

**EXENSA ET LES DATA**  
**INPI :**  
**PRÉSENTATION DE**  
**BOMERCE, MOTEUR**  
**D'ANALYSE DES**  
**DONNÉES BREVETS**

**VI**  
**inno**  
**-VA**  
**tion**

Date 15/06/2017

# A propos de cette présentation

- On parle de modèles prédictifs
  - Le **sens** des mots, des documents
  - Le **comportement** des gens
  - Les **liens** entre entités
- Cas d'usage
  - Exploration documentaire
  - Recommandation d'auteurs
  - Tagging

# Petite histoire

- 2009 : découverte d'une méthode d'analyse (**NCISC**)
- 2009-2010 : expérimentations, améliorations, extensions
- 2011 : création d'eXenSa sur recommandation e-commerce
- Fin 2013 : R&D sur moteur d'analyse eXenGine
  - Texte
  - Relations (hyperliens, par exemple)
  - Comportements
- 2014 : premier client (Augure : marketing d'influence)
- 2016 : Version Webscale Online – analyse de CommonCrawl



# Modèles prédictifs

- **Objectif :**
  - Donner des documents similaires
  - Classifier / Segmenter
  - Recommander des produits
  - Prédire des personnes intéressantes
- **Moyen :**
  - Créer une synthèse numérique des données

# Modèles prédictifs

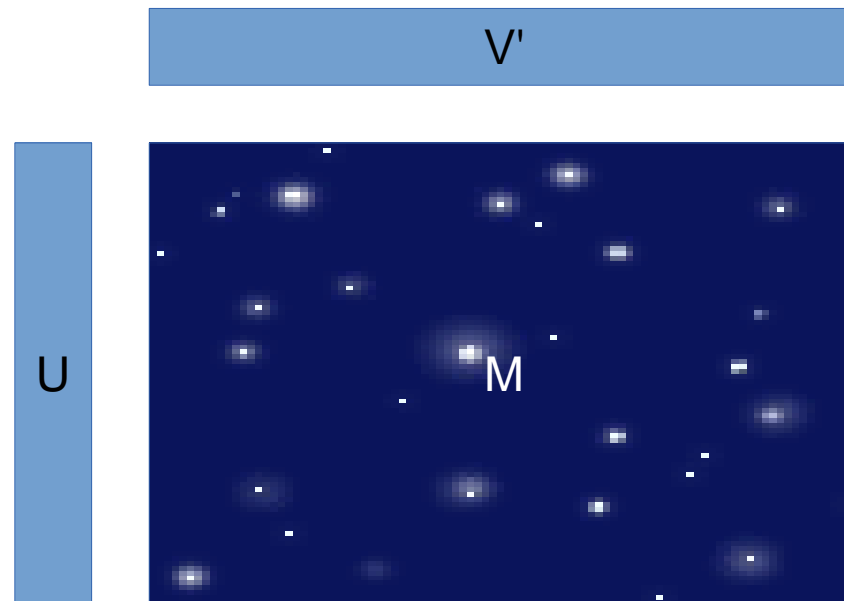
- **Données :**

- Matrices de co-occurrence (mots, événements). Par ex: 2 produits ont été achetés par la même personne
- Matrices de mots/documents (modèle Bag of Words) : un document est représenté par un vecteur des mots contenus.
- Matrice d'adjacence pour un graphe

Note : ces matrices peuvent être explicites ou implicites (dans ce cas on a les données brutes qui arrivent en flux)

# Factorisation de matrice

- En entrée, une grosse matrice creuse  $M$  issue d'un échantillonnage
- Le but est de trouver  $U$  et  $V$  tel que  $U^*V' \approx M$



# Factorisation de matrice

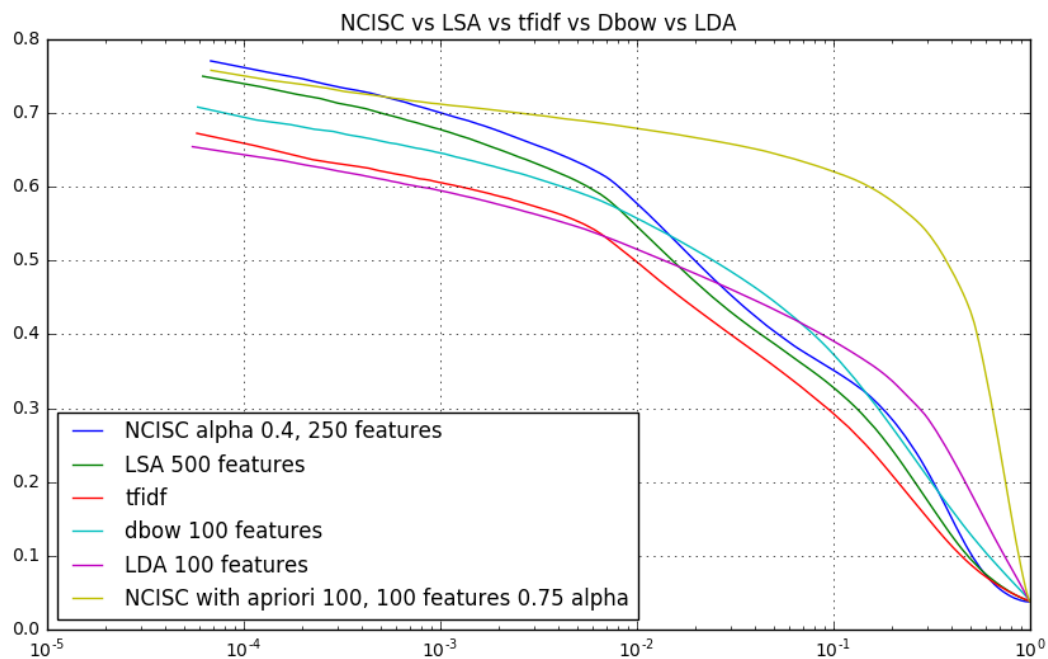
- Au final, on passe
  - d'une représentation creuse en très haute dimension :  $\text{docI}(0,0,0,0,\dots,I,\dots,4,\dots)$  où les valeurs non nulles représentent les occurrences d'un mot particulier dans le document
  - à une représentation dense avec quelques centaines de dimensions:  $\text{docI}(0.14329,0.0871,-0.1889,\dots)$  où les dimensions représentent une direction synthétique

# NCISC

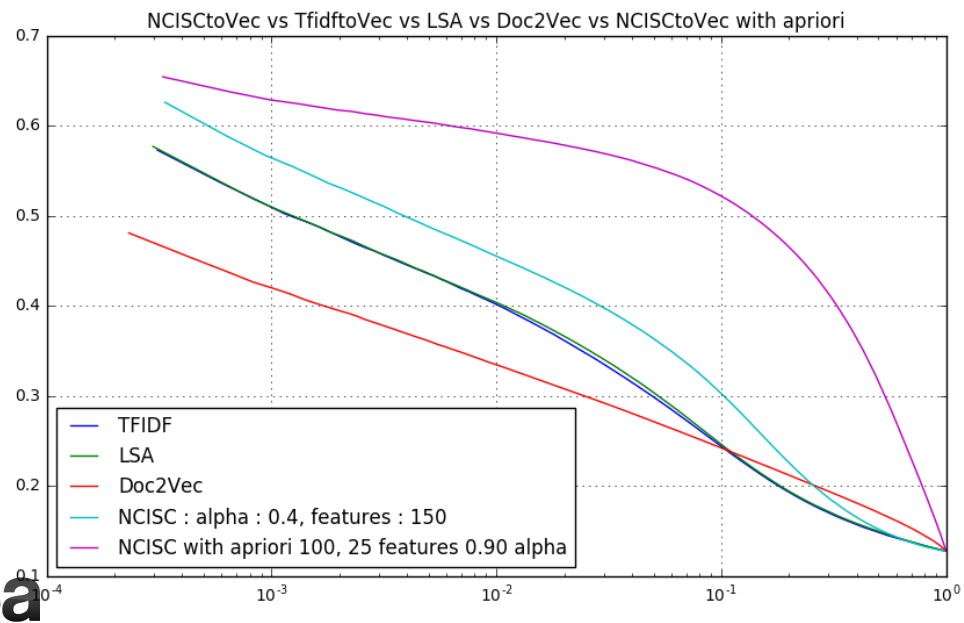
- Propriétés intéressantes
  - Très rapide (10x plus rapide que LSA optimisé, 50x plus rapide que Doc2Vec, plus de 500x plus rapide que LDA)
  - Scalable (distribuable)
  - Robuste (peu de préparation nécessaire)
  - Excellente qualité
- Propriété rare
  - Permet d'injecter des connaissances externes



# RCVI (reuters)



# Ohsumed (abstract médicaux)



# Cas d'usage

- Exploration documentaire
  - Grande quantité de documents
  - On cherche des informations
- Autres cas d'usage
  - E-commerce (recommandation)
  - Marketing réseaux sociaux

# Détail d'un cas d'usage

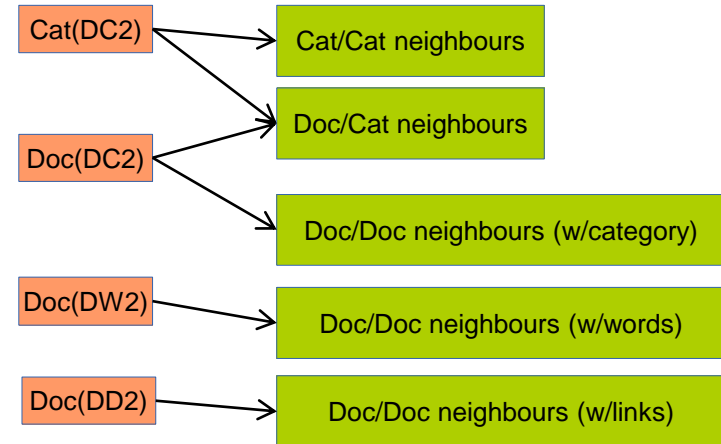
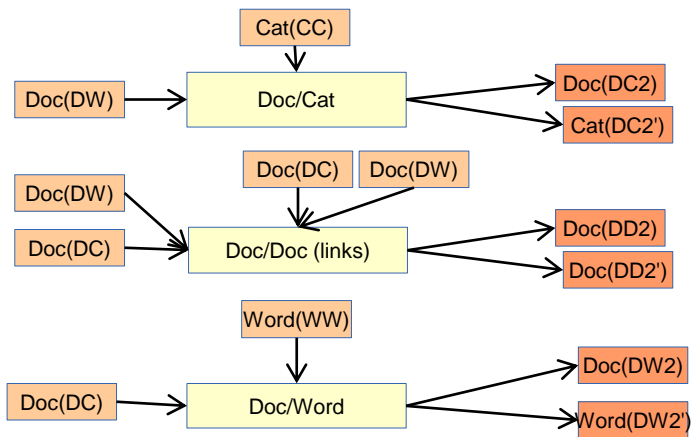
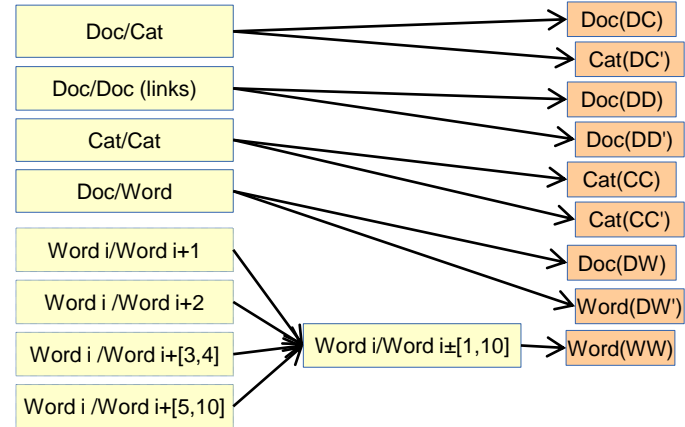
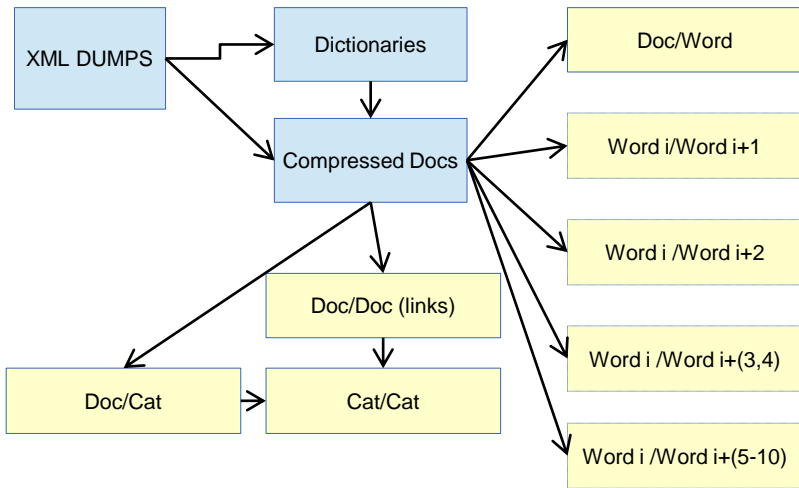
- Moteur d'exploration Wikipedia
  - 4.5M articles
  - 1M categories
  - 113M links
  - 2M words



# Détail d'un cas d'usage

- Dans Wikipedia, actuellement:
  - Se promener dans le graphe des pages (très gros)
  - Trouver la liste des contributeurs, catégories
- Une autre vision :
  - Utiliser les similarités prédites entre pages, catégories, etc. basées sur le contenu texte, les liens, les categories, ...

# Workflow



# wikinsights.org

WikInsights - powered by eXenGine ©



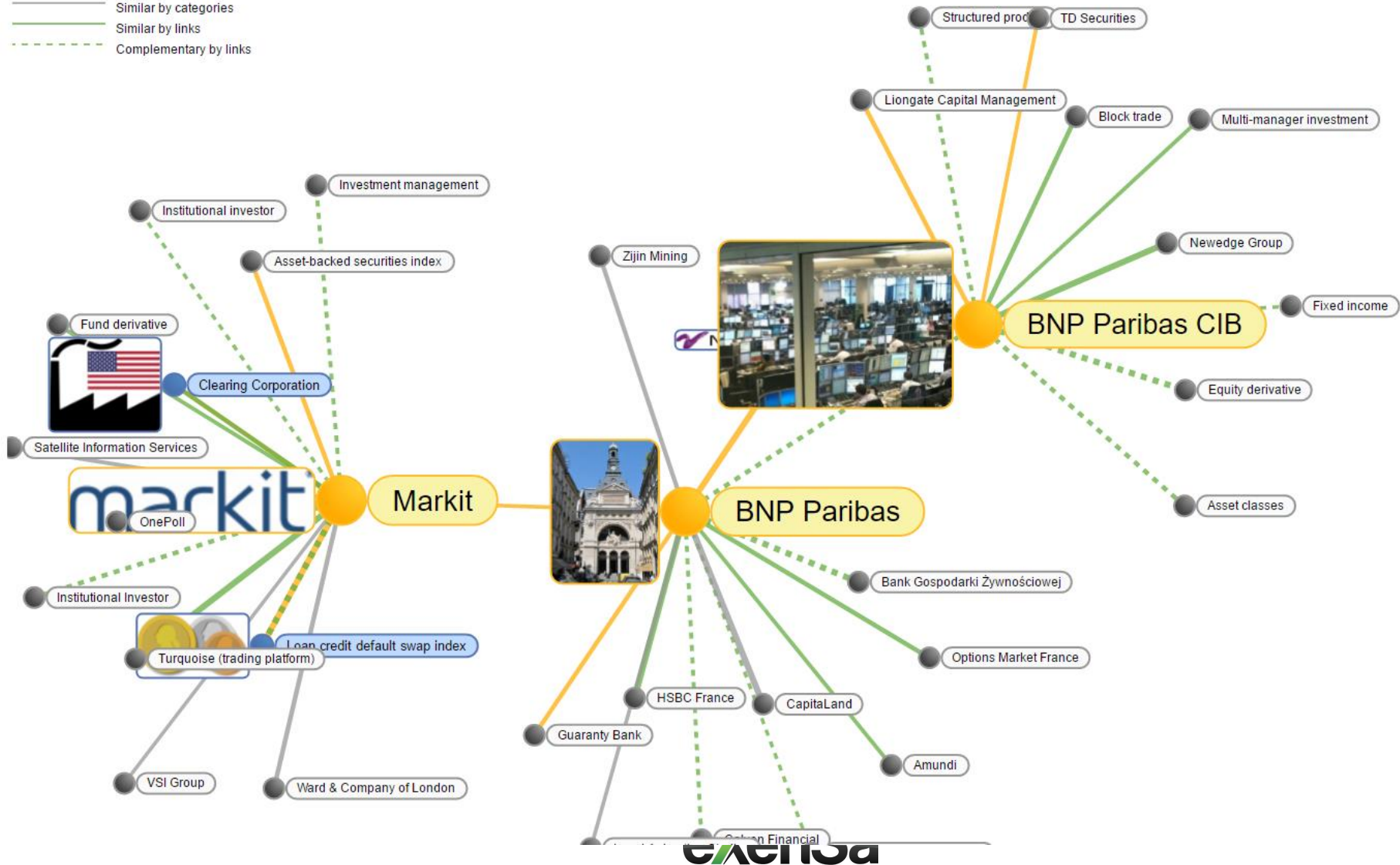
BNP Paribas

eXenSa

2014

WikInsights.org powered by eXenSa eXenGine (c) 2014

- Similar by content
- Similar by categories
- Similar by links
- - - Complementary by links



# Brevets / INPI

- 7.5Go de data, environ 1/2 million de brevets
- Pas de liens entre brevets
- Très peu de structure, variabilité du format
- Informations sémantiques souvent noyées dans le texte

# Brevets / INPI

- Créations de quelques modèles
  - Par “intro” (début du brevet)
  - Complets
  - Par revendications
- Service en ligne : <http://patents.exensa.net>
- API dispo à la demande ([guillaume.pitel@exensa.com](mailto:guillaume.pitel@exensa.com))



# Patents.exensa.net

Patent Demo - powered by eXenGine ©

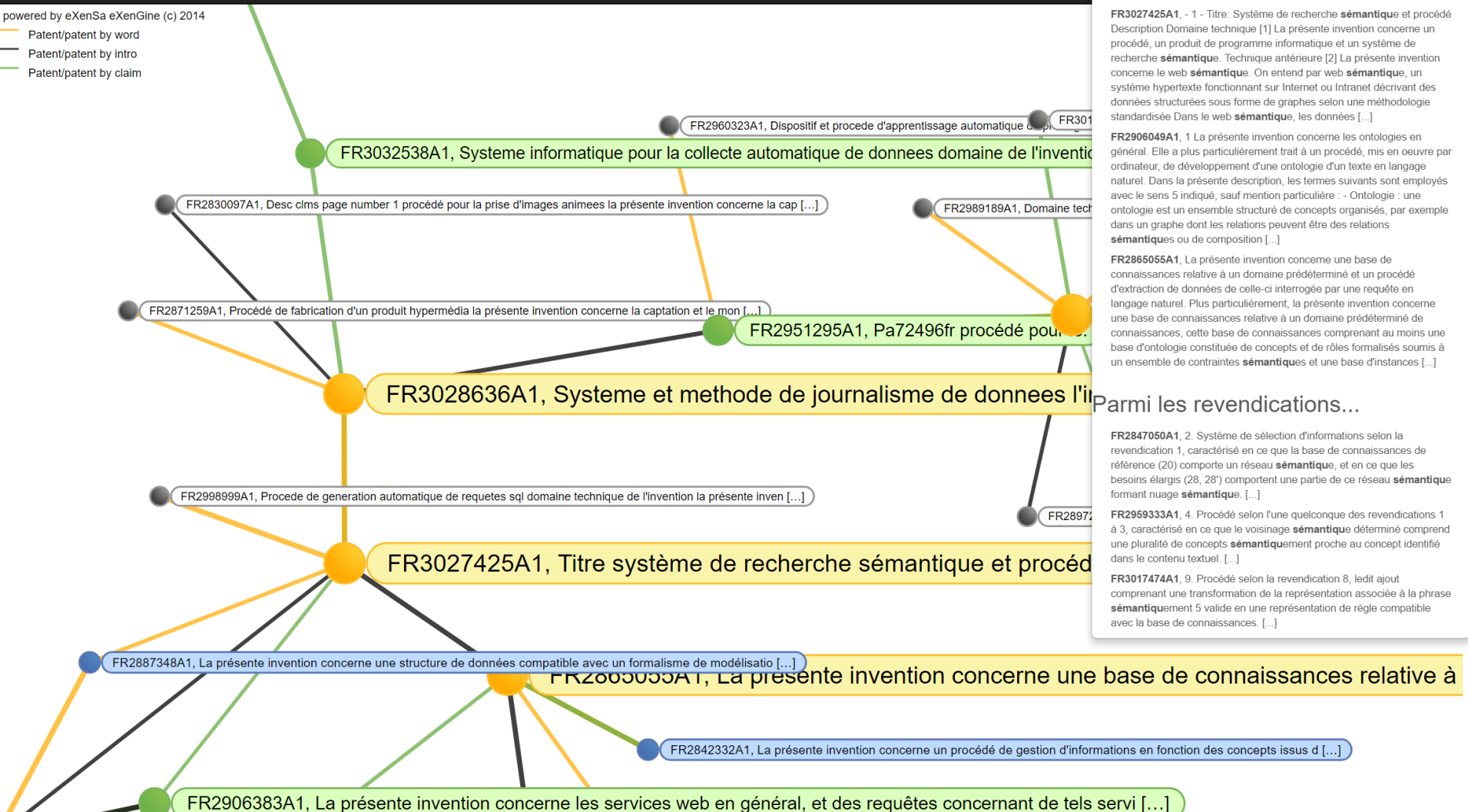
eXenSa  
2014



sémantiqu

Patent Demo powered by eXenSa eXenGine (c) 2014

- Patent/patent by word
- Patent/patent by intro
- Patent/patent by claim



## Parmi les introductions...

**FR3027425A1**, - 1 - Titre: Système de recherche **sémantique** et procédé  
Description Domaine technique [1] La présente invention concerne un procédé, un produit de programme informatique et un système de recherche **sémantique**. Technique antérieure [2] La présente invention concerne le web **sémantique**. On entend par web **sémantique**, un système hypertexte fonctionnant sur Internet ou Intranet décrivant des données structurées sous forme de graphes selon une méthodologie standardisée Dans le web **sémantique**, les données [...]

**FR2906049A1**, 1 La présente invention concerne les ontologies en général. Elle a plus particulièrement trait à un procédé, mis en oeuvre par ordinateur, de développement d'une ontologie d'un texte en langage naturel. Dans la présente description, les termes suivants sont employés avec le sens 5 indiqué, sauf mention particulière : - Ontologie : une ontologie est un ensemble structuré de concepts organisés, par exemple dans un graphe dont les relations peuvent être des relations **sémantiques** ou de composition [...]

**FR2865055A1**, La présente invention concerne une base de connaissances relative à un domaine prédéterminé et un procédé d'extraction de données de celle-ci interrogée par une requête en langage naturel. Plus particulièrement, la présente invention concerne une base de connaissances relative à un domaine prédéterminé de connaissances, cette base de connaissances comprenant au moins une base d'ontologie constituée de concepts et de rôles formalisés soumis à un ensemble de contraintes **sémantiques** et une base d'instances [...]

## Parmi les revendications...

**FR2847050A1**, 2. Système de sélection d'informations selon la revendication 1, caractérisé en ce que la base de connaissances de référence (20) comporte un réseau **sémantique**, et en ce que les besoins élargis (28, 28') comportent une partie de ce réseau **sémantique** formant nuage **sémantique**. [...]

**FR2959333A1**, 4. Procédé selon l'une quelconque des revendications 1 à 3, caractérisé en ce que le voisinage **sémantique** déterminé comprend une pluralité de concepts **sémantiquement** proche au concept identifié dans le contenu textuel. [...]

**FR3017474A1**, 9. Procédé selon la revendication 8, ledit ajout comprenant une transformation de la représentation associée à la phrase **sémantiquement** 5 valide en une représentation de règle compatible avec la base de connaissances. [...]