INSTITUT DES SCIENCES DE L'HOMME

# Analyses sémantiques en SHS et extraction de connaissance :

→ « Capture intelligente de réseaux d'auteurs »

→ « Analyses synchroniques et diachroniques des thématiques dans un corpus textuel»

*8ème journée annuelle ARAMIS, 07 avril 2016*

# 1- Capture intelligente de réseaux d'auteurs

***Sonia Guérin-Hamdi (sonia.guerin-hamdi@ish-lyon.cnrs.fr)***

- Contexte et Objectif
- Méthodologie
- Mise en œuvre
- Pistes

# Contexte et objectif

- **Contexte**: Rôle des communautés de recherche dans la politique environnementale.

- **Objectif** :Synthèse de réseaux d'auteurs hétérogènes à travers des publications liées à la crise écologique des années 1930 provoquée par le « Dust Bowl » aux USA

# Méthodologie

- Moissonnage ciblé de métadonnées et de documents
  - Protocoles : API Rest, OAI, scraping RDF/XML/HTML
  - Sources : theses.fr; sudoc; hal; wos, open data; jstor. scholar …
- Indexation des informations structurées( métadonnées des documents)
- Construction d'un référentiel d'auteurs
- Enrichissement par l'extraction des relations de citations :
  - Citations « explicites » issues de la bibliographie
  - Citations « implicites » issues du corps des documents.
- Construction du réseau hétérogène d'auteurs

# Mise en œuvre

- Indexation de données de sources diverses
  - Collecte avec l'utilisation du composant *DomCrawler* pour l'extraction
  - Indexation Solr avec l'utilisation de la librairie *Solarium* pour interfacer Symfony et SolR
- Indexation de données brutes
  - Utilisation de l'import massif de données de solR
    - **DataImportHandler**

  issues d'une arborescence **FileListEntityProcessor**

  issues de fichiers
    - **XML : XPathEntityProcessor**
    - **TXT / CSV : LineEntityProcessor**

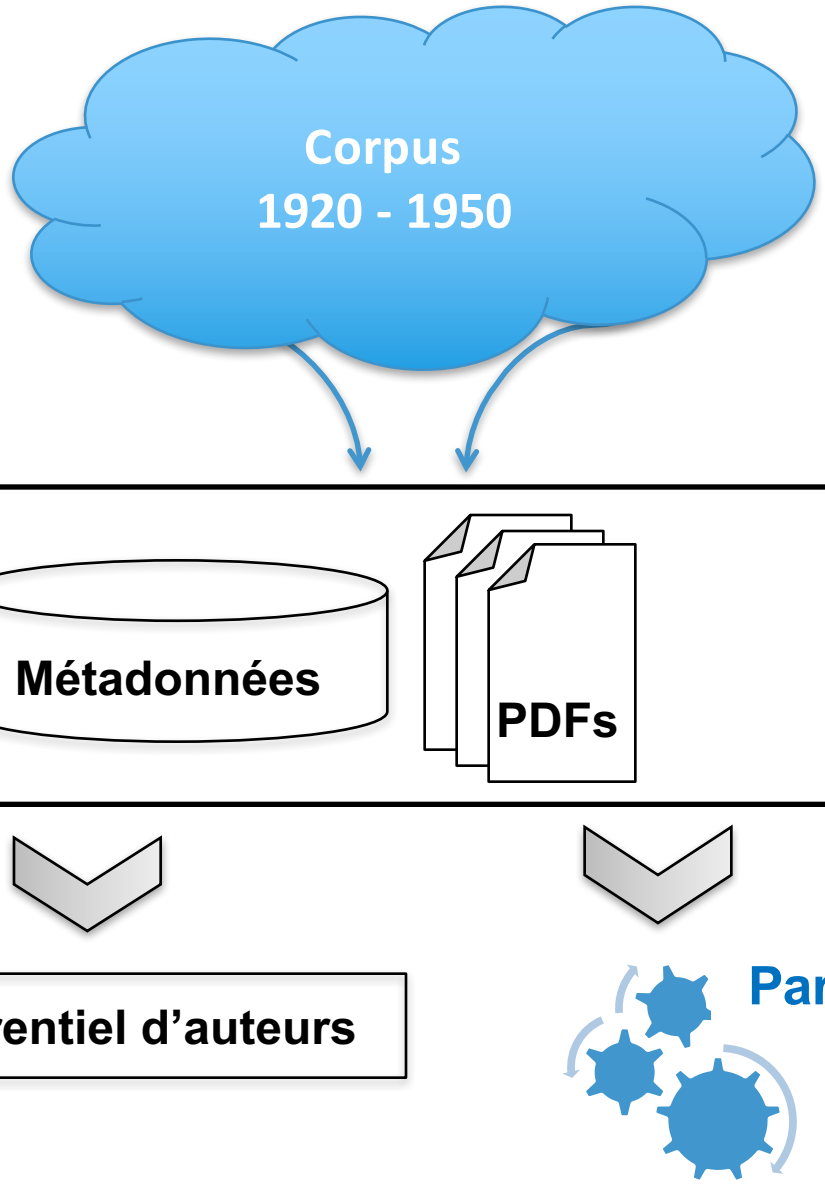
# Mise en œuvre

- Visualisation de données
  - Extraction des informations indexées par SolR et construction du graphe d'auteurs au format JSON;
  - Représentation graphique

  *D3.js* mettant en œuvre les technologies SVG, Javascript, Css.

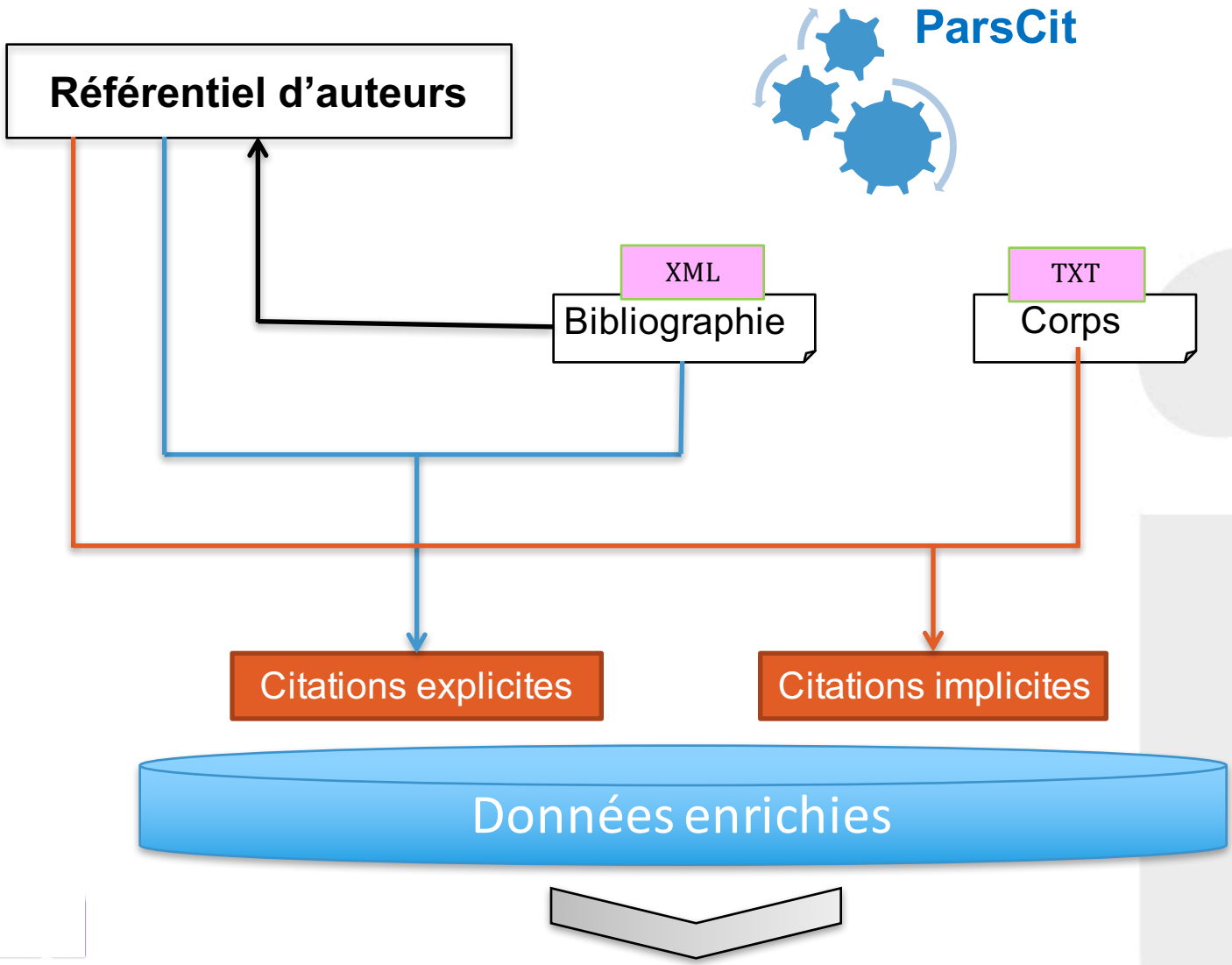
# Mise en œuvre - Schéma

**Crawler Symfony**
*Services Web,*
*Html, RSS,*
*API Rest, OAI,*
*…*



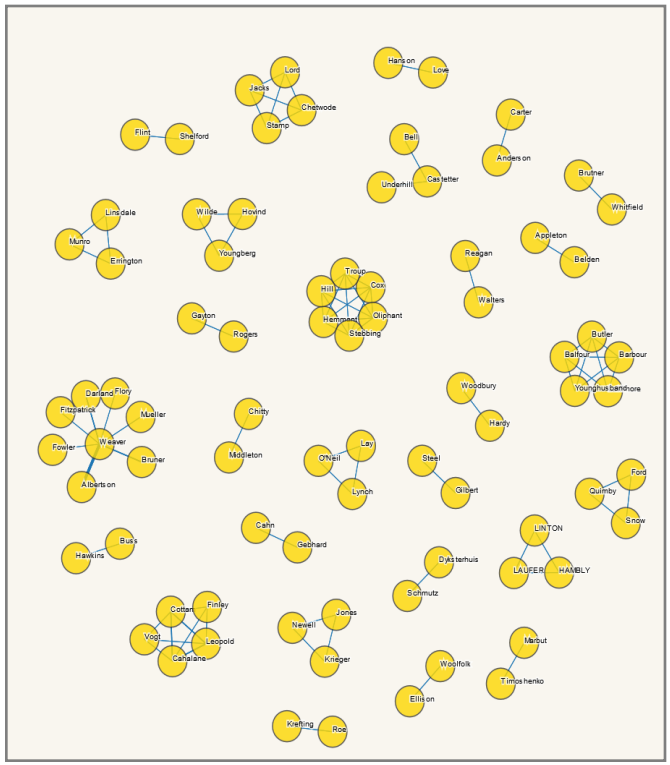

**DIH (DataImportHandler)**
*Données Brutes :*
*xml, txt, csv, json*

# Mise en œuvre - Schéma



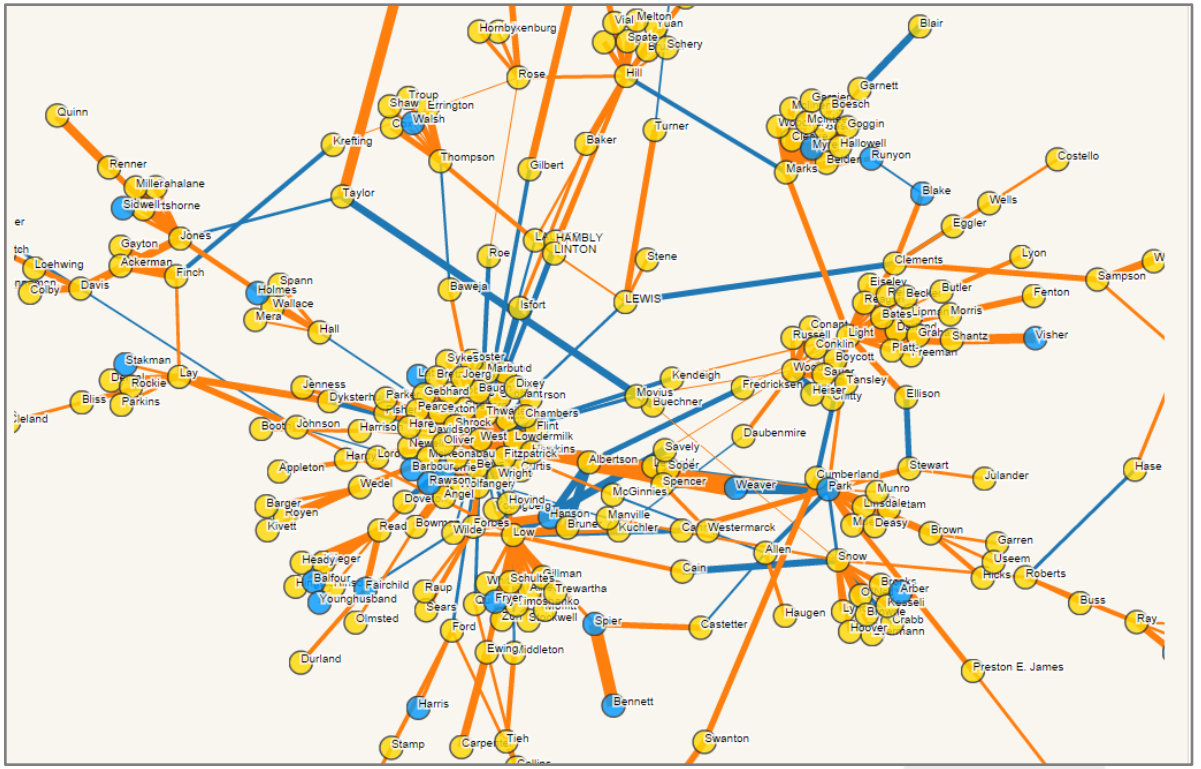
# Réseaux d'auteurs

331 auteurs,
81 relations de co-écriture



**Réseau de collaboration (relation de co-écriture)**

331 auteurs,
5332 relations de citations



**Réseau de citations d'auteurs ( explicites et implicites )**

# Pour aller plus loin

- Détection automatique d'entités désambigüisées non présentent dans la bibliographie et dans les métadonnées initiales dans le corpus textuel via des extracteurs sémantiques.

- Utilisation des facettes pour la
  - Construction de réseaux de thématiques,

# 2- Analyses synchroniques et diachroniques des thématiques


- ***Sofiane Bouzid (sofiane.bouzid@ish-lyon.cnrs.fr)***


- Contexte
- Synchronie : points communs, spécificités et différences
- Extraction et analyse des thématiques (proximités sémantiques des thématiques)
- Diachronie des thématiques
- Conclusion

# Contexte et objectifs

- L'association lance le premier défi EGC

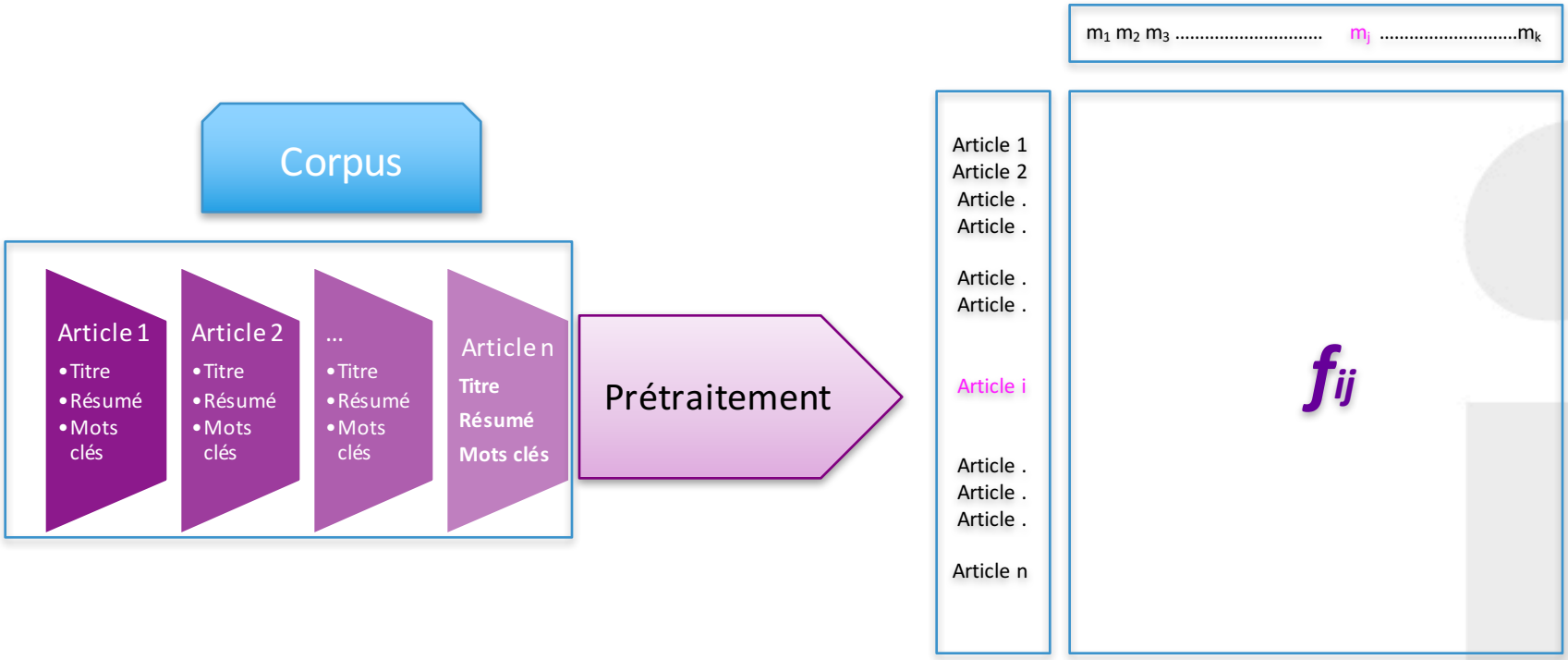
- L'objectif est d'expliquer la structure et l'évolution de l'ensemble de la production scientifique au fil des éditions
  - le contenu textuel décrit dans les articles en français (titre, résumé) entre 2004 et 2015

# Prétraitements du corpus

- Input : csv, txt, htm, pdf xml, etc.
- Package textcat (Hornik et al. (2013)) → détection de la langue (69 langue, SpamAssassin (filtre de spam))
- Package tm
  - pré-traitements (omission des mots-outils, des ponctuations, etc.)
  - Possibilité de faire
    - lémmatisation Package koRpus (TreeTagger)
    - stemming (racinisation) library(SnowballC → Porter Stemming Algorithm)

# Passage du corpus à une représentation matricielle

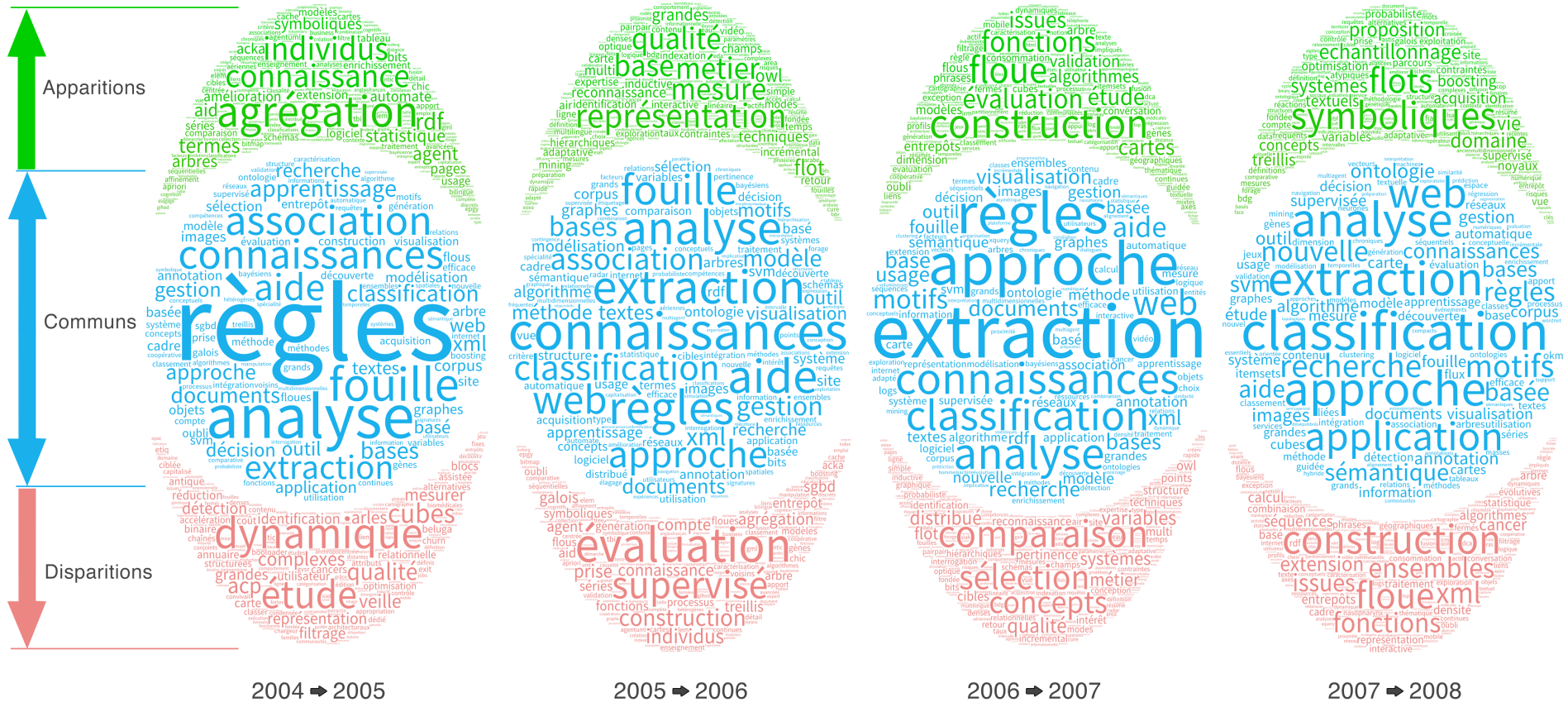
- modèle vectoriel de (Salton et al., 1975)
  - Construction de la Matrice (**DocumentTermMatrix**) [ DOCUMENT * TERMES ] TDM ou DTM → TF, TFIDF




$f_{ij}$ : *fréquence du mot* $m_j$ *dans un* ***article i***

# Visualisation 1

- Corpus par période
- analyse visuelle orientée mots-clés
- évolution de termes
- orientations qui apparaissent, disparaissent et constantes



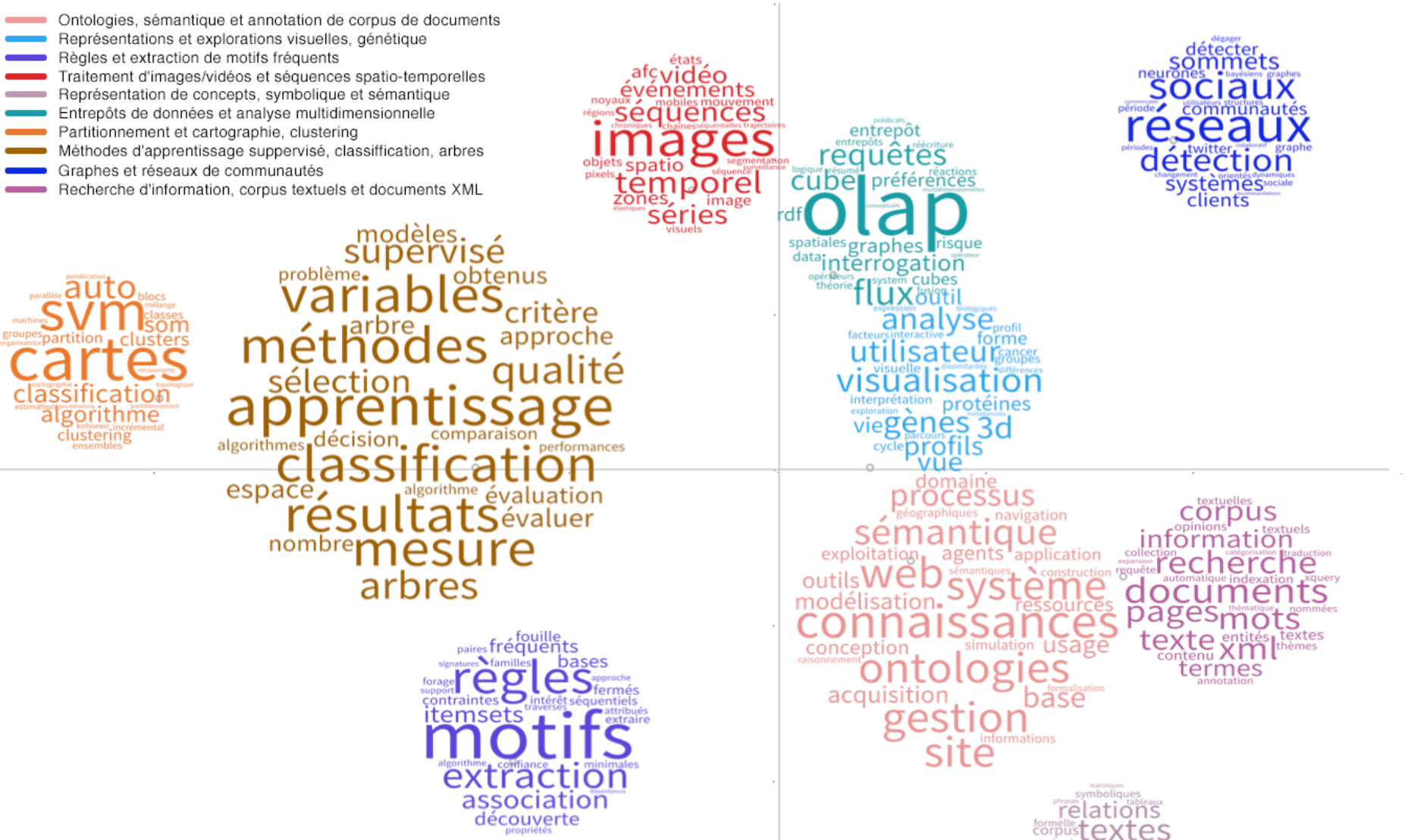

- les représentations visuelles mettent en évidence :
  - les principaux termes communs sous forme de disques bleus, les principaux termes qui sont apparus dans l'édition suivante en vert, enfin les principaux termes ayant disparus depuis l'édition précédente en rouge.
- les termes communs et les nouveaux termes définissent l'année N+1 alors que les termes communs et les termes disparus décrivent l'année N

# Extraction des Thématiques

- Identifier les axes de recherches de la revue
- Utilisation du modèle probabiliste LDA , Latent Dirichlet Allocation (Blei et al. (2003)
  - Package (lda)
  - Utilisation de l'algorithme « Collapsed Gibbs Sampler » (Griffiths and Steyvers, 2004) pour l'estimation des paramètre de probabilité
  - Sievert et Shirley (2014)

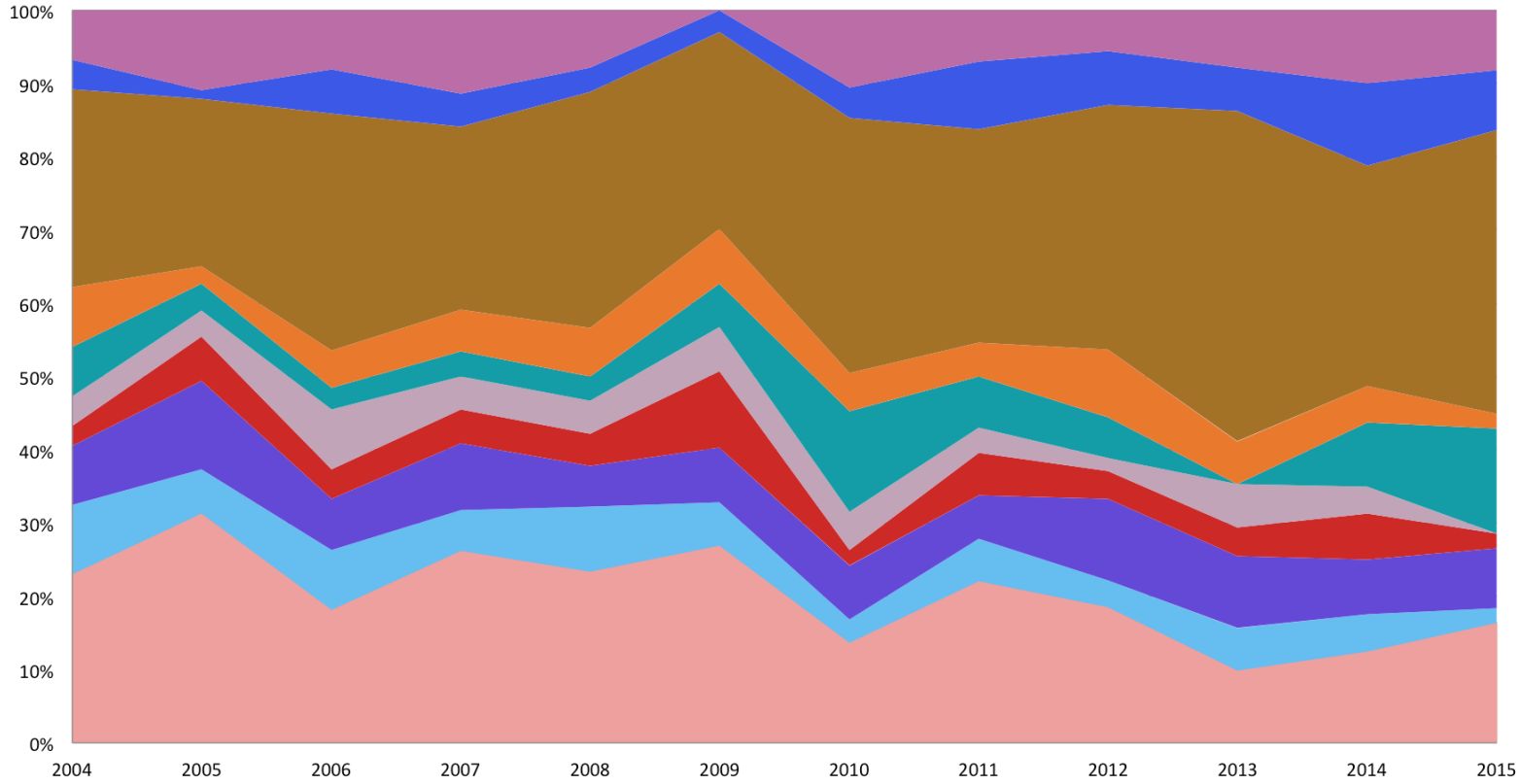
- Thématique 1 : Ontologies, sémantique et annotation de corpus de documents ;
- Thématique 2 : Représentations et explorations visuelles, génétique ;
- Thématique 3 : Règles et extraction de motifs fréquents ;
- Thématique 4 : Traitement d'images/vidéos et séquences spatio-temporelles ;
- Thématique 5 : Représentation de concepts, symbolique et sémantique ;
- Thématique 6 : Entrepôts de données et analyse multidimensionnelle ;
- Thématique 7 : Partitionnement et cartographie, clustering ;
- Thématique 8 : Méthodes d'apprentissage supervisé, classification, arbres ;
- Thématique 9 : Graphes et réseaux de communautés ;
- Thématique 10 : Recherche d'information, corpus textuels et documents XML ;

# Proximités sémantiques

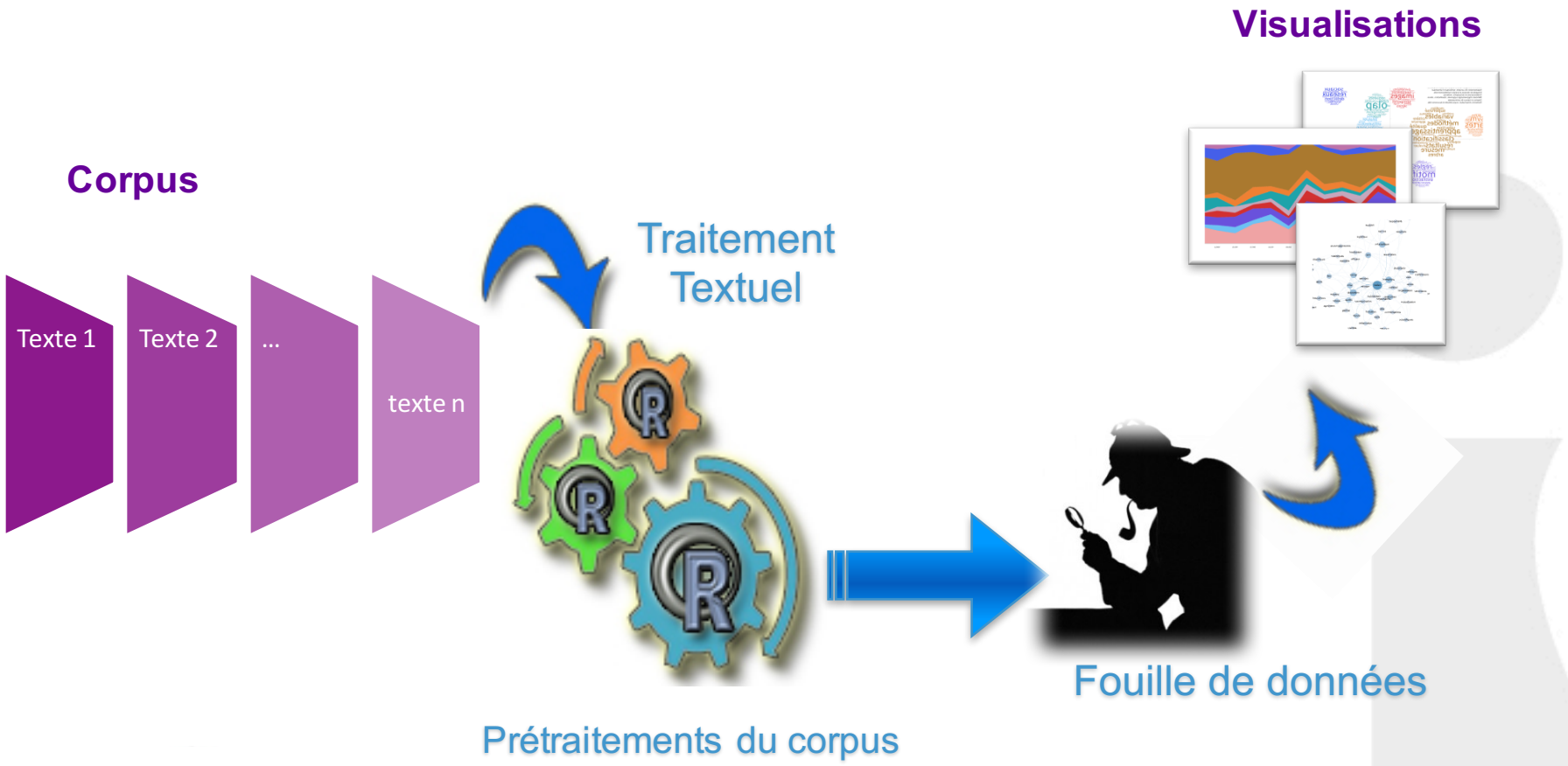



→ approximation de la proximité sémantique entre les thématiques
→ mesure divergence de Kullback-Leible;
→ matrice de dissimilarité entre thématiques.
→ Multidimensional scaling

# Diachronie des thématiques



# Plateforme de visualisation



# Conclusion

- Passage du corpus à une représentation matricielle en utilisant la librairie « TM » de R.

- Analyse et visualisation des proximités sémantiques en utilisant les méthodes (LDA, MDS (mesure de divergence de Kullback-Leibler)) utilisation des librairies « NLP » et « LDA » sous R.

- Analyse de la pertinence et la cohérence des thématiques extraites avec des approches probabilistes basées sur la distribution des termes dans le corpus et utilisation de la pondération proposée par (Sievert et Shirley (2014)) et développé dans la librairie « LDAvis ».

Pistes :

Changements de thématiques est liés à de nouveaux auteurs
Construction d'un réseau basé sur les thématiques extraites

# Merci


RÉPONSES
SYNTHÈSE
QUESTIONS
PRÉCISIONS
DISCUSSION
CLARIFICATIONS
MERCI
DÉTAILS
ÉVALUATION
PRÉSENTATION
PROBLÉMATIQUE
CONCEPT
TECHNIQUE
PROBLÈME
PERSPECTIVES
EXPOSÉ
BILAN
SUGGESTIONS
NOTIONS
CONCLUSION
COMMENTAIRES
RÉSUMÉ
DÉBAT
EXPLICATIONS
SESSION
RECUL
MOTS-CLÉS
QUESTIONNEMENTS
ARGUMENT
ÉCLAIRCISSEMENTS
RÉSULTATS
REMARQUES
INTERROGATIONS
CRITIQUES