Freins et potentiels d'analyse de corpus textuels issus d'un dispositif de veille en agriculture biologique

Guillaume Ollivier

▶ **To cite this version:**

Guillaume Ollivier. Freins et potentiels d'analyse de corpus textuels issus d'un dispositif de veille en agriculture biologique. RMT devAB, Oct 2011, Paris, France. 17 p. hal-02804087

HAL Id: hal-02804087
https://hal.inrae.fr/hal-02804087v1

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Freins et potentiels d'analyse de corpus textuels issus d'un dispositif de veille en Agriculture Biologique

Ollivier, Guillaume

Sociologue, INRA Ecodéveloppement

RMT DevAB, Paris, 20 octobre 2011

Plan

- Objectif : illustrer l'intérêt de l'analyse de corpus documentaires (issus de veille ou non) tout en illustrant les problèmes génériques que cela pose

1- Eléments généraux sur la veille

2- Exemples d'application :

  – Analyse de production académique
  – Analyse de presse

Eléments généraux sur la veille

- Une infinité de configurations de dispositifs de veille possible

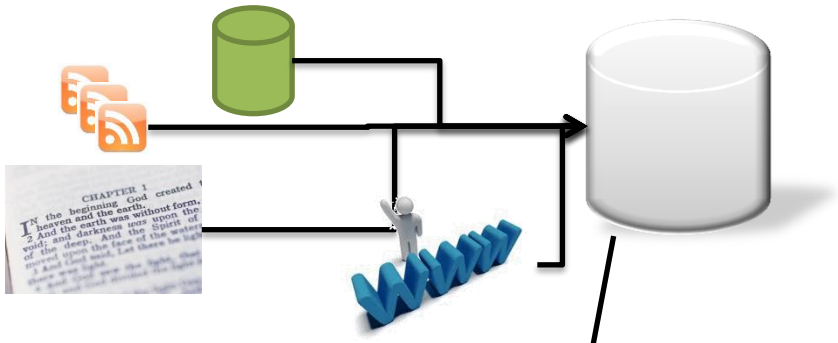
- Forte dépendance des analyses à la nature des données collectées :
  - Couvertures géographiques, temporelles, linguistiques
  - Formats
  - Variables prises en compte…
- Nombreux outils disponibles (gratuits/payants, dédiés à la veille ou non) **mais pas de solutions universelles**

=> nécessité d'identifier les grandes questions prioritaires à traiter en laissant ouvert à des questionnements ultérieurs moins prioritaires

Eléments généraux sur la veille

La veille peut permettre la constitution de corpus textuels de volume conséquent et capitalisés

collecte → (capitalisation) → analyses → diffusion



Numérisation
Variables pertinentes ?
Standardisation ?
Dépôt de données :
Dimensionnement,
Maintenance… ?

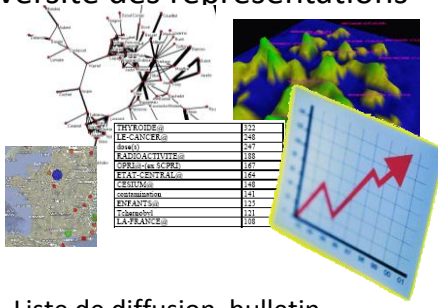
Les connaissances produites :
Qui ? : quels sont les acteurs importants
Quoi ? : quels sont les thèmes chauds
Quand ? : la dynamique d'une controverse
Où ?

Analyse en flux continu :
- synthèse de l'actualité / revue de presse
- Analyse continue de tendances (tableau de bord)
- Catégorisation automatique des contenus

Analyses rétrospectives, ponctuelles et ciblées
- Analyse de controverse (sociologie)
- Préparation d'appels à projets

Outils : la jungle
Dédiés+payants : Matheo-Analyser, LexiQuest, MindSet InfoExtract…
Généraliste+payant : Alceste, Sphinx…
Généraliste+gratuit : Prospero,

Diversité des représentations



- Liste de diffusion, bulletin
- Site web
- …

- Site web
- Rapports
- Usage interne

Degré d'automatisation du processus
- nécessité d'un interprète humain connaisseur du domaine
- nécessite le développement d'applications dédiées

Exemple d’analyses de corpus textuels

(issus de veille ou d’extraction ponctuelle)

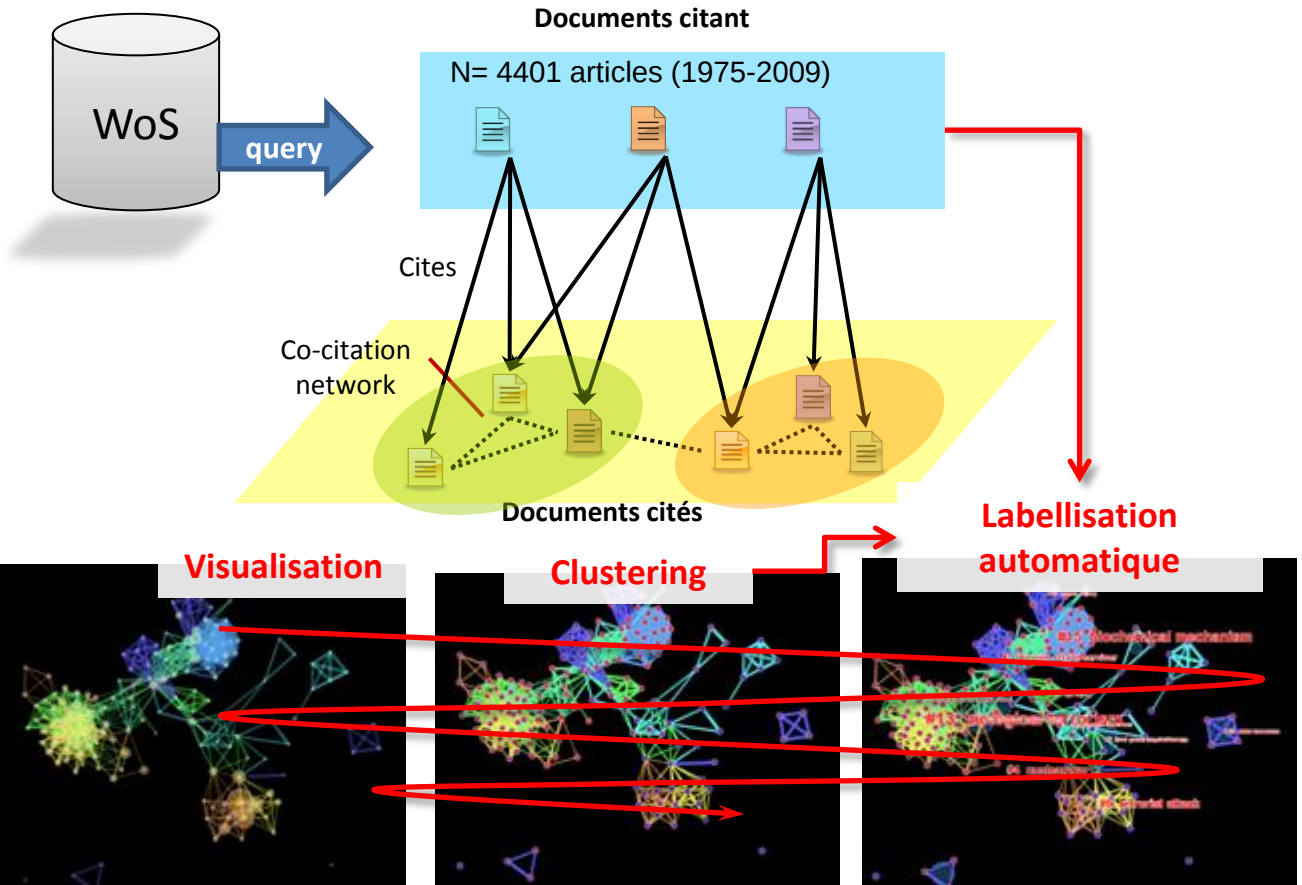
Production scientifique et AB

Presse et AB

Production scientifique et AB

- **Scientométrie :**
  - Web of Science : Base de donnée internationale généraliste de référence
  - Notice reflet de l'activité scientifique et de la production de connaissance
  - Construction d'une requête complexe pour couvrir l'ensemble du domaine scientifique de l'agriculture et l'alimentation biologique

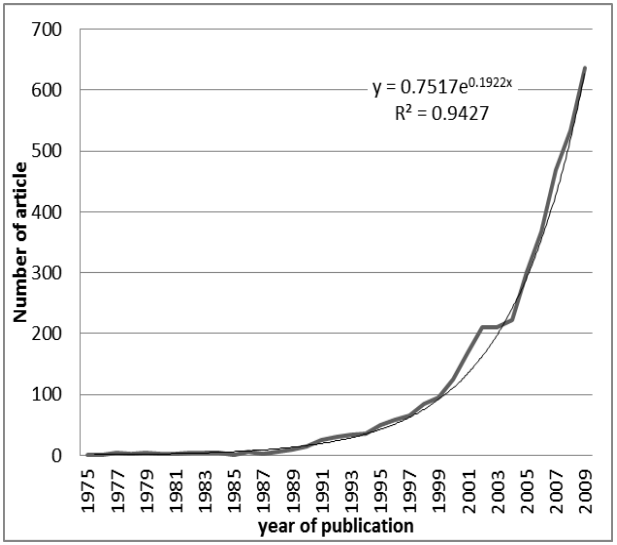
- **Analyse dynamique de co-citation CiteSpace (Chen et al. 2010)**



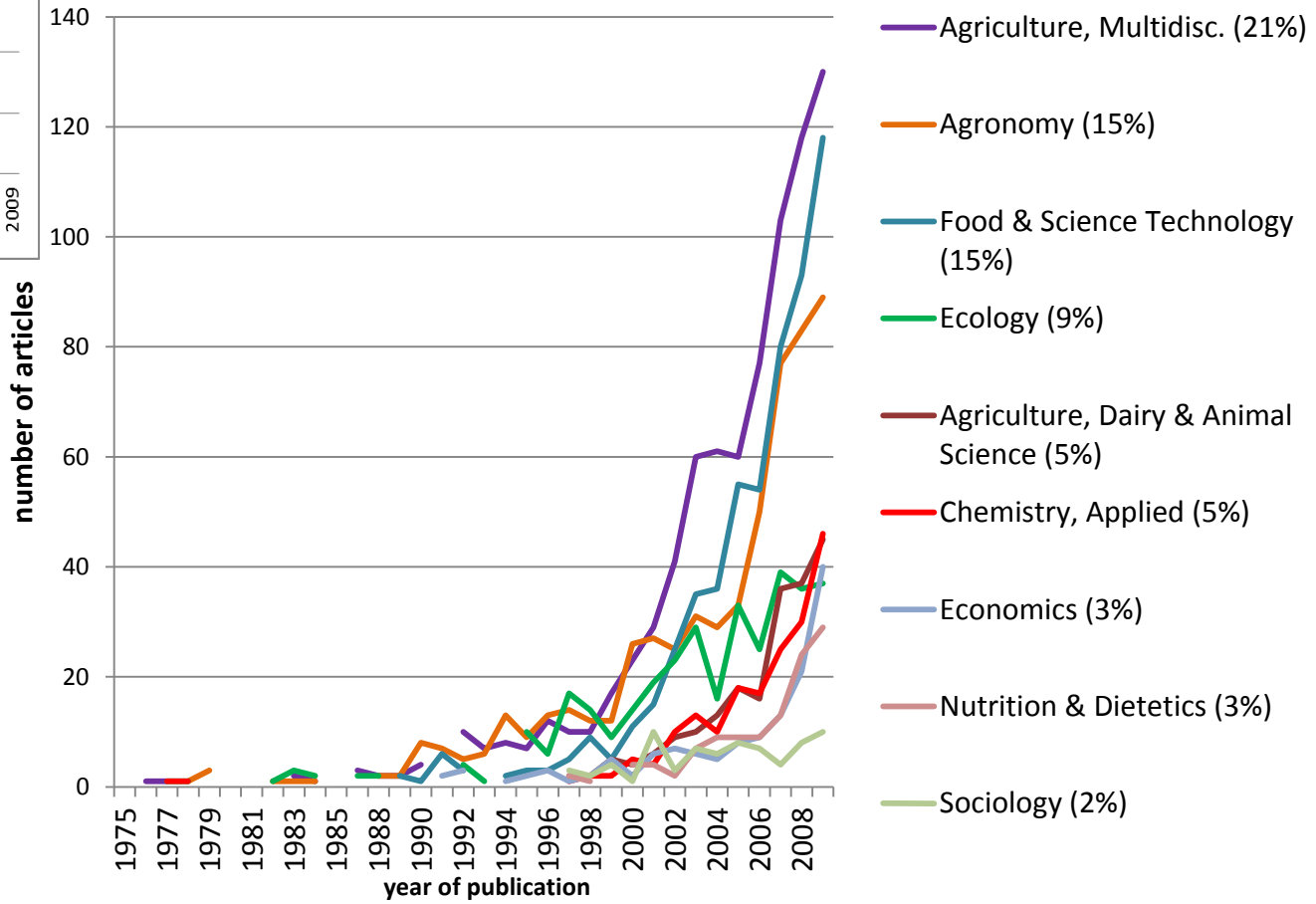
Interprétation itérative pour détection les spécialités fortement structurées.
Variation des paramètres + recherche de convergences

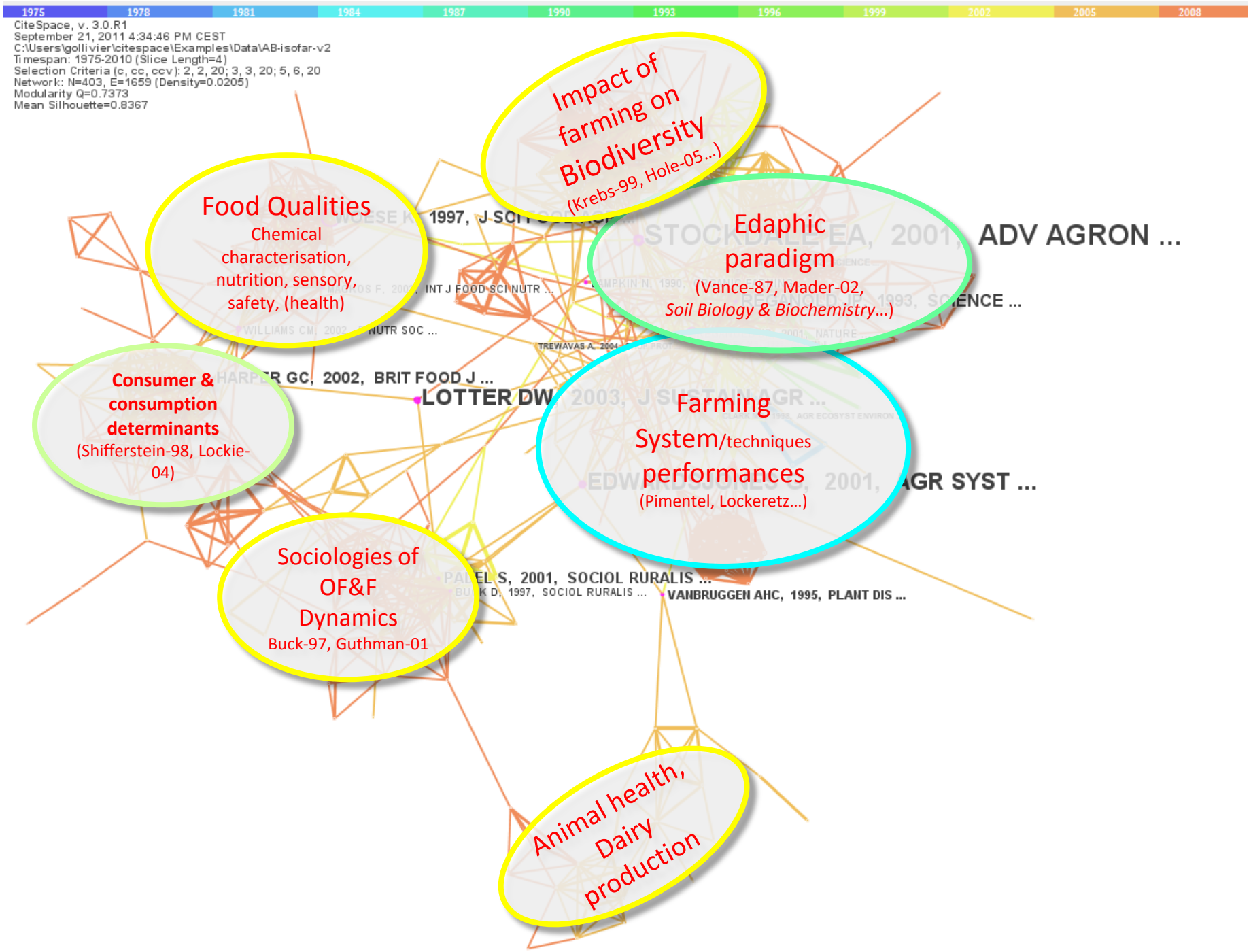
3. Résultats

1) Dynamique d'ensemble



2) Dynamique des principaux sujets d'études (*Subject Areas)*




1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005 2008
CiteSpace, v. 3.0.R1
September 21, 2011 4:34:46 PM CEST
C:\Users\gollivier\citespace\Examples\Data\AB-isofar-v2
Timespan: 1975-2010 (Slice Length=4)
Selection Criteria (c, cc, ccv): 2, 2, 20; 3, 3, 20; 5, 6, 20
Network: N=403, E=1659 (Density=0.0205)
Modularity Q=0.7373
Mean Silhouette=0.8367
Impact of farming on Biodiversity
(Krebs-99, Hole-05…)
Food Qualities
Chemical characterisation, nutrition, sensory, safety, (health)
Edaphic paradigm
(Vance-87, Mader-02, Soil Biology & Biochemistry…)
Consumer & consumption determinants
(Shifferstein-98, Lockie-04)
Farming System/techniques performances
(Pimentel, Lockeretz…)
Sociologies of OF&F Dynamics
Buck-97, Guthman-01
Animal health, Dairy production
1997, J SCI
ADV AGRON ...
INT J FOOD SCI NUTR ...
NUTR SOC ...
TREWAVAS A, 2004
GC, 2002, BRIT FOOD J ...
LOTTER DW
2003, J SUSTAIN AGR ...
2001, AGR SYST ...
EL S, 2001, SOCIOL RURALIS ...
BUCK D, 1997, SOCIOL RURALIS ...
VANBRUGGEN AHC, 1995, PLANT DIS ...

Presse et AB, exemple

- **Intérêt du genre**
  - Repérage d'informations factuelles : évènements, chiffrages…
  - Presse = prescripteur d' « opinion publique »
  - Reflet de représentations / positions des acteurs dans des situations +/- controversées

- **Base de donnée : Factiva**
  - accès : abonnements universitaires ms sans les possibilité de veille/alerte (flux RSS)
  - enjeu juridique / réutilisation des textes
  - sources : presse nationale, quotidienne régionale, spécialisées françaises / internationales depuis parfois les années 90 (très variable selon les titres)

- **Construction de la requête** : étape délicate à ne pas sous-estimer :
  - Industrie = « Agriculture biologique » : N= 2115 articles
  - Texte libre :
    - **Simple ("agriculture biologique") : N=17938 articles**
    - **Requête Élaborée (RE): N= 81832 articles !!**
      (dont environ 20 % de "faux" résultats et doublons)

Pls jours de w :
- Inclusion des multiples désignations (le/la bio, agribio…)
- Inclusion des segments de la filière (« fourche=>fourchette »)
- Inclusion des différents productions
- Exclusion des bruits (bio carburants?)

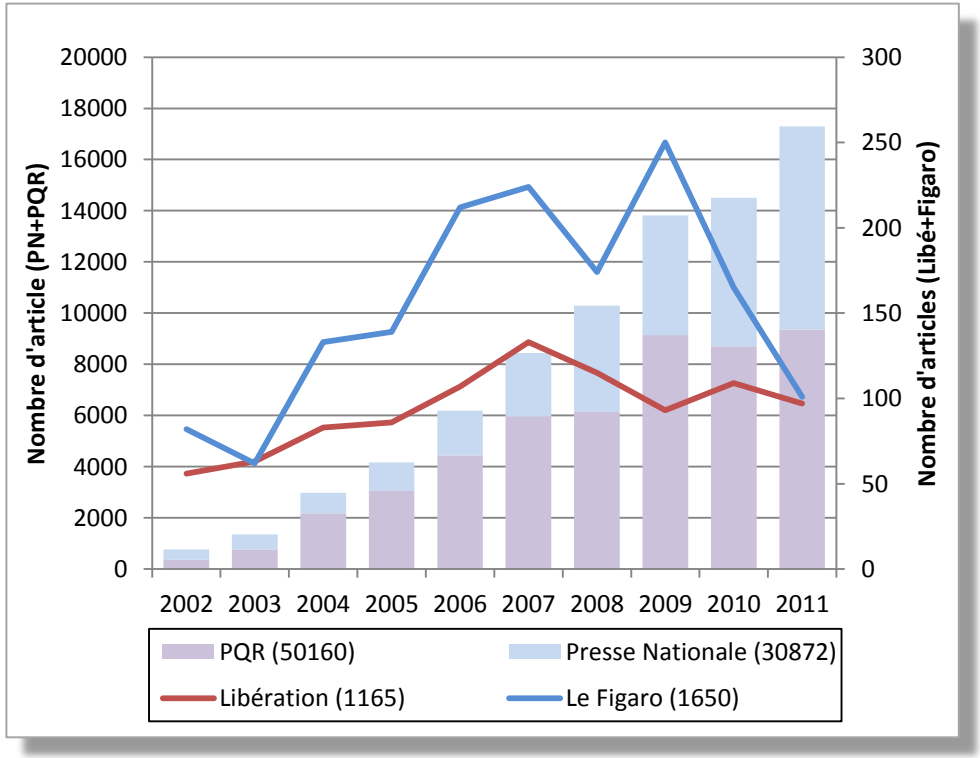
⇒ **Nécessité de bien définir les critères du domaine d'intérêt**
⇒ **Nécessité d'un contrôle humain pour assurer la qualité du corpus**

Presse et AB, exemple

Possibilité de filtrer dans Factiva (ou en aval) :

Requête	N	Remarque
RE x PQR	59302	Beaucoup d'usages « anecdotiques », évènementiels pouvant malgré tout entrer dans l'analyse
RE x Presse nationale (quotidienne, hebdo, spécialisée	30872	Le Monde exclus
RE x (Libération ou Le Figaro)	2765	Depuis 1996, gradient droite-gauche

Dynamiques d'ensemble



- Cinétique d'extension globale
- Principalement portée par la PQR et accroissement récent dans la PN **=> institutionnalisation de l'AB ?**
- Traitements différenciés, au plan temporel du moins, selon les journaux **=> des modes de convocation de l'AB différents ?**

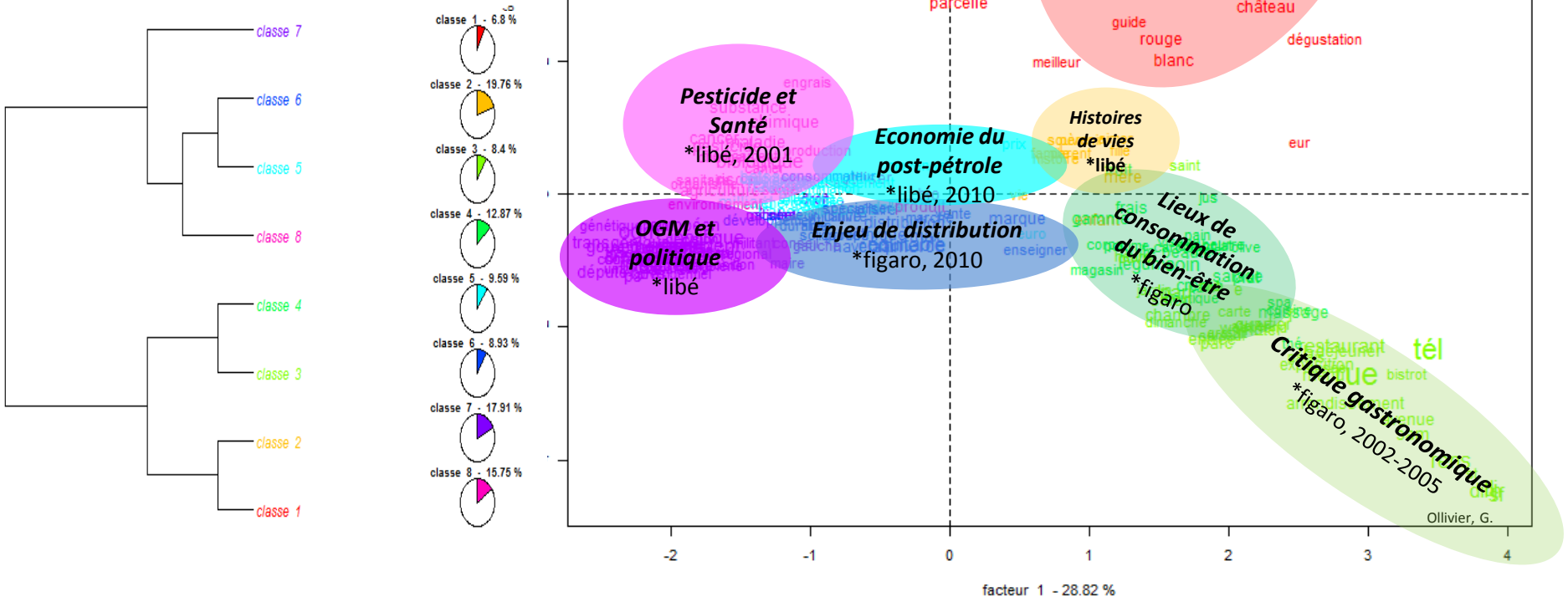
Presse et AB, exemple

Analyse exploratoire (corpus Libé-Figaro)

1- Identification des principaux univers de discours où s'inscrit l'AB

- IraMuteQ : logiciel de lexicométrie

- Classification des termes très cooccurents



Presse et AB, exemple

Analyse exploratoire (corpus Libé-Figaro)

2- Identification des spécificités des différentes sources

- *IraMuteQ : logiciel de lexicométrie*
- *Calcul des surreprésentations statistiques*

Libération	Le Figaro
paysan	marque
Etat	soin
Bové	vin
Confédération	spa
agriculture	prix
politique	restaurant
écologie	carte
militant	massage
McDo	domaine
loi	capitale

Producteur politisé en prise avec l'Etat ⟵ Modes clivés de représentation de l'AB ⟶ Consommateur Chic

Presse et AB, exemple

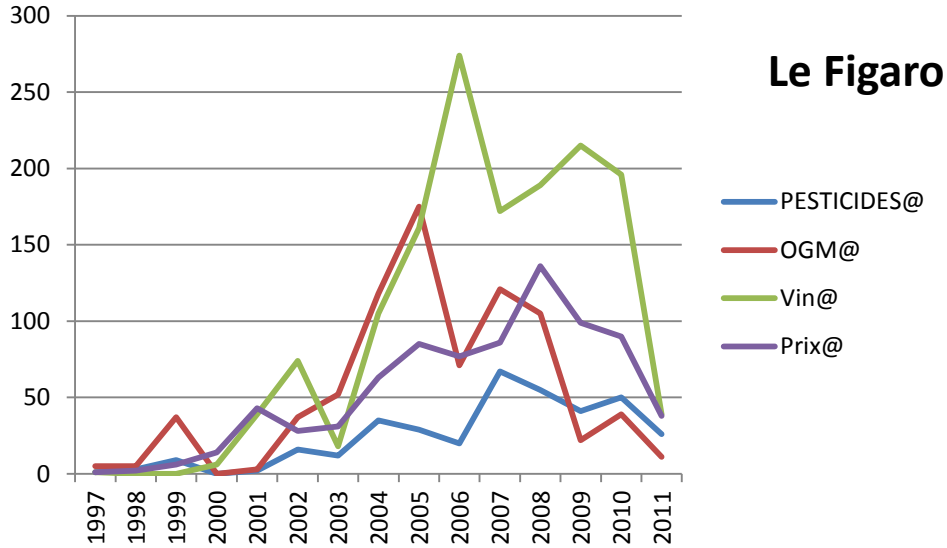
Analyse exploratoire (corpus Libé-Figaro)

3- Identification des dynamiques des principaux objets de controverses

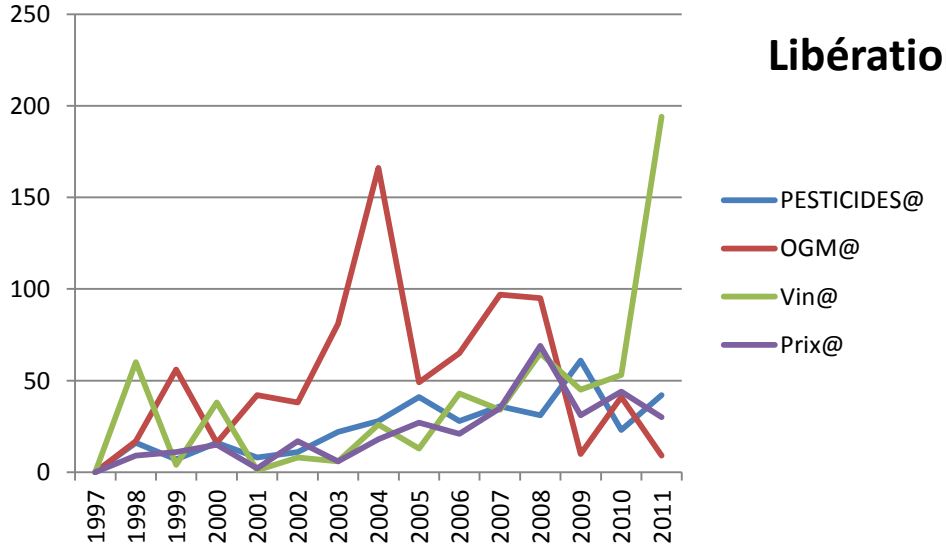
- Outil d'analyse Prospéro (Chateauraynaud, 2003) :

- logiciel d'inspiration sociologique autour de la compréhension des dynamiques de controverses

- développement continu en lien avec projet de recherche (ex : observatoire de veille sociologique des dossiers sanitaires et environnementaux – ANSES, http://gspr.ehess.free.fr/documents/rapports/RAP-2011-AFSSET.pdf)



Le Figaro



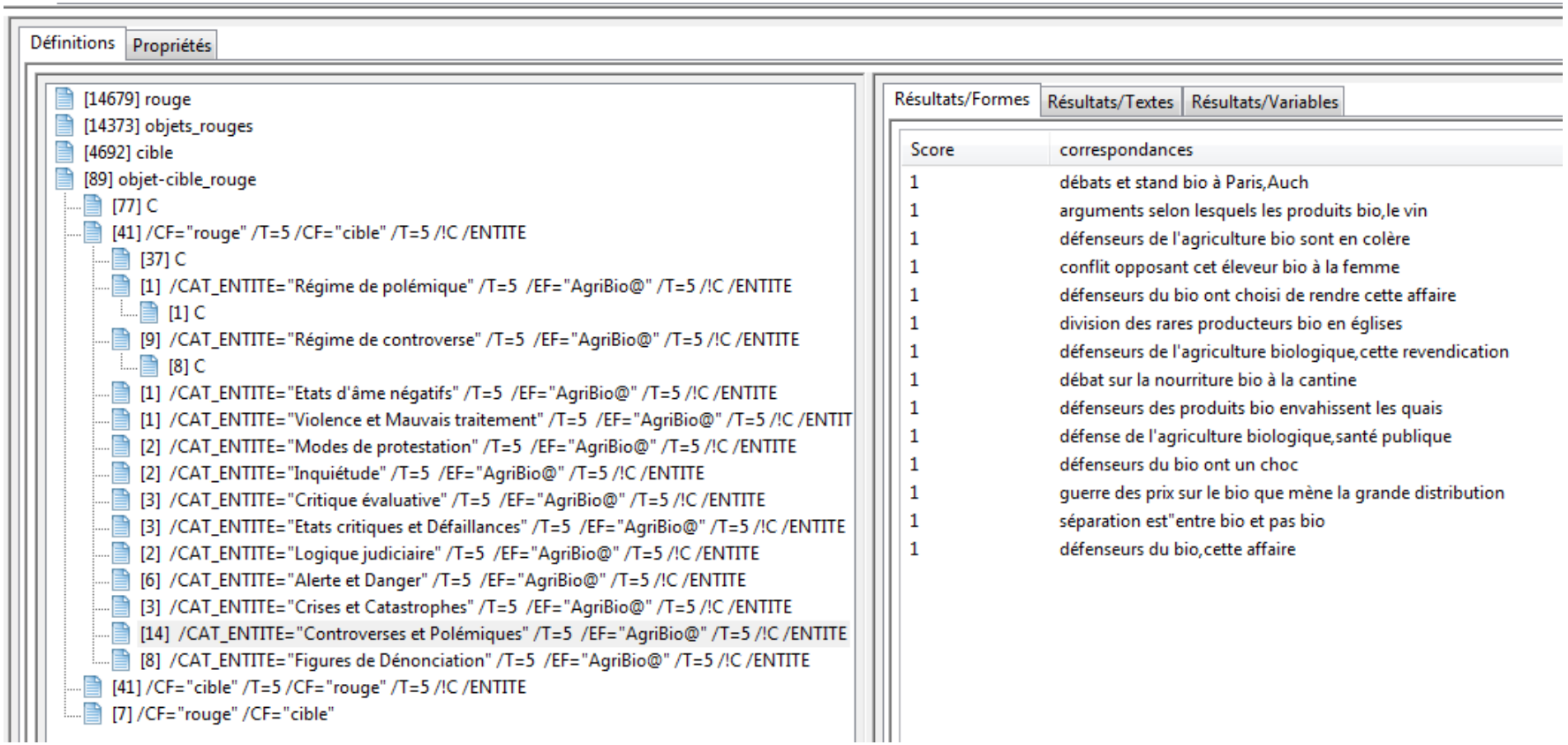
Libération

Presse et AB, exemple

Analyse exploratoire (corpus Libé-Figaro)

3- Identification des dynamiques des principaux objets de controverses

- recherche de 'formules'



Conclusion

- Identification délicate du périmètre (sources – mots clés – critères)
- Nécessité d'un contrôle humain tout au long du processus d'analyse
- Analyse : que cherche t'on en priorité à savoir ?
- Production de connaissance : des outils certes mais surtout des **questions** !