



# Structures hybrides : l'apport des infrastructures libres aux moteurs de recherche sémantiques



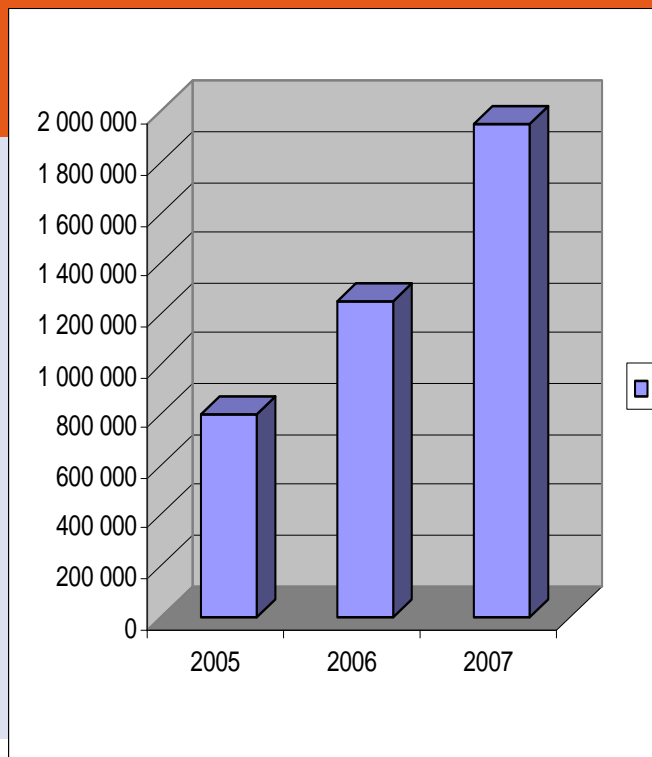
Des moteurs  
de recherche  
qui comprennent  
votre métier



## > Contexte

- 2001 Création par une équipe de 12 spécialistes du Traitement automatique des langues (venant de Erli–Lexiquest)
- Choix de se baser sur des composants open-source dès le départ
- Afin de se concentrer sur notre cœur de compétence
  - NLP, analyse sémantique, extraction et catégorisation
- Tendances aujourd'hui se généralise
  - 80% des logiciels commerciaux contiendront des composants open source d'ici 2012 (Gartner)

# > Une centaine de clients



## > Quels moteurs pour quels usages

### > Moteurs généralistes

- Indexent tout : Web, Intranet, Desktop
- De manière indifférenciée
  - Tous les documents et tous les utilisateurs sont traités de la même manière
  - Exemples : Google, Exalead. ... Vista

### > Moteurs spécialisés (vertical search)

- Ont une connaissance de la nature des documents
  - Outils de structuration / indexation dédiés
- Ont une connaissance des besoins de l'utilisateur
  - Adaptation du comportement rappel / précision
- Ont une connaissance du domaine d'application
  - Dictionnaires adaptés

## > Le cœur de métier de Lingway

- **NLP natural language processing**
  - Equipe mixte informaticiens / linguistes
- **Tenir compte de la nature des documents**
  - Objectif : « Rendre le texte calculable »
  - Transformer en une structure XML enrichie de nombreuses méta-données
- **Tenir compte des besoins de l'utilisateur**
  - Les parties importantes dans un texte
  - Les critères de recherche et de navigation adaptés
- **Tenir compte du domaine d'application**
  - Des dictionnaires pour chaque langue et pour chaque métier

# > L'offre LINGWAY

## 1. Spécialisations métiers



LINGWAY  
HR Suite



LINGWAY  
Patent Suite



LINGWAY  
e-commerce Suite



LINGWAY  
Medical Suite



LINGWAY  
Custom Search

## 2. Moteurs sémantiques Lingway



Lingway KM  
(Plateforme linguistique et sémantique)



Dictionnaires métiers:  
Médical, TIC...

## 3. Infrastructure

### Open Source



### Propriétaire



## > Open source dans notre domaine

- Logiciels généraux
- Logiciels documentaires ou moteurs de recherche
- Logiciels NLP
- Dictionnaires et réseaux sémantiques
- Corpus d'apprentissage

## > Logiciels généraux

- Linux Diverses distributions
  - Tomcat (serveur d'applications) Apache
  - Maven, ANT (outils de développement) Apache
  - CXF (génération WS Java) Apache
  - Groovy (langage de scripts) Code Haus
  - My-Sql My-SQL AB
  - Open Office Sun
  - FLEX Adobe
  - Spring (framework de développement) Interface 21
- 
- LINGWAY utilise largement ces outils
  - Interactions régulières avec ces fondations et communautés
  - Réactions variables ( CodeHaus beaucoup plus réactif que Apache)

## > Logiciels moteurs de recherche

- Lucene
  - Très largement répandu, diffusé par Apache
  - API et non produit complet
  - Utilisation à un niveau « rudimentaire » facile
  - Utilisation « évoluée » plus délicate
  - Initiatives complémentaires: SOLR, Nutch
- Il y en a d'autres
  - MG4J (Université de Milan)
  - Swish (C++,PHP)
- Lingway utilise Lucene
  - Avec nombreuses modifications (par surcharges)
  - Prépare des versions basées sur d'autres moteurs

## > Logiciels NLP Open source

- **Nombreuses initiatives**
  - Voir par exemple le site [OpenNLP](#)
- **Outils divers**
  - Analyseurs, taggers, gestion de corpus, etc.
  - Assez parcellaire et hétérogène
  - Pour spécialistes, monde plutôt universitaire
- **Lingway n'utilise pas ces outils**
  - C'est notre cœur de métier
- **Cas particulier**
  - Framework UIMA (issu d'IBM, distribué par Apache)

## > Interêt pour un (petit) éditeur

- On est trop petit pour tout faire
  - Se concentrer sur nos spécificités
- Permet l'indépendance
  - Offre autonome et complète
  - Mais reste compatible avec des solutions propriétaires
- Réduit les coûts et délais de développement

## > Interêt pour un (petit) éditeur

- Petite équipe dans une grande communauté
  - Permet le développement de l'expertise
  - Dialogue et partage avec des développeurs partout dans le monde
- Vitrine pour les experts et pour les sociétés
  - Exemple Doug Cutting chez Yahoo
  - Exemple article « Moving Lucene a step forward » de Cédric Champeau --> afflux de visiteurs sur le site Lingway

## > Interêt pour nos clients

- Réduction des risques
  - Infrastructure largement partagée
- Qualité de l'open source largement confirmée
  - Cf enquête Coverity de mai 2008
- Non intrusif
  - Permet d'étendre facilement des infrastructures existantes sans tout refaire
- Complétude de la solution
  - Open Source + Lingway équivalents aux meilleures solutions propriétaires

## > Conclusion

- Une évolution majeure
  - Permet le développement rapide de très nombreux éditeurs logiciels spécialisés
  - Avec la garantie apportée par les grands éditeurs
- Il faut « jouer le jeu »
  - Accepter de contribuer aux communautés
  - Encourager la participation, afficher ses choix