CILLEX

Aider l'utilisateur dans sa recherche en rendant lisible la structure des résultats

Emmanuel Navarro, Pierre Magistry, Bruno Gaume, Yann Desalle, Yannick Chudy

CLLE/CNRS & Université de Toulouse

Séminaires ISTEX

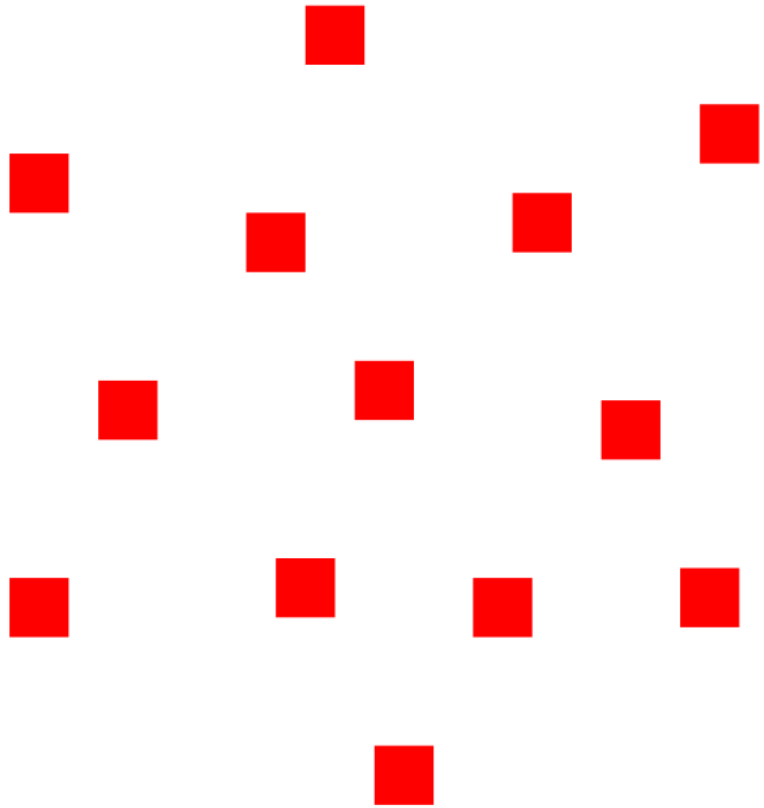
18/19 mars 2015

Plan de l'exposé

Objectifs

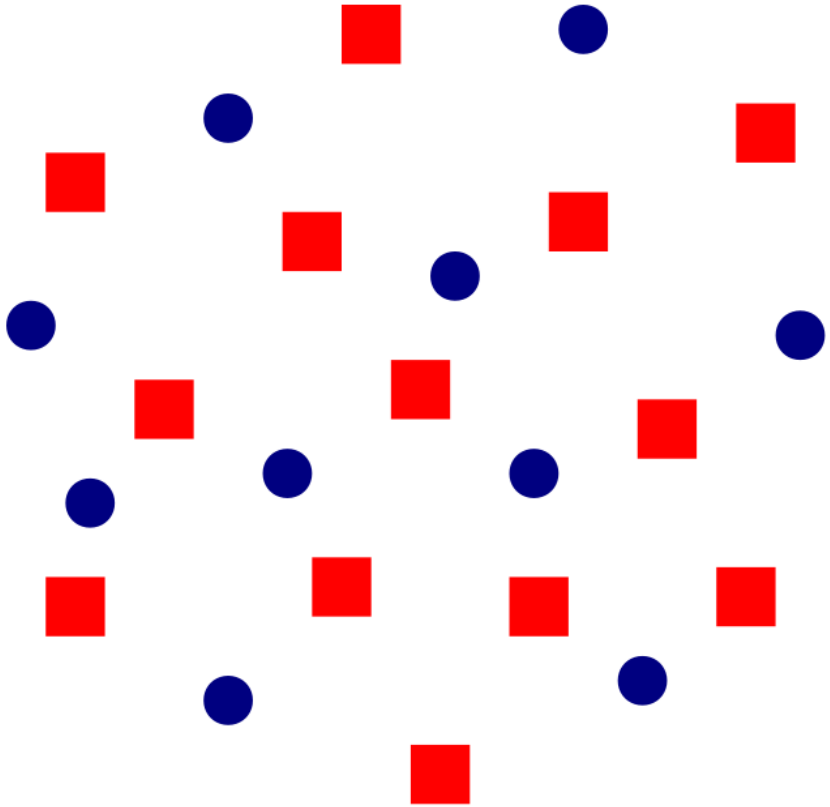
Objectifs

Des Documents



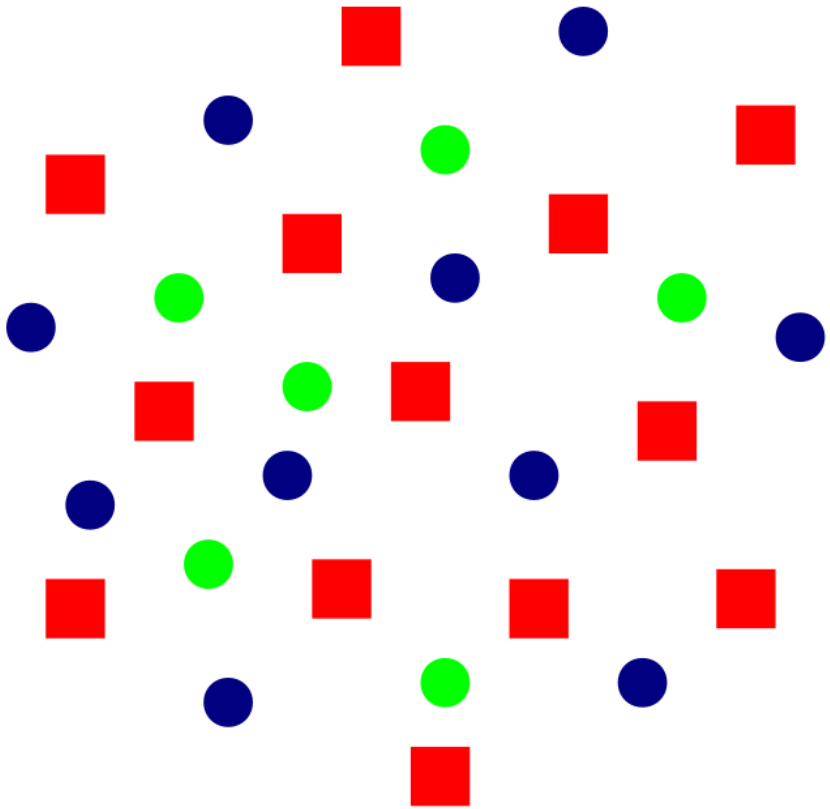
Objectifs

Des Documents, Des Mots



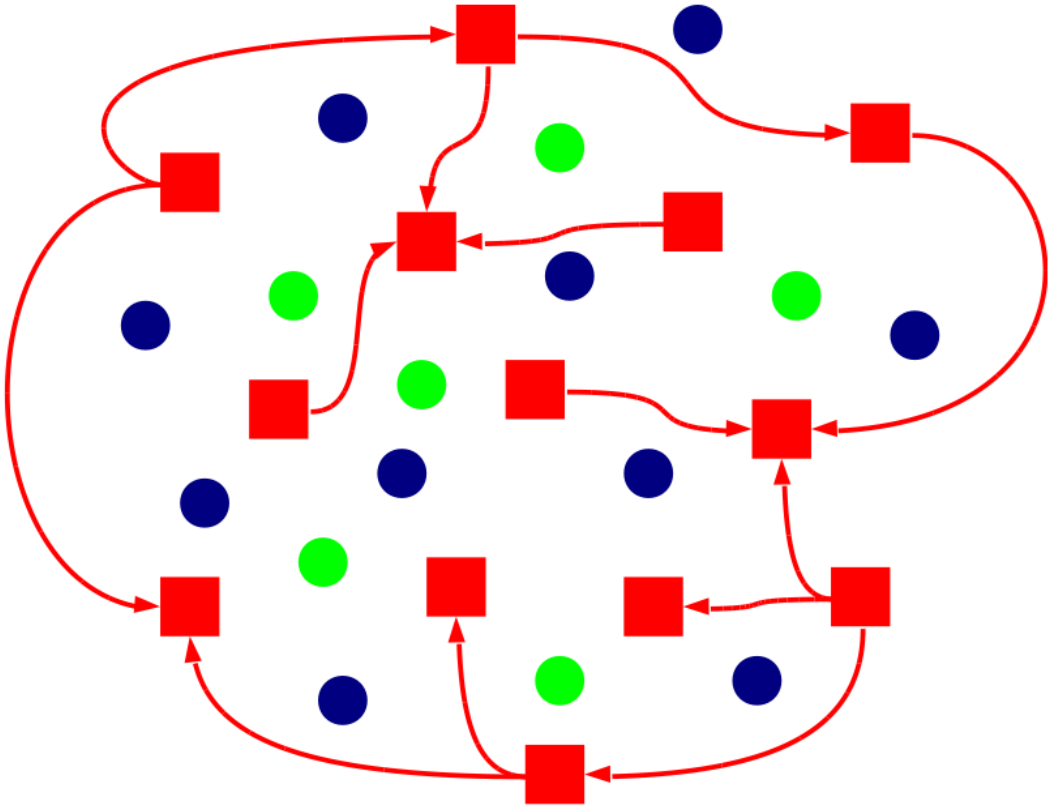
Objectifs

Des Documents, Des Mots, Des Auteurs



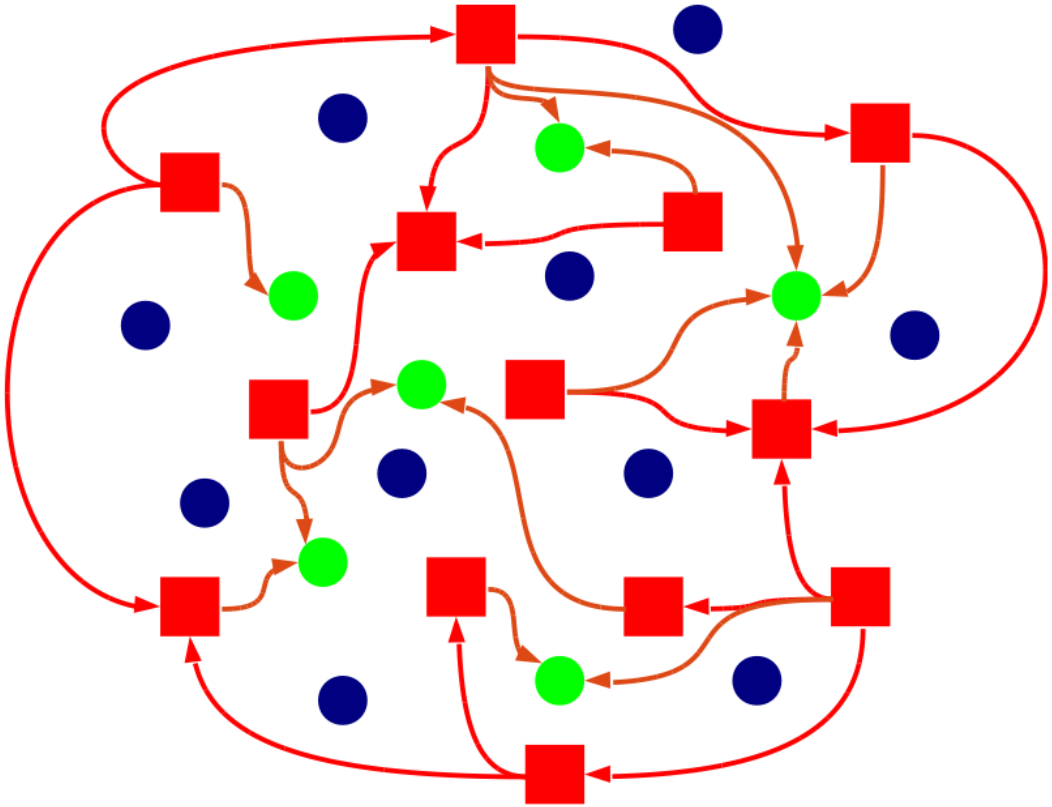
Objectifs

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



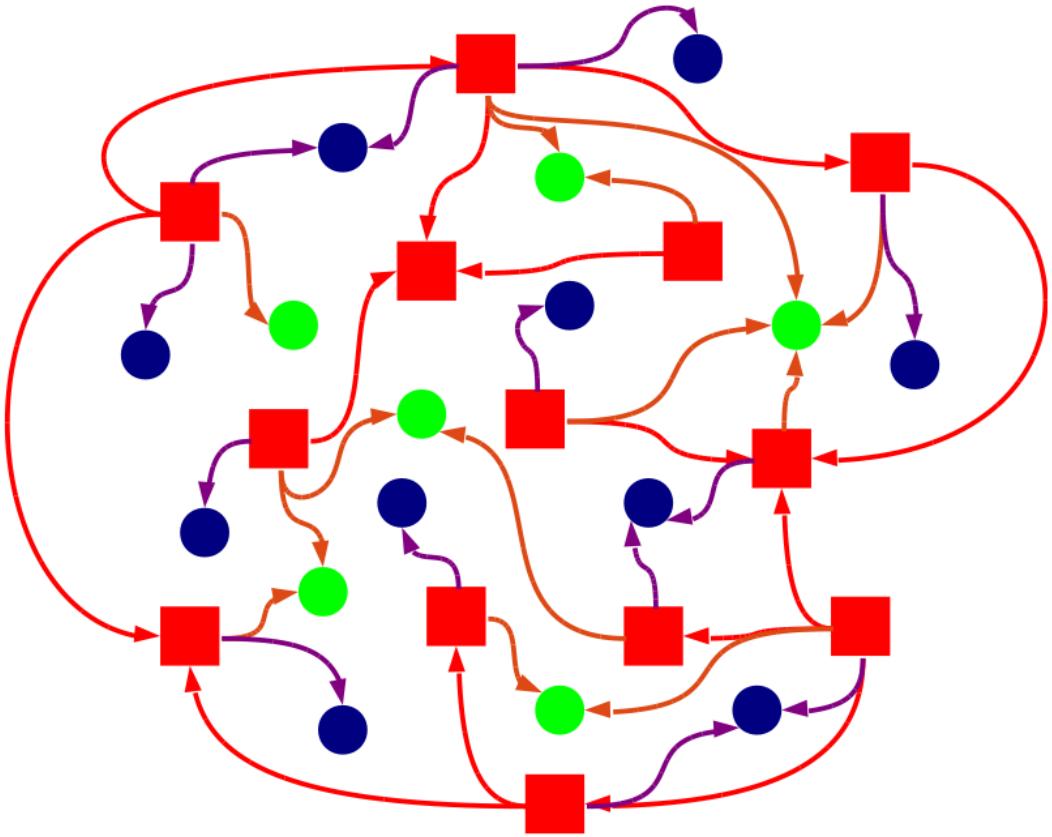
Objectifs

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



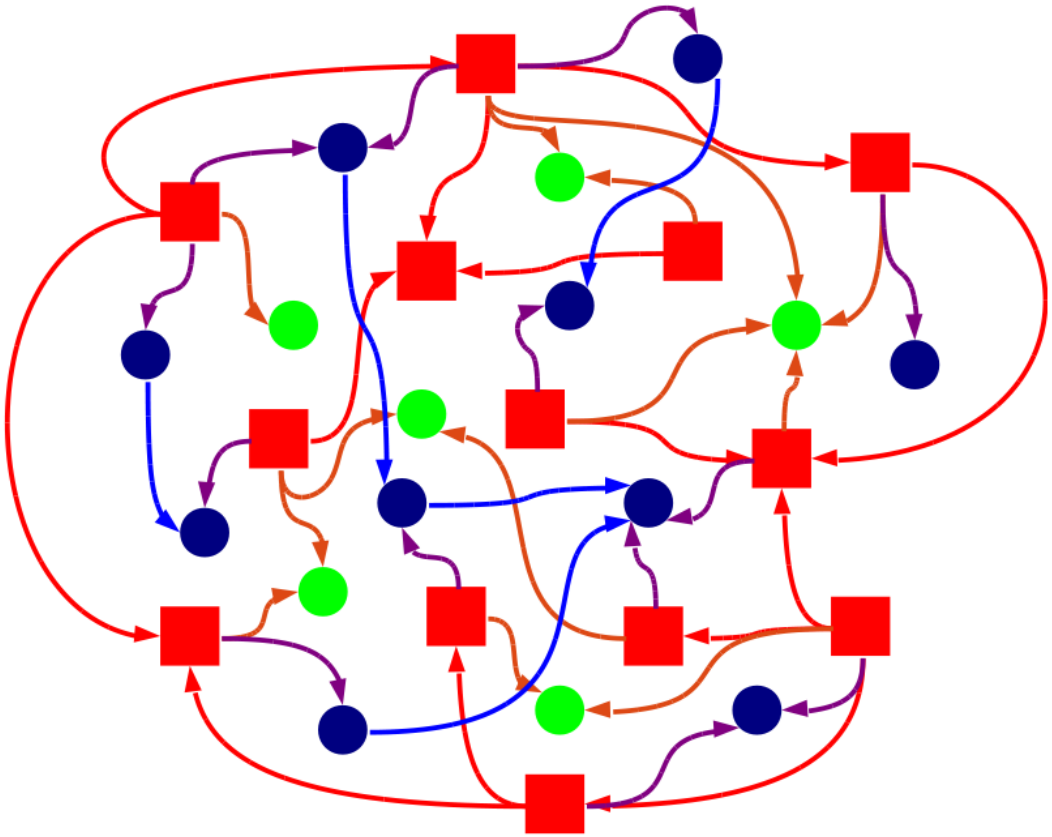
Objectifs

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



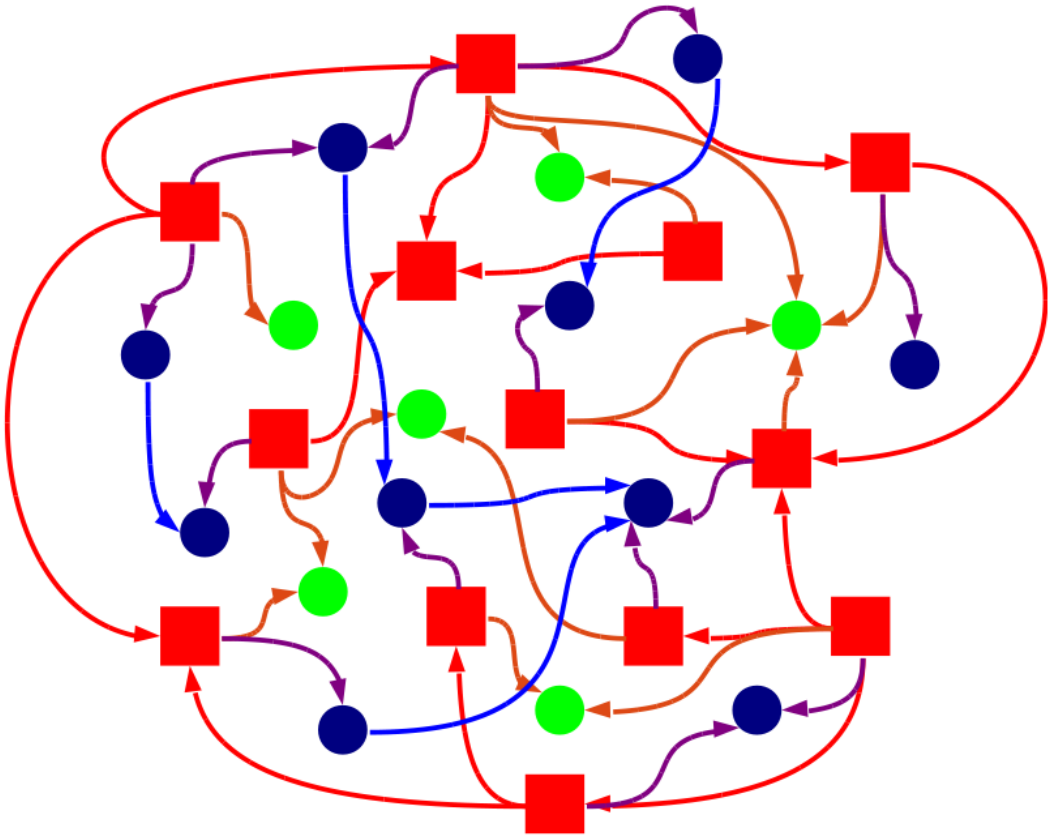
Objectifs

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



Objectifs

Graphe de terrain



Objectifs

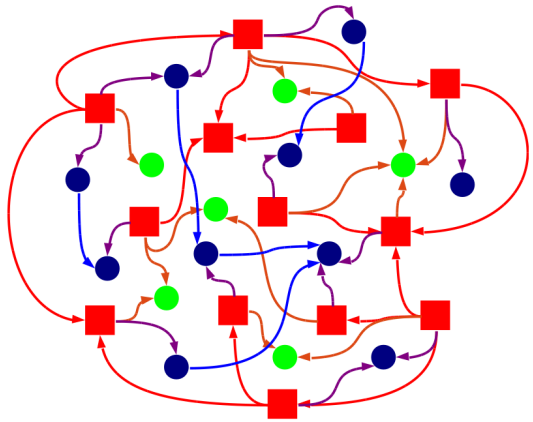
Graphe de terrain

- 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

- 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

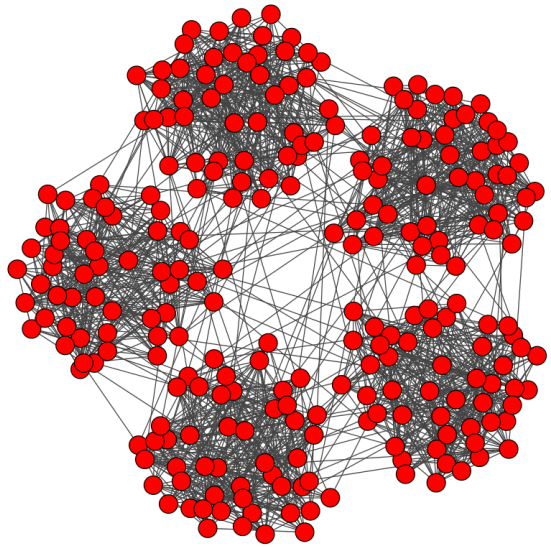
- Faible densité
- Chemins courts
- Distribution des degrés à queue lourde (loi de puissance) $\Rightarrow$ Ranking
- Fort coefficient de clustering : zones denses en arêtes $\Rightarrow$ Sens

Objectifs

Graphe de terrain

Quatre propriétés fondamentales

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



japan search options

30 results for "japan"

Japan - Wikipedia
Find information about the history, politics, geography, economy, and culture of **Japan**.
http://en.wikipedia.org/wiki/Japan

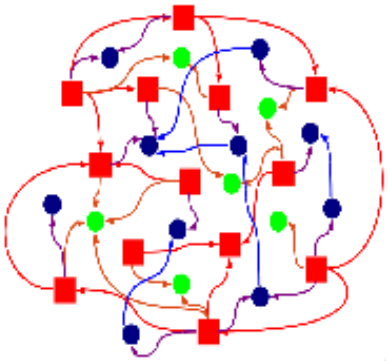
Japan National Tourism Organization |
Japan is situated in northeastern Asia between the North Pacific and the Sea of **Japan**. ... **Japan** consists of four major islands, surrounded by more than 4,000 ...
http://www.jnto.go.jp/eng/

Japan Today
Japan Today is an international news network covering news, politics, business, sports, technology, and more.
http://www.japantoday.com/

VISIT **JAPAN** 2011
It presents **Japan's** history and culture, covers events, festivities, tourism, food and shopping, and provides other kinds of helpful information. Come...
http://www.visitjapan.jp/

japan-guide.com
Everything about modern and traditional **Japan** with emphasis on travel and living related information.
http://www.japan-guide.com/

Official Tourism Guide for **Japan** Travel
Japan National Tourist Organization offers information on transportation, lodging, restaurants, tourist attractions, culture, history, festivals, and ...
http://www.japantravelinfo.com/



japan search options

30 results for "japan"

Japan - Wikipedia
Find information about the history, politics, geography, economy, and culture of **Japan**.
http://en.wikipedia.org/wiki/Japan

Japan National Tourism Organization |
Jap
four
http://

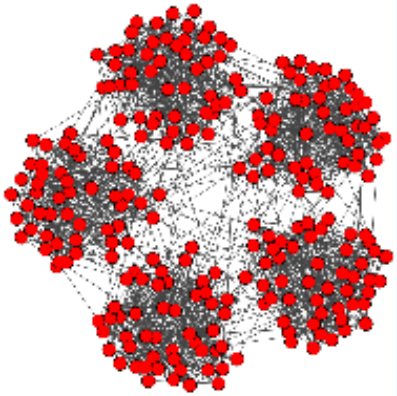
Japan Today
Japan Today is an international news network covering news, politics, business, sports, technology, and more.
http://www.japantoday.com/

VISIT **JAPAN** 2011
It presents **Japan's** history and culture, covers events, festivities, tourism, food and shopping, and provides other kinds of helpful information. Come...
http://www.visitjapan.jp/

Japan-guide.com
Everything about modern and traditional **Japan** with emphasis on travel and living related information.
http://www.japan-guide.com/

Official Tourism Guide for **Japan** Travel
Japan National Tourist Organization offers information on transportation, lodging, restaurants, tourist attractions, culture, history, festivals, and ...
http://www.japantravelinfo.com/

Construire le sous graphe local



japan search options

30 results for "japan"

Japan - Wikipedia
Find information about the history, politics, geography, economy, and culture of **Japan**.
http://en.wikipedia.org/wiki/Japan

Japan National Tourism Organization |
Jap
four ts of
http://

Japan Today
Japan Today is an international news network covering news, politics, business, sports, technology, and more.
http://www.japantoday.com/

VISIT **JAPAN** 2011
It presents **Japan's** history and culture, covers events, festivities, tourism, food and shopping, and provides other kinds of helpful information. Come...
http://www.visitjapan.jp/

japan-guide.com
Everything about modern and traditional **Japan** with emphasis on travel and living related information.
http://www.japan-guide.com/

Official Tourism Guide for **Japan** Travel
Japan National Tourist Organization offers information on transportation, lodging, restaurants, tourist attractions, culture, history, festivals, and ...
http://www.japantravelinfo.com/

Clusteriser ce sous graphe local

japan search options

30 results for "japan"

japanese, china, asia, history, government,

travel, guide, sightseeing, tourism, information,

news, business, network, photo, sports,

airlines, flight, purchase, japan domestic flight, jal group,

nuclear, tsunami, crisis, nuclear power plant, earthquake,

Japan - Wikipedia
Find information about the history, politics, geography, economy, and culture of **Japan**.
http://en.wikipedia.org/wiki/Japan

Japan National Tourism Organization |
Jap
four
http:

Japan
Japan Today is an international news network covering news, politics, business, sports, technology, and more.
http://www.japantoday.com/

VISIT **JAPAN** 2011
It presents **Japan's** history and culture, covers events, festivities, tourism, food and shopping, and provides other kinds of helpful information. Come...
http://www.visitjapan.jp/

Japan-guide.com
Everything about modern and traditional **Japan** with emphasis on travel and living related information.
http://www.japan-guide.com/

Official Tourism Guide for **Japan** Travel
Japan National Tourist Organization offers information on transportation, lodging, restaurants, tourist attractions, culture, history, festivals, and ...
http://www.japantravelinfo.com/

Labeliser les clusters

L'utilisateur est alors informé de la structure

japan search options

30 results for "japan"

-->

japanese, china, asia, history, government,

travel, guide, sightseeing, tourism, information,

news, business, network, photo, sports,

airlines, flight, purchase, japan domestic flight, jal group,

nuclear, tsunami, crisis, nuclear power plant, earthquake,

Japan.org
SENDAI, **Japan** — A strong new earthquake rattled **Japan's** northeast Monday as the government urged more people living near a tsunami-crippled nuclear **...**
http://www.japan.org/

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)
Wo… …nd the
nuc…
http:/…

L'utilisateur peut choisir un focus

Embassy of **Japan**, Addis Ababa, ETHIOPIA
Japan has the experience of a relatively fast recovery from the devastation of the war and has overcome various natural disasters in the past. **...**
http://www.et.emb-japan.go.jp/

Japan: News & Videos about **Japan** - CNN.com
The nuclear crisis wreaks havoc on one of **Japan's** prized exports: green tea. **...** Matador's destination expert on **Japan** lays out the country's avoidable attractions ...
http://topics.cnn.com/topics/Japan

Special report: **Japan's** throwaway nuclear workers | Reuters
FUKUSHIMA, **Japan** (Reuters) - A decade and a half before it blew apart in a hydrogen blast that punctuated the worst nuclear accident since Chernobyl, **...**
http://www.reuters.com/article/2011/06/24/us-japan-nuclear-idUSTRE75N18A20110624

japan search options

30 results for "japan"

-->

japanese, china, asia, history, government,

travel, guide, sightseeing, tourism, information,

news, business, network, photo, sports,

airlines, flight, purchase, japan domestic flight, jal group,

nuclear, tsunami, crisis, nuclear power plant, earthquake,

Japan.org

SENDAI, **Japan** — A strong new earthquake rattled **Japan's** northeast Monday as the government urged more people living near a tsunami-crippled nuclear **...**

http://www.japan.org/

Japan News - Earthquake, Tsunami and Nuclear Crisis (2011)

Wo… …nd the

nuc…

http…

Em…

Jap…

vari…

http…

Japa…

The nuclear crisis wreaks havoc on one of **Japan's** prized exports: green tea. **...** Matador's destination expert on **Japan** lays out the country's avoidable attractions ...

http://topics.cnn.com/topics/Japan

Special report: **Japan's** throwaway nuclear workers | Reuters

FUKUSHIMA, **Japan** (Reuters) - A decade and a half before it blew apart in a hydrogen blast that punctuated the worst nuclear accident since Chernobyl, **...**

http://www.reuters.com/article/2011/06/24/us-japan-nuclear-idUSTRE75N18A20110624

L'utilisateur peut choisir un focus ou optimiser sa requête

...

...

...

Etat courant au 19 mars 2015

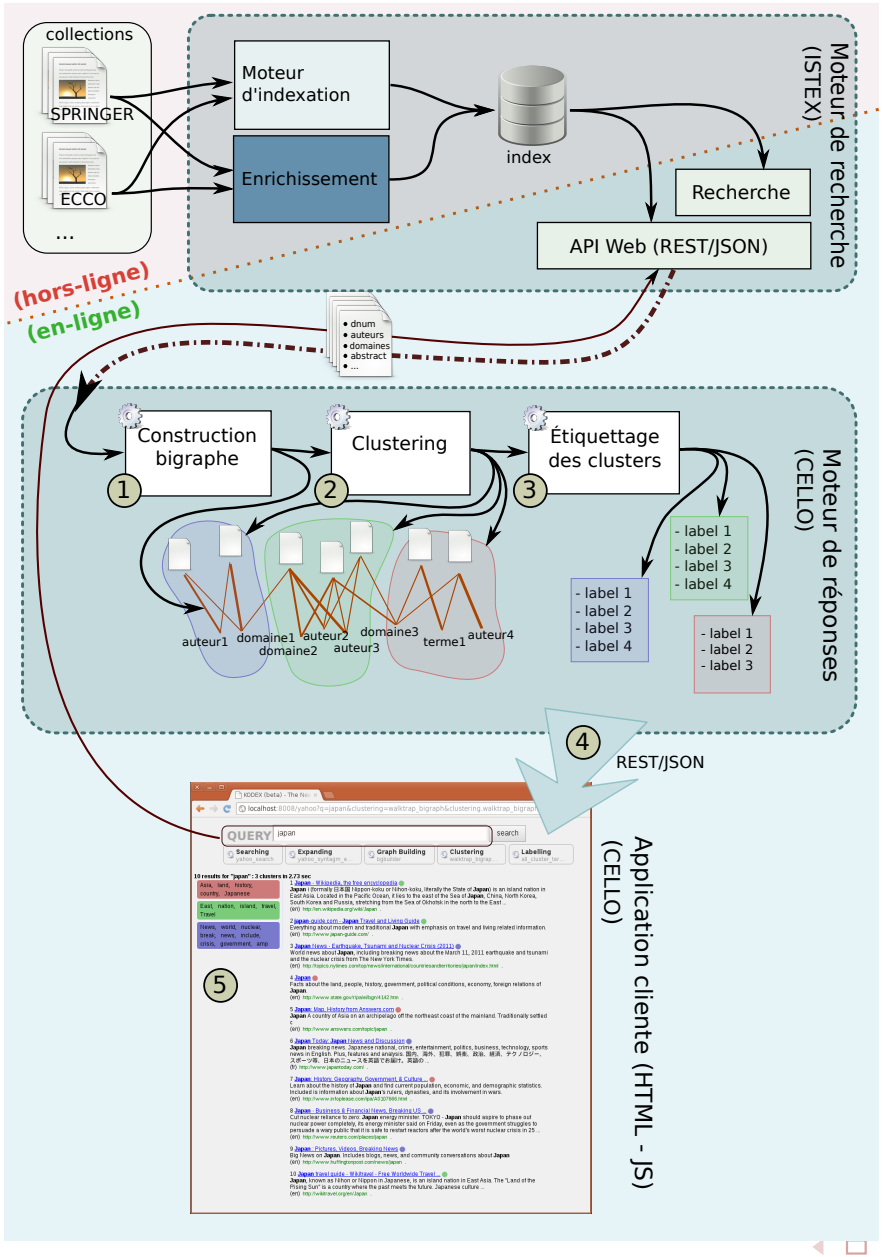
Etat courant : Résumé

- Moteur de réponse en place,
- Systeme de sauvegarde/annotation en place,

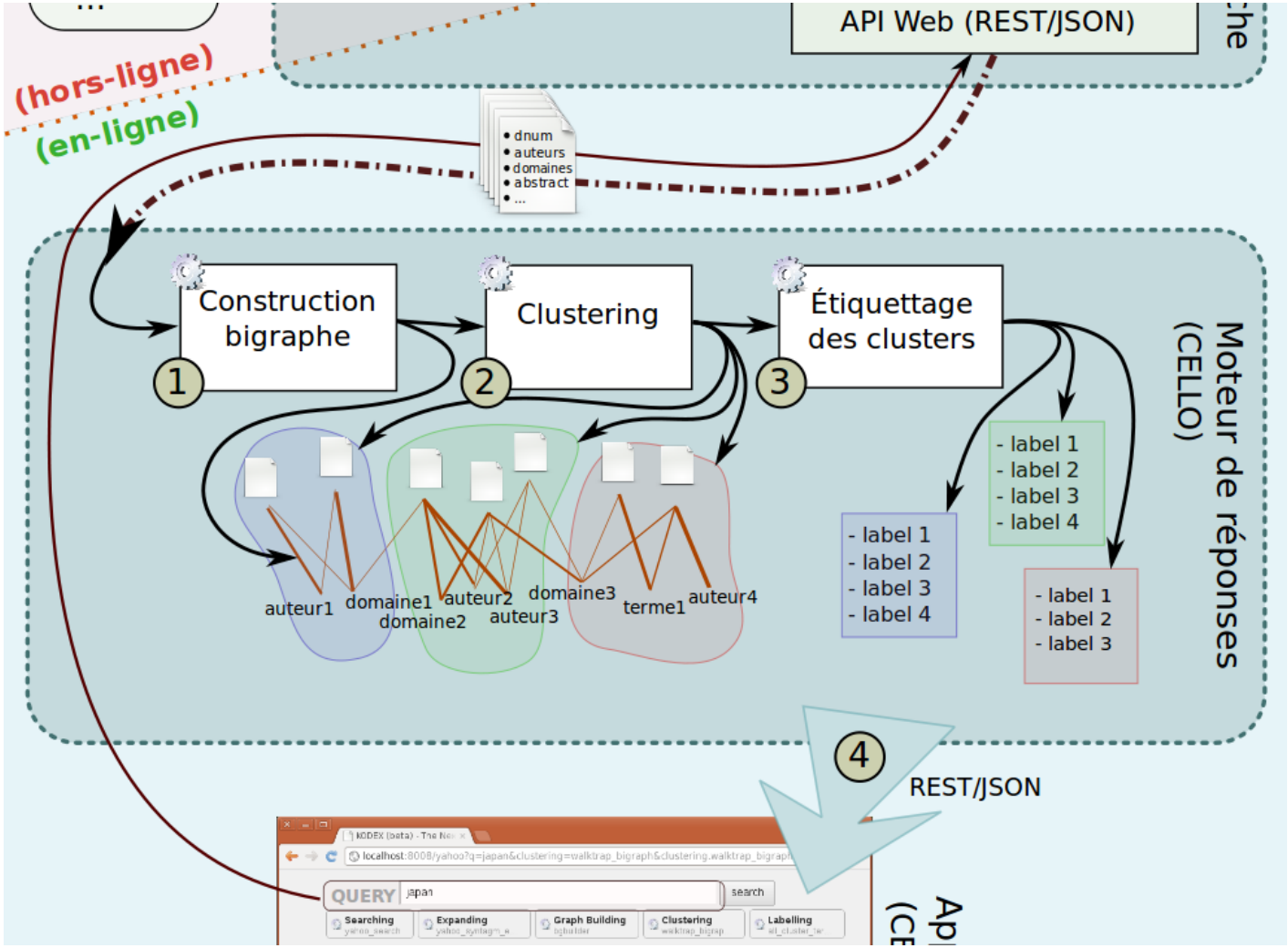
Etat courant : Résumé

- Moteur de réponse en place,
  - chaine de traitement,
  - quelques détails techniques...
  - données prise en compte aujourd'hui,
- Systeme de sauvegarde/annotation en place,
  - méthodologie de mise au point;

CILLEX : la chaine de traiment



CILLEX : la chaine de traiment



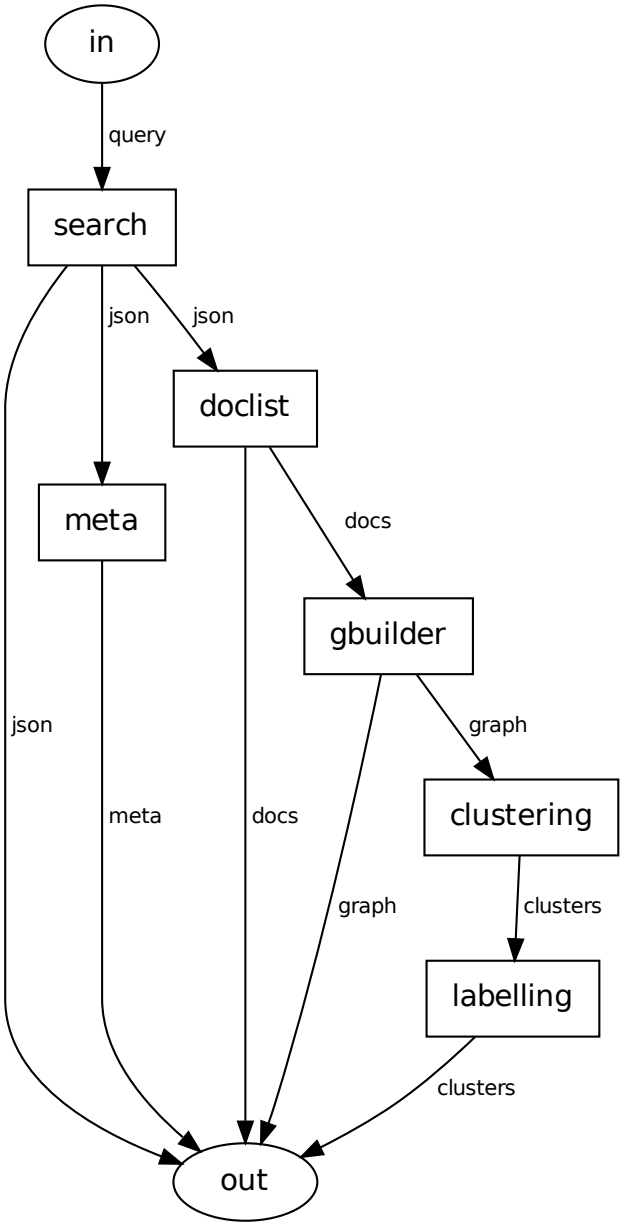
CILLEX : sous le capot

- serveur *Python* fournit une API REST/JSON
- application cliente *Javascript* (*backbone*, *semantic-ui*)

Chaine de traitement coté serveur : framework ***reliure***

- système modulable de chaines de composants,
- découverte des composants et des options,
- release opensource à venir (LGPL);

CILLEX : chaine de traitement (détail)



Etat courant au 19 mars 2015

Données utilisées pour le graphe local (au 19 mars 2015)

- mots des *abstract*
- mots du *title*
- *subject*
- *subject serie*
- *authors*

Méthodologie de mise au point : le constat !

- chaine de traitement complexe,
- combinatoire de paramétrage important,

=> Besoin de tester automatiquement différentes configurations !

Campagnes évaluations classiques :

- *dataset* existant :
  - pas les mêmes données,
  - mal adaptées à ce que l'on cheche à évaluer;
- jeu d'éval sur mesure :
  - coute cher !

Constat : avec ou sans évaluation...

on passe beaucoup de temps à annalyser à la main des résultats !

... mais ce travail n'est pas capitalisé.

Méthodologie de mise au point : non-régression ?

Changement de point de vue :

$$\text{évaluation} \longrightarrow \text{non-régression}$$

- se baser sur les résultats du systeme :
  - résultats mauvais $\Rightarrow$: (
  - résultats bon $\Rightarrow$ corrections à la main à faible coût
  - résultats très bon $\Rightarrow$ on enregistre juste !
- itératif : le jeu d'annotation n'est pas figé !

On à donc construit :

- Systeme d'annotation intégré à l'interface CILLEX (demo) :
  - enregistrement d'un résultat,
  - modification à la main le clustering,
  - rechargement/modification d'une précédente annotation;
- Données exploitables en bash pour des expés;

Développement à venir

Développement à venir

Le point faible : qualité du graphe !

- *sparse* (subject, auteurs),
- meta-données absentes (auteurs, résumés)
- non-homogénéité (subject)
- bruit (termes des titres et résumés)

Pistes d'améliorations ?

- Traitement en ligne des titres et résumés (amélioration),
- Données des projets *enrichissement du texte plein* :
  - Entités nommées,
  - Termes et variantes,
  - Références bibliographiques,

Autres pistes envisageables :

- Enrichissement (en ligne) avec une ressource externe (*RLF*?),
- Folksonomies alimentées par les usagers d'ISTEX (... ?)

Développement à venir : en pratique !

1. Amélioration du système :
   - limité sur une sous-collection,
   - ref. bib. ? entités nommés ? traitement en ligne titres et résumés ?
2. Annotations de résultats sur un 1er (petit) jeu de requêtes,

Puis on itère...

A chaque itération:

- augmenter le nombre de recherche annotés,
- étendre la participation a un public plus large,
- améliorer le système...

(Autre point : log des recherches sur l'API ISTEX ?)

Merci