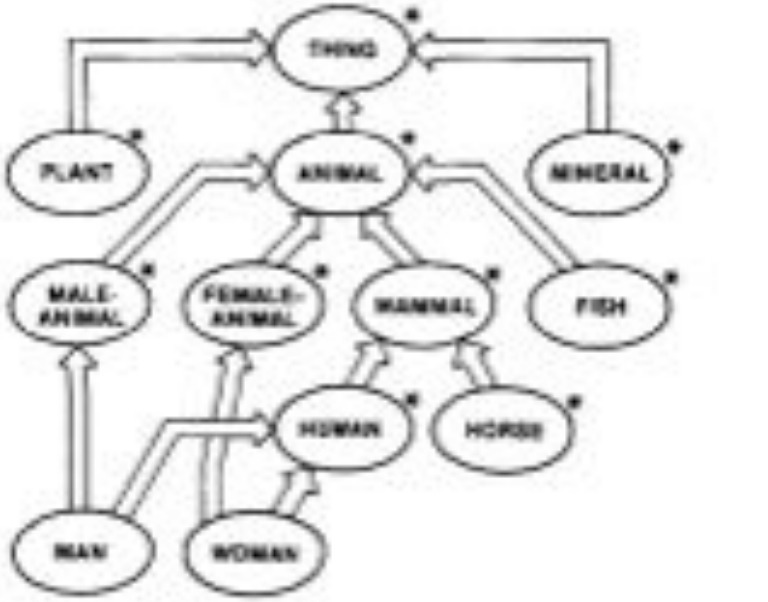


Figure 1. A simple KL-ONE network of Generic Concepts.

[Brachman and Schmolze 1985]

Les logiques descriptives

Phase 1 (1980-1990)

Algorithmes de subsomption structurelle

KL-One, K-Rep, Krypton, Back, Loom

Très efficaces, mais **non complets**,
voire même **indécidables**
sauf pour des fragments très pauvres

English	FOL	DL
Dog with a Spot (DWS)	DWS(x) ⇔ Dog(x) ^ (∃y.has(x,y) ^ Spot(y))	DWS ⇔ Dog ⊓ ∃has.Spot
Large Dog with a Dark Spot (LDWDS)	LDWDS(x) ⇔ (Dog(x) ^ Large(x)) ^ (∃y.has(x,y) ^ (Spot(y) ^ Dark(y))	LDWDS ⇔ Dog ⊓ Large ⊓ ∃has.(Spot ⊓ Dark)

Les logiques descriptives

Phase 2 (1990-1995)

Introduction d'algorithmes fondés sur les **tableaux**.

Permet de décider de la **cohérence**
Premiers systèmes utilisant ces méthodes: Kris et Krack.
Implémentations efficaces, même si la complexité dans le pire des cas n'est plus polynomiale (**TBox/ABox**)

TBox	*ABox*
$Femelle \sqsubseteq \top \sqcap \neg Mâle$	$Humain(Anne)$
$Mâle \sqsubseteq \top \sqcap \neg Femelle$	$Femelle(Anne)$
$Animal \equiv Mâle \sqcup Femelle$	$Femme(Sophie)$
$Humain \sqsubseteq Animal$	$Humain(Robert)$
$Femme \equiv Humain \sqcap Femelle$	$\neg Femelle(Robert)$
$Homme \equiv Humain \sqcap \neg Femelle$	$Homme(David)$
$Mère \equiv Femme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Sophie, Anne)$
$Père \equiv Homme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Robert, David)$
$MèreSansFille \equiv Mère \sqcap$ $\forall relationParentEnfant.\neg Femme$	
$relationParentEnfant \sqsubseteq \top_R$	

[Baader and B. Hollunder, 1991]

1ère définition moderne d'ontologie

Thomas Gruber

"A Translation Approach to Portable Ontology"
1993

(1959,)

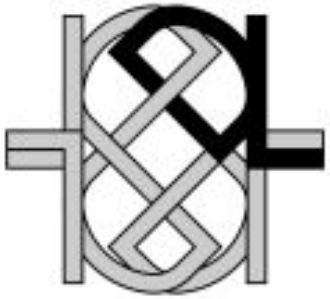
Concepts, propriétés relations, fonctions, contraintes, axiomes, sont définis

Sans ambiguïté

Une ontologie est une spécification explicite d'une conceptualisation

Modèle abstrait et vue simplifiée d'un phénomène réel que nous souhaitons représenter

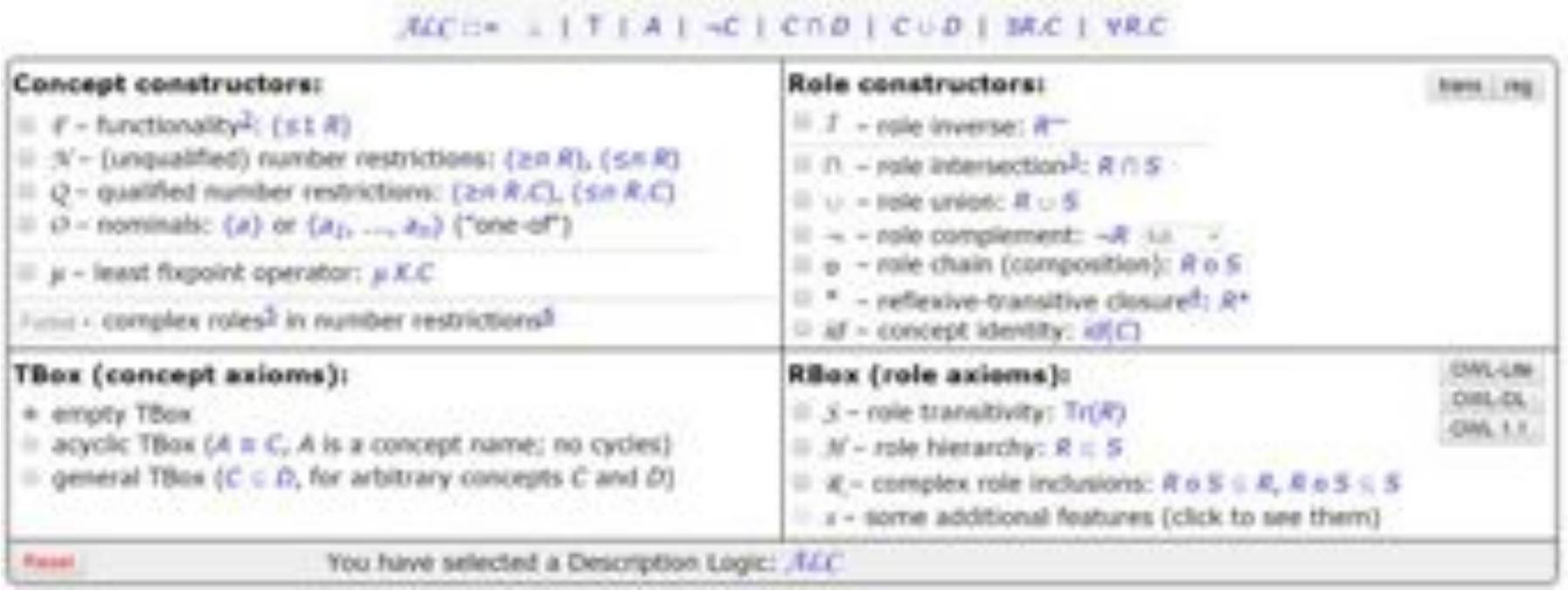
Les logiques descriptives



Phase 3 (1995-2000)

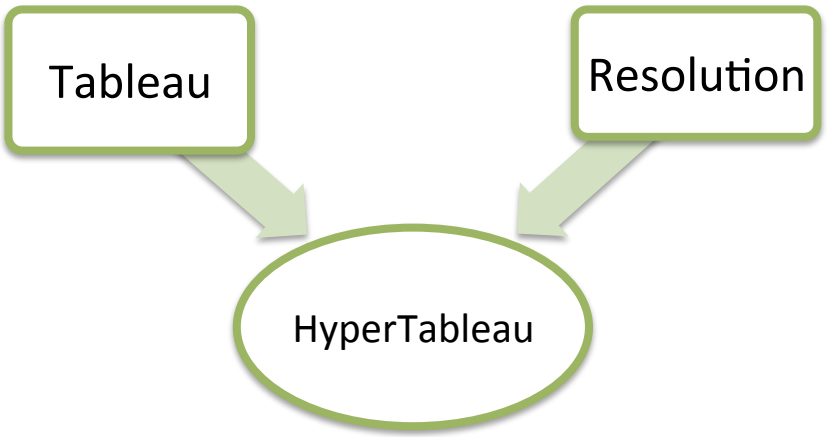
LD très expressives basées sur les tableaux

Exploration des liens avec les logiques modales



Raisonnement avec Les LD

Systems	Algorithm	Sound	Complete	Rule Support	Expressivity
Pellet	Tableau	Yes	Yes	Yes (SWRL)	SROIQ (D)
FaCT++	Tableau	Yes	Yes	No	SROIQ (D)
Racer	Tableau	Yes	Yes	Yes (SWRL)	SHIQ
Kaon2	Resolution	Yes	Yes	Yes (SWRL)	SHIQ
HermiT	HyperTableau	Yes	Yes	Yes (SWRL)	SROIQ (D)



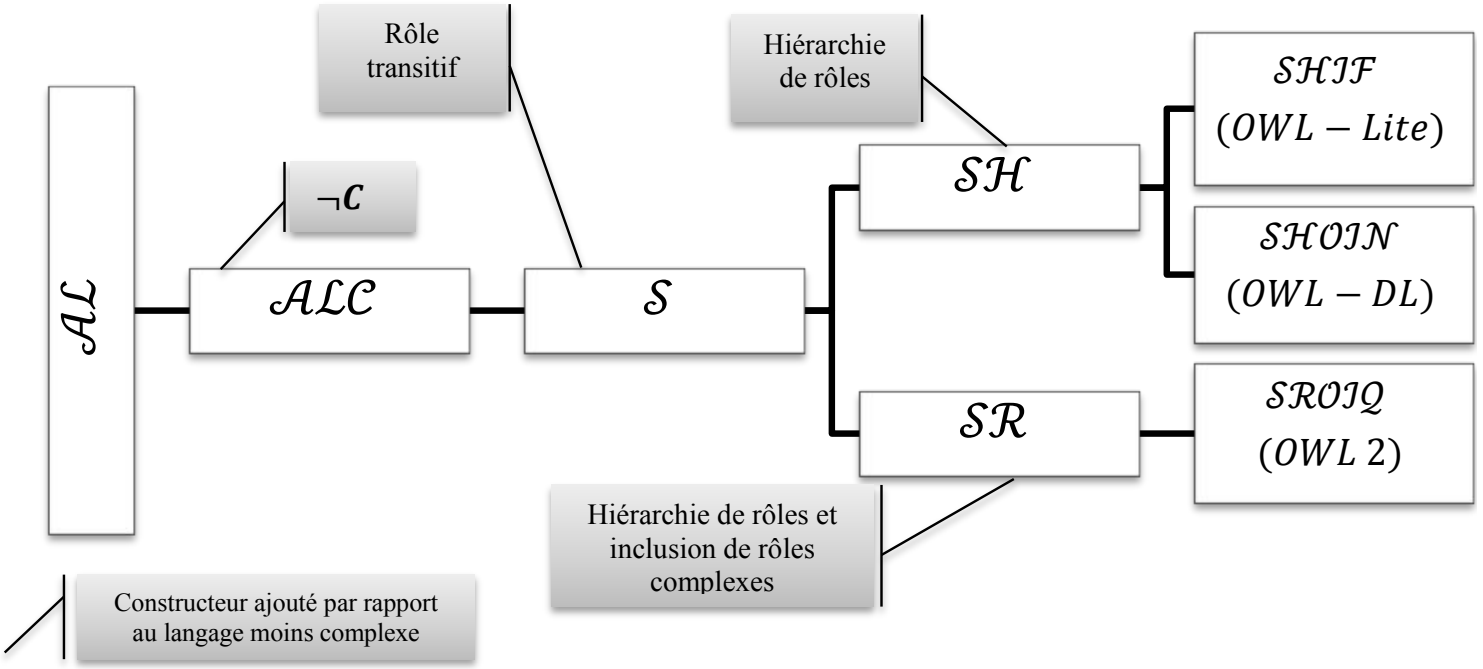
Motik, B. (2009). Hypertableau Reasoning for Description Logics

Les logiques descriptives

Phase 4 (2000-2012)

Web sémantique

Systèmes d'information, …
Utilisation DL moins expressives…



Le Web sémantique

« Le web sémantique ne fonctionnera jamais ! »

James Hendler « Why the semantic Web will nerver work », ESWC'11

Le Web sémantique

Numérique

Théologie

Botanique

Physique

Politique

Sciences de la nutrition

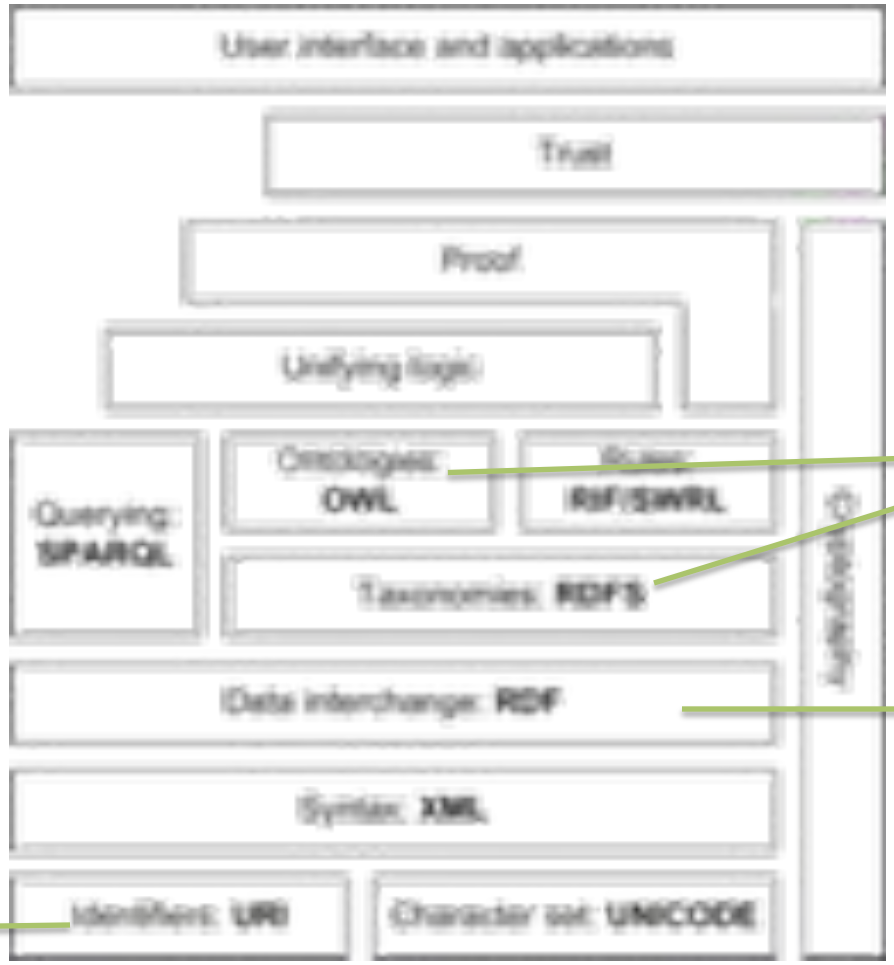
Agriculture

Culture

Semantic Error 409 – Ontology Not Found

You've encountered a " Ontology Not Found" error while trying to access a semantic term grounded on the University of Burgundy Laboratory LE2I Cheksem web server.

Le Web sémantique



Approche formelle

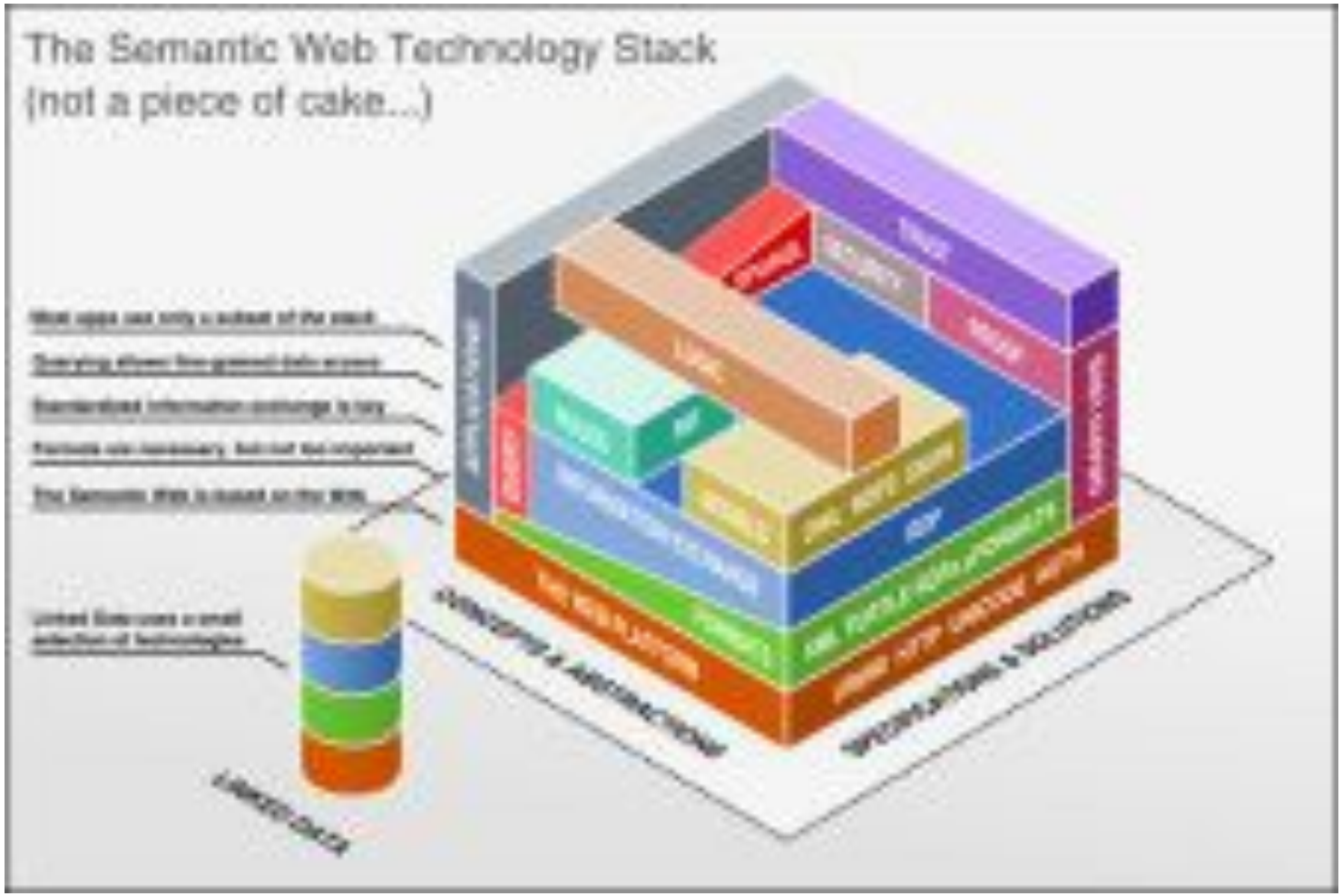
Annotation, Triplestore, BD graphe Scalable

id

14 ans après …

« La pile technologique du Web sémantique »

Le Web sémantique



14 ans après …

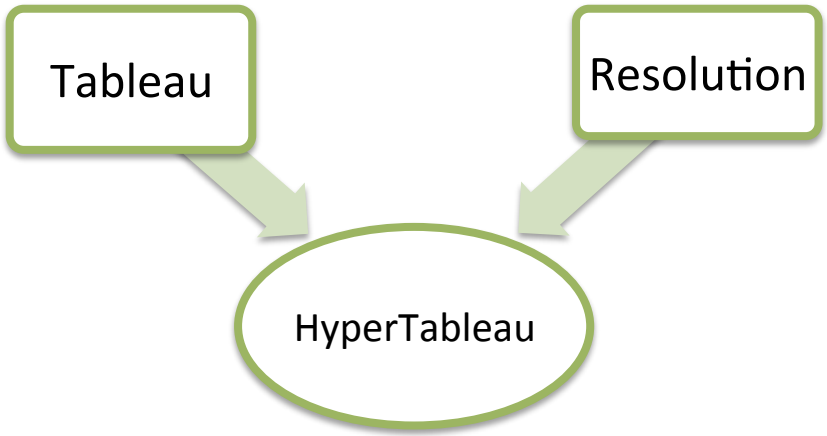
« La pile technologique du Web sémantique »

Les logiques descriptives

"I'm sorry Dave,
I'm afraid I can't do that."

Phase 5 (2012-20..)

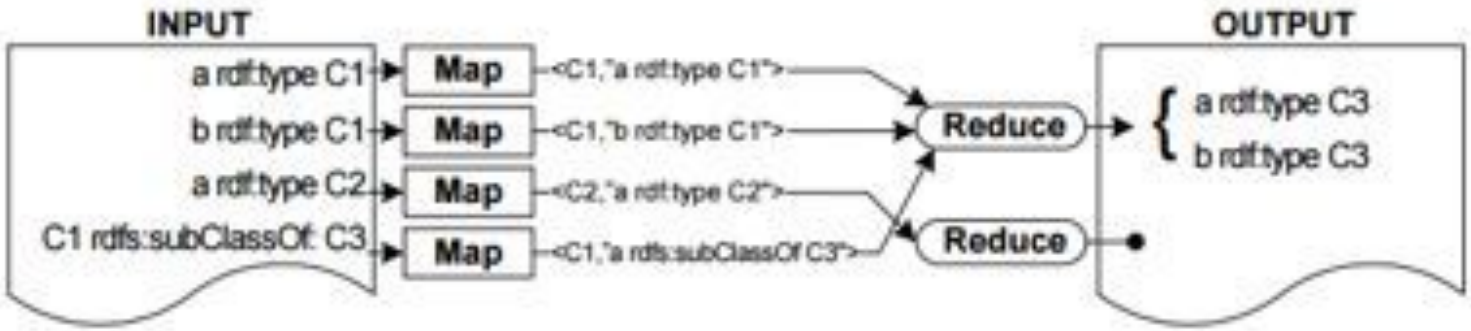
Application à l'échelle du web (sémantique)
« Web Reasonning »



Passage à l'échelle difficile …

Raisonner à grande échelle

WebPIE - Raisonnement basé sur les règles …



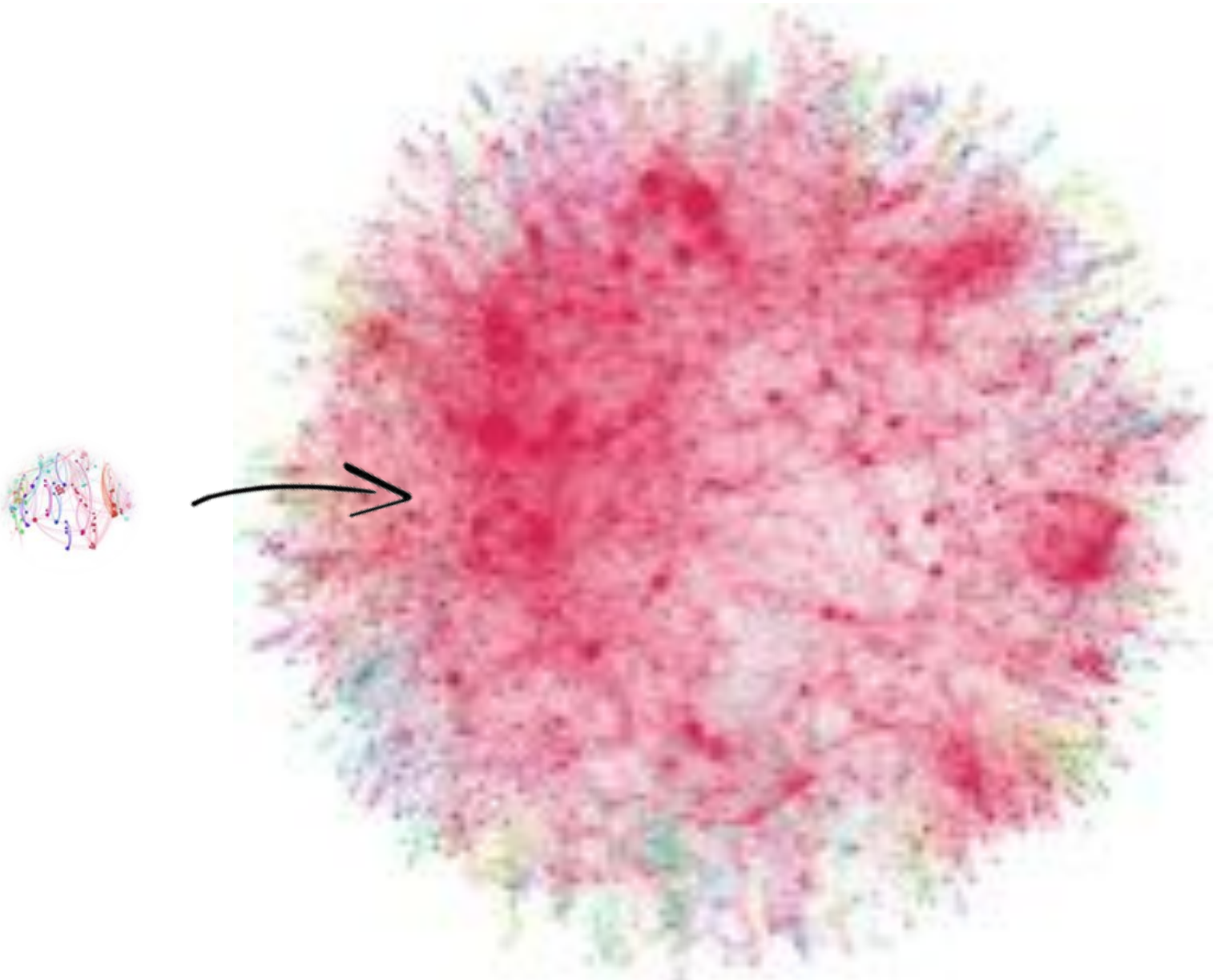
et Map Reduce

Urbani, J., Kotoulas, S., Maassen, J., Van Harmelen, F., & Bal, H. (2012). WebPIE: A Web-scale parallel inference engine using MapReduce. *Web Semantics: Science, Services and Agents on the World Wide Web.*

Urbani, J., Kotoulas, S., Oren, E., & Harmelen, F. (2009). **Scalable Distributed Reasoning** Using MapReduce. In - *ISWC 2009 SE - 40* (Vol. 5823, pp. 634–649).

CHECKSEM
SEMANTICINTELLIGENCERESEARCH

Raisonner à grande échelle



Raisonner à grande échelle

Forward-chaining (Materialization)

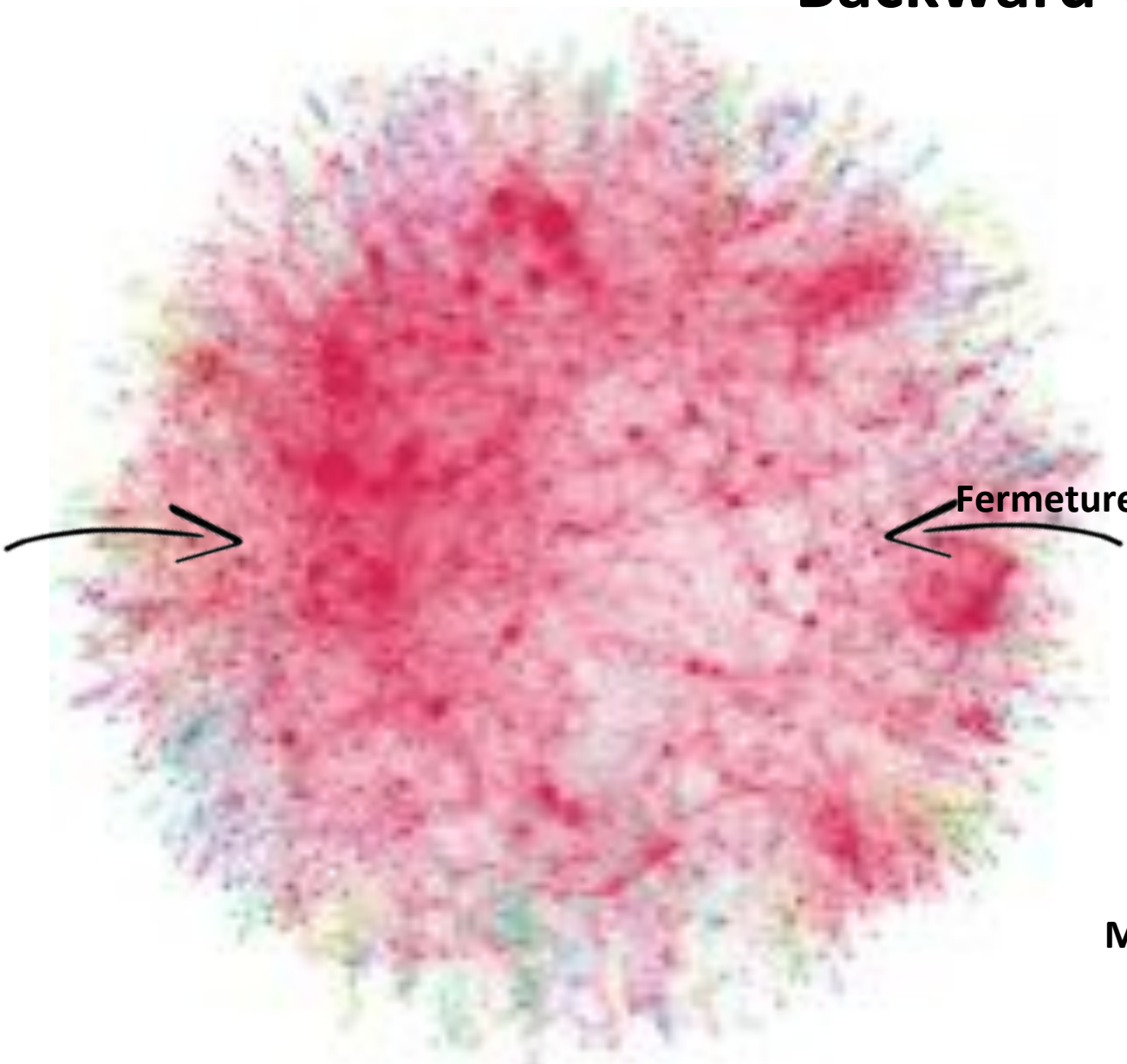
Fermeture

Matérialisation

Requêtes

Raisonner à grande échelle

Backward-Chaining



Fermeture

Requêtes

Matérialisation

Approches hybrides

QueryPIE

Matérialisation
Backward-chaining
OWL-RL

Règles pré-matérialisées

Urbani, J., Piro, R., van Harmelen, F., & Bal, H. (2013). Hybrid reasoning on OWL RL. *Semantic Web*.

Approches hybrides

Conférence ESWC'15

Prototype sur OWL 2 EL

Distributed and Scalable OWL EL Reasoning

Approche incrémentale

Slider: an Efficient Incremental Reasoner

J. Chevalier (Fragments RDFS et pD*) - **streamming**

Raisonner à grande échelle

Raisonner en OWL DL à l'échelle du Web

Problème non résolu

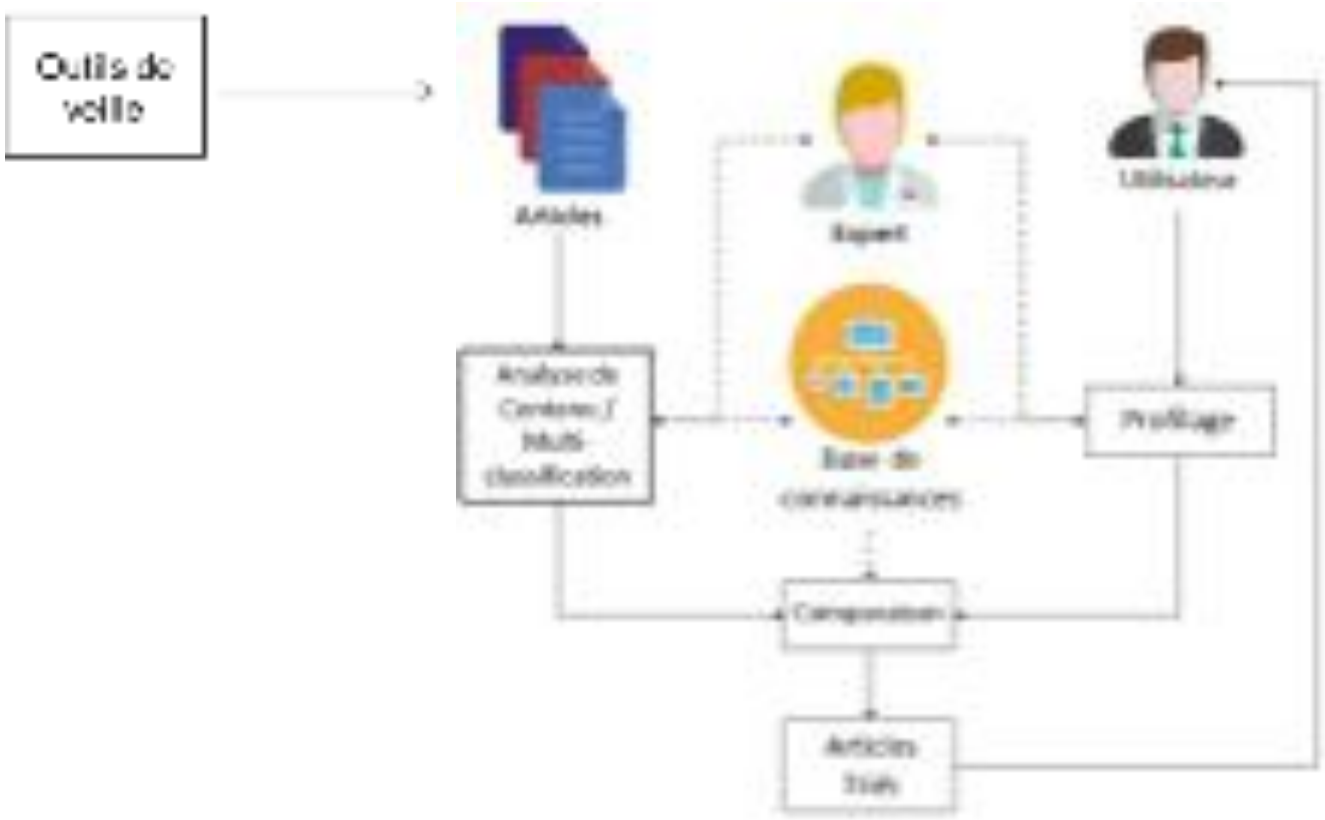
Urbani, J. (2013). Three Laws Learned from Web-scale Reasoning. In *2013 AAAI Fall Symposium Series*.

Le projet de l'entreprise …

La sémantique, la veille et le web

Architecture actuelle

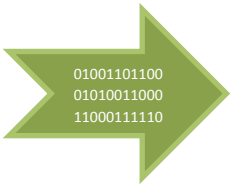
Outils actuels de veille



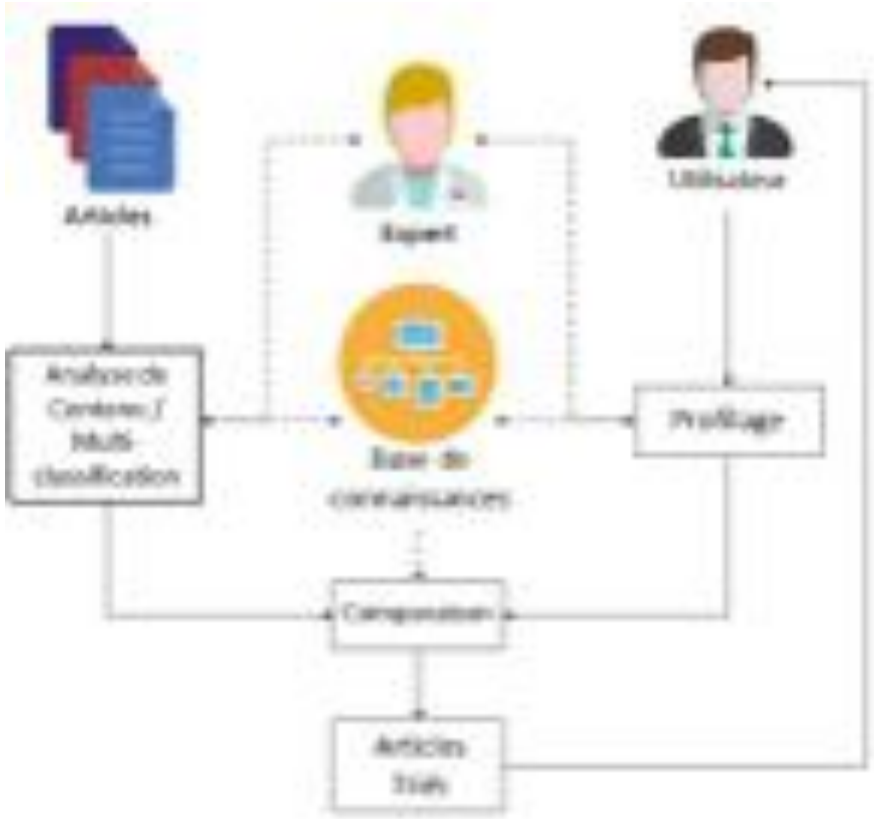
Architecture future

Quels sont les leviers pour faciliter cette étape ?

Web et Flux



Volume !
!!



Contexte scientifique – Big Data

Sourcing sémantique

Croisement de l'information

Contexte scientifique – Big Data

Processus de **classification multi-label hiérarchique sémantique** pour le **Big Data**

Contexte scientifique – Big Data

Processus de **classification multi-label hiérarchique sémantique** pour le **Big Data**

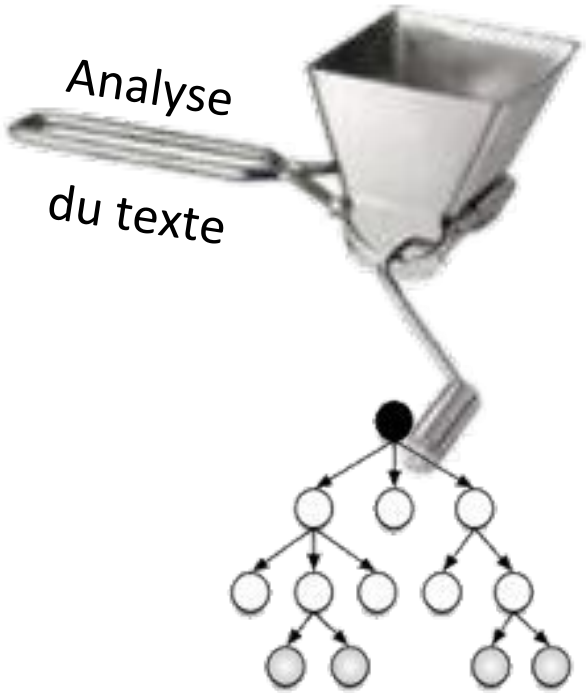
Contexte scientifique – Big Data

Processus de **classification multi-label hiérarchique sémantique** pour le **Big Data**

=

Comment générer cette taxonomie à l'échelle du Web ?

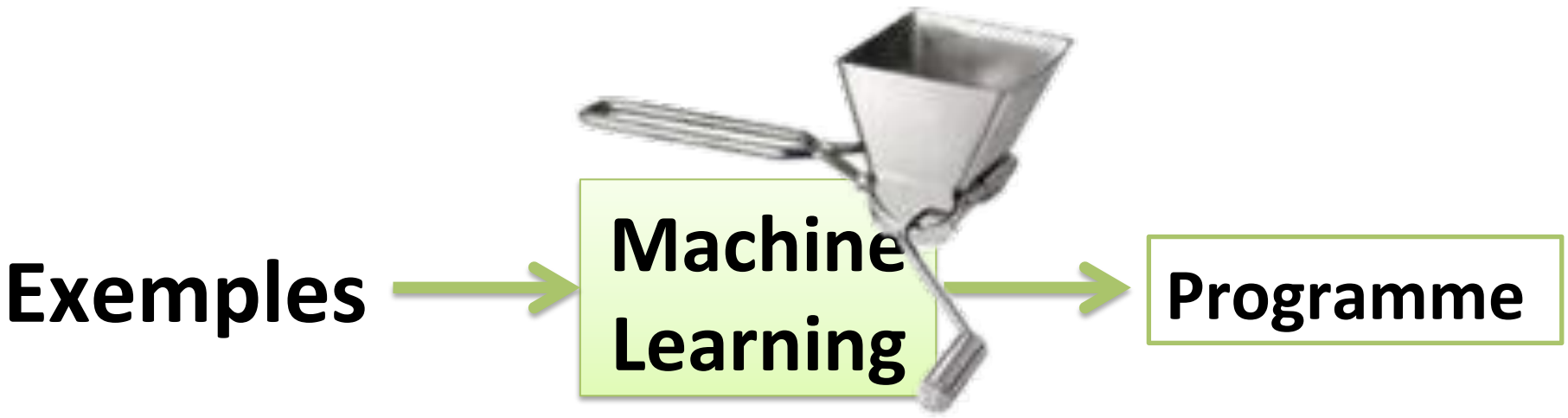
Analyse des données



Modèle prédictif

Le problème …

Comment coder cette moulinette ?


Exemples
Machine Learning
Programme

Par exemple …

Donner un sens au sens des mots !

- La **veille** sanitaire est l'action de surveiller l'état de santé d'une population afin de prévenir des menaces d'épidémies.

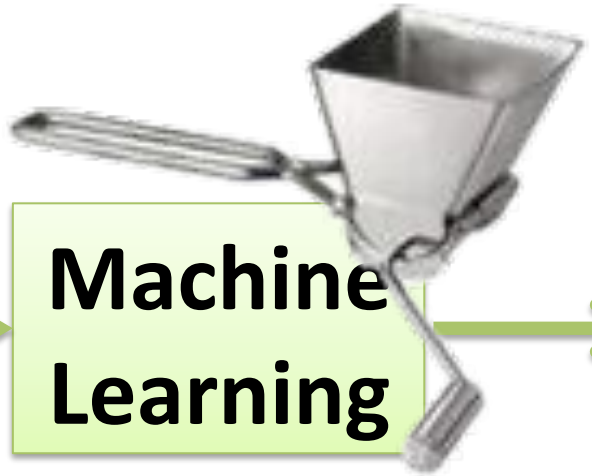
- La **veille** des jours fériés, la journée de travail finit une heure plus tôt.

- J'ai entendu plusieurs versions sur les conditions pour avoir le jour férié rémunéré : il faut travailler la **veille** du jour et le lendemain et avoir un ...

- Ainsi la garde de nuit doit être apte au cours de sa **veille** nocturne, …

- La garde de nuit **veille** sur la ville. Archivé. Quatre agents, deux chiens : la police municipale veillera sur la tranquillité publique le ...

Phrases
+
Dictionnaires
& Définitions



Machine Learning

Décision du sens du mot

« La **veille** concurrentielle est l'activité continue et en grande partie itérative qui vise à une surveillance active de ... »


jour
1


jour
1
éveil
2


jour
1
éveil
2
garde
3

jour, précédent, un, autre

fait, de, ne, pas, dormir, éveil

garde, fait, pendant, la, nuit

Le café noir très fort qu'il a pris au début de la **veille**, empêche Arsène André de dormir.

jour, précédent, un, autre

fait, de, ne, pas, dormir, éveil

garde, fait, pendant, la, nuit

jour, précédent, un, autre

fait, de, ne, pas, **dormir**, éveil

garde, fait, pendant, la, nuit

Le café noir très fort qu'il a pris au début de la **veille**, empêche Arsène André de **dormir**.

jour, précédent, un, autre

Le café noir très fort qu'il a pris au début de la **veille**, empêche Arsène André de **dormir**.

fait, de, ne, pas, dormir, éveil, *café, noir, très, fort, début, la, empêche, Arsène, André*

garde, fait, pendant, la, nuit

George E. P. Box

Essentially, all models are wrong, but some are useful.

Empirical Model-Building and Response Surfaces **(1987) p. 424**

jour (35), précédent (62), un (36), autre (21), finis (3), feux(1), longues(33), parlottes (2), précèdent(40), immédiatement(5), avant(37), etc.

fait (3), de (21) , ne (10), pas(30), dormir(64), éveil(45), café(2), noir(10), très (3), fort(2), début(1), la(21), empêche(40), Arsène(1), André(2), etc.

garde(50), fait(12), pendant(23), la(37), nuit(15), préparez(22), concurrentielle(6), à(38), rester(9), éveillé(11), difficile(17), alerte(55), surveillance(19), etc.

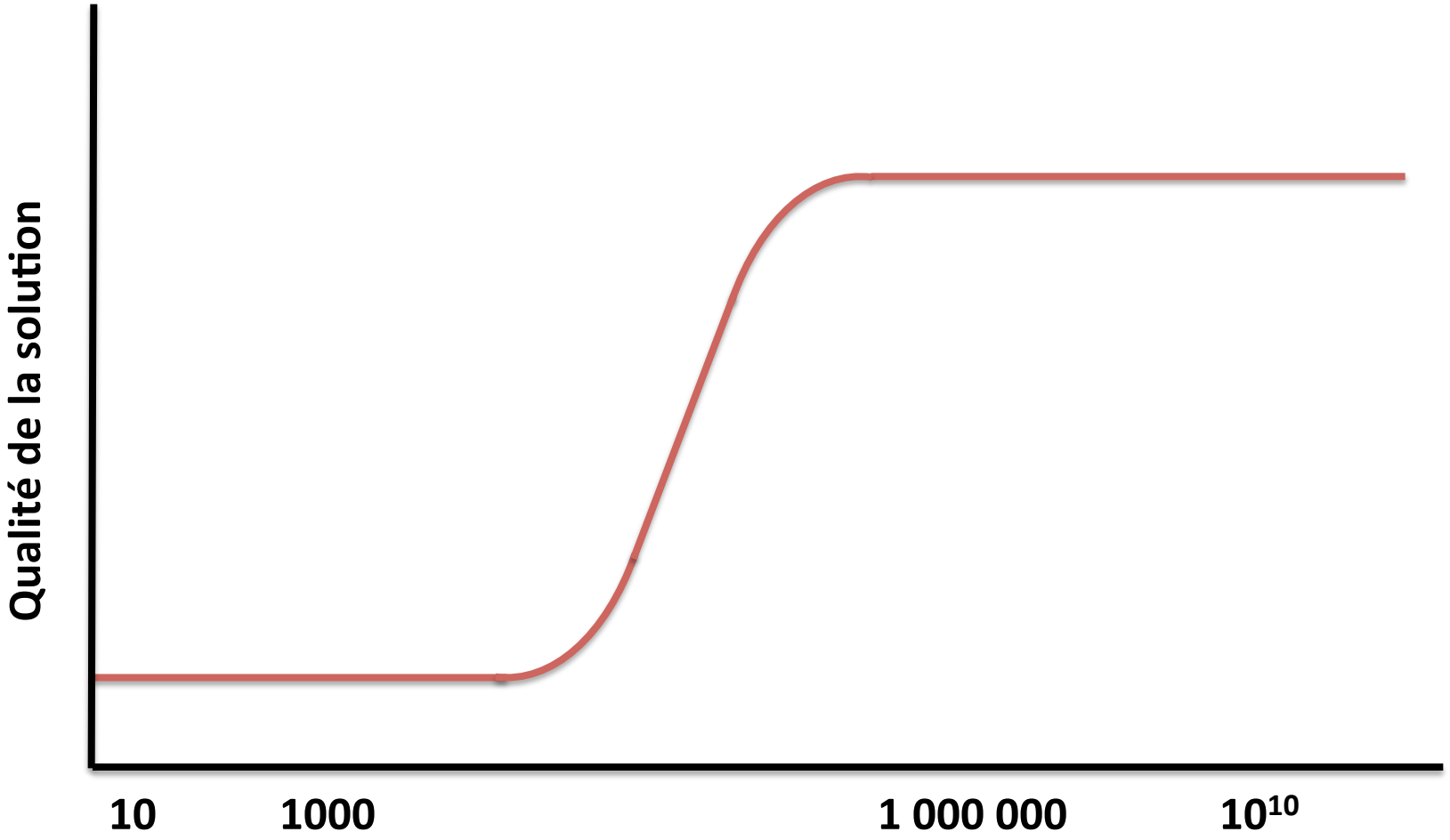
« La **veille** **concurrentielle** est l'activité continue et en grande partie itérative qui vise à une **surveillance** active de ... »

Décision du sens du mot

Veille[3]

garde(50), fait(12), pendant(23), la(37), nuit(15), préparez(22), **concurrentielle**(6), à(38), rester(9), éveillé(11), difficile(17), alerte(55), **surveillance**(19), etc.

Seuils des données



Et notre problème …

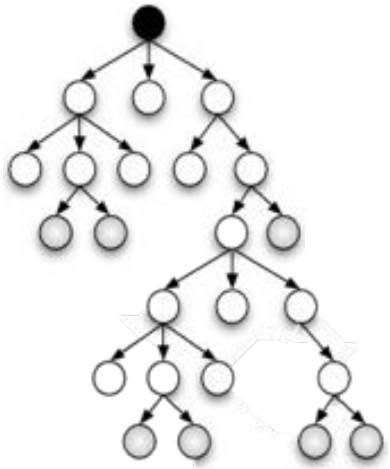
Développement d'une solution Big Data

50 000 **sacs** de 50 000 **mots**
$>10^{10}$ **cases mémoires**

10^{10} = 10 x 10 x 10 x 10 x 10 x 10 x 10 x 10 x 10 x 10

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

=>



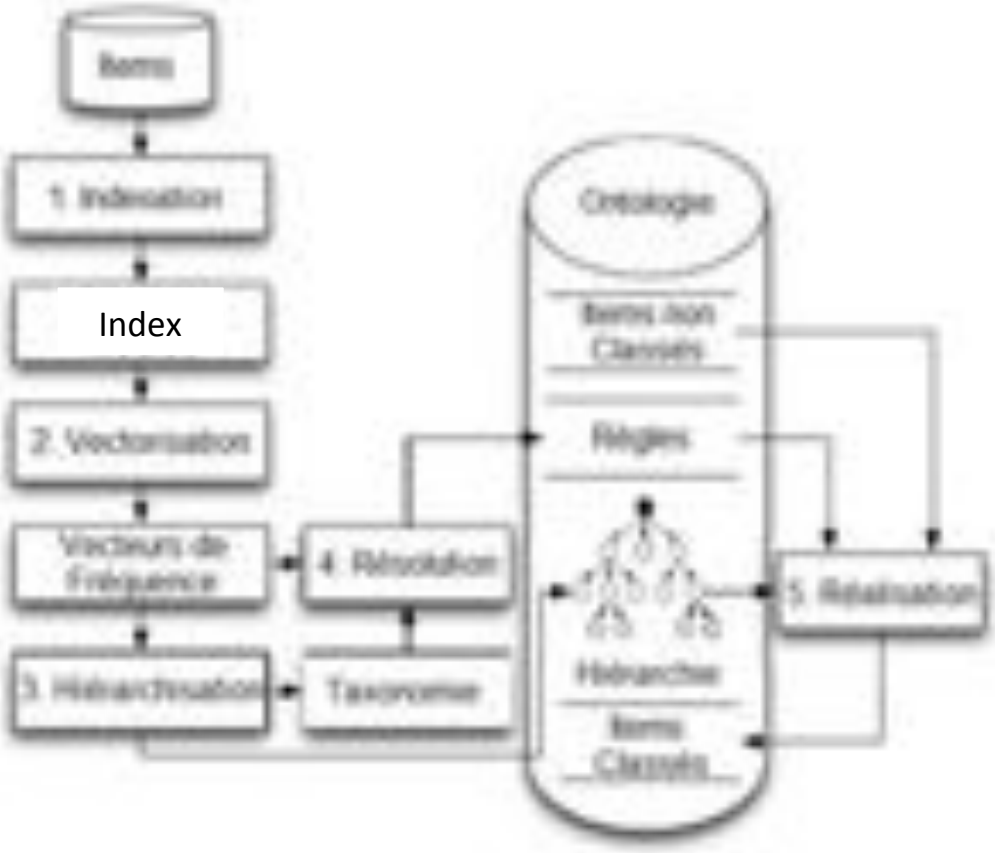
La méthode …

En cinq phases

Usage des logiques descriptives

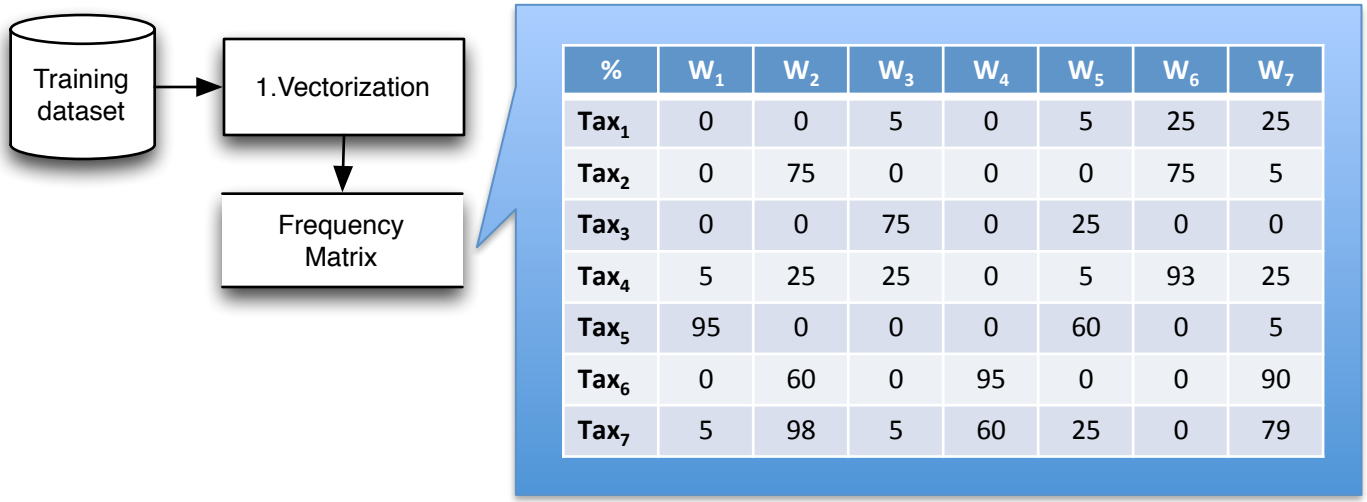
Prototype **vs 1.0**

Processus de traitement des données



Ontologies OWL et moteur d'inférence

1. Vectorisation

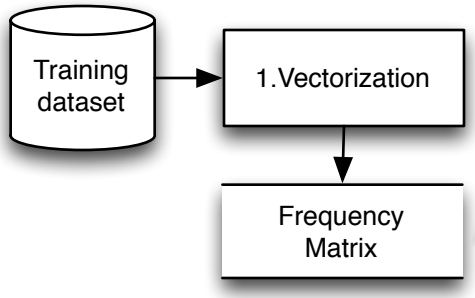


%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

Tax_i sont les termes de la taxonomie

W_i sont les mots fréquents dans les textes

1. Vectorisation

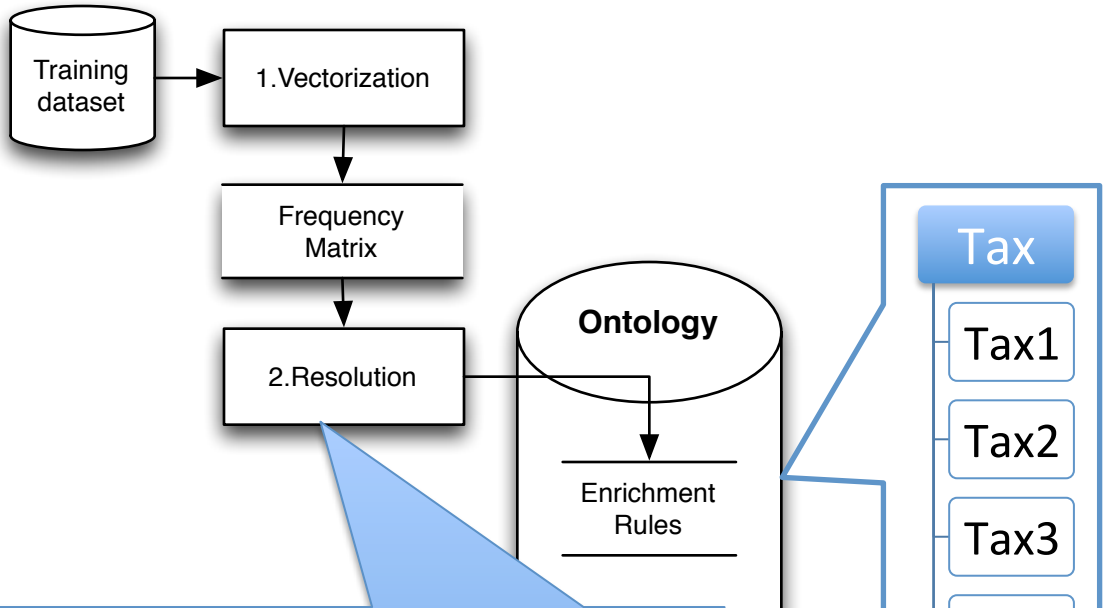


%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

91

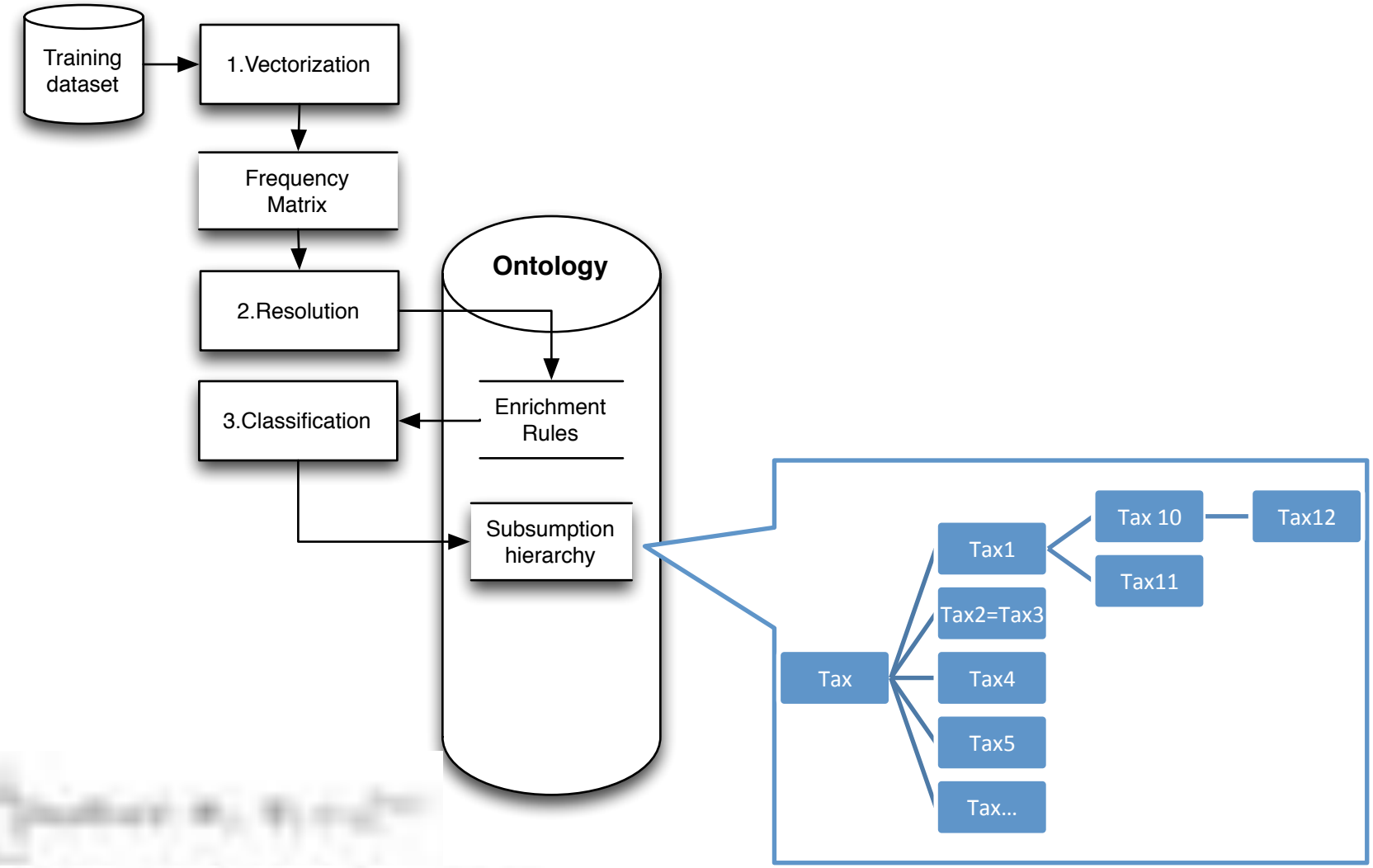
70

2. Résolution

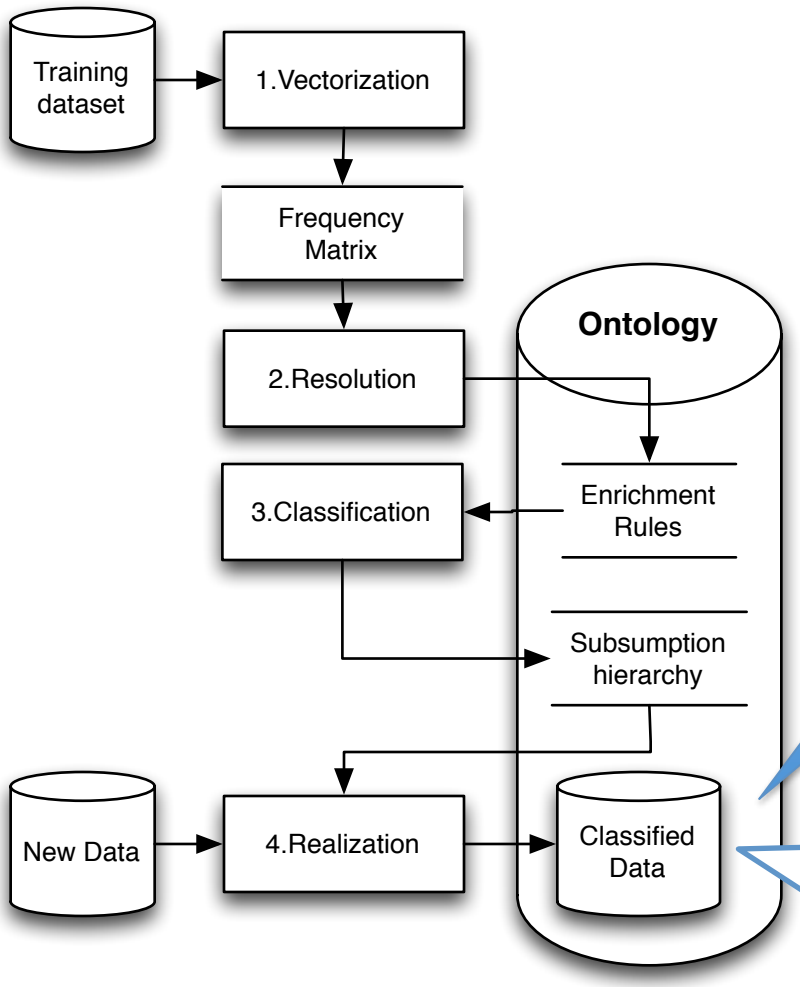


%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

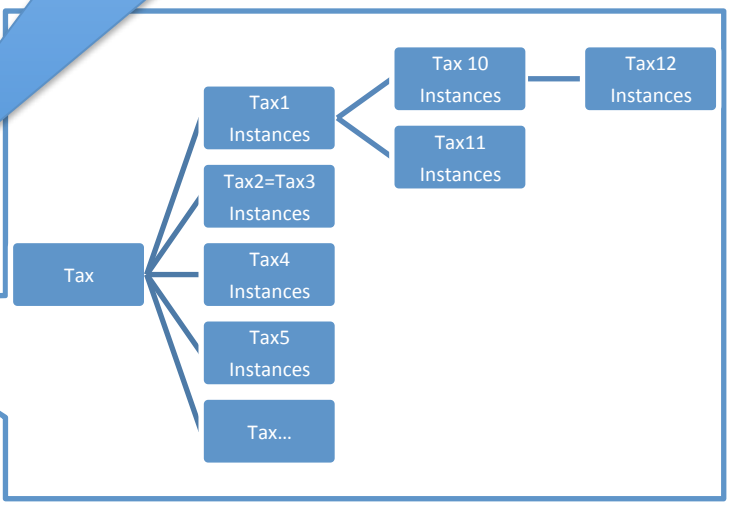
3.Classification



4. Réalisation



	L_1	L_2	L_3	L_4	L_5
Doc_1	x	-	-	-	x
Doc_2	-	x	x	-	x
Doc_3	x	-	-	x	-
Doc_4	x	-	-	-	-
Doc_5	-	x	-	x	-
Doc_6	-	-	x	-	x
Doc_7	x	-	-	x	-



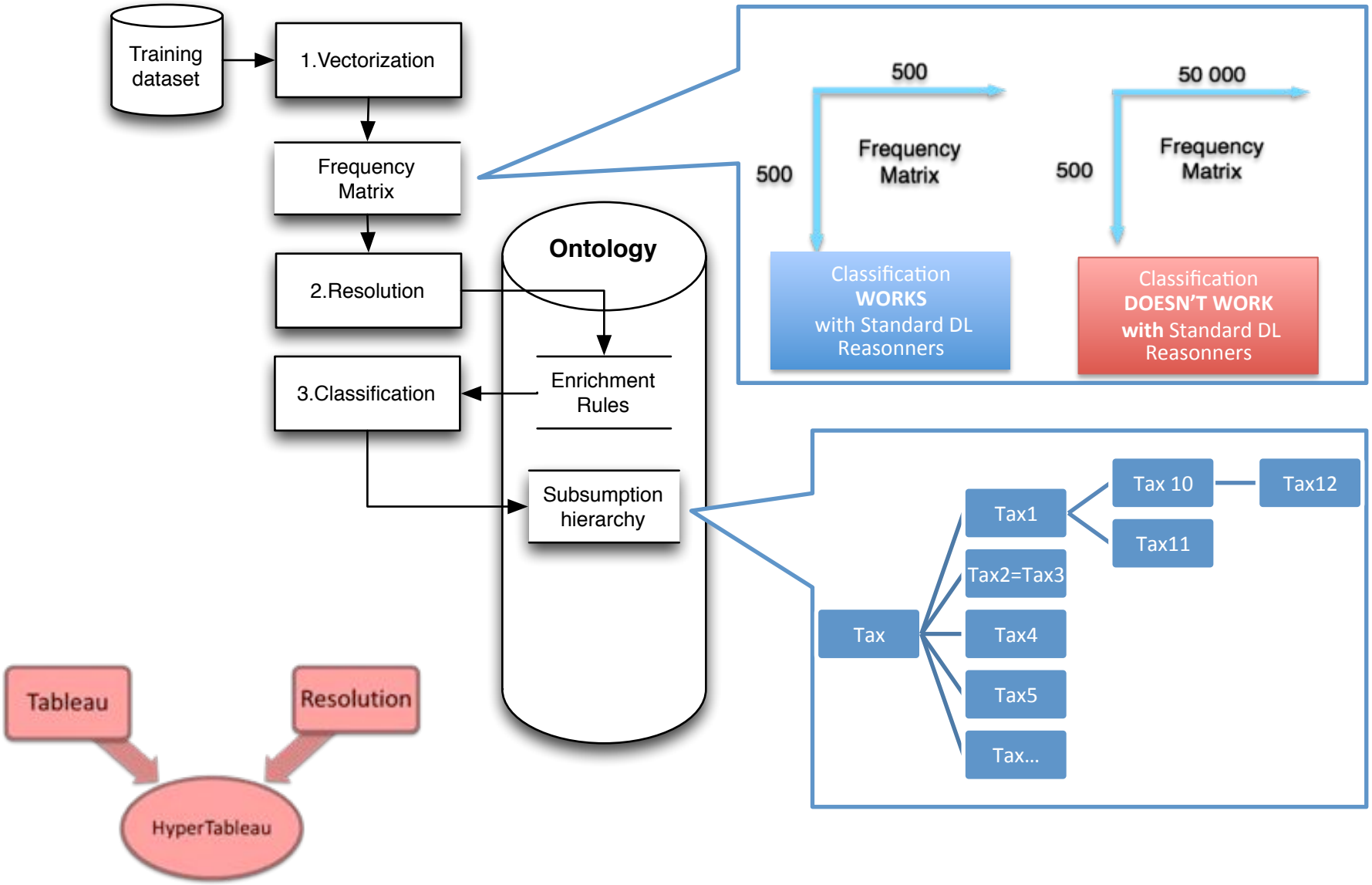
Résultats – Delicious dataset

α-rules	FaCT++	HermiT	Pellet
i7 4Go DDR3	50 s	n.e.m.[1]	n.e.m.
Xeon E3 24Go DDR3	-	8 h	18 h
$\alpha\beta$-rules	**FaCT++**	**HermiT**	**Pellet**
i7 4Go DDR3	n.e.m.	n.e.m.	n.e.m.
Xeon E3 24Go DDR3	n.e.m.	out[2]	out
Xeon E5 128Go DDR3	2 h / out[3]	out	out

Evaluation	Precision	Recall	F1-Mesure
Proposal with A-type rules	30%	6%	10%
HOMER [16]	-	-	25%

Our way to create rules (i.e. with an average of 10 terms for all rules) has the consequence the creation of some ruleless classes. With this method, for our 983 classes, only 427 have rules. There are 556 classes without labeling rule (obviously, theses classes should have had β rules). So there are classes that the predictive model can not affect. This impacts very negatively the Recall.

Résultats – Delicious dataset



La méthode Big Data

En cinq phases

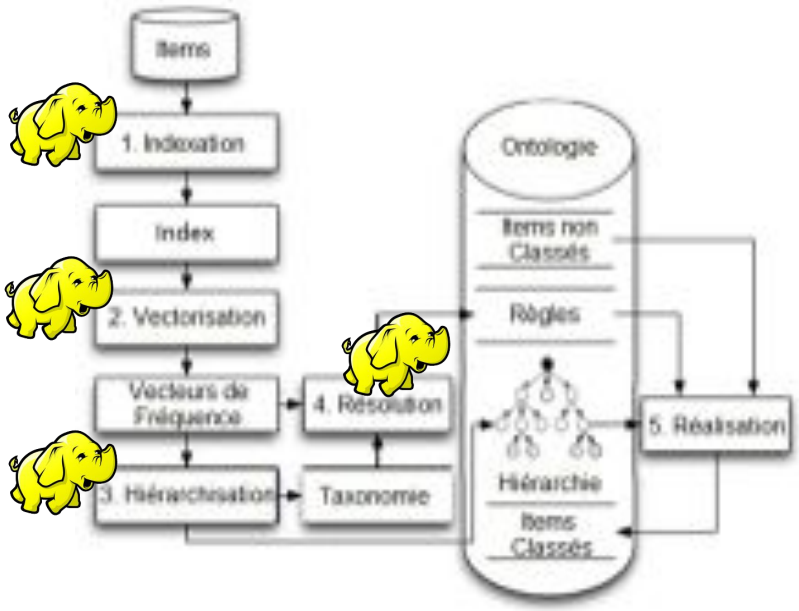
Usage des LD et des règles logiques

Prototype **vs 2.0**

Nouvelle architecture BigData

Montée en charge

Utilisation du **framework Hadoop** et du modèle **MapReduce**



Comment distribuer chacune des phases du processus ?

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

1. Indexation

Génération d'un index des termes

Parsing,
Traitement du
Langage Naturel*

* Tokenisation, Lemmatisation, Suppression des mots vides

2. Vectorisation

Détection des termes pertinents basée sur l'index

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

Les termes dont la fréquence est inférieure à un certain seuil sont rejetés

- Uni-grams (termes)
- N-grams (collocations) - Amélioration

2. Vectorisation - Vecteurs

Génération des vecteurs de fréquences

Liste de collocation
{ uni-grams + n-grams }

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5

Apache Solr

%
Tax_1
Tax_2
Tax_3
Tax_4
Tax_5
Tax_6
Tax_7

Seuil de pertinence

Algorithme des Stripes

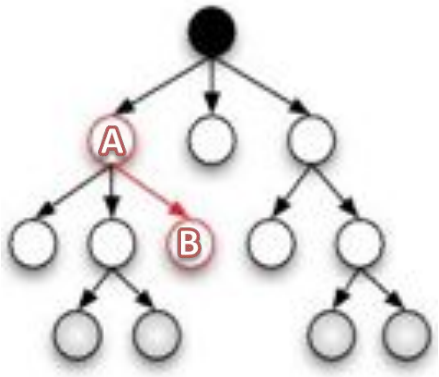
Liste des fréquences

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

3. Hiérarchisation

Algorithme de subsomption

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

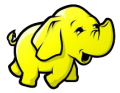


$$P(p|x) \geq t, P(x|p) < t$$

De Knijff, Jeroen and Frasincar, Flavius and Hogenboom, Frederik, 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering, Data & Knowledge Engineering, pp.54-69

3. Hiérarchisation

Algorithme de subsomption MapReduce



Map **Reduce**

Liste de cooccurrences

Label_A, (Label_B, Nb_Cooc)
Label_A, (Term_1, Nb_Cooc)
Label_B, (Label_D, Nb_Cooc)
Label_B, (Term_2, Nb_Cooc)
Term_2, (Label_E, Nb_Cooc)
…

<k,v> Paire

Parentés potentielles

Label_A, (Label_B, **Score**)
Label_B, (Label_D, **Score**)
Label_B, (Label_E, **Score**)
Label_C, (Label_A, **Score**)
Label_D, (Label_A, **Score**)
…

<k,v> Paire

Parentés validées

Label_A, Label_B
Label_B, Label_D
Label_B, Label_E
Label_C, Label_E
Label_D, Label_A
…

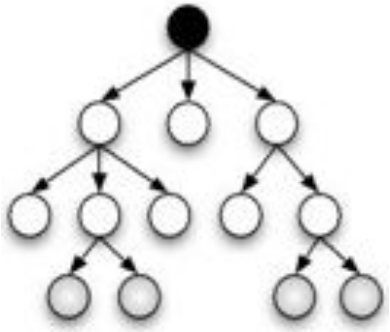
<k,v> Paire

$$P(p|x) \geq t, P(x|p) < t$$

Score calculé à partir des cooccurrences

3. Hiérarchisation

Hiérarchie chargée dans un triplestore



Hiérarchie

Stardog

4. Résolution

Génération de règles de classification

Clauses de Horn (SWRL et Stardog)

%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

Seuils α et β

α : $Item(?d) \wedge Word(?w1) \wedge hasWord(?d,?w1) \rightarrow Tax(?d1)$

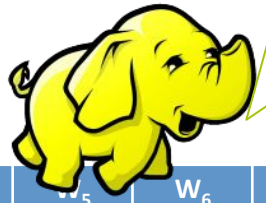
β : $Item(?d) \wedge Word(?w1) \wedge hasWord(?d,?w1) \wedge Word(?w2) \wedge hasWord(?d,?w2) \rightarrow Tax(?d1)$

Item: *nouveau document*

4. Résolution

Génération de règles de classification

Approche MapReduce
« diviser pour régner »



%	W_1	W_2	W_3	W_4	W_5	W_6	W_7
Tax_1	0	0	5	0	5	25	25
Tax_2	0	75	0	0	0	75	5
Tax_3	0	0	75	0	25	0	0
Tax_4	5	25	25	0	5	93	25
Tax_5	95	0	0	0	60	0	5
Tax_6	0	60	0	95	0	0	90
Tax_7	5	98	5	60	25	0	79

Tax_2, (Term_2, Nb_Cooc)
Tax_2, (Term_6, Nb_Cooc)
Tax_2, (Term_8, Nb_Cooc)
Tax_2, (Term_9, Nb_Cooc)
...

<k,v>
Paire

4. Résolution

Génération de règles de classification

Item*(?d)*∧*Word(?w1)*∧*hasWord(?d,?w1)*
→ *Tax(?d1)*

Règles

Stardog

5. Réalisation

Utilisation des règles pour classer

Règles

Item*(?d)*∧*Word(?w1)*∧*hasWord(?d,?w1)*
→ *Tax(?d1)*

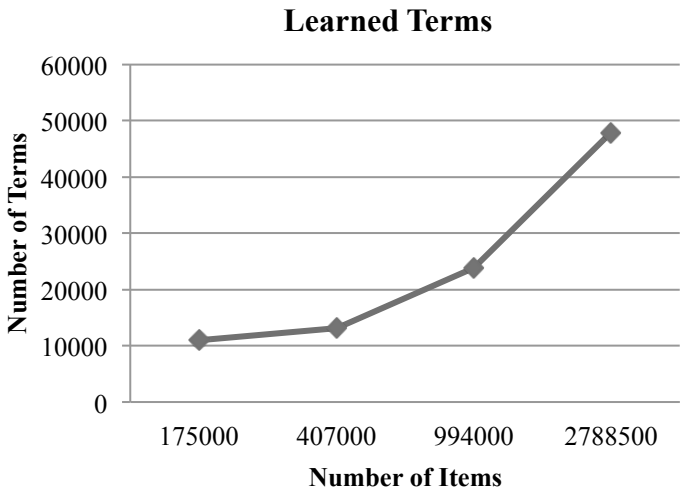
Classés dans Stardog

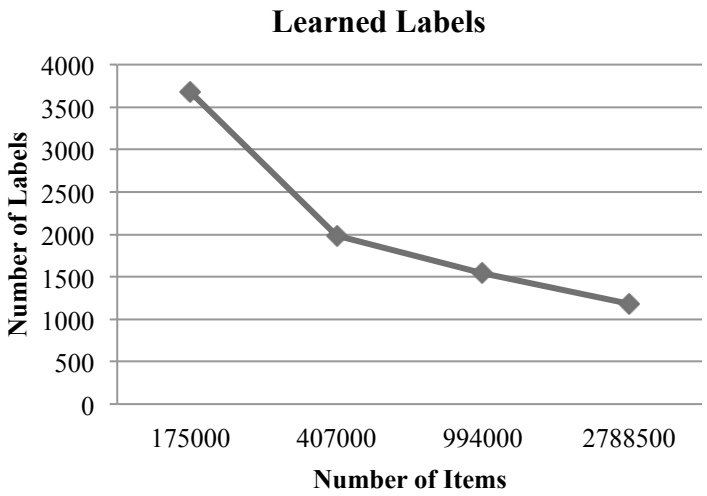
	L_1	L_2	L_3	L_4	L_5
Doc_1	x	-	-	-	x
Doc_2	-	x	x	-	x
Doc_3	x	-	-	x	-
Doc_4	x	-	-	-	-
Doc_5	-	x	-	x	-
Doc_6	-	-	x	-	x
Doc_7	x	-	-	x	-

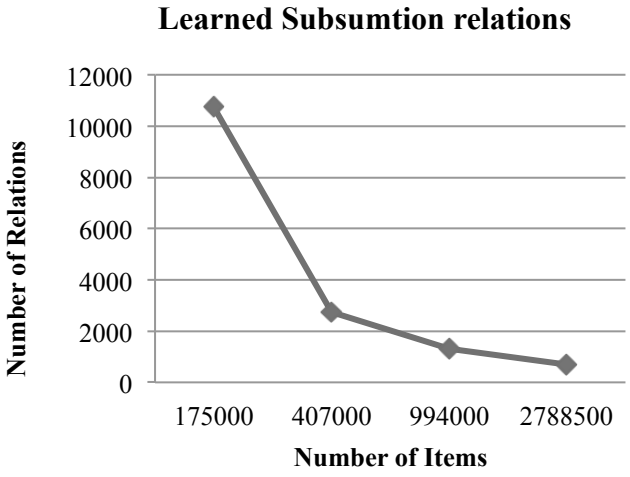
Vecteurs « document »

Résultats

Evaluation de l'apprentissage

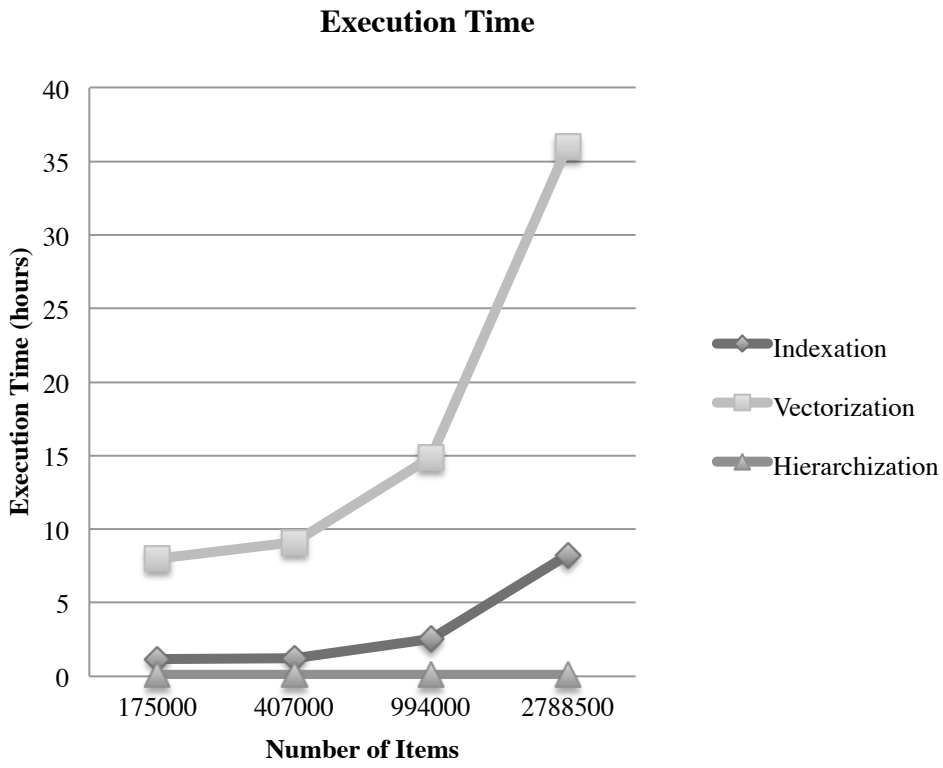






Résultats préliminaires

Evaluation de la performance du processus (4 nœuds):



Conclusion

Sémantique et connaissances

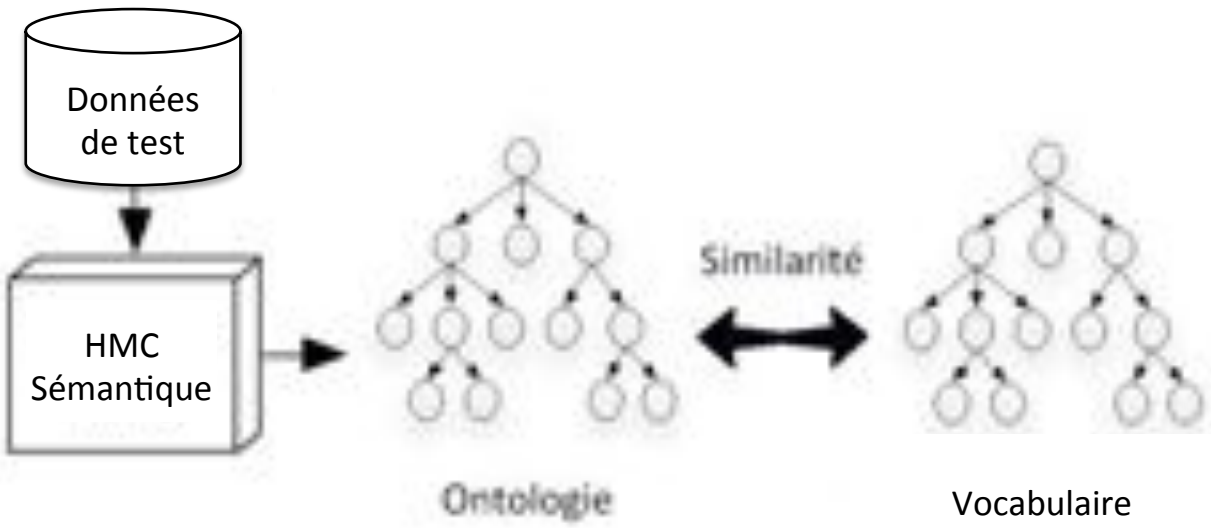
- La sémantique
- Problématique
- Application au Big Data

Merci …

Questions ?

Travaux en cours

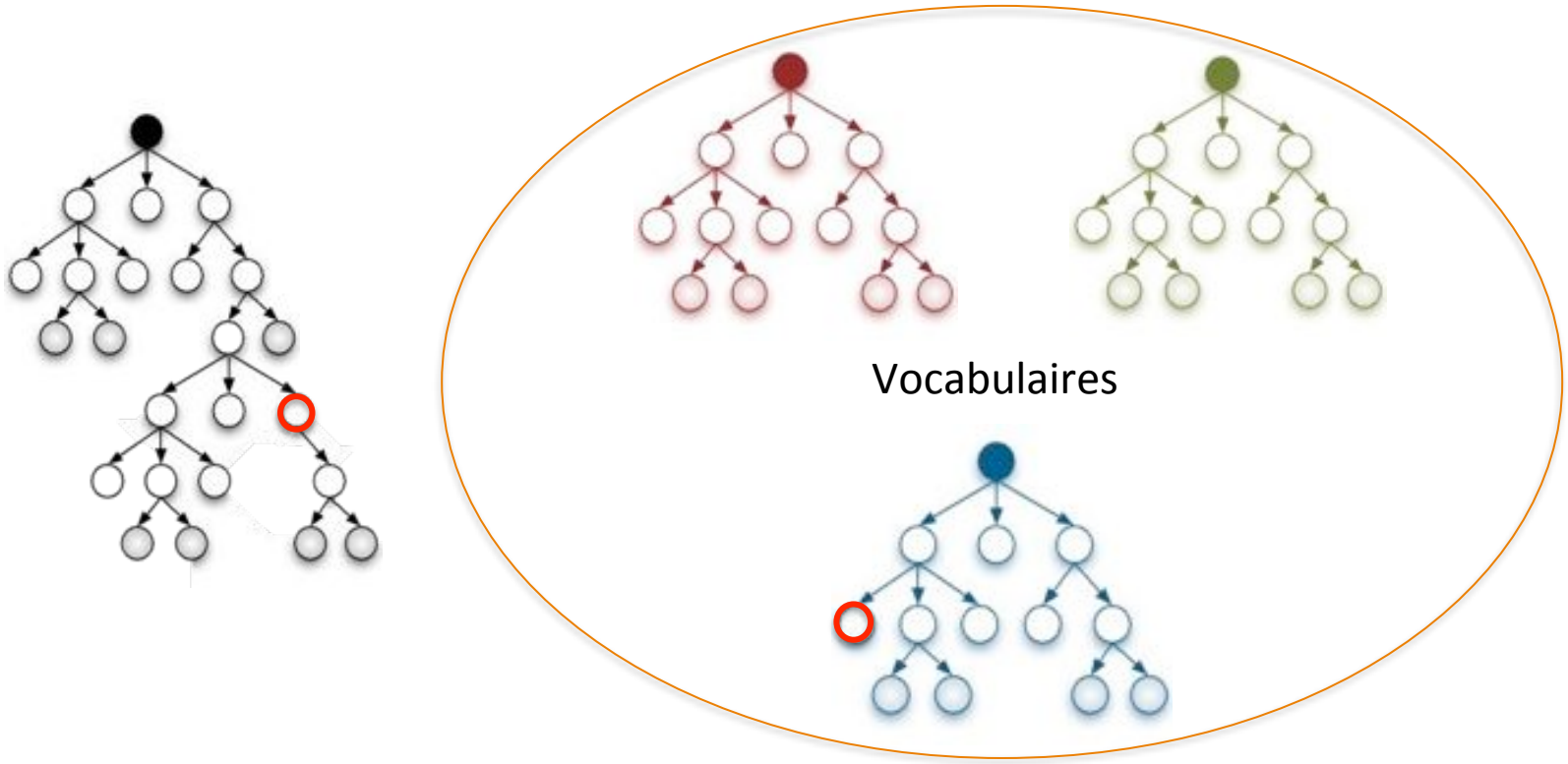
Evaluation de la qualité de la hiérarchie :



Profondeur, similarité des noeuds, topologie

Travaux en cours

Evaluation de la qualité de la classification :



Précision, rappel, F1

L'architecture …

Ensemble d'outils logiciels

Architecture

Experimentale

FIRSTECOPRO'FIL
BOOSTEZ VOTRE BUSINESS AVEC LA VEILLE PERSONNALISÉE

Fondamentale

Classification
Hiérarchique Multi-label Sémantique

Technologique

Semantic

Stardog

Indexing

Apache Solr

Distributed

hadoop

Streaming

APACHE STORM™
Distributed · Resilient · Real-time