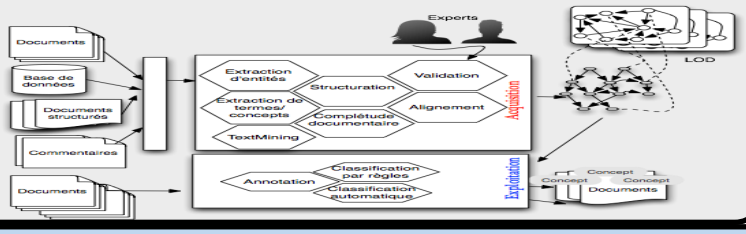
Plateformes sémantiques : une étude empirique

Fatma Chamekh
Léonard de Vinci Pôle Universitaire, Research Center, 92916, Paris La Défence. France

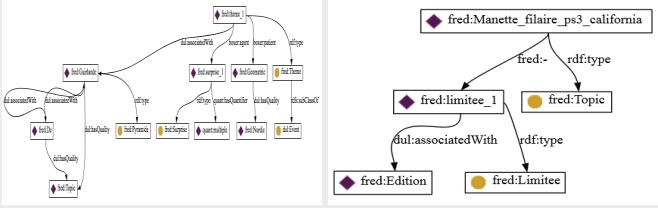
Contexte : Le projet PCU (plate-forme de connaissances unifiées) vise à développer une plateforme open source de valorisation des données d'entreprise. Un scénario d'application est celui de moteur de recherche e-commerce (Magento Elastic Suite). Cette solution utilise un thesaurus pour optimiser la recherche d'information. Ce thesaurus est plus proche d'une liste de mots et ne suit pas un standard de formalisation .

Proposition : Un framework sémantique pour créer une ressource sémantiques à partir des fiches produits. Ce framework est modulaire, deux phases distinctes (exploitation, acquisition), validation par des experts et liens avec des ressources du LOD.



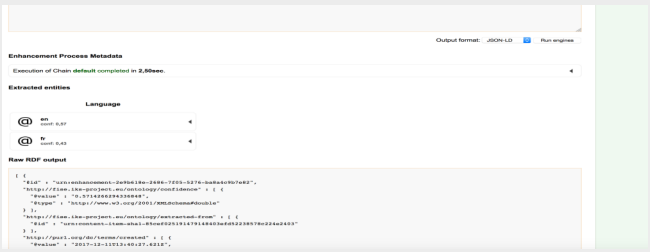
Hypothèse de travail : Repérer les plateformes / les composants pour les réutiliser dans le framework sémantique PCU.
Les critères de choix : Licence open source, les technologies utilisées (nlp, machine learning, ..), les tâches (extraction, annotation ..)

FRED : génére un graphe RDF à partir d'un texte et les ressources de DBpedia en utilisant des technologies de traitement automatique des langues.

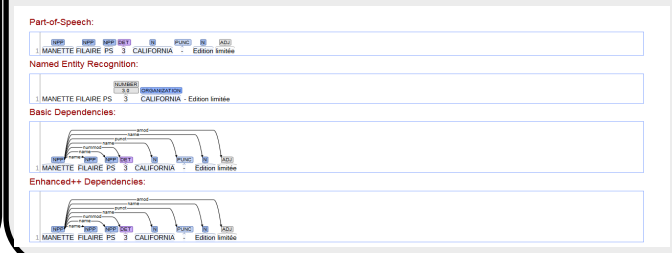


DBpedia Spothlight : annote automatiquement des textes à partir des entités de DBpedia .

Apache Stanbol : est un framework de gestion de contenue sémantique. Il est composé de plusieurs composantes.



StanfordNLP : est un ensemble d'outils pour le traitement automatique des langues. Il utilise les technologies tel que : deep learning, statistical NLP….



Comparaison :

	FRED	Apache Stanbol	DBpediaSpothlight	Stanford NLP
Langue française	Utilise la langue anglaise comme une langue pivot	Annotation Opencalais et DBpedia.	Annote des texte en français	Un modèle pour le français
Annotation	Non	Oui	Oui	Non
Acquisition	Lien avec DBpedia	Oui mais pas de modèle nlp pour le français	Non	Oui
Structure modulaire	Non	oui	Non	Oui
Complémentarité entre les outils	Non	Stanford nlp	Intégré dans Apache Stanbol	Apache Stanbol

Conclusion et perspectives :

- Le Choix des plateformes à réutiliser : Apache Standbol et Stanford NLP.
- Pour choisir les modules de Stanford nlp à intégrer : une étude empirique pour l'acquisition (par apprentissage, terminologique, conceptuel, sémantique).
- Les données : les fiches produits et les données entreprise (cv, rapport…).
- Les fonctionnalités à évaluer : extraction de termes (forme canonique, variantes lexicales/terminologiques/sémantiques).