

# **CAMPUS CONDORCET** Paris–Aubervilliers

Cité des humanités et des sciences sociales

## **Expérimentation de text-mining (TDM) au sein du LaDéHiS-CRH (CNRS/EHESS) : faciliter la recherche d'information**

Francine FILOCHE, chargée de mission services aux publics  
Laura PAGÈS, chargée de mission ressources et innovation numériques



# 1. Origines de l'expérimentation

## La rencontre des besoins du chercheur et du bibliothécaire

L'expérimentation de fouille de textes qui est menée depuis janvier 2016 par le Campus Condorcet est née de la rencontre entre deux besoins:



- Côté LaDéHiS: améliorer et systématiser les processus d'extraction de l'information dans le cadre de ses travaux de recherche
- Côté Grand équipement documentaire: expérimenter sur le terrain un service de préfiguration innovant fondé sur la manipulation de documents numériques



# 2. Déroutement de l'expérimentation

## Étape 1 - Hiérarchisation des textes et arborescence

- Hiérarchisation et classement par similarité de segments de texte
- Proposition d'une arborescence construite par grappes de mots clés générés automatiquement et modifiables

Application à un volume de textes important en un temps réduit



The screenshot shows the OntoGen interface. On the left, a tree structure lists concepts such as 'political', 'urban', 'medical', and 'women'. The main window displays a list of documents with columns for document ID, title, and similarity score. A search bar at the bottom shows the query 'tel-00769798 2000 Between'.

The screenshot shows a hierarchical tree of concepts. The root node is 'SUPPLIQUES'. It branches into 'petitioning', 'politics', 'letter', and 'letter'. Each node contains a list of associated terms and phrases, such as 'rhetoric, people, king', 'groups, power, relations', 'social, practice', 'poor', 'justice, royal', 'modern, early\_modern, early', 'authority, social', 'history, culture', 'written, english, process', 'poor, pauper', 'law', and 'strategies'.

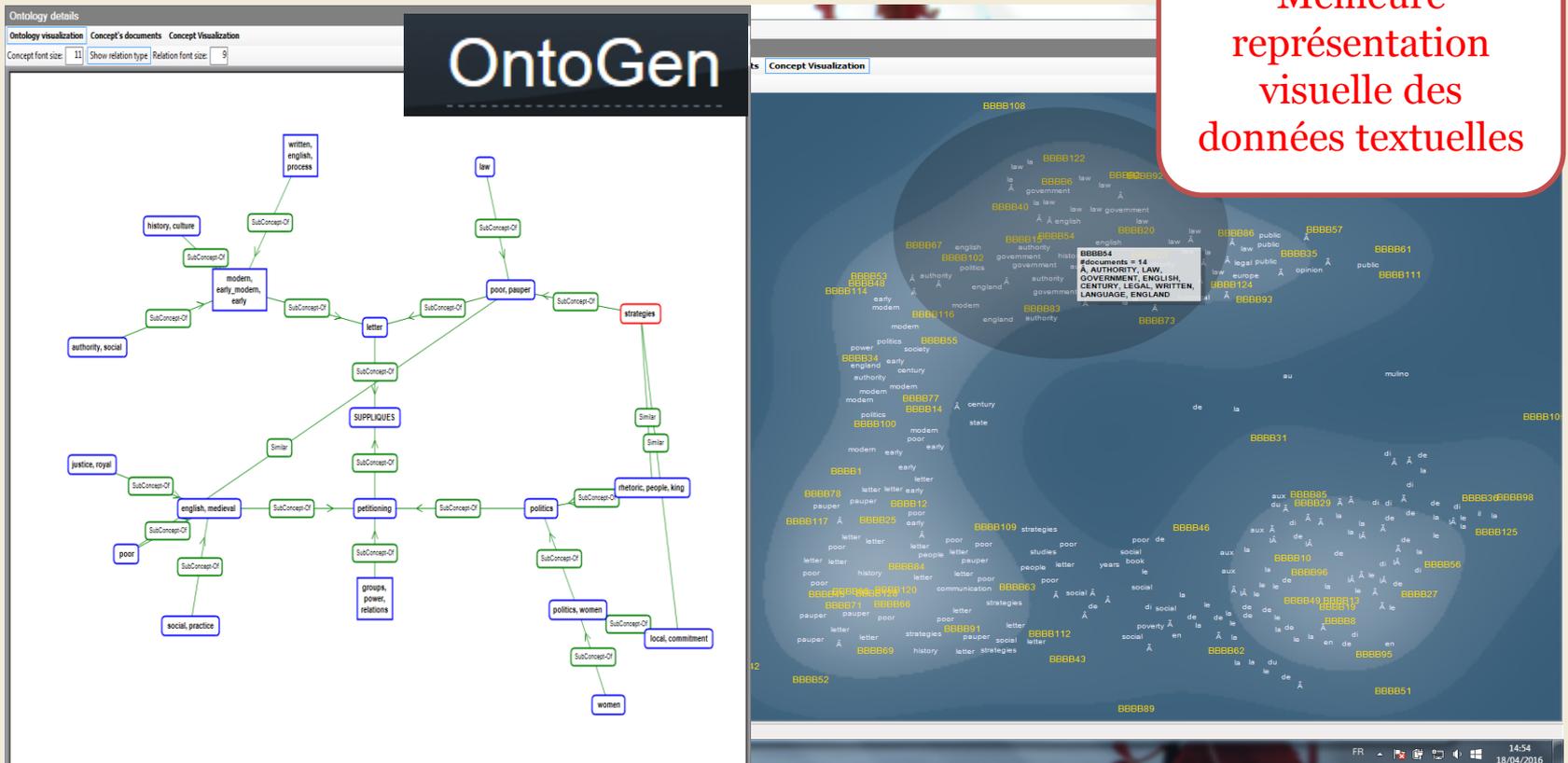


## 2. Déroulement de l'expérimentation

### Étape 2 - Classement et approche ontologique

- Transformation des grappes de mots clés en concepts
- Facilitation de la recherche de similarité de groupes de textes
- Renforcement de l'analyse par la mise en relation des concepts

Meilleure  
représentation  
visuelle des  
données textuelles



## 2. Déroulement de l'expérimentation

### Étape 3 - Orientation ontologique

- Traitement à partir d'une base de connaissance de classes sémantiques de l'ontologie créée avec le 1<sup>er</sup> outil
- Confirmation des nouveaux résultats et création de scénarios de mise en relation de concepts

3111 list of classes

- 0136 aristocracy
  - 0029 important person
- 0046 authority
- 0153 behavior authorities claimers
  - 0038 goodness
  - 0009 language as weapon
- 0013 citizenship
- 0031 colleges and universities
- 0054 communication
- 0058 difference
- 0099 discours en jeu
- 0295 ecrits, exchanges, circulation
  - 0011 circulation
  - 0009 exchanges
  - 0093 letters
  - 0182 petitions
- 0011 family
- 0101 fight, conflicts
- 0143 historians
- 0251 justice, law
  - 0176 law
  - 0023 law (moins science)
- 0024 life
- 0065 outside europe
  - 0020 asia
  - 0008 middle east
  - 0037 u s a
- 0151 politics
- 0027 power
- 0130 regime
  - 0111 state instituc
  - 0063 governr
  - 0019 instituc
  - 0029 officials
- 0113 religion
- 0374 social group
  - 0050 groups
  - 0026 peasant
  - 0091 people
  - 0118 poverty, poor
  - 0089 woman
- 0056 society
- 0289 through centuries
- 0491 west europe
  - 0052 europe
  - 0286 united kind

Apparition  
de nouveaux  
matériaux  
de recherche

Identification de  
thématiques  
émergentes



Tropes

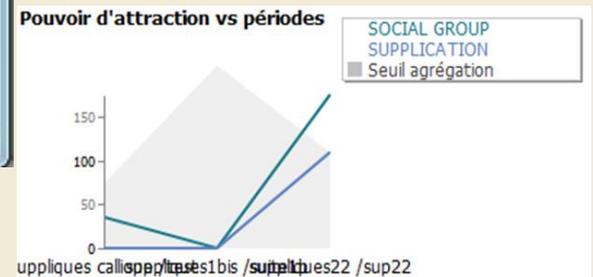
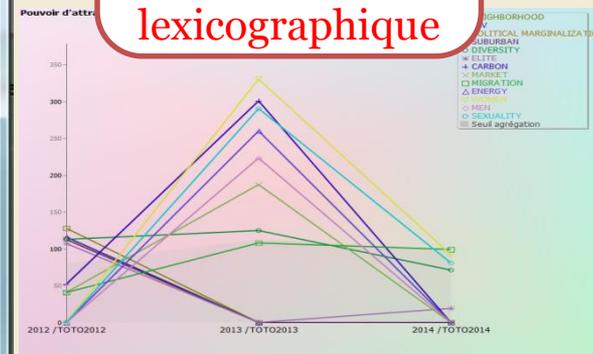
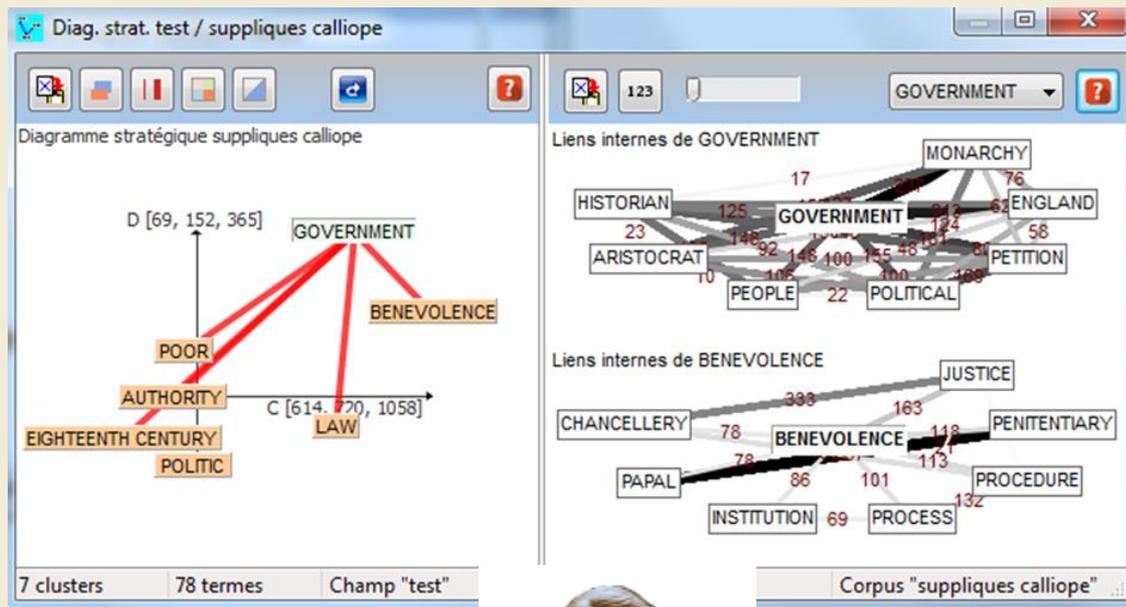
<http://www.tropes.fr>

## 2. Déroulement de l'expérimentation:

### Étape 4 - Indexation automatique et extraction terminologique

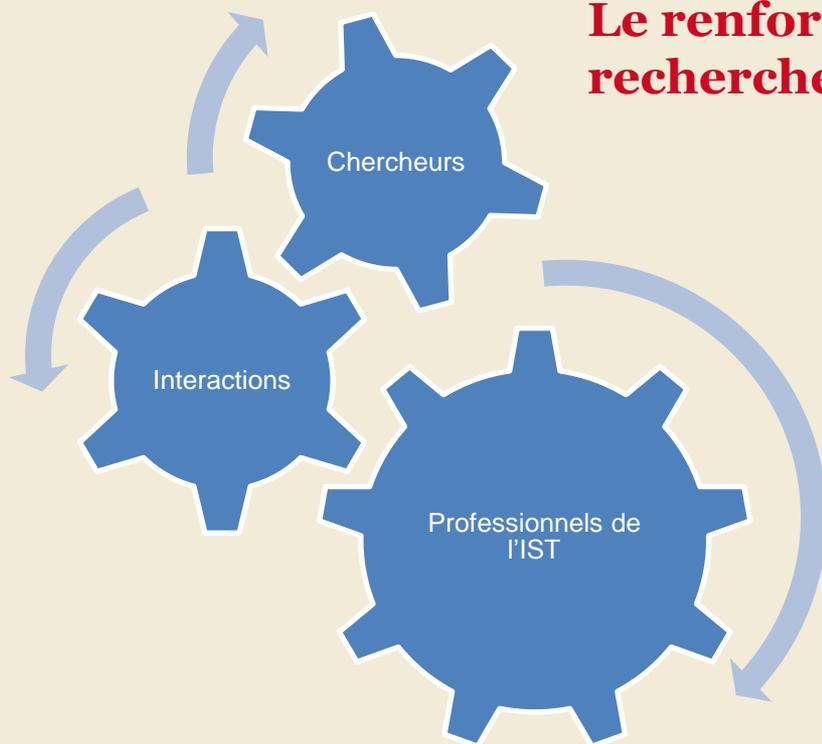
- Analyse avec le chercheur de l'environnement sémantique
- Création de lexiques de report de références générés automatiquement selon les concepts validés par le 2ème outil
- Fonction de comparaison chronologique de corpus

Enrichissement  
des thésaurus  
grâce à  
l'extraction  
lexicographique



Calliope™

### 3. Bilan de l'expérimentation



#### Le renforcement des synergies entre recherche et documentation

- Les retours positifs du chercheur du LaDéHiS
- La nécessité d'affiner notre périmètre d'intervention
- Une expérience riche pour mieux penser les futurs services du Grand équipement documentaire

*« Le TDM m'a permis de questionner mon corpus et plus largement mon propre travail »*

#### L'expérimentation et après ?

- Un travail de veille à poursuivre dans un contexte florissant de nouvelles initiatives, de nouvelles pratiques et d'évolutions juridiques et techniques prometteuses



## 4. Pour aller plus loin...

- **Articles 30 et 38**, [Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.](#)
- **CNRS**, Livre blanc, [Une science ouverte dans une République numérique: études et propositions en vue de l'application de la loi](#), 2016.
- **Commission européenne/DG Research and Innovation**, [TDM Report from the Expert Group. Study on the legal framework of text and data mining \(TDM\)](#), Jean-Paul Triaille, Jérôme de Meeüs d'Argenteuil et Amélie de Francquen, 2014.
- **EPRIST**, [Le projet de nouvelle directive « droit d'auteur » présenté par la Commission européenne prévoit une exception obligatoire au droit d'auteur en faveur du TDM scientifique](#), in Analyse I/IST-n°21-septembre 2016.
- **Hakim Hachour**, *De la fouille à la visualisation de données : un processus interprétatif*, I2D – Information, données & documents 2015/2 (Volume 52), p. 42-43.
- **Johann Gillium**, Mémoire d'étude Enssib, [Big data et bibliothèques](#), 2015.
- **Ligue des Bibliothèques Européennes de Recherche (LIBER)**, [Text and Data Mining : the need for change in Europe](#), 2015.
- **Lionel Maurel**, [L'exception TDM dans la loi numérique: mérites, limites et perspectives](#), 2016.
- **London School of Economics Library**, [Liberating Data: How libraries and librarians can help researchers with TDM](#), 2016.

