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

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

eXenSa

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

eXenSa

Modèles prédictifs

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

eXenSa

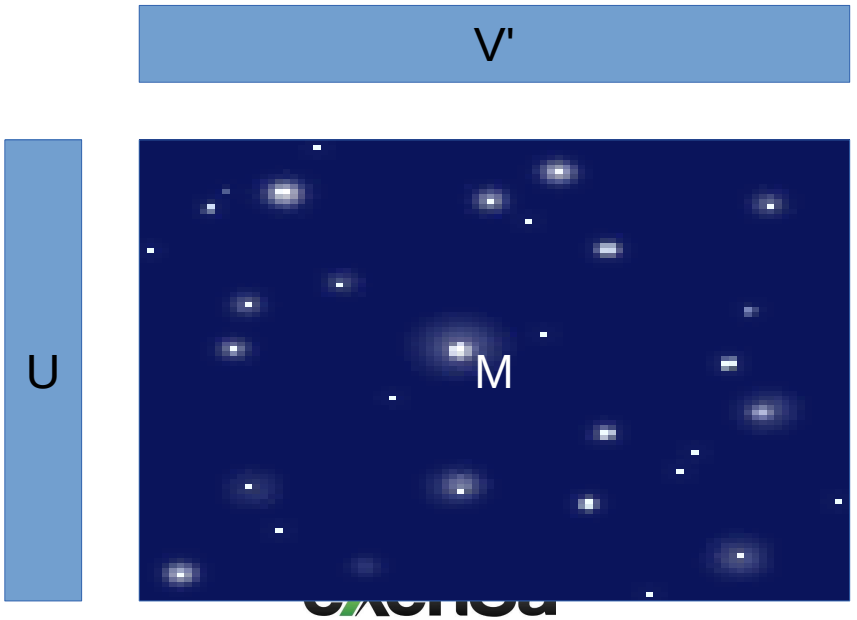
Modèles prédictifs

- **Données :**
  - Matrices de co-occurrence (mots, evénéments). 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' ≈ M



Factorisation de matrice

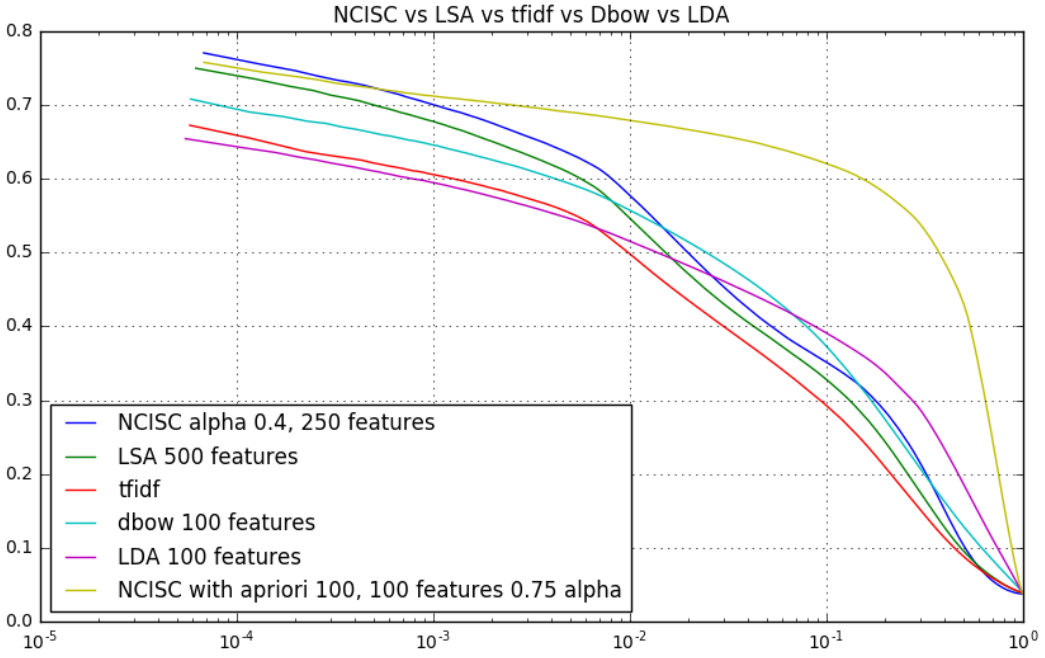
- Au final, on passe
  - d'une representation creuse en très haute dimension : doc1(0,0,0,0,…,1,….,4,…) ou les valeurs non nulles représentent les occurrences d'un mot particulier dans le document
  - à une representation dense avec quelques centaines de dimensions: doc1(0.14329,0.0871,-0.1889,…) où les dimensions représentent une direction synthétique

NCISC

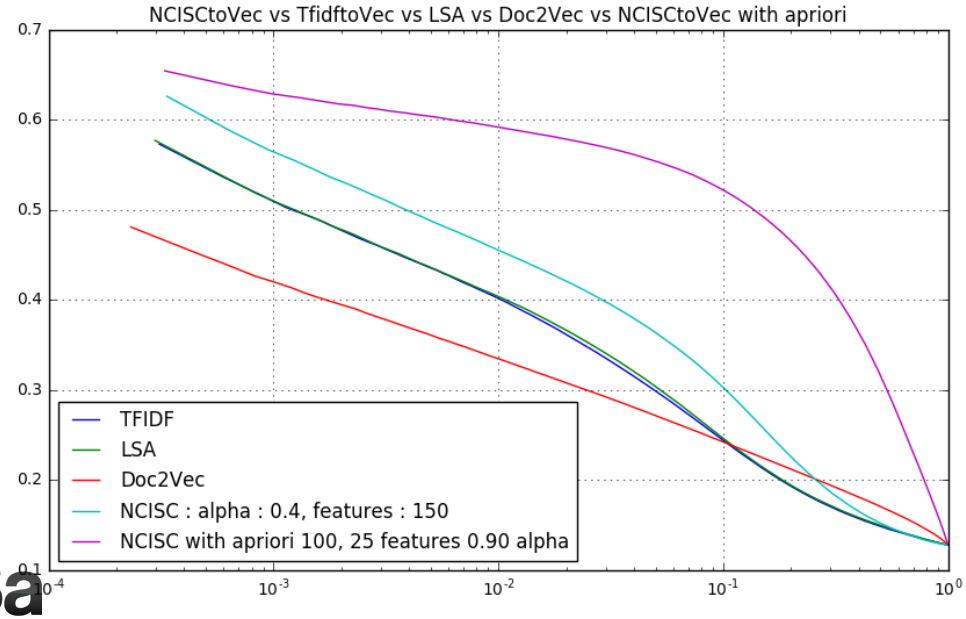
- Propriétés intéressantes
  - Très rapide (10x plus rapide que LSA optimisé, 50x plus rapide que Doc2Vec, plus de 500x plus rapide que LDA)
  - Scalable (distribuable)
  - Robuste (peu de préparation nécessaire)
  - Excellente qualité
- Propriété rare
  - Permet d'injecter des connaissances externes

eXenSa

RCV1 (reuters)



Ohsumed (abstract médicaux)



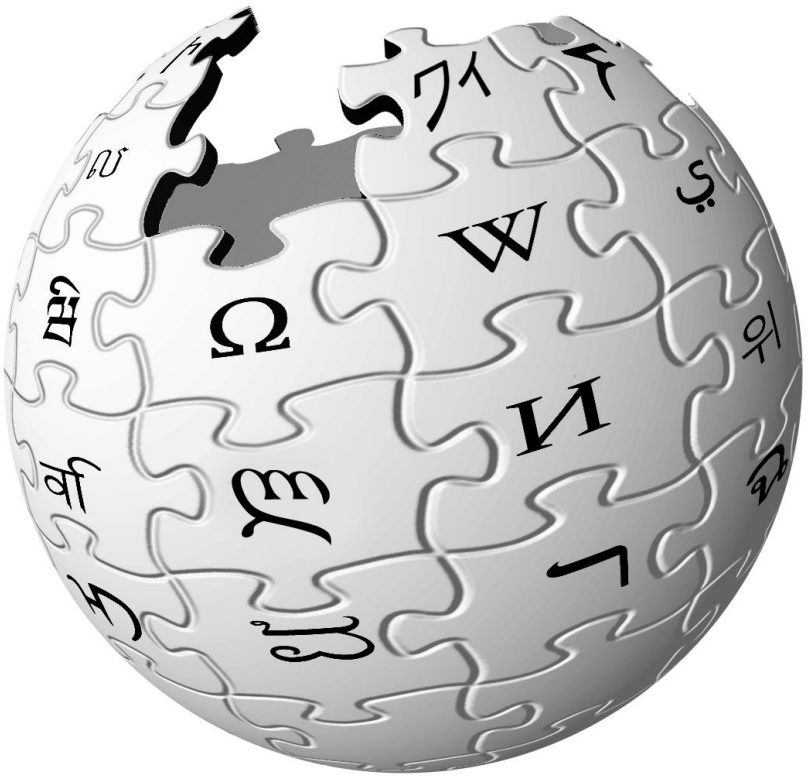
Cas d'usage

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

eXenSa

Détail d'un cas d'usage

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



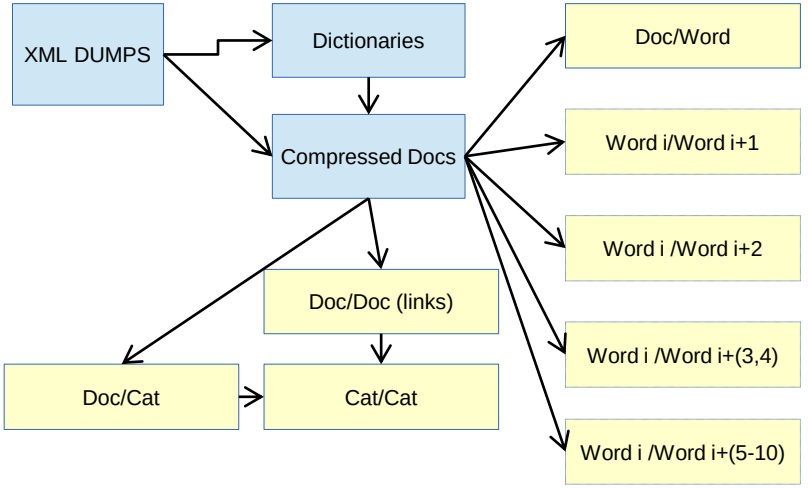
eXenSa

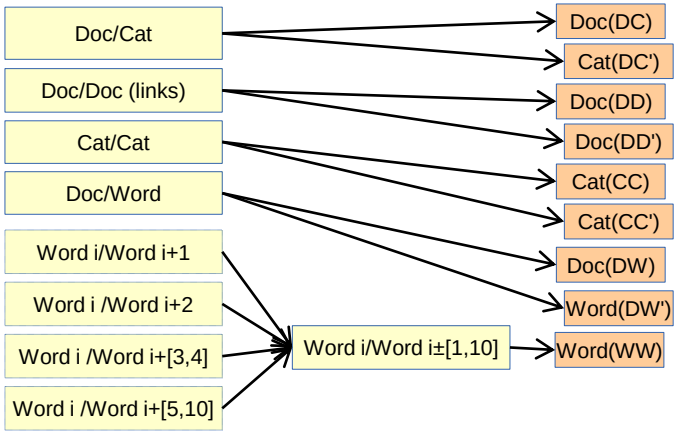
Détail d'un cas d'usage

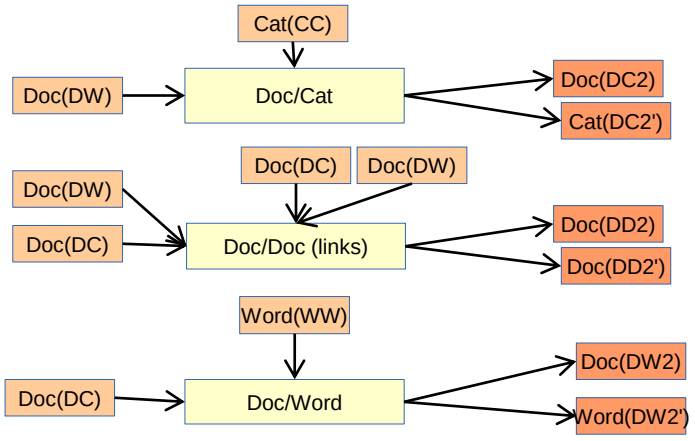
- Dans Wikipedia, actuellement:
  - Se promener dans le graphe des pages (très gros)
  - Trouver la liste des contributeurs, catégories
- Une autre vision :
  - Utiliser les similarités prédites entre pages, catégories, etc. basées sur le contenu texte, les liens, les categories, …

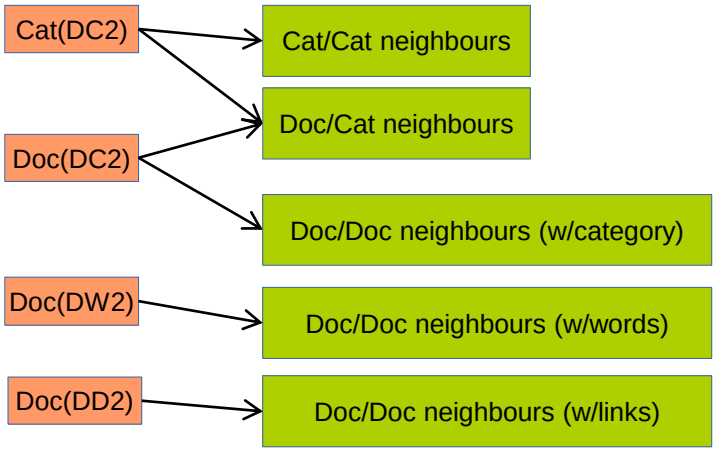
eXenSa

Workflow



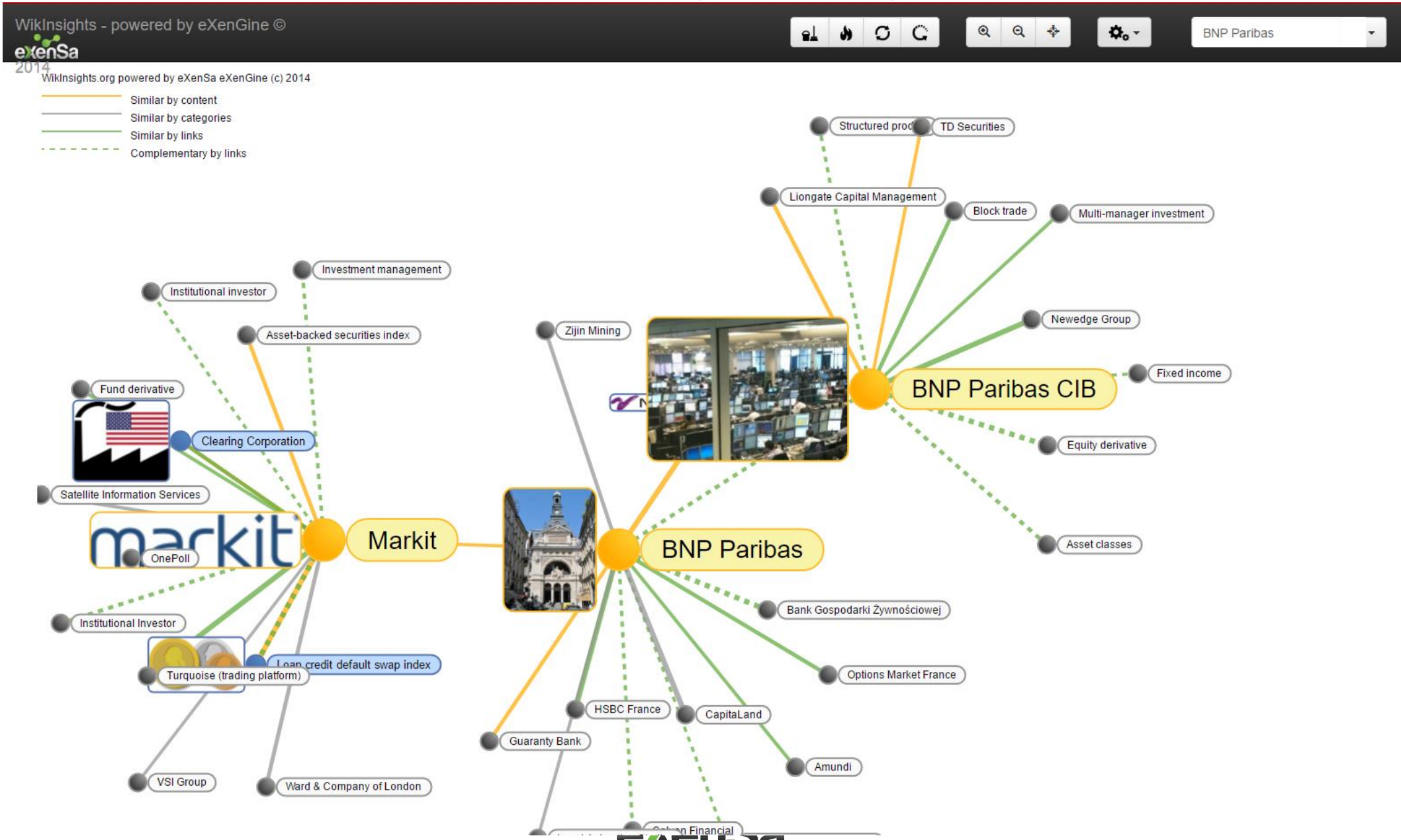






eXenSa

wikinsights.org



Brevets / INPI

- 7.5Go de data, environ ½ million de brevets
- Pas de liens entre brevets
- Très peu de structure, variabilité du format
- Informations sémantiques souvent noyées dans le texte

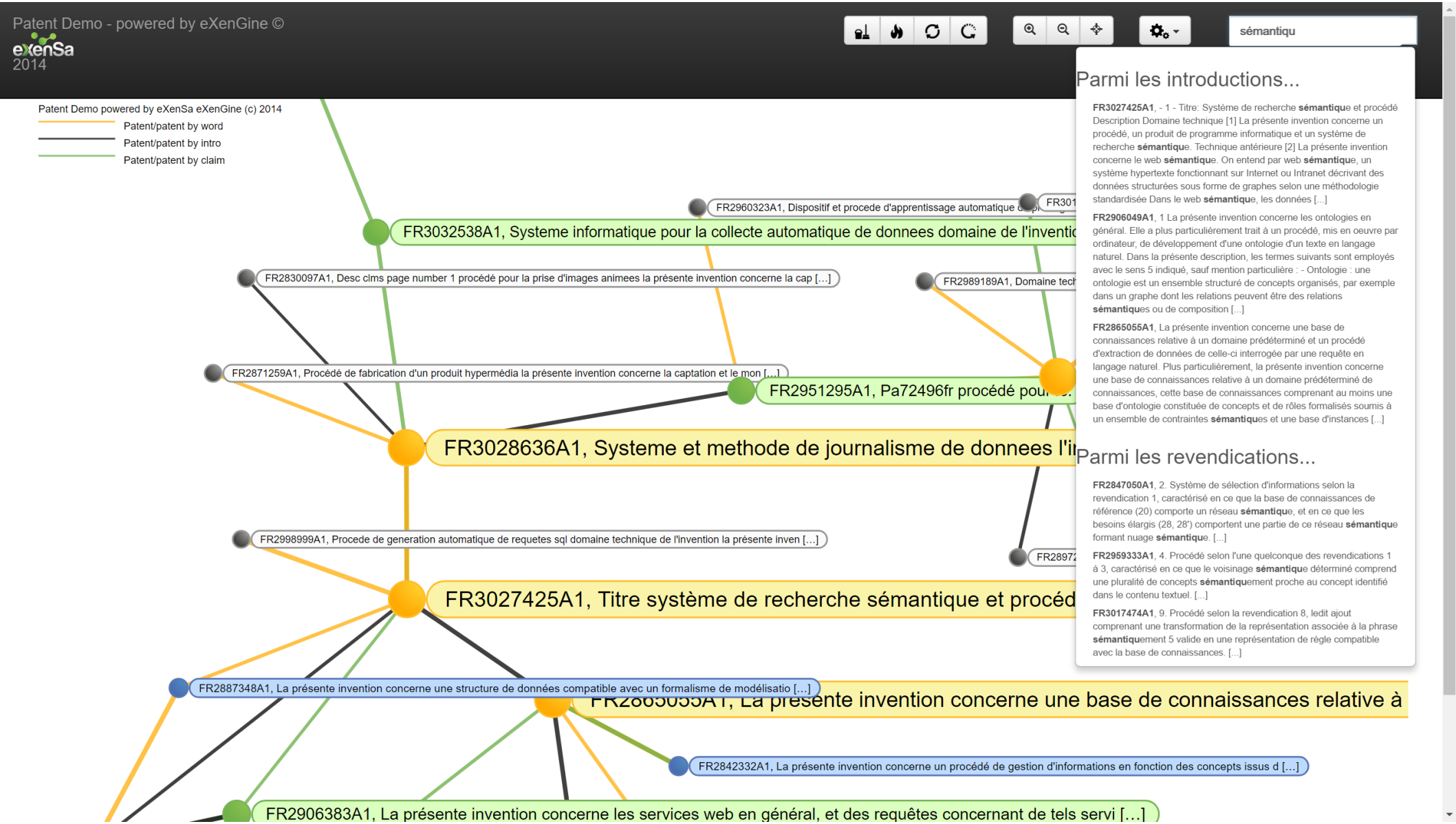
eXenSa

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)

eXenSa

Patents.exensa.net



eXenSa