3ST
Surligneur Sémantique de Textes Scientifiques

Séminaire technique
« Chantiers d'usage » d'ISTEX
26 avril 2016

Coordinateur du projet : Fabrice MUHLENBACH
courriel : fabrice.muhlenbach@univ-st-etienne.fr

Membres du projet

Laboratoire Hubert Curien, UMR CNRS 5516 (UJM)
Équipe *Connected Intelligence*

Fabrice Muhlenbach → fouille de données, fouille de textes, systèmes de recommandation, représentation des connaissances, transfert de connaissances

Pierre Maret → communautés virtuelles, modélisation de connaissances, *web* sémantique

Dennis Diefenbach → reconnaissance des entités nommées, étiquetage morpho-syntaxique, *parsers*, appariement de phrases, désambiguïsation

Membres du projet

Laboratoire ERIC, EA 3083 (Univ. Lyon 2 et Lyon 1)

Djamel A. Zighed → apprentissage automatique, fouille de texte, désambiguïsation d'auteurs

Hussein Al-Natsheh → fouille de données, fouille de textes, humanités numériques

Lucie Martinet → théorie des graphes, réseaux sociaux, systèmes complexes, sciences du langage, fouille de données

Fabien Rico → fouille de données, fouille de texte, apprentissage automatique, recherche opérationnelle

Contexte : l'**infonomie**

Infonomie ?

- projet d'une nouvelle science s'intéressant à la manière de produire, de diffuser et d'utiliser des contenus immatériels : données, informations, connaissances, savoir-faire…
- étude des conditions dans lesquelles l'information, en tant qu'objet et produit socio-économique, scientifique et culturel, se crée, se transforme, se diffuse et agit à son tour pour créer d'autres informations

  → question de la valeur économique de l'information

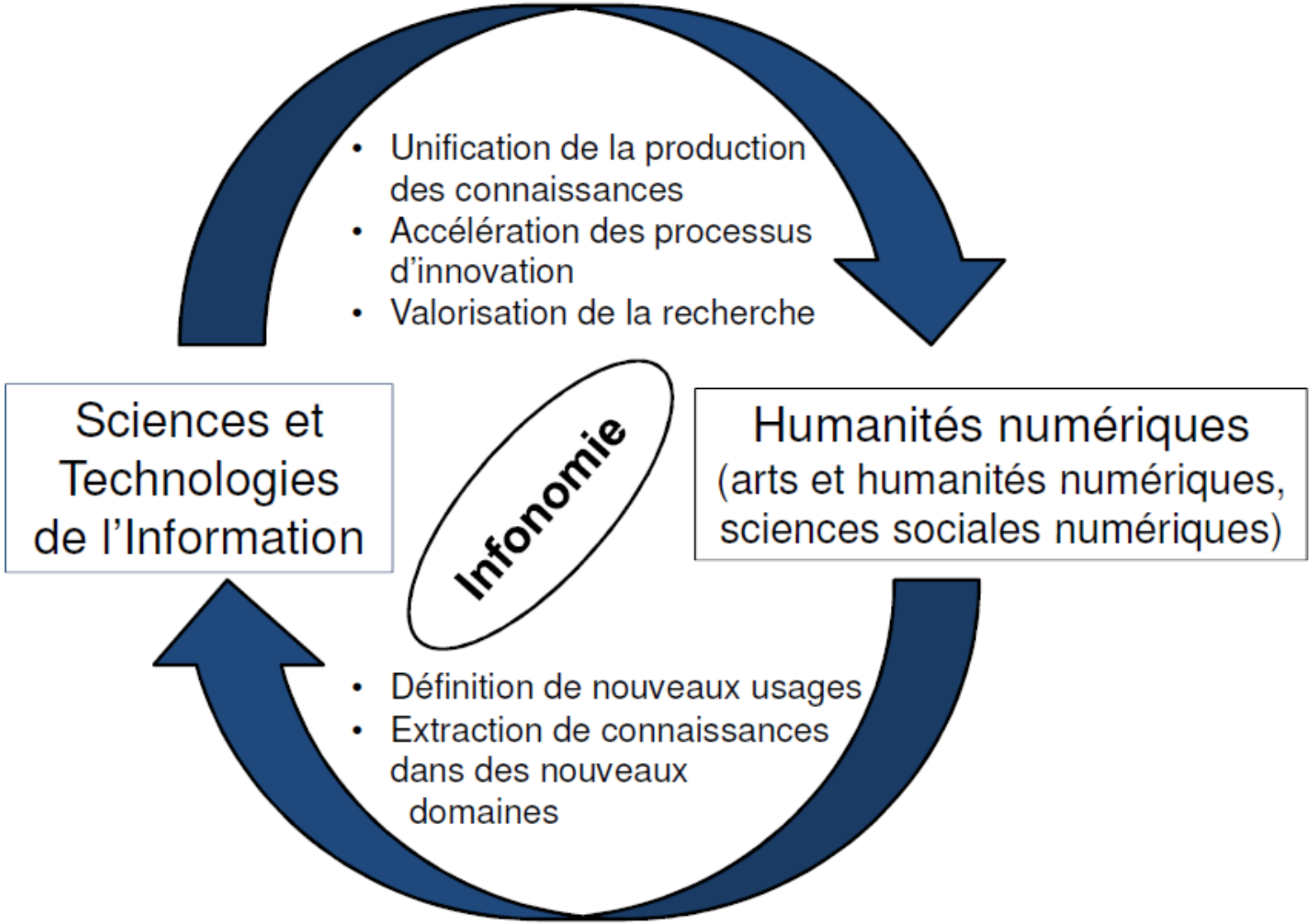
  → question de la mesure du sens

Contexte : l'**infonomie**

Infonomie ?

- l'**infonomie** se situe à la confluence des Arts et Humanités numériques, des Sciences sociales numériques et des Sciences et Technologies de l'Information
- **problème** : les disciplines scientifiques, et particulièrement en sciences humaines et sociales, mais aussi dans le domaine des arts et des lettres, sont le plus souvent déconnectées des autres disciplines
- pourtant : richesse des collaborations pluridisciplinaires p. ex., les linguistes travaillant sur les origines du langage ont bénéficié des apports des travaux réalisés par les archéologues, anthropologues ou biologistes (génétique)

Contexte : l'**infonomie**



Contexte : l'**infonomie**

Projets

- Plan d'Avenir Lyon Saint-Étienne sur l'infonomie (2014) : développement d'un réseau intéressé par les humanités numériques au sein des différents établissements de l'Université de Lyon
- Appel à projet Région Rhône-Alpes (2015) :
  ARC6 - TIC et Usages Informatiques Innovants
  « Mondes numériques pour l'humain et la société : conception, comportements et usages »
  Projet de thèse :
  Investigation et exploitation de corpus textuels scientifiques
- ISTEX (2016) : projet **3ST**

Objectifs

Constat

- accès des chercheurs à des masses d'informations (bibliothèques numériques d'articles scientifiques en ligne)
- exploration des documents scientifiques limitée à la communauté d'appartenance de chaque chercheur

Proposition

- extension de l'exploration bibliographique au-delà de la communauté d'appartenance
- → contexte pluri- et trans-disciplinaire

Objectifs

Problèmes

- grande taille des bibliothèques numériques
- hétérogénéité des données
- complexité du langage naturel
- limitations cognitives et manque de temps :
  → incapacité à pouvoir embrasser des concepts issus :
    - d'articles anciens pourtant pertinents (focalisation sur l'axe diachronique limitée aux articles scientifiques les plus récents)
    - d'articles venant de disciplines complémentaires (focalisation sur l'axe synchronique limitée à la communauté scientifique d'appartenance)

Objectifs

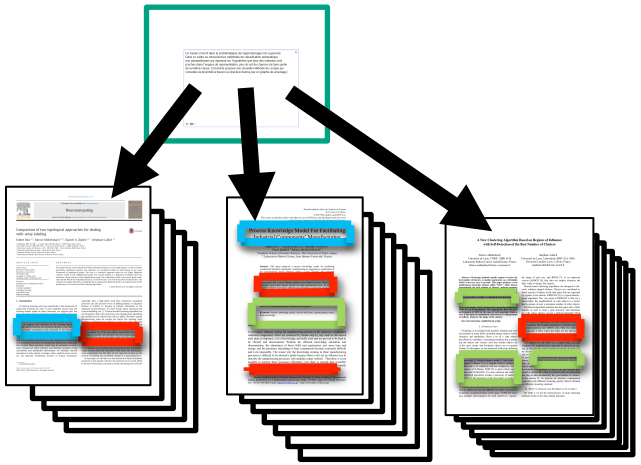
Conséquence

→ le saut **quantitatif** en masse d'information apportée par les bibliothèques numériques ne se traduit pas vraiment en saut **qualitatif** pour le chercheur qui souhaite exploiter ces documents

Objectifs

Proposition pratique

- projet de recherche appliquée
- construction d’un outil de lecture assistée par ordinateur
- surlignage sémantique des textes scientifiques



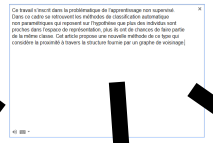
Objectifs

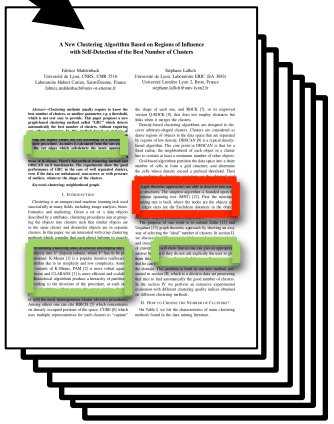
Le surligneur sémantique 3ST

en **entrée** : une requête composée d'une ou plusieurs phrases cibles portant sur un sujet d'intérêt de l'utilisateur

en **sortie** : présentation d'un ensemble d'articles du même domaine d'appartenance que cet utilisateur mais aussi d'autres domaines scientifiques

aide apportée à travers un système de **surlignage**





Objectifs

Le surligneur sémantique 3ST

- **moyen** : spécification de mesures de similarité entre noyaux de phrases, agrégation et extension aux textes entiers
- **but** : identification et visualisation des passages d'un texte qui expriment les idées proches d'une requête
- **principe** : textes cibles surlignés selon les proximités trouvées avec chaque phrase du texte de la requête
  - jeu de couleurs → rôle dans la signification
  - intensité de couleur → niveau de similarité
  - passages surlignés ou non surlignés

  → index sémantique permettant d'accélérer les recherches d'information dans les contenus de grands corpus

Objectifs

Comment estimer la correspondance sémantique ?

- Soient les phrases suivantes :

1. un père sur une chaise à bascule sommeille avec son bébé sur les genoux
2. un homme avec son fils se repose sur un rocking-chair

- Comment établir la relation sémantique entre ces phrases ?

→ pas de termes communs

→ pourtant les deux phrases décrivent la même image

Objectifs

Comment estimer la correspondance sémantique ?

Notre approche :

- transformation d'une paire de phrases en un ensemble de caractéristiques appariées :

1. application de l'étiquetage morpho-syntaxique (*Part-of-Speech tagging*)
2. transformation de chaque *tag* en représentation vectorielle de mots (*Word Embedding*)

- utilisation de la représentation vectorielle de mots pour :

1. aligner les mots tagués par étiquetage morpho-syntaxique
2. calculer les distances vectorielles entre mots alignés

Objectifs

Comment estimer la correspondance sémantique ?

Notre approche :

- agrégation des distances de mots par apprentissage supervisé :

1. utilisation de milliers de paires de phrases étiquetées avec un score de relation sémantique (de 0 à 5) [*SemEval*]
2. ces paires étiquetées manuellement proviennent de différents domaines comme des grands titres de journaux, des forums de type questions-réponses, des annotations…
3. entraînement d'un modèle de régression pour apprendre comment estimer le score donné par une paire de phrases

Objectifs

Comment estimer la correspondance sémantique ?

- études de mesures de similarité entre textes

→ base de tests : *Exercices de Style* de Raymond Queneau

- même histoire racontée de 99 façons différentes (« Litotes », « Rétrograde », « Surprises », « Rêve », « Hésitations »)
- textes de l'essai de Queneau mélangés avec un corpus d'histoires courtes non pertinentes
- comparaison de différentes méthodes (mesures de similarité / indices de RI)

Objectifs

Comment estimer la correspondance sémantique ?

- modèle suffisamment générique pour être appliqué sur des documents très variés (textes littéraires, articles scientifiques)

1. "First, we build a geometrical connected graph like Toussaint's Relative Neighbourhood Graph on all examples of the learning set."
2. "At first, they build a multidimensional neighbourhood structure by using some particular models like the Toussaint's Relative Neighbourhood Graph (Toussaint 1980)."

- Sent_relatedness score: '3.85 out of 5.0'

Objectifs

Comment estimer la correspondance sémantique ?

First/ADV ,/PUNC we/PRON build/VERB a/DET geometrical_n connected graph/NOUN like Toussaint's Relative Neighbourhood/NOUN Graph/NOUN on all examples/NOUN of the learning/NOUN set/VERB.

At first ADV ,/PUNC they/PRON build/VERB a/DET multidimensional neighbourhood/NOUN structure/NOUN by using/VERB some particular models/NOUN like the Toussaint's Relative Neighbourhood/NOUN Graph/PROP (Toussaint/PROP 1980).

- appariement des annotations morpho-syntaxiques des phrases sur la base de la plus grande similarité cosinus des représentations vectorielles de mots :

  we/PRON, they/PRON: 0.8023 et set/VERB, using/VERB: 0.3978
- agrégation moyenne de tous les mots appariés
- utilisation de la valeur d'étiquetage morpho-syntaxique agrégée comme caractéristique du modèle de régression
- obtention d'un score de correspondance sémantique

Besoins

Accès à des corpus de données scientifiques

- nécessité d'accéder à des données non limitées
  → à une discipline scientifique
  → à un éditeur scientifique
- nécessité de traiter des gros volumes de données
- nécessité de travailler avec des documents transformés : pas de fichiers PDF bruts, mais du plein texte ou des documents XML (enrichissement par des méta-données)
- ISTEX : 16 millions de textes scientifiques

→ en pratique : journée *hackathon ISTEX* prévue demain

Besoins

Ressources matérielles pour le passage à l'échelle

- nécessité d'une connexion rapide avec les données ISTEX (a priori OK avec une API efficace)
- nécessité de disposer d'un espace de mémoire conséquent (mémoire de travail / mémoire de stockage)
- nécessité de puissance de calcul

→ serveur dédié virtuel ?

Besoins

Ressources financières pour valoriser le travail

- mettre en place des réunions régulières sur les sites de Lyon et Saint-Etienne
- extension du réseau par d'autres collaborations (locales sur l'UdL, régionales, nationales, internationales…)
- crédits de fonctionnement pour missions / séminaires / conférences (publications)

Besoins

Renforcement de l'équipe

- recrutement d'une post-doctorante : Lucie Martinet
- compétences scientifiques et techniques
- études appliquées au langage naturel
- expertise en étude de graphes et réseaux sociaux
- → possibilité de traiter deux types de problèmes
    - focalisation sur l'axe **diachronique** :
      étude de la dynamique des réseaux
    - focalisation sur l'axe **synchronique** :
      trouver d'autres communautés liées à une communauté donnée (recherche de complémentarité)

Besoins : apports au projet

Aspects sémantiques

- représentation en langage naturel (techniques de TALN)
- matrices de co-occurrences réduites
  → identifier les contextes les plus pertinents
  → représentation en vecteurs (*Word Embedding*) et apprentissage supervisé
- mesures de similarité entre documents
  → amélioration des mesures de similarité sémantique entre mots
  → élargissement aux phrases
  → extension aux documents

Besoins : apports au projet

Aspects réseaux sociaux

- recherche de communautés de documents (communautés d'auteurs / auteurs d'une même communauté)
  - → transfert de technologie : quand des auteurs co-publiants appartiennent à des communautés différentes
  - → réseaux pondérés basés sur la classification des documents par des mesures de similarité
  - → communautés hiérarchiques / communautés recouvrantes de documents par apprentissage non supervisé
  - → étude de l'évolution des communautés au cours du temps (aspect dynamique)

Résultats envisagés

Construction du surligneur sémantique 3ST

- résultats attendus en sortie

  → liste des documents pertinents par identification thématique
  → surlignage des éléments à l'intérieur des documents, identification de concepts pertinents pour la requête

- améliorations possibles

  → distinguer visuellement des éléments appartenant à une même communauté de ceux appartenant à des communautés voisines ou complémentaires
  → identifier les communautés de co-citation afin de mieux mettre en avant de nouvelles communautés détectées

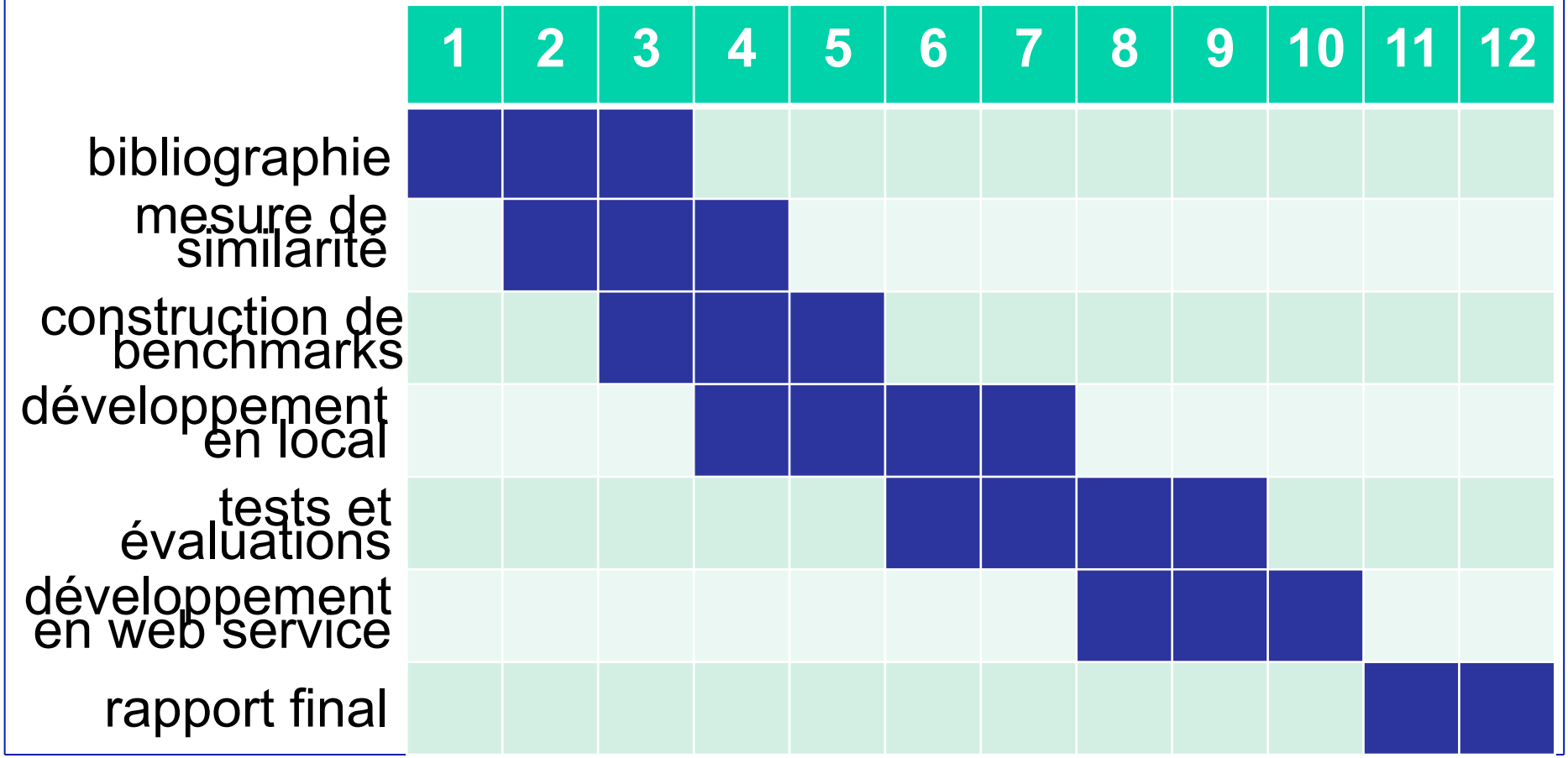
Résultats envisagés

Apports théoriques (complément du surligneur 3ST)

- fouille de données
- apprentissage automatique
- représentation des connaissances
- système de recommandation
- *web* sémantique
- communautés virtuelles

Résultats envisagés

Calendrier (à partir du 1er avril 2016)



Résultats envisagés

Conclusion et problématiques de recherche

- représentation des mots et des phrases
- matrices de co-occurrence
- réduction de la dimension
- représentation vectorisée condensée
- conservation de l'information sémantique
- avancement sur la thématique de l'infonomie

UNIVERSITÉ DE LYON

LABORATOIRE HUBERT CURIEN

UMR • CNRS • 5516 • SAINT-ETIENNE

CONNECTED INTELLIGENCE

eric

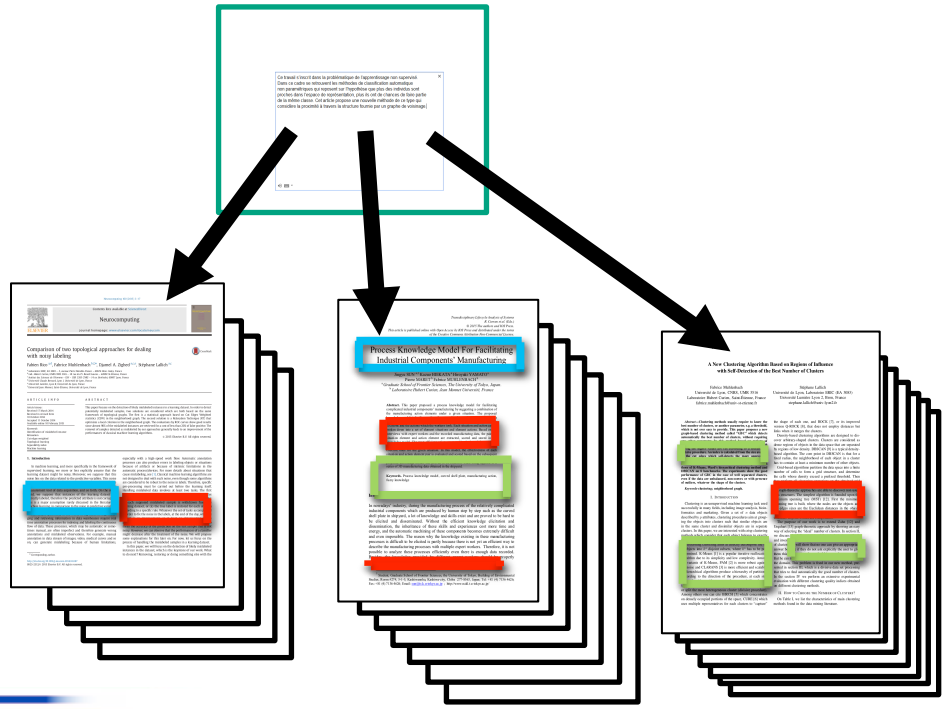
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2

UNIVERSITÉ DE LYON

3ST

Surligneur Sémantique de Textes Scientifiques



Coordinateur du projet : Fabrice MUHLENBACH
courriel : fabrice.muhlenbach@univ-st-etienne.fr