Extraction automatique et harmonisation de données web

Retour sur une étude bibliométrique à partir de sommaires en ligne de revues de sciences sociales

Julien Boelaert

Université de Lille, CERAPS
Rencontres de statistiques appliquées, INED

12/02/2021

"Les aléas de l'interdisciplinarité, *Genèses* et l'espace des sciences sociales françaises (1990-2014)", J. Boelaert, N. Mariot, É. Ollion, J. Pagis, *Genèses* 100-101, 2015, p. 20-49.

Introduction

Genèses et l'interdisciplinarité :

- une revue pensée dès l'origine comme interdisciplinaire (sociologie, histoire, anthropologie, sciences politiques)
- dans un champ de recherche qui promeut depuis longtemps l'interdisciplinarité
- "discours omniprésent et incantatoire sur l'interdisciplinarité" (Heilbron et Gingras 2015), mais peu de mesure.

Numéro anniversaire (25 ans), occasion de revenir sur l'interdisciplinarité, dresser une cartographie des sciences sociales françaises entre 1990 et 2014.
Genèses comme site d'observation stratégique pour saisir les aléas de l'idée interdisciplinaire en France sur vingt-cinq ans.

Introduction

Méthode :

- bibliométrie comme outil de connaissance
- entrée par les articles.

Première source : la revue elle-même

- les auteurs : qui écrit dans *Genèses* ?
  - source : tables des matières, section "auteurs", codage manuel (685 auteurs)
  - caractérisation : sexe, statut, rattachement géographique, discipline
- les citations : qui est cité dans *Genèses*, et par qui ?
  - source : notes de bas de page, bibliographie, extractions par expressions régulières.
  - évolution des auteurs les plus cités, et caractérisation des citations selon quelles disciplines les citent.

Introduction

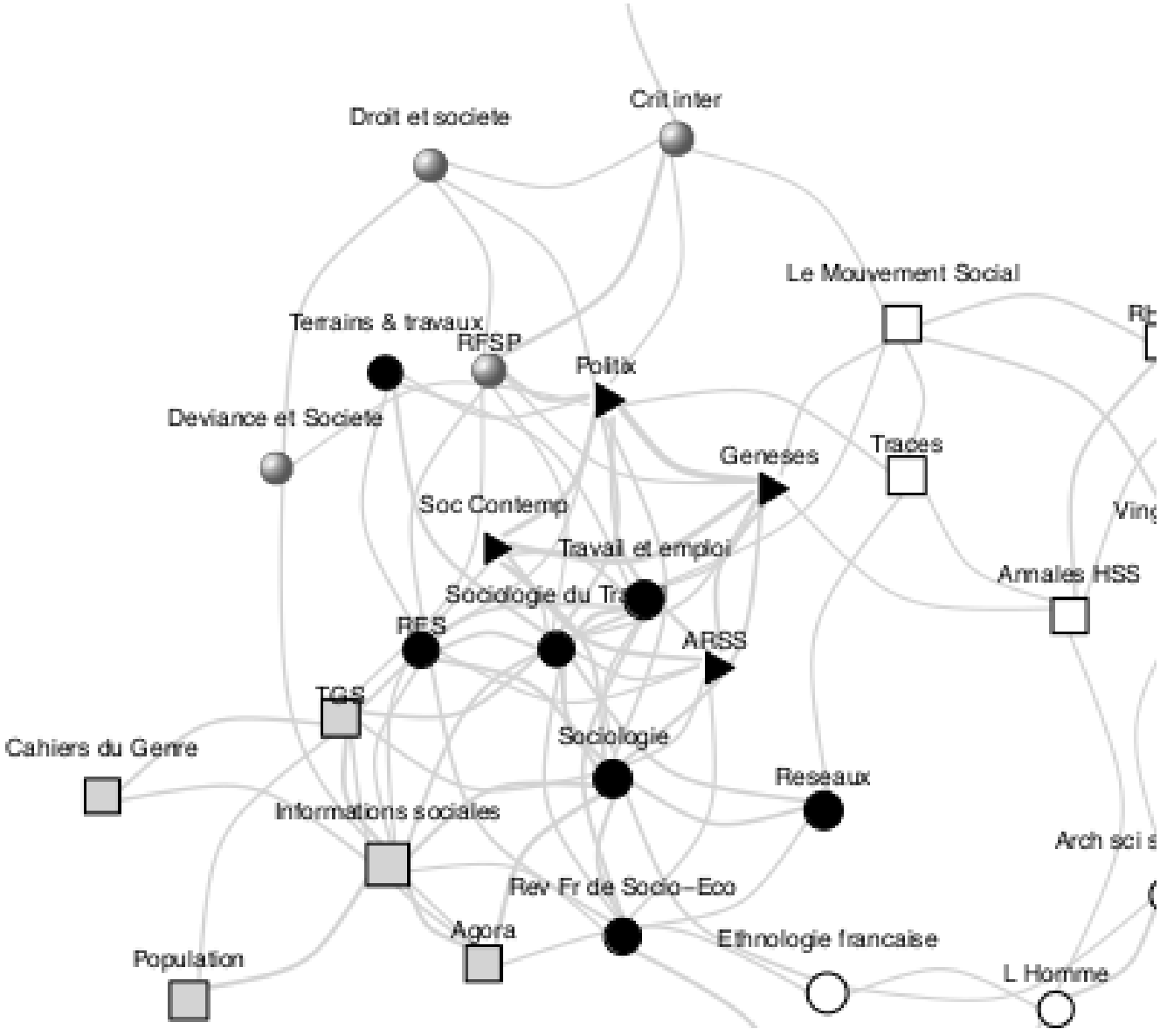
Au-delà, nécessité de saisir la place de *Genèses* parmi ses pairs

- idée de récolter les tables de matières en ligne des principales revues françaises de SHS.

Corpus :

- 38 revues françaises de SHS (sociologie, histoire, anthropologie, sciences politiques, démographie)
- source : sommaires de revues en ligne (cairn, persee, jstor), exhaustivité (parmi les revues choisies)
- idée : deux revues sont proches si elles partagent des auteurs, traitement par méthodes d'analyse de réseaux.
- base à construire : une ligne par auteur-article, trois colonnes (journal, auteur, année)

Introduction



Construction de la base

Etapes de construction de la base de données :

- webscraping :
  - téléchargement des pages web
  - extraction des informations : xpath
- nettoyage :
  - sélection des "vrais" articles
  - harmonisation de base : extraction des années, harmonisation du nom des revues
  - harmonisation avancée : nom des auteurs.

Base finale : plus de 15 000 auteurs, plus de 29 000 couples auteur-revue.

1/ Webscraping

- Source : sommaires de revues (persee, cairn, jstor)
- Téléchargement "respectueux", semi-manuel
- Environ 7 500 pages

1/ Webscraping

Extraction d'informations : xpath

- Xpath : langage de requêtes pour données XML
- Exploite la structure en arbre des langages à balise
- Chemins à trouver :
  - unité d'analyse : articles
  - informations par article : revue, titre, auteurs, année, nombre de pages

1/ Webscraping

Exemple sur CAIRN :

- article : "//div[contains(@class, 'article-list-item')]"
- auteur : ".//span[@class='auteur']"

1/ Webscraping

Traitement toujours le même :

- package R XML, fonction xpathSApply
- boucles imbriquées : par page, par article
- attention aux valeurs manquantes / valeurs multiples

Par après, fonction mise en ligne : package R scraEP, fonction xscrape.

```
library(scraEP)
extract <- xscrape(pages.cairn,
                   row.xpath= "//div[contains(@class, 'article-list-item')]",
                   col.xpath= c(auteur= ".//span[@class='auteur']",
                                titre= "//li[@class='titre-article']"))
```

2/ Nettoyage

Trois bases fusionnées, à nettoyer :

- harmonisation initiale : extraction d'informations par expressions régulières : année de publication, pages, harmonisation du nom des revues
- sélection des "vrais" articles : suppression des recensions d'ouvrage, des notices nécrologiques... (à partir des titres, sections, nombre de pages)
- harmonisation avancée : nom des auteurs
  - "loic wacquant", "wacquant loic", "loic j. d. wacquant"...

2/ Nettoyage

L'harmonisation du nom des auteurs :

- trop complexe à automatiser
- traitement semi-automatique avec openrefine
- résultat : plus de 15 500 auteurs au départ, finalement 15 182.

Outil "text facet" pour le clustering :

- fingerprint key-collision
- n-gram fingerprint key-collision
- k plus proches voisins sur distance de Levenshtein
- k plus proches voisins sur distance PPM

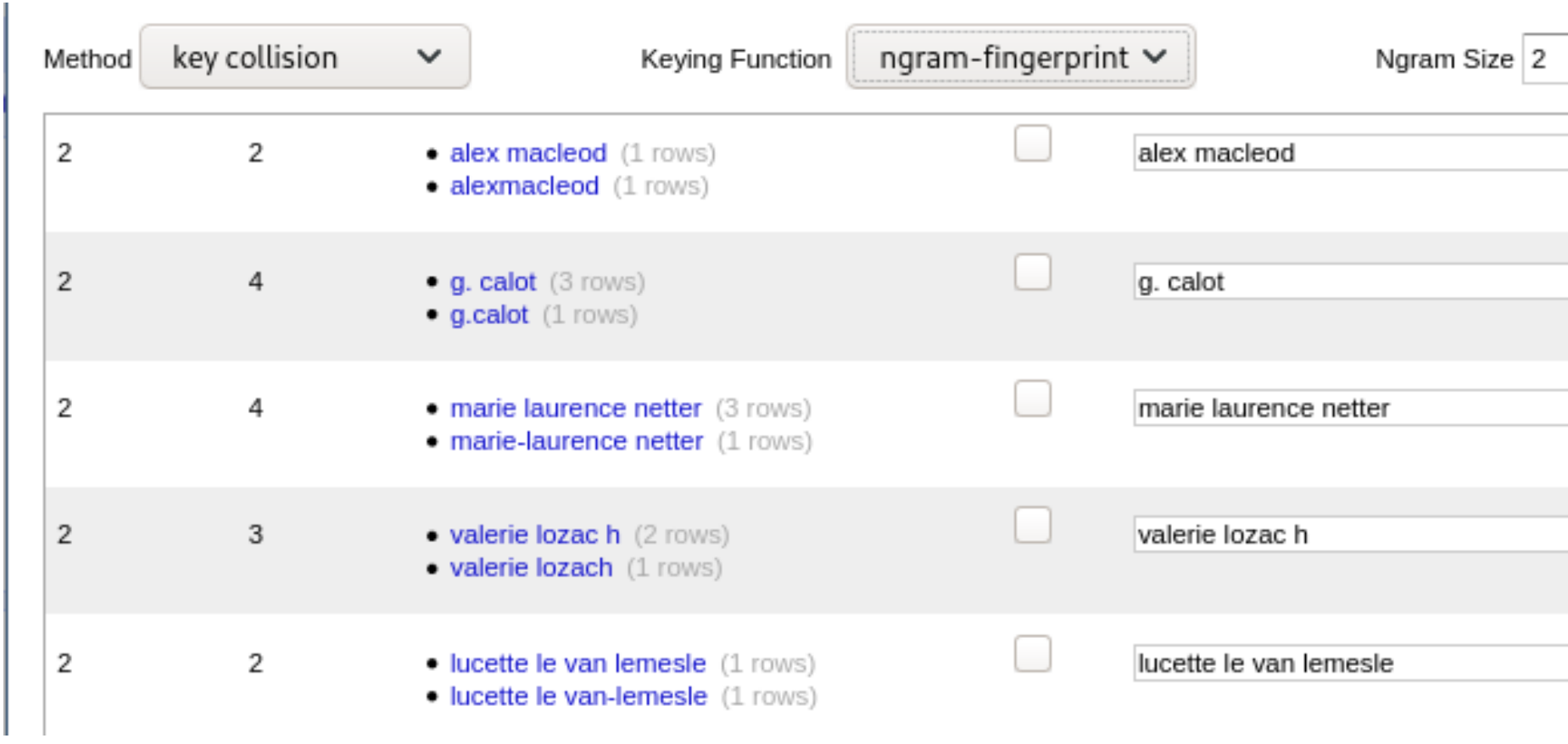
Après tous ces traitements, restaient quelques corrections à faire : transformation des initiales en prénoms complets.

2/ Nettoyage

Method: key collision ⌄ Keying Function: fingerprint ⌄

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	5	• benoit ceroux (4 rows) • ceroux benoit (1 rows)	☐	benoit ceroux
2	3	• yves-antoine flori (2 rows) • yves-antoine flori (+) (1 rows)	☐	yves-antoine flori
2	10	• marie-laure raynaud (8 rows) • raynaud marie-laure (2 rows)	☐	marie-laure raynaud
2	10	• willy gianinazzi (9 rows) • >willy gianinazzi (1 rows)	☐	willy gianinazzi
2	3	• danielle rozenberg (2 rows) • danielle rozenberg. (1 rows)	☐	danielle rozenberg

2/ Nettoyage



Method key collision ⌄ Keying Function ngram-fingerprint ⌄ Ngram Size 2

2	2	• alex macleod (1 rows) • alexmacleod (1 rows)	☐	alex macleod
2	4	• g. calot (3 rows) • g.calot (1 rows)	☐	g. calot
2	4	• marie laurence netter (3 rows) • marie-laurence netter (1 rows)	☐	marie laurence netter
2	3	• valerie lozac h (2 rows) • valerie lozach (1 rows)	☐	valerie lozac h
2	2	• lucette le van lemesle (1 rows) • lucette le van-lemesle (1 rows)	☐	lucette le van lemesle

2/ Nettoyage

Method: key collision | Keying Function: ngram-fingerprint | Ngram Size: 1

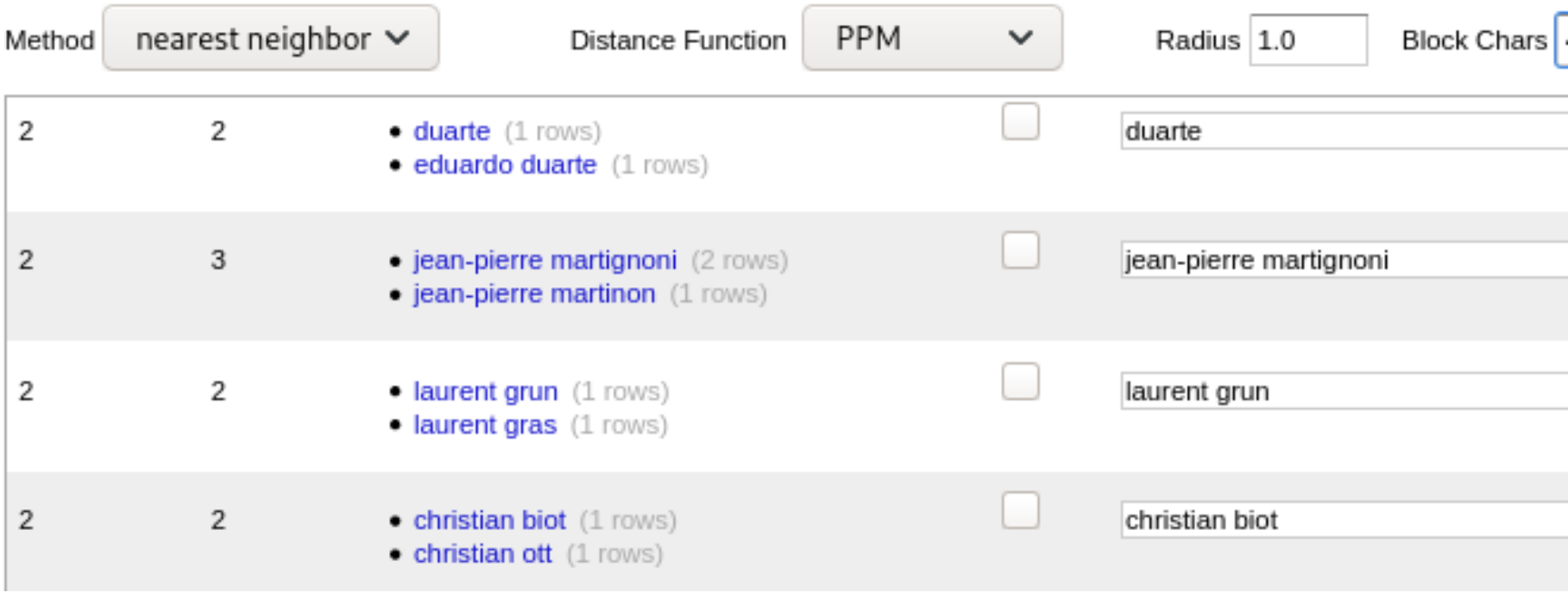
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
10	14	• francoise orlic (3 rows) • florence lianos (2 rows) • francois sellier (2 rows) • claire lefrancois (1 rows) • francis calcoen (1 rows) • francoise collin (1 rows) • francoise enel (1 rows) • francoise lorcerie (1 rows) • francoise sellier (1 rows) • nicolas ferran (1 rows)	☐	francoise orlic
9	15	• etienne francois (4 rows) • francesco fistetti (3 rows) • francis conte (2 rows) • france tissot (1 rows) • francisco santana ferra (1 rows) • franco cesetti (1 rows) • francois foret (1 rows) • francois terre (1 rows) • francoise tristani (1 rows)	☐	etienne francois

2/ Nettoyage

Method nearest neighbor ⌄ Distance Function levenshtein ⌄ Radius 1.0 Block Chars 3

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	3	• michel picard (1 rows) • michel sicard (1 rows) • michele picard (1 rows)	☐	michel picard
2	21	• philippe robert (20 rows) • phlippe robert (1 rows)	☐	philippe robert
2	2	• anne-marie rieu (1 rows) • anne-marie dieu (1 rows)	☐	anne-marie rieu
2	2	• emmanuel dumont (1 rows) • emmanuel dupont (1 rows)	☐	emmanuel dumont
2	4	• wolfgang gaiser (3 rows) • wolfgang kaiser (1 rows)	☐	wolfgang gaiser

2/ Nettoyage



- openrefine permet de garder la trace des opérations effectuées, au format JSON

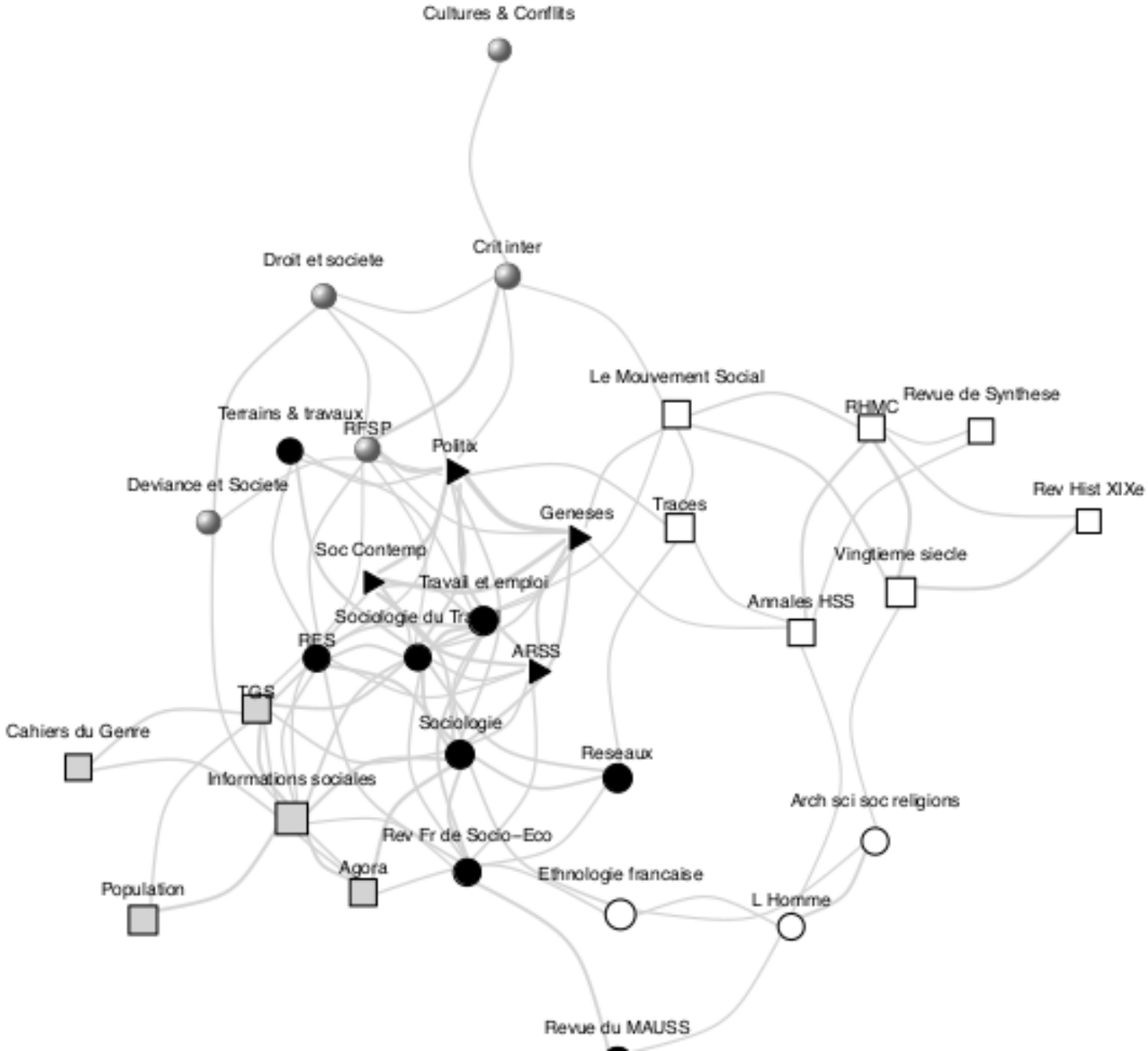
3/ Résultats

Analyse de réseau (one-mode) :

- lien entre deux revues : partage d'auteurs
- réseau par période : 1990-1994, 2000-2004, 2010-2014
- détéction de classes (Louvain), mesures de centralité, mesures de cohésion (densité du graphe, moyenne des plus courts chemins)

Résultats :

- *Genèses* au centre des revues françaises, à l'interface des disciplines
- re-disciplinarisation des revues de SHS entre les années 1990 et 2010.


Cultures & Conflits
Droit et societe
Crit inter
Le Mouvement Social
Revue de Synthese
RHMC
Terrains & travaux
RFSP
Politix
Deviance et Societe
Rev Hist XIXe
Traces
Geneses
Soc Contemp
Vingtieme siecle
Travail et emploi
Annales HSS
Sociologie du Tra
RES
ARSS
TGS
Cahiers du Genre
Sociologie
Reseaux
Informations sociales
Arch sci soc religions
Rev Fr de Socio-Eco
Agora
Ethnologie francaise
Population
L Homme
Revue du MAUSS