CILLEX

Exploiter les réseaux documentaires pour mieux comprendre les résultats d'une recherche

Emmanuel Navarro, Bruno Gaume

CLLE, CNRS & Université de Toulouse

Séminaires ISTEX

25 avril 2016

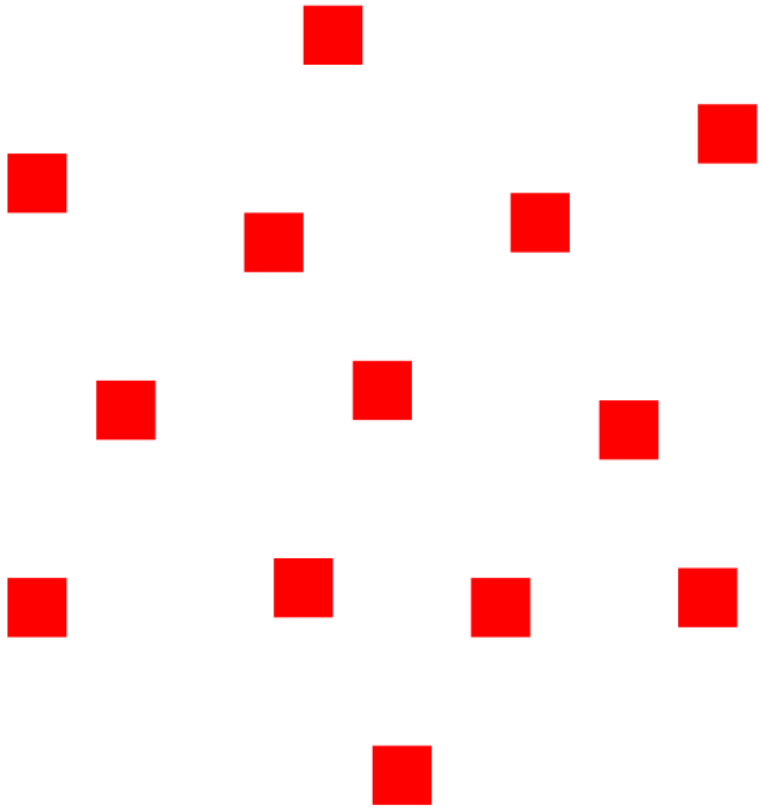
Plan

1. Objectifs
2. Etat Courant
3. Résultats exploitables pour les « Chantiers d'usage »
   - API Cillex
   - Annotations et système d'annotation
   - Graphes globaux
   - Bibliothèque Python, client API ISTEX (et plus)

Objectifs

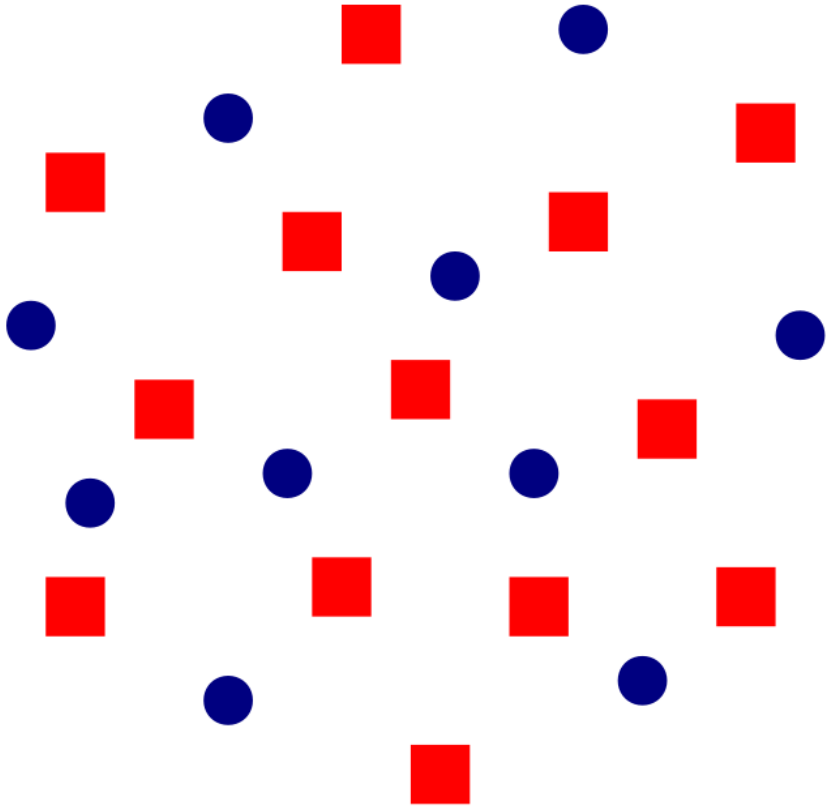
Objectifs

Des Documents



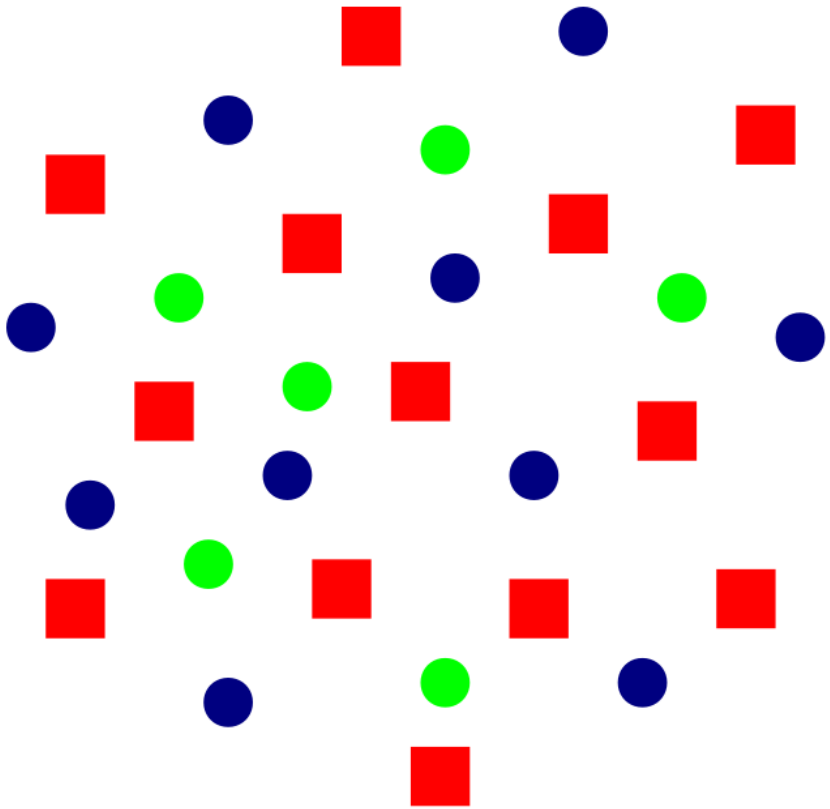
Objectifs

Des Documents, Des Mots



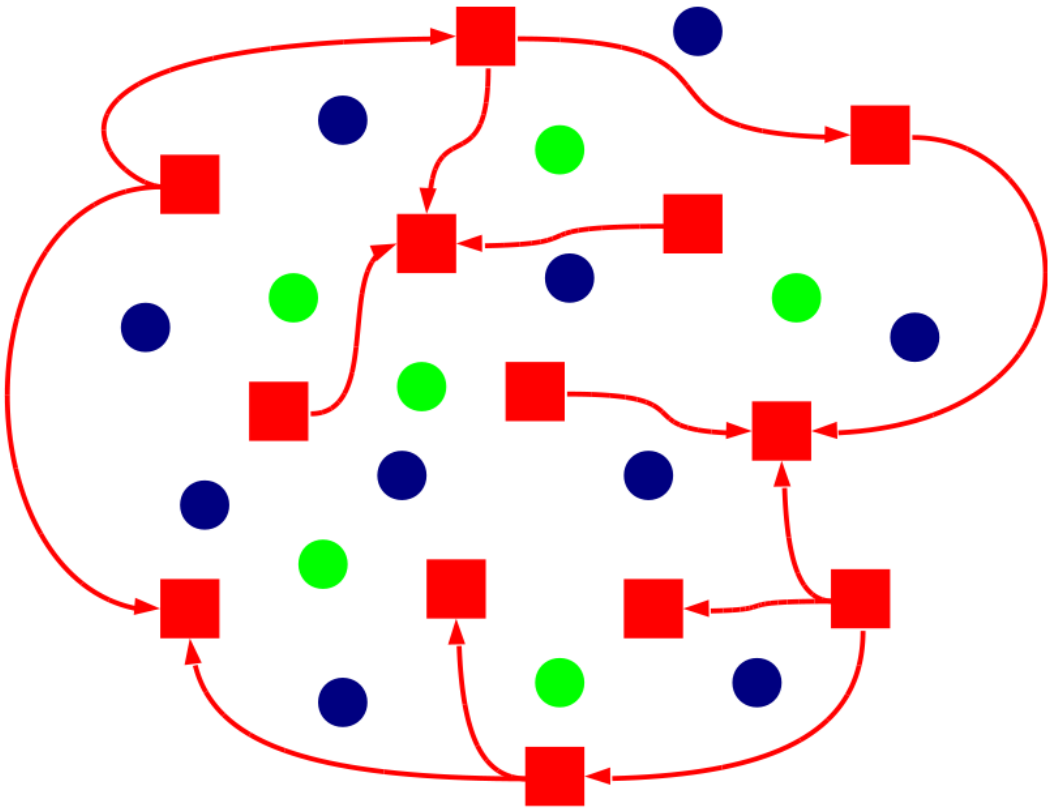
Objectifs

Des Documents, Des Mots, Des Auteurs



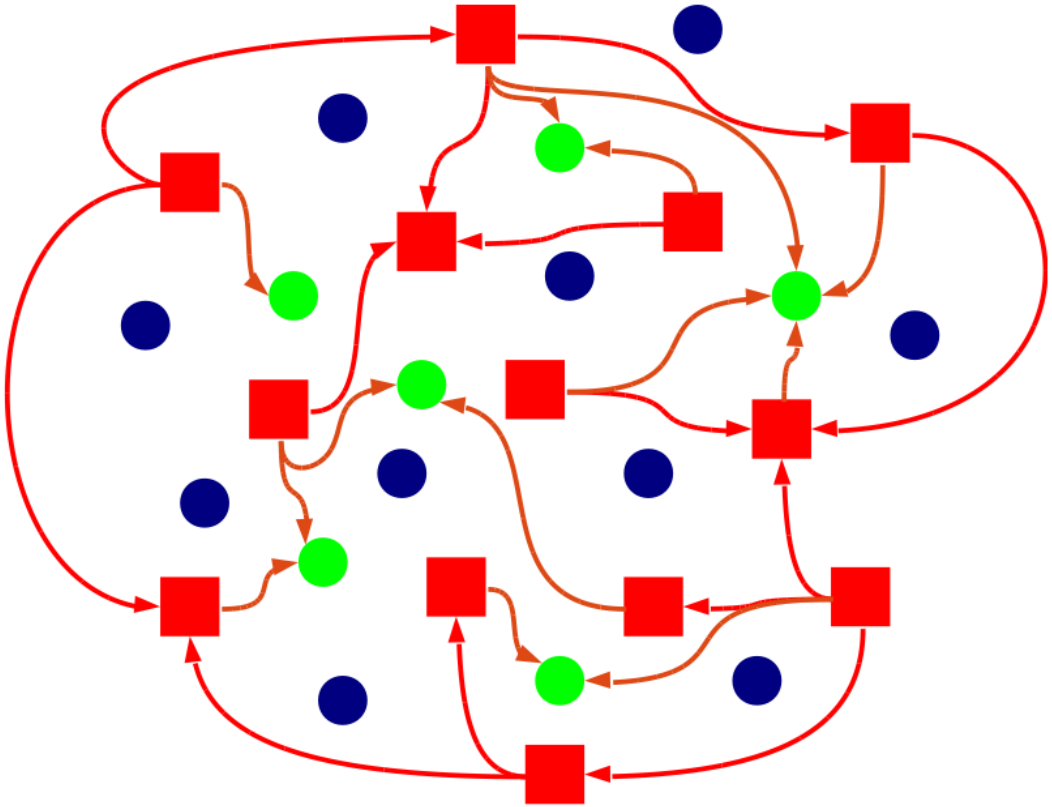
Objectifs

Des Documents, Des Mots, Des Auteurs, ... Des Liens



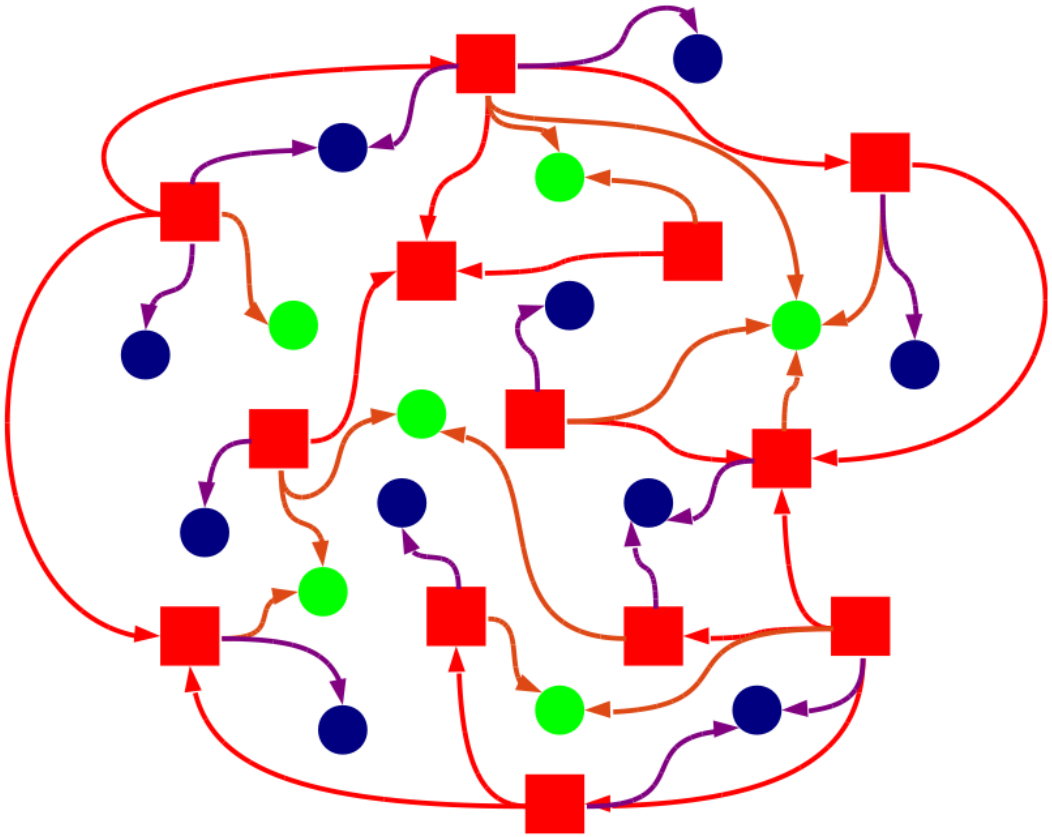
Objectifs

Des Documents, Des Mots, Des Auteurs, ... Des Liens, Des Liens



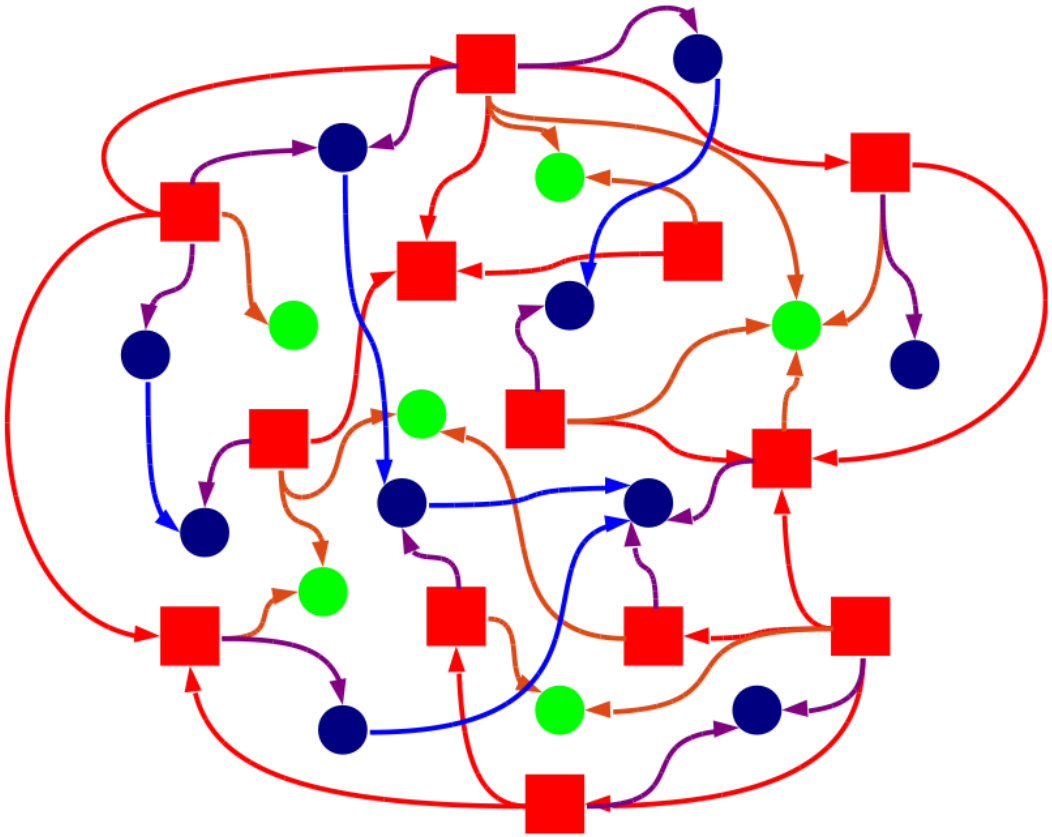
Objectifs

Des Documents, Des Mots, Des Auteurs, ... Des Liens, Des Liens



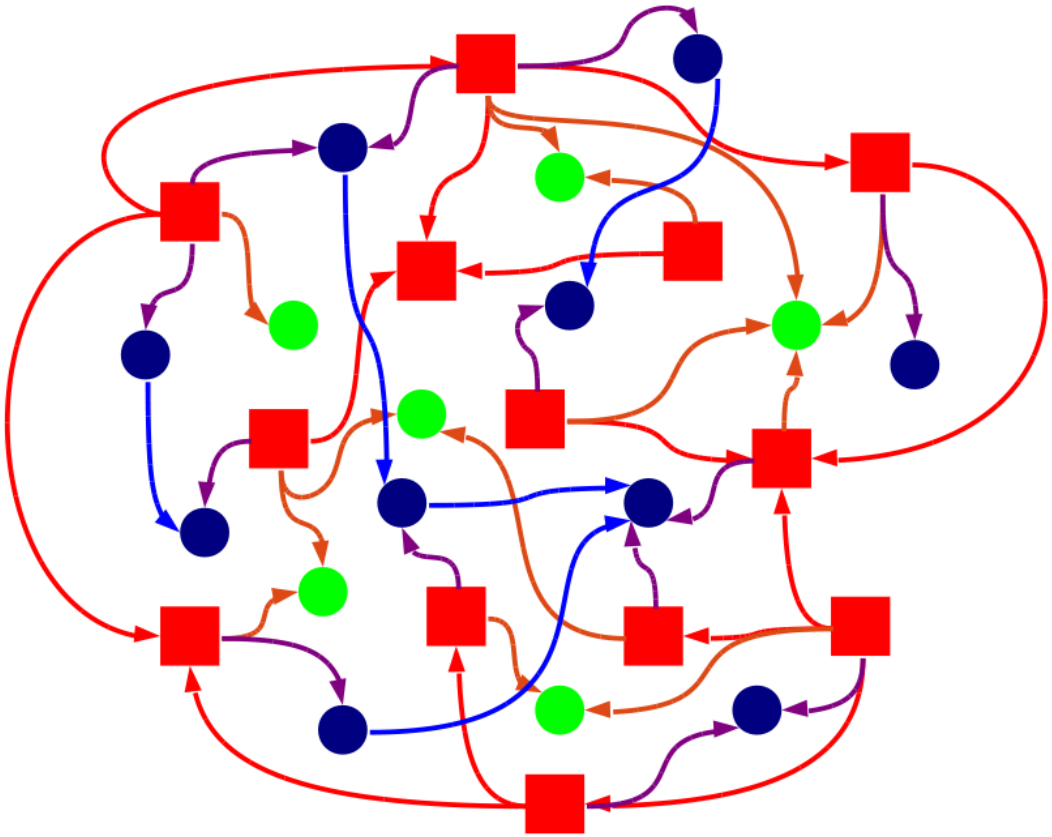
Objectifs

Des Documents, Des Mots, Des Auteurs, ... Des Liens, Des Liens



Objectifs

Graphe de terrain



Objectifs

Graphe de terrain

- Les graphes d'accointance d'un groupe d'humains
- Le graphe du World Wide Web
- Le graphe de Caenorhabditis elegans
- Les graphes Lexicaux
- Les graphes extraits des bases documentaires

Objectifs

Graphe de terrain

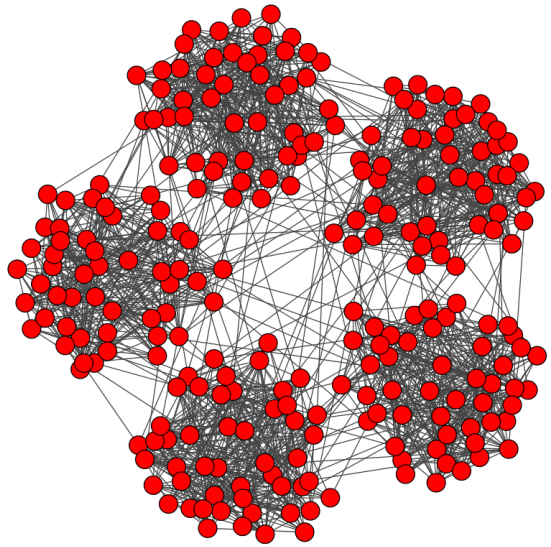
Quatre propriétés fondamentales

- Faible densité
- Chemins courts
- Distribution des degrés à queue lourde (loi de puissance)
- Fort coefficient de clustering : zones denses en arêtes

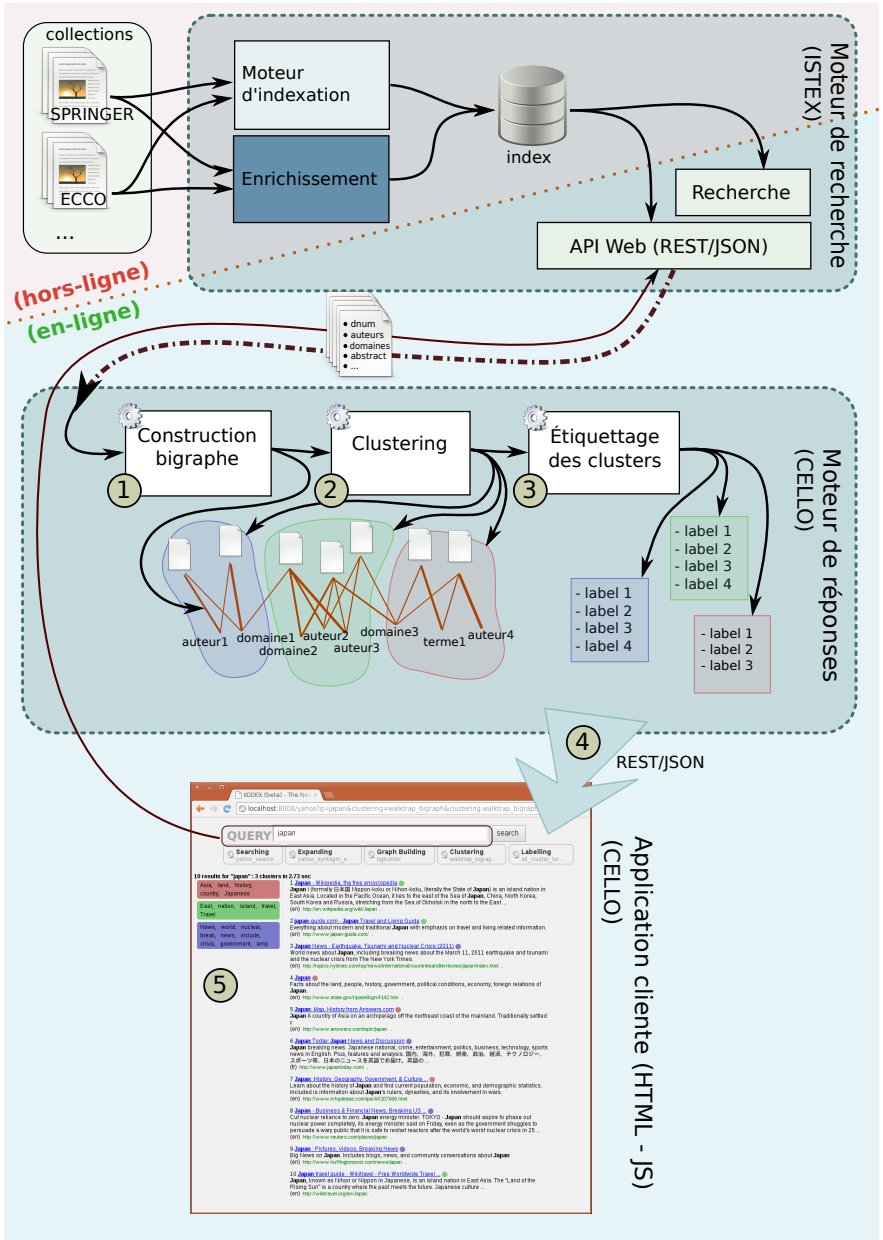
Graphe de terrain

Quatre propriétés fondamentales

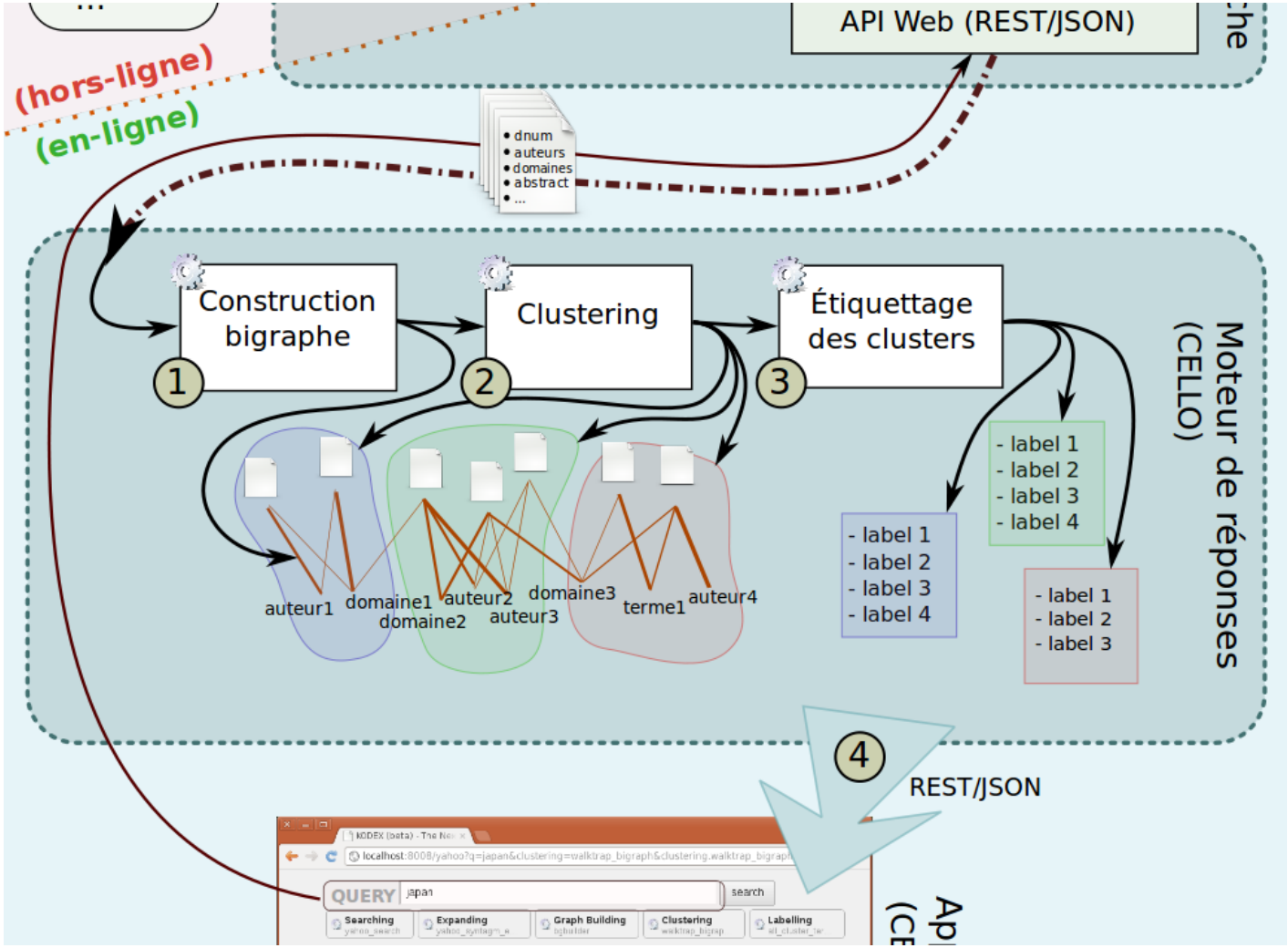
- Faible densité
- Chemins courts
- Distribution des degrés à queue lourde (loi de puissance)
- **Fort coefficient de clustering : zones denses en arêtes**



Schema général



Schema général



Etat Courant

Demo !

http://cillex.kodexlab.com/

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Un point sur les données disponibles/exploitables

- Cat. WOS
  - Résultats intéressants,
  - Couverture partielle,
  - PB : liées aux « host » non aux articles.
- Mots des titres et abstracts
  - Couverture de 100% (ou presque...),
  - Mais, sélection/extraction nécessaire... (travail en cours)
- « Subjects »
  - Couverture partielle,
  - PB : peu de consistance, données éparses,
  - Utilisation du graphe global (travail en cours).
- Ref. bib.
  - Couverture partielle,
  - PB de match/normalisation,
  - On a uniquement les citations sortantes...
- Auteurs, affiliations,
  - PB : données éparses, normalisation.
- « Host », revue, série,
- Enrichissement par les autres projets, à venir...

Résultats exploitables pour les « Chantiers d'usage »

API Cillex (1/2)

- API HTTP/JSON, avec « en plus » de l'API ISTEX :
  - Clustering
  - Labelling
  - Graphe biparti (absent par défaut)

```
$ curl -GET http://cillex.kodexlab.com/api/search/brain
{
  "meta": {...},
  "results": {
    "clusters": {
      "clusters": [...],
      "labels": [...],
    },
    "docs": [...],
    "meta": {...},
    "query": "br  ain"
  }
}
```

Listing 1 – Exemple simple avec requete GET

API Cillex (2/2)

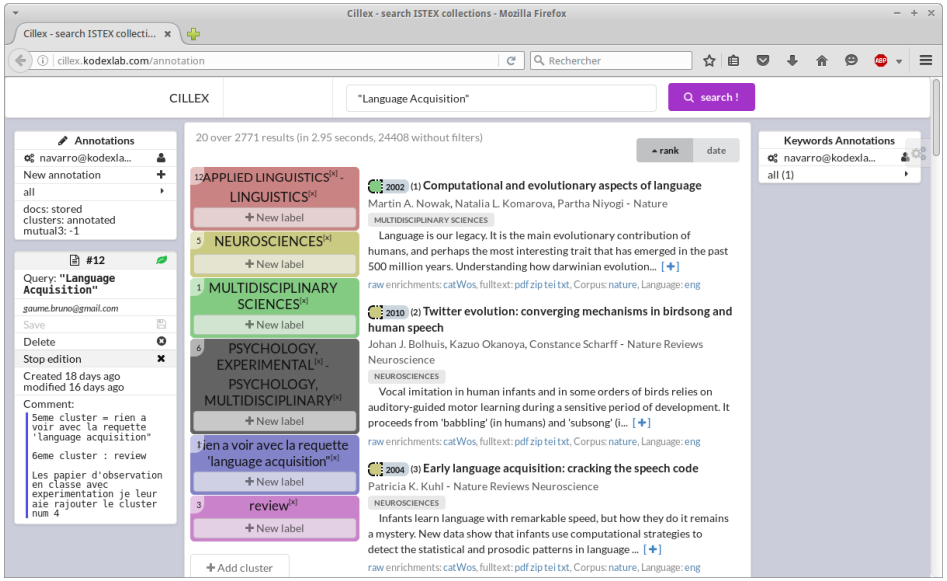
Déjà en place, mais :

- Peu documentée,
- Pas encore de version stabilisée.

Tout ça se corrige si il y en a la besoin...

Annotations et système d'annotation

- Approche hybride entre non-regression & campagne d'éval
- Sys. en trois parties
  - Serveur d'annotation (API JSON)
  - Client annotation des clusters
  - Client annotation des titres & abstracts (mots-clés)



Potentiellement réutilisable :

- Réutilisation du système (serveur) pour construire d'autres annotations,
- Jeu d'annotation de clusters ($\sim$ 15 requêtes pour le moment)
- Jeu d'annotation des titres & abstracts (à venir...)

Graphes globaux

- Possible avec l'API ISTEX de construire des graphes globaux :
  - Subject-Subject,
  - Auteurs-revues,
  - Cat. WOS-subjects,
  - Cat. WOS-revues,
  - *Ref. Bib. (pb de match)*,
  - ...

  (lien si présent ensemble sur -au moins- un document)
- En fct des données demandées, de quelques heures à quelques jours.
- Pour nous : expérimentation (en cours) utilisation du graphe global de *Subjects* pour répondre au problème de données clairsemées.

Envisageable de diffuser :

- Des graphes (format ? licence ?),
- Les scripts de constructions.

Reliex : client Python pour l'API ISTEX (et plus)

- Basé sur *Reliure*[1], framework minimal pour construire des chaines de traitement de données,
- Jeu de composants pour :
  - Construire des requêtes (filtres),
  - Intéroger l'API,
  - Post-Traitements minimaux.

Fonctionnalités principales :

- Système modulable, enchainement de composants (opérateur « | »)
- Déclaration/découverte d'options :
  - construction auto. d'API HTTP/JSON,
  - construction auto. de parser de param. en ligne de cmd (basé sur *argparse*)
- Système de cache HTTP local très simple avec *requests-cache*.

1. `http://reliure.readthedocs.org/`

```python
from reliex.search import ISTEXScanQuery
from reliex.query import LanguageFilter, HasSubject

# Composant d'enrichisement de la requete
query_builder = LanguageFilter() | HasSubject(default=True)
# Composant complet de recherche
search = query_builder | ISTEXScanQuery(outputs="subject")

if __name__ == '__main__':
    from argparse import ArgumentParser
    from reliure.utils.cli import arguments_from_optionable,
        get_config_for
    parser = ArgumentParser()
    arguments_from_optionable(parser, search, prefix="search-")
    parser.add_argument('QUERY', type=str, help='The search query')

    args = parser.parse_args()
    search_config = get_config_for(args, search, prefix="search-")
    query = args.QUERY

    # Run it
    for doc in search(query, **search_config):
        pprint(doc)
```

Listing 2 – simple_example.py

```
$ python simple_example.py -h
usage: simple_example.py [-h]
                         [--search-language
                             {ALL,eng,fre,deu,lat,spa,ita,dut,rus,wel}]
                         [--search-not-has_subject]
                         [--search-has_host_subject]
                         [--search-operator {AND,OR}]
                         QUERY

positional arguments:
  QUERY                 The search query

optional arguments:
  -h, --help            show this help message and exit
  --search-language {ALL,eng,fre,deu,lat,spa,ita,dut,rus,wel}
                        Documents language
  --search-not-has_subject
                        'subject' needed
  --search-has_host_subject
                        'host_subject' needed
  --search-operator {AND,OR}
                        Default search operator
```

Listing 3 – Usage

```
 $ python simple_example.py --search-language fre brain |head -20
{'id': 'DFBCC1D6D7A029F5B26CC3BEE331559893293220',
 'rank': 0,
 'subject': [{'lang': ['fre'], 'value': 'Anura'},
             {'lang': ['fre'], 'value': 'ontogeny'},
             {'lang': ['fre'], 'value': 'heterochrony'},
             {'lang': ['fre'], 'value': 'variation'},
             {'lang': ['fre'], 'value': 'Ceratophryinae'}]}
{'id': 'CA69329F917E7FE9694CB50C047AA711ACAB3781',
 'rank': 1,
 'subject': [{'lang': ['fre'], 'value': 'phylogeny'},
             {'lang': ['fre'], 'value': 'Leptodactylus fuscus
                 group'},
             {'lang': ['fre'], 'value': 'osteology'},
             {'lang': ['fre'], 'value': 'morphology'}]}
{'id': '4335E3679C9B3B080D158F3DC2EEA07656A32ED8',
 'rank': 2,
 'subject': [{'lang': ['fre'], 'value': 'AUFSÄTZE'}]}
{'id': 'E60B5E6F0BBFDE96BB245A09D7C160FB6F0C6072',
 'rank': 3,
 'subject': [{'lang': ['fre'], 'value': 'Escherichia coli'},
             {'lang': ['fre'], 'value': 'facteurs de virulence'},
```

Listing 4 – Exemple

Merci

Questions ?