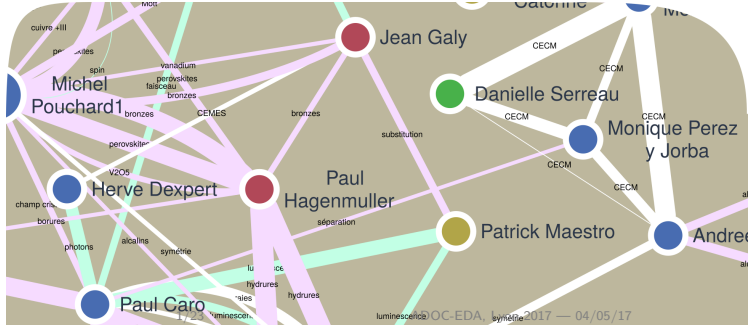
ADOC - EDA, Lyon 2017

Extraction d'expressions et mise en réseau d'un corpus

Matthieu Quantin[1,2], **Benjamin Hervy**[3], **Pierre Teissier**[2]

[1] LS2N / Ecole Centrale de Nantes
[2] Centre François Viète / Université de Nantes
[3] CERHIO / Université d'Angers



Genèse

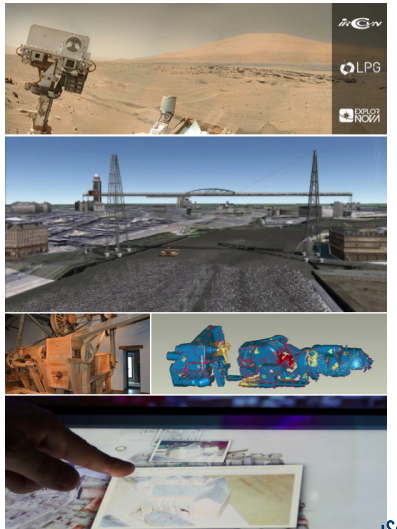
Interdisciplinaire?

- CFV: histoire des sciences et des techniques et épistémologie
- LS2N: 3D et analyse de données

Projets majeurs

- 2007: Machine à laver le sel de Batz-sur-Mer
- 2010: Cintreuse à membrure Benni
- 2014: Nantes1900
- 2015: Pont transbordeur de Nantes
- 2015: Curiosity

→ Patrimoine scientifique / technique / industriel



Usages

- O L'analyse (ex: Topic-modeling)
- O La valorisation (ex: data-visualisation)
- O La description (ex: DCterms)
- O L'indexation (ex: OpenTheso)
- X Le partage (ex: OAI)
- X La formalisation (ex: CIDOC)
- X La structuration (ex: TEI)
- ...

Corpus de textes et numérique

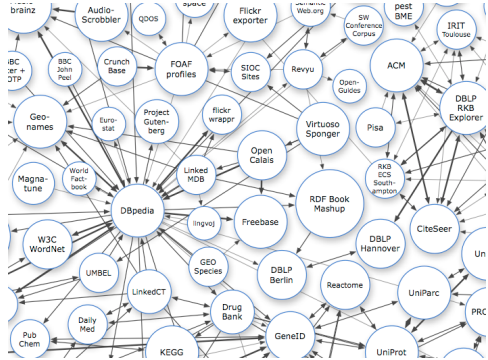


Figure: Rentrer dans le monde des LOD...
...le Graal des humanités numériques?

Cahier des charges d'une solution

Corpus *construit*

- Composants choisis. Homogénéité potentielle.
- Connaissance qualitative du corpus $\rightarrow$ volume $\leq$ 5M mots
- Textes bruts, faiblement ou non-structurés

Cahier des charges d'une solution

Corpus *construit*

- Composants choisis. Homogénéité potentielle.
- Connaissance qualitative du corpus $\rightarrow$ volume $\leq$ 5M mots
- Textes bruts, faiblement ou non-structurés

Corpus *unique*

- Minimiser l'apprentissage
- Pas de modèle de données à priori (ni volonté/ ni besoin de s'y référer)
- Pas de restriction du champs lexical des textes
- Logique floue

Cahier des charges d'une solution

Corpus *construit*

- Composants choisis. Homogénéité potentielle.
- Connaissance qualitative du corpus $\rightarrow$ volume $\leq$ 5M mots
- Textes bruts, faiblement ou non-structurés

Corpus *unique*

- Minimiser l'apprentissage
- Pas de modèle de données à priori (ni volonté/ ni besoin de s'y référer)
- Pas de restriction du champs lexical des textes
- Logique floue

Corpus *connu*

- Interfaçage avec le spécialiste
- Apport de nouvelles informations

Objectif

Pragmatique

Créer nouvelles connaissance historiques (heuristique)

- Dépasser la juxtaposition d'informations
- Ouvrir de nouvelles pistes de recherches

Objectif

Pragmatique

Créer nouvelles connaissance historiques (heuristique)

- Dépasser la juxtaposition d'informations
- Ouvrir de nouvelles pistes de recherches

→ Un graphe

Objectif

Pragmatique

Créer nouvelles connaissance historiques (heuristique)

- Dépasser la juxtaposition d'informations
- Ouvrir de nouvelles pistes de recherches

→ Un graphe

Épistémologique

Penser les relations historien / algorithme (réflexif)
→ L'historien supervise

Objectif

Pragmatique

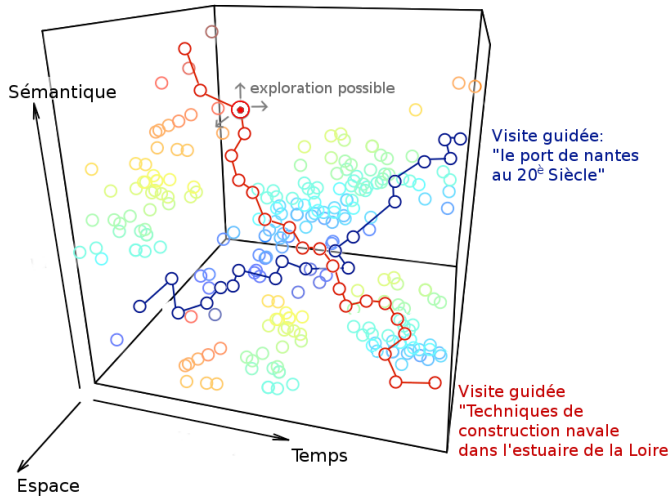
Créer nouvelles connaissance historiques (heuristique)

- Dépasser la juxtaposition d'informations
- Ouvrir de nouvelles pistes de recherches

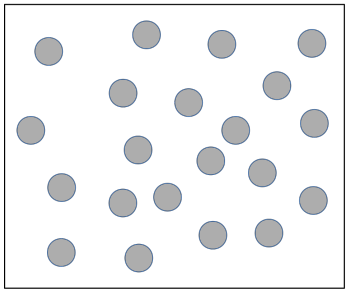
→ Un graphe

Épistémologique

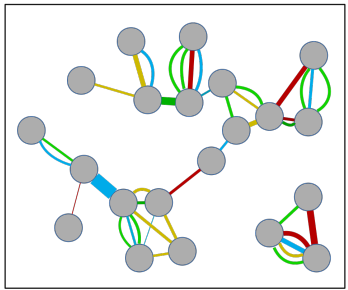
Penser les relations historien / algorithme (réflexif)
→ L'historien supervise



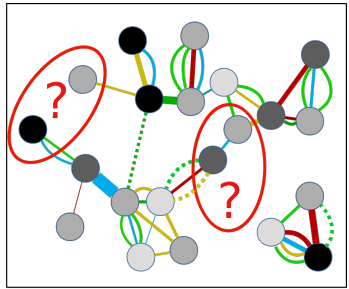
Objectif



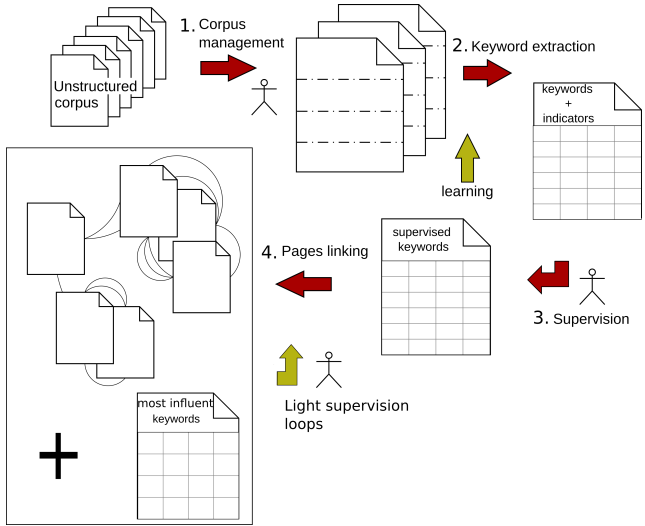
Corpus



Réseau pondéré

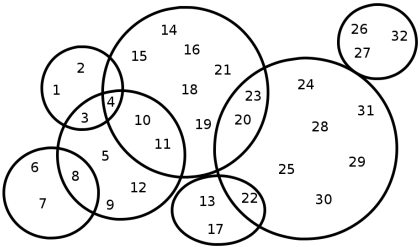


Modération, identification de chaînons singuliers


Unstructured corpus
1. Corpus management
2. Keyword extraction
keywords + indicators
learning
supervised keywords
3. Supervision
4. Pages linking
Light supervision loops
most influent keywords

1. Découpage du corpus

- Lemmatisation [Schmid, 1994]
- Topic-modeling: NMF [Berry et Browne, 2005]
- Suggestion d'erreurs



Si fichier $\neq$ document

Découpage selon paramètres

Objectifs

- Sous-corpus homogènes
- Filtrage de clusters (eg. partiellement bilingue)
- Pré-analyse

2. Extraction d'expressions-clé

Design

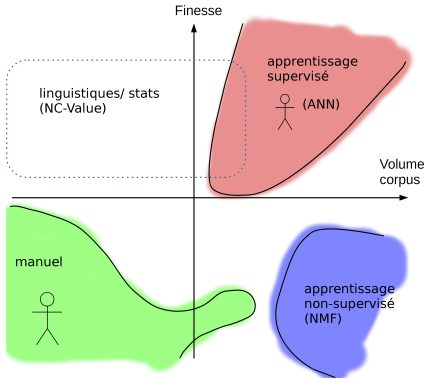
- Discriminer → skip ngrams (MWE)
- Exhaustif/Supervisé → *Recall* privilégié
- Indicateurs de supervision

Algorithme

- Proche C-Value [Frantzi, 1998]
- Inspiré de ANA [Enguehard, 1995]

Indicateurs

- Compositionnalité (MED [Bu, 2011])
- POS
- Anomalie de fréquence (wrlog [Cram, 2016])
- Ascendance



Éviter les faux-positif *a posteriori*
(sémantiquement pauvre AND discriminant)

Exemples d'expressions extraites

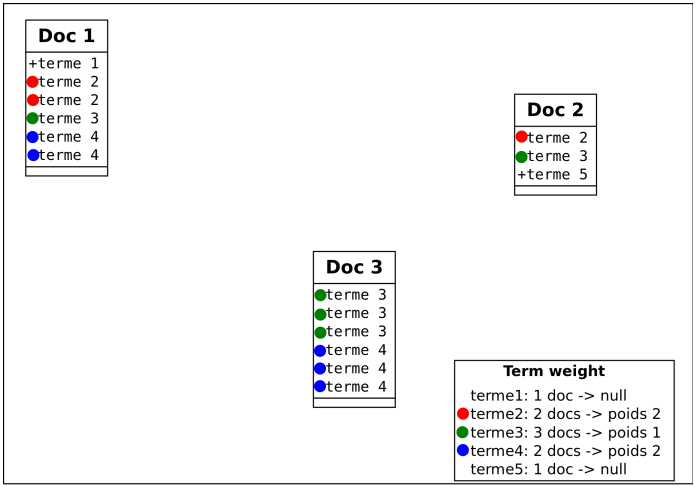
id	Forme extraite	occ	inDoc	theme	mark	merge
1	thèse	186	43	Éducation	0.8	-
2	chimie de coordination	22	8	Chimie	0.81	-
3	thèse de troisième cycle	16	7	France, Éducation	0.7	-
4	année de thèse	78	34	Éducation	0.44	1
5	microscopie électronique	63	20	Physique	0.86	-
6	microscopie électronique à transmission à haute résolution	3	2	Physique	0.58	5, 17
7	four solaire d'Odeillo	3	4	Industrie, Sciences	0.91	-
8	bronzes de vanadium	26	5	Chimie	0.89	-

4. Création de liens

Idée

Liens entre fiches, basés sur la cooccurrence de MWE

- **idf** du terme : générique vs spécifique

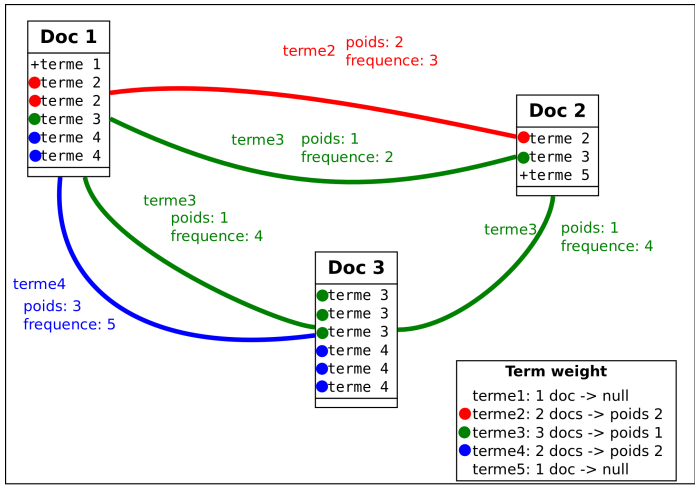


4. Création de liens

Idée

Liens entre fiches, basés sur la cooccurrence de MWE

- **idf** du terme : générique vs spécifique
- **tf** terme / paire de fiches



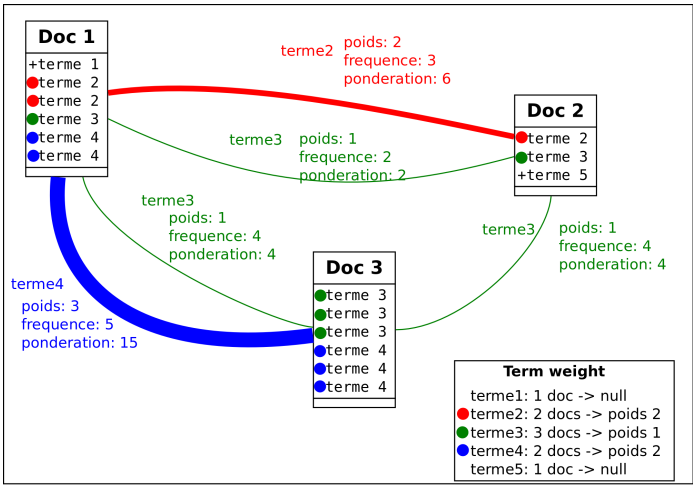
4. Création de liens

Idée

Liens entre fiches, basés sur la cooccurrence de MWE

- **idf** du terme : générique vs spécifique
- **tf** terme / paire de fiches

→ **pondération du lien** : $w = f(volume, poids)$



4. Création de liens

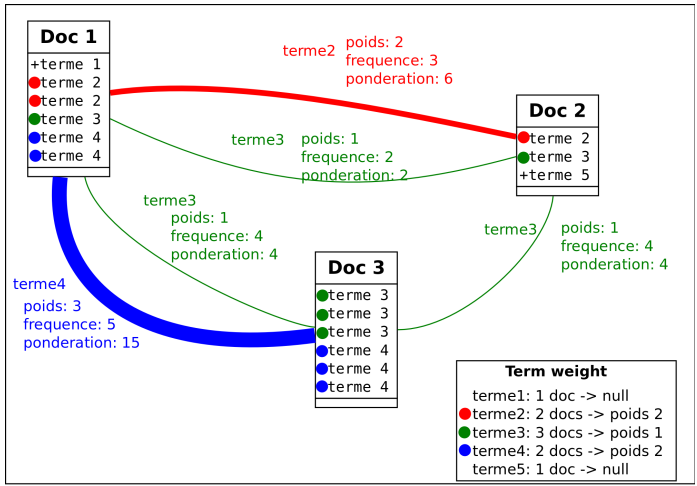
Idée

Liens entre fiches, basés sur la cooccurrence de MWE

- **idf** du terme : générique vs spécifique
- **tf** terme / paire de fiches

→ **pondération du lien** :
$w = f(volume, poids)$
On parle de graphe multiple flou

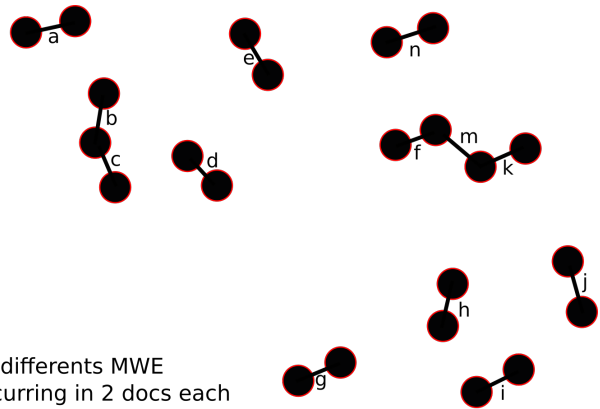
[Blue, 2002]



4. Création de liens

Explosion combinatoire

- Spécifique

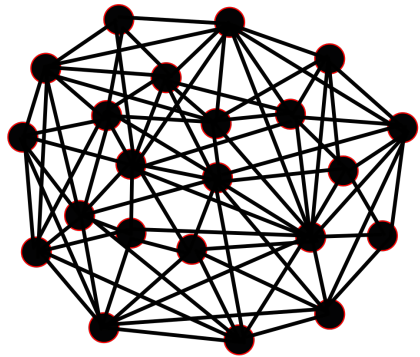


21 differents MWE
occurring in 2 docs each
→ 21 links

4. Création de liens

Explosion combinatoire

- Spécifique
- Générique



One single MWE
occurring in 21 documents
→ 253 links

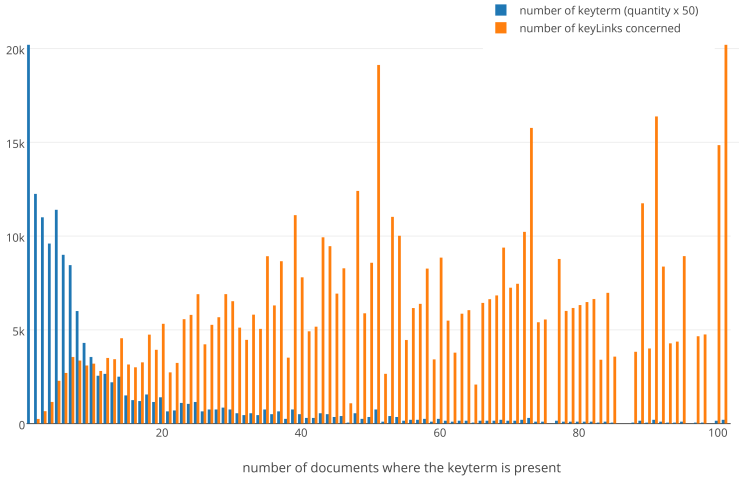
4. Création de liens

Explosion combinatoire

- Spécifique
- Générique

→ pondération avec effet de seuil

few keyLinks concerned by high specific keyterms

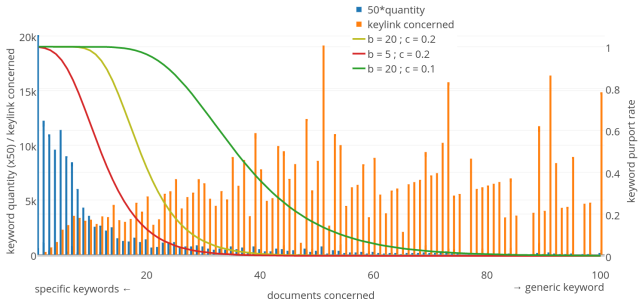


4. Création de liens

Inverse Document Frequency

- Non linéaire
- Stable sur les spécifiques
- Paramétrable

→ Sigmoïde: fonction de Gompertz



$IDF_{t_i} = 1 - \exp^{-b * \exp^{(-c*(D^{t_i}-2)*100/|D|}}$
avec
$|D| = m$: le nombre de document dans le corpus
$D^{t_i} = |\{d_j : t_i \in d_j\}|; j \in [0, m]$ le nombre de document contenant au moins 1 occurrence du terme i
$b \in [5, 20]$ la zone stable
$c \in [0.1, 0.5]$ la pente du seuil

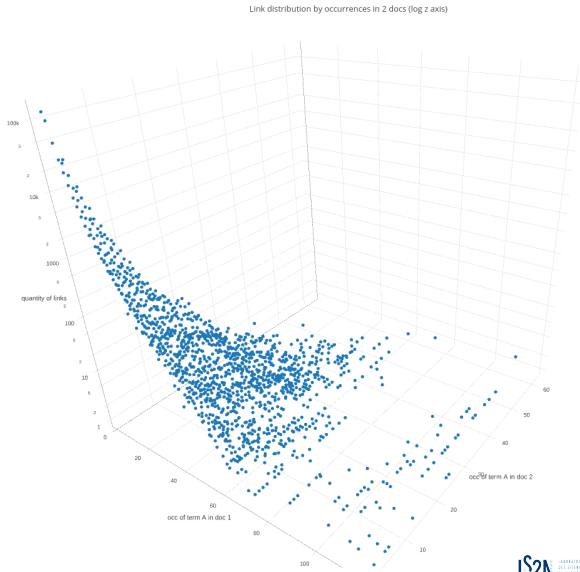
4. Création de liens

Term Frequency

- Non linéaire
- Favorise les équi-répartitions
- Stable en fortes fréquences

$$TF^{t_i}_{d_j;d_k} = \log((O^{t_i}_{d_j}; O^{t_i}_{d_k}) \times min^3(O^{t_i}_{d_j} + O^{t_i}_{d_k}))$$

avec
$O^{t_i}_{d_j} = |\{\{t_i : t_i \in d_j\}\}|$ nombre d'occurrences du terme i dans le document j



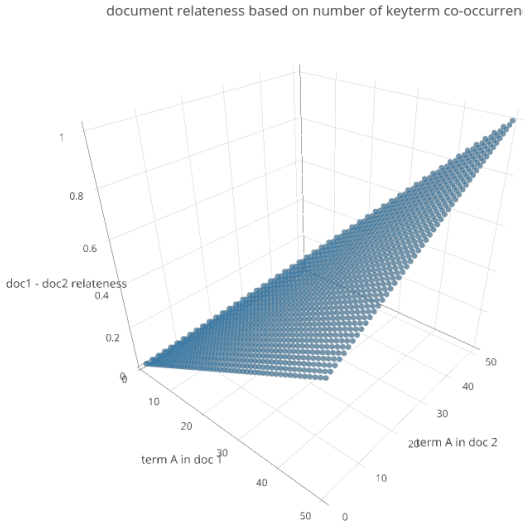
4. Création de liens

Term Frequency

- Non linéaire
- Favorise les équi-répartitions
- Stable en fortes fréquences

$$TF^{t_i}_{d_j;d_k} = \log((O^{t_i}_{d_j}; O^{t_i}_{d_k}) \times min^3(O^{t_i}_{d_j} + O^{t_i}_{d_k}))$$

avec
$O^{t_i}_{d_j} = |\{\{t_i : t_i \in d_j\}\}|$ nombre d'occurrences du terme i dans le document j



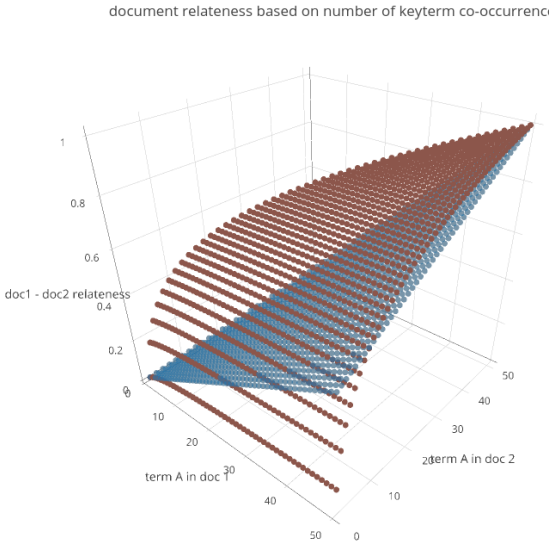
4. Création de liens

Term Frequency

- Non linéaire
- Favorise les équi-répartitions
- Stable en fortes fréquences

$$TF^{t_i}_{d_j;d_k} = \log((O^{t_i}_{d_j}; O^{t_i}_{d_k}) \times min^3(O^{t_i}_{d_j} + O^{t_i}_{d_k}))$$

avec
$O^{t_i}_{d_j} = |\{\{t_i : t_i \in d_j\}\}|$ nombre d'occurrences du terme i dans le document j



Chimie du solide

Le corpus

- Entretiens
- Chimie du solide / *Materials Research*
- Thèse (2007), livre(2013), articles.
  Pierre Teissier

Problématique générale

Identifier la naissance et la dispersion d'une communauté scientifique.

Table: Caractéristiques du corpus étudié

format de fichier	Open Document
structure interne (titre, sous-titre, etc.)	non
nombre de documents	41
nombre de mots	339k
nombre de mots après filtre	87 316
nombre de lemmes différents	9 884
moyenne du nb de mots par document	8268
écart-type du nb de mots par document	4837

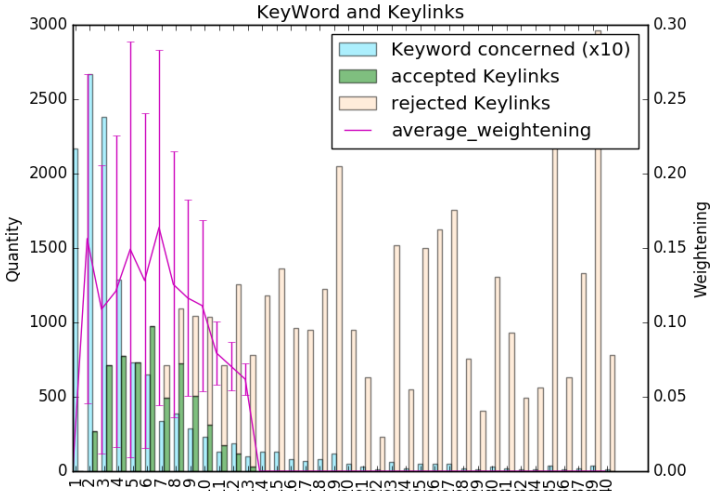
2 pics

Création de liens

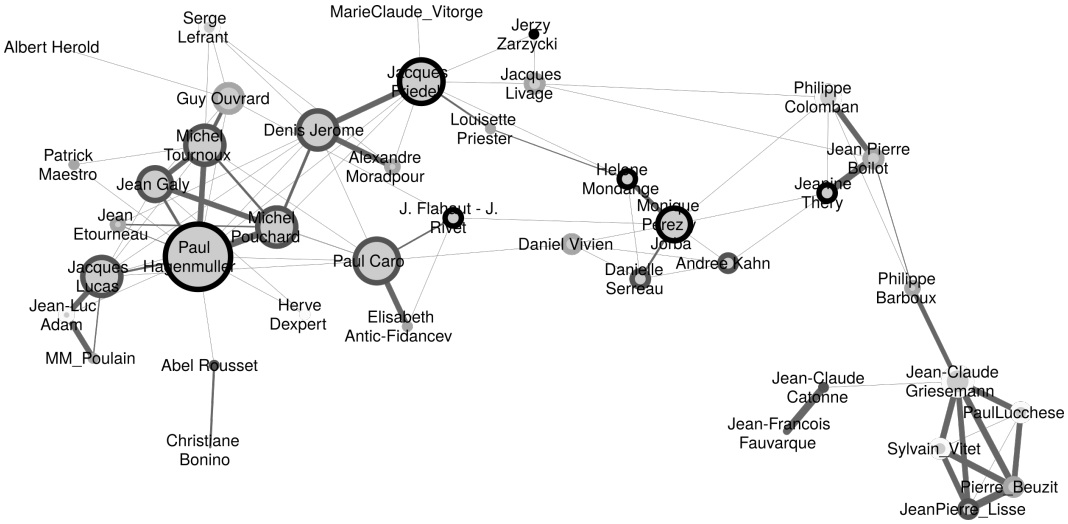
- Expressions $\in$ 7 docs
- Expressions $\in$ 2 ou 3 docs

2 vitesses

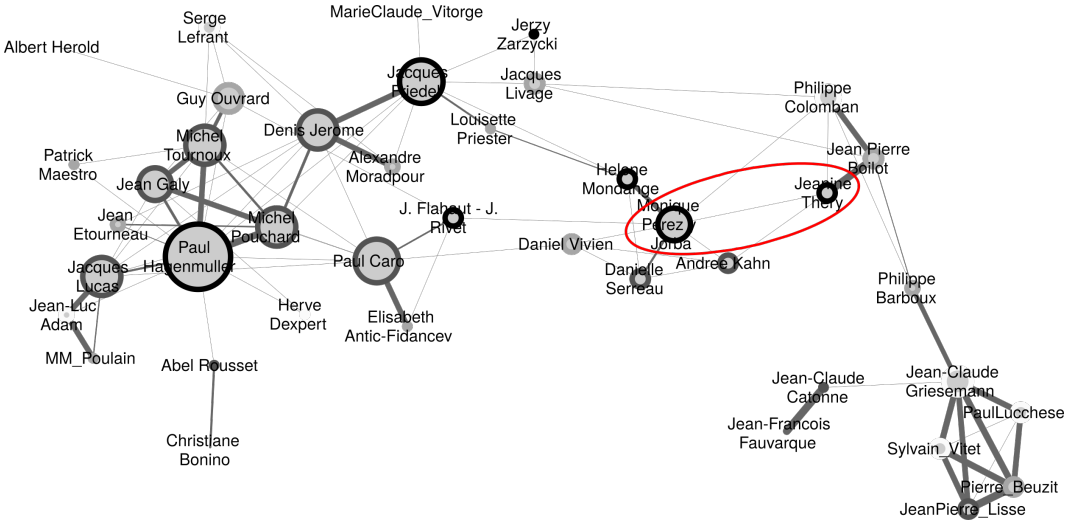
- Groupes de 2 ou 3 chercheurs dont 1 est détenteur de la majorité des occurrences
- Groupes de 7 chercheurs sans prépondérance



Le point d'entrée



Le point d'entrée



Point surprenant

Point surprenant

Table: Liste des expressions clés les plus représentatives liant Monique Pérez et Jeanine Théry.

Mot-clé	Poids	tot	A	B
alumine	0.414	30	4	4
mat. réfract.	0.197	23	3	3
Daniel Vivien	0.093	28	2	2
Vitry-sur-Seine	0.062	50	3	6
Perez	0.052	11	1	1
ferrites	0.052	10	1	1

Point surprenant

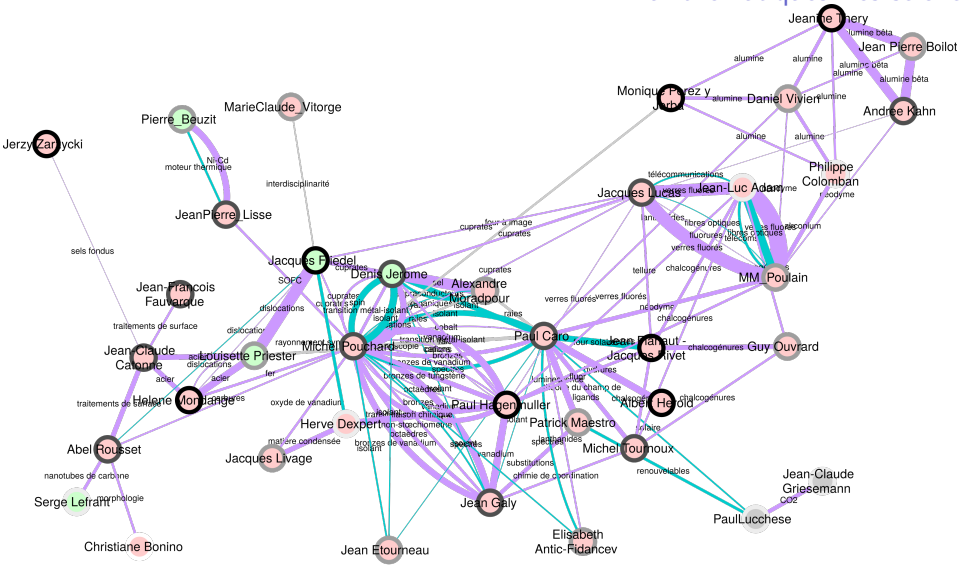
Table: Liste des expressions clés les plus représentatives liant Monique Pérez et Jeanine Théry.

Mot-clé	Poids	tot	A	B
alumine	0.414	30	4	4
mat. réfract.	0.197	23	3	3
Daniel Vivien	0.093	28	2	2
Vitry-sur-Seine	0.062	50	3	6
Perez	0.052	11	1	1
ferrites	0.052	10	1	1

Table: Liste des expressions clés les plus représentatives liant Paul Hagenmuller à ses voisins

Mot-clé	Poids	tot	A	B	Interview associée
vanadium	0.606	23	6	8	J. Galy
vanadium	0.606	23	6	8	M. Pouchard
Trombe	0.533	66	4	45	Paul Caro
bronzes de tungstène	0.511	9	4	5	M.Pouchard
John Goodenough	0.509	19	4	9	M. Pouchard
Jacques Lucas	0.472	24	4	4	JL Adam
fluor	0.441	12	3	6	J. FJ Rivet
verres fluorés	0.421	55	2	23	MM_Poulain
Trombe	0.413	66	4	4	J.J. Rivet
transition métal-isolant	0.409	6	3	3	D. Jerome
verres fluorés	0.408	55	2	19	Jean-Luc Adam
bronzes de vanadium	0.395	22	2	15	Michel Pouchard1
octaèdres	0.375	19	2	12	Michel Pouchard1
théorie des bandes	0.330	17	4	4	Guy Ouvrard

Par thématiques: les sciences



Organisation du process

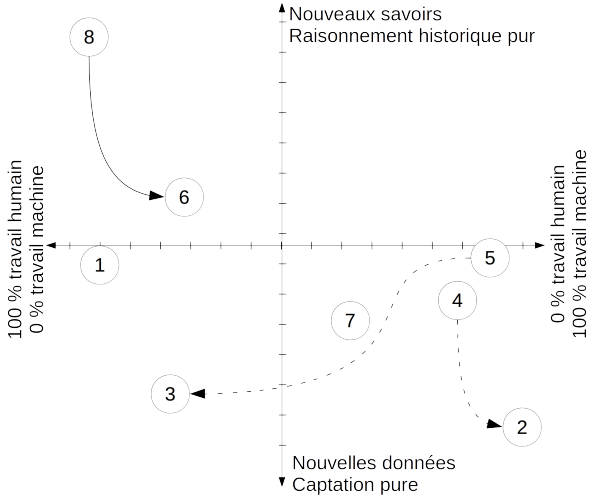


Figure: 1: Constitution du corpus; 2: Extraction d'expression-clés; 3: Supervision des expressions clés; 4: Apprentissage de la supervision; 5: Création de liens entre documents; 6: Construction d'une requête; 7: Construction du graphe; 8: Interprétation du graphe

Conclusion

Retour sur les objectifs

- **Pragmatique**: Apport de connaissances historiques
- **Épistémologique**: Outil construit avec les historiens pour mesurer leurs corpus

merci pour votre attention

matthieu.quantin@ec-nantes.fr