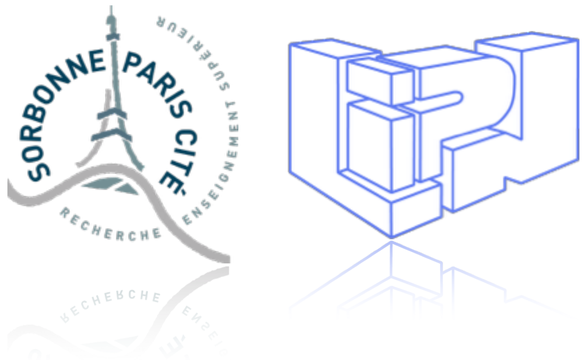


Classification of Keyphrases from Scientific Publications using WordNet and Word Embeddings

Davide Buscaldi, Simon David Hernandez-Perez, Thierry Charnois
LIPN - Université Paris 13
Equipe Représentation de Connaissances et Langage Naturel (RCLN)

davide.buscaldi@lipn.univ-paris13.fr

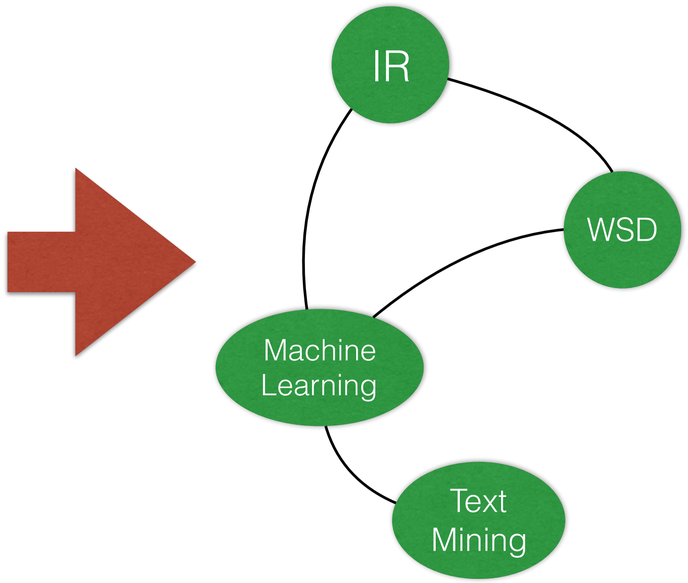
Atelier VADOR, Toulouse
31/05/2017

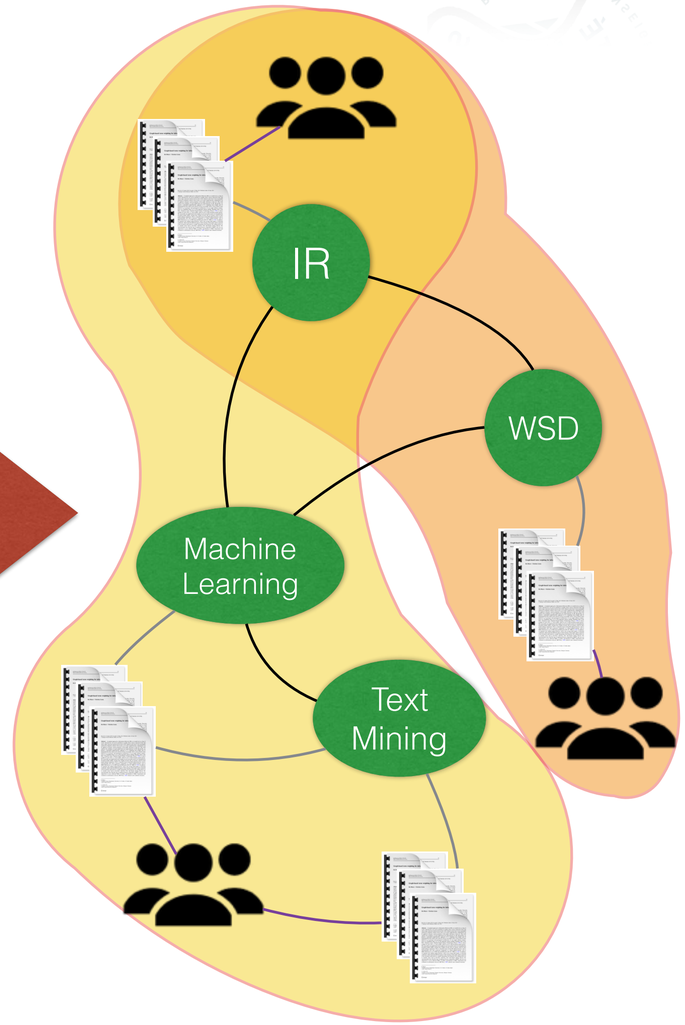
Contexte

- Analyse automatique (*machine reading*) de textes scientifiques
  - En gros: transformation de textes en graphes concepts-relations (+auteurs, articles, etc.)
- Buts:
  - Améliorer l’accès à la littérature scientifique
  - Permettre d’identifier plus facilement les experts de domaine
  - Résumer le contenu des publications scientifiques
    - Etats de l’art automatiques
    - +Dimension temporelle: Construction de *phylomémies* (Chavalarias et Cointet, 2013)

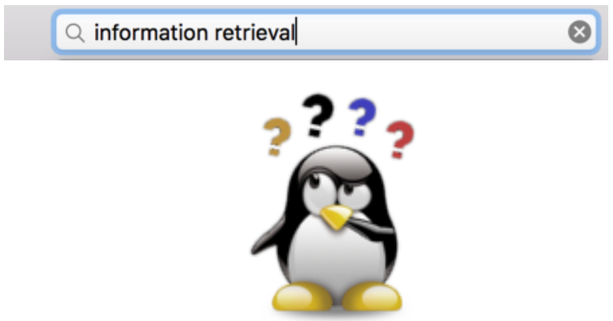
Exemple: recherche d'experts

Who can review it?



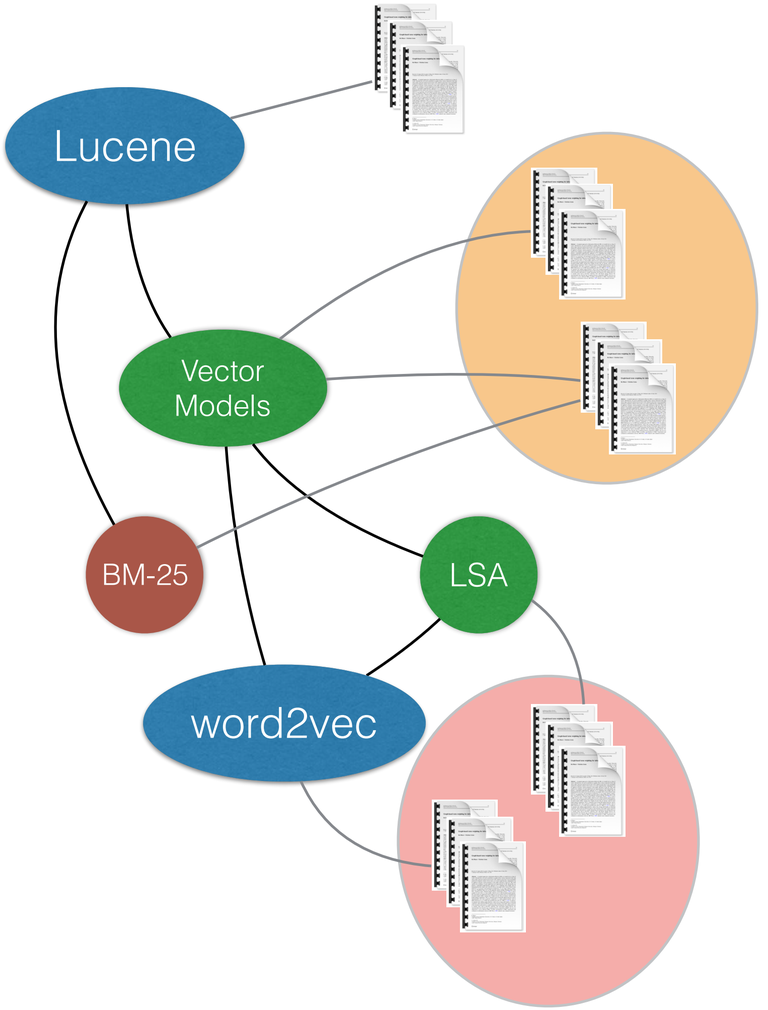


Exemple: états de l'art



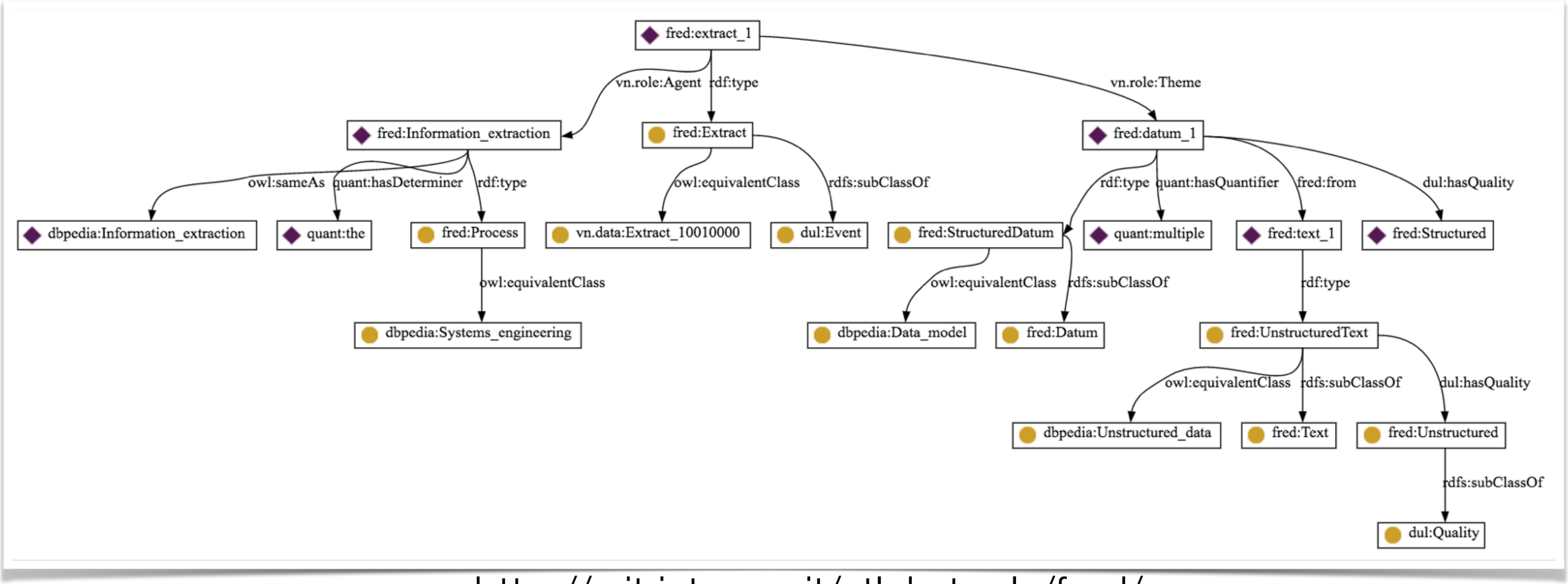
Find the state of the art
for a given subject/topic

Semantic graph with
key concepts and papers



FRED + textes scientifiques

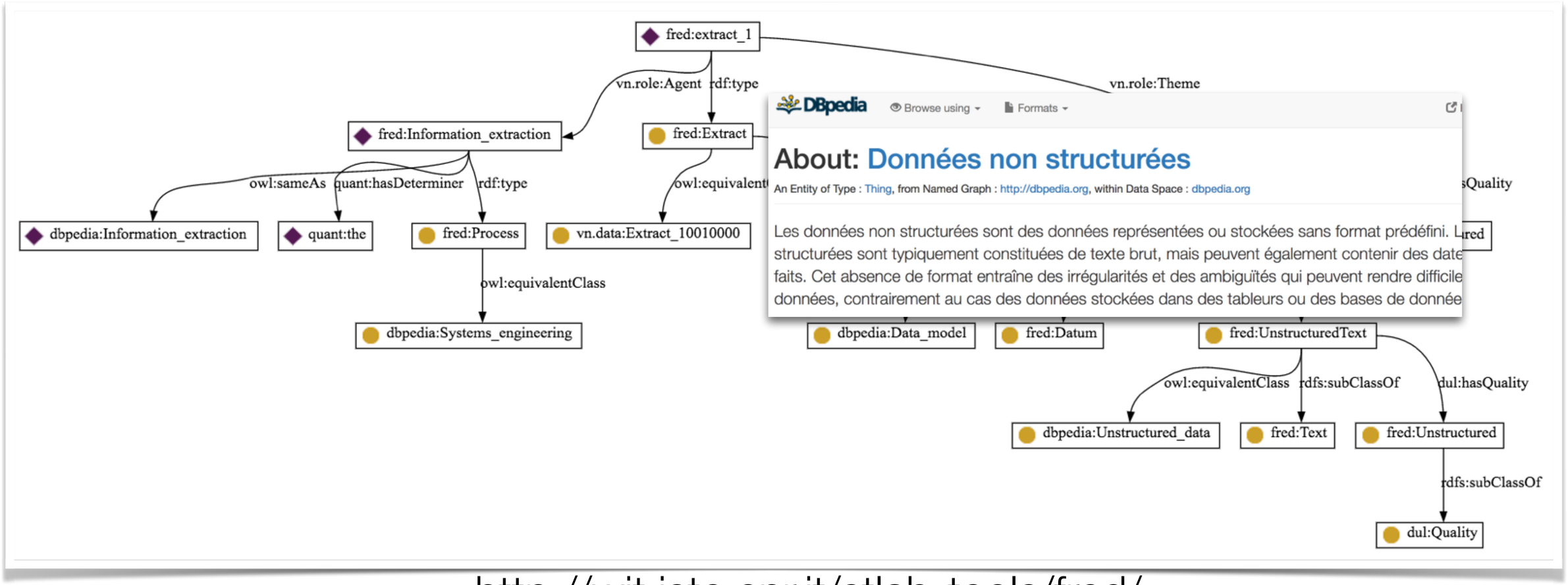
- “Information extraction is the process of extracting structured data from unstructured text”



http://wit.istc.cnr.it/stlab-tools/fred/

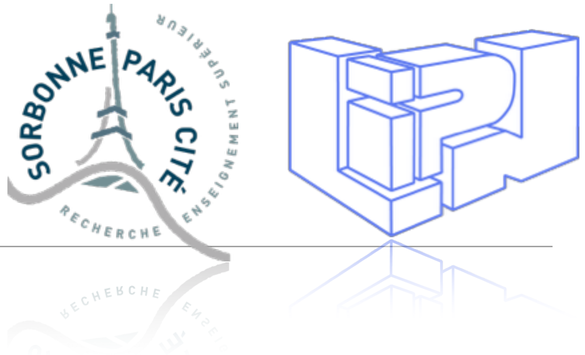
FRED + textes scientifiques

- “Information extraction is the process of extracting structured data from unstructured text”



http://wit.istc.cnr.it/stlab-tools/fred/

Problèmes



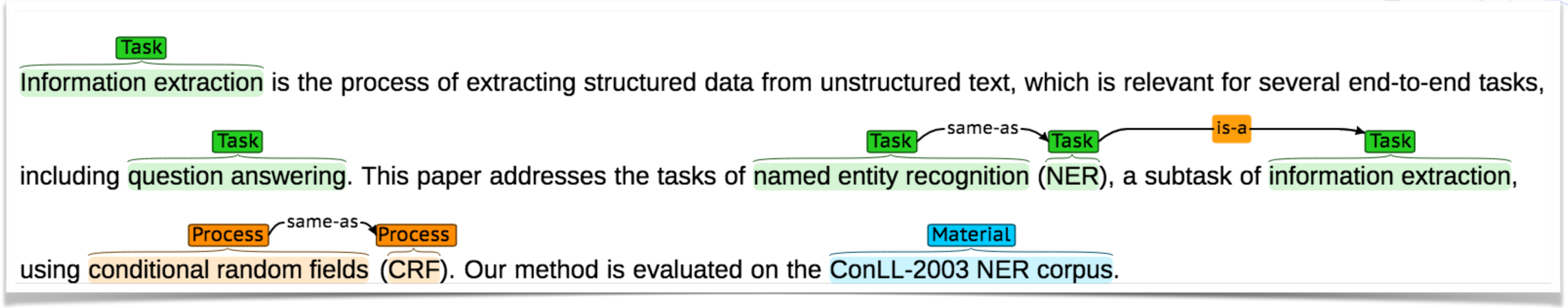
- Identifier correctement les concepts du discours scientifique
  - Problèmes d'ambiguïté entre domaines différents
    - ex: SVM: Support Vector Machines (Apprentissage Automatique)
    - SVM: Secure Virtual Machine (AMD technology)
  - Différentes façons d'exprimer le même concept (ou très similaire)
    - ex: Sentiment Analysis vs. Opinion Mining
- Trouver les relations entre ces concepts
  - Les relations sont définies entre catégories de concepts (domaine et co-domaine)
  - -> il faut catégoriser le concepts

Science IE @ SemEval 2017

- Définition de trois catégories fondamentales
  - PROCESS, TASK, MATERIAL
- Questions typiques:
  - which papers have *addressed* a specific TASK ?
  - which papers have *studied* a PROCESS or variants ?
  - which papers have *utilized* such MATERIALS ?
  - which papers have *addressed* this TASK *using* variants of this PROCESS ?

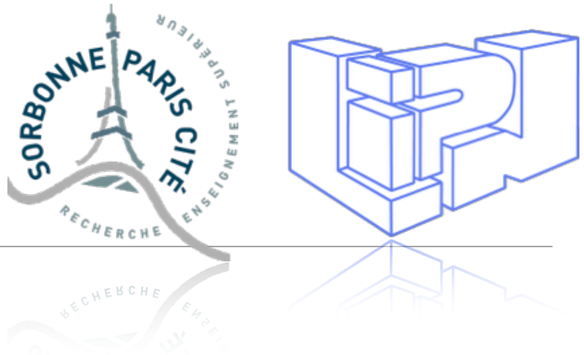
https://scienceie.github.io

Science IE @ SemEval 2017



- Défi A: trouver les mots clés
- Défi B: classifier les mots clés entre PROCESS, MATERIAL, TASK
- Défi C: trouver des relations (same-as, is-a) entre des mots clés antérieurement identifiés
- Evaluation de chaque scénario ou des combinaisons A+B ou A+B+C

Science IE @ SemEval 2017



- Training set: 400 articles dans les domaines:
  - Computer Science
  - Physics
  - Material Science
- Test set: 100 articles

Méthode base pour le défi B

- Classificateur SVM
- Caractéristiques de base:
  - Prefixes et suffixes de taille 3, 4, 5 (vus dans le training) de la séquence (*keyphrase*) candidate
  - Capitalisation de la keyphrase
  - Nombre de chiffres dans la keyphrase
  - Nombre de tirets dans la keyphrase
  - Nombre de mots dans la keyphrase

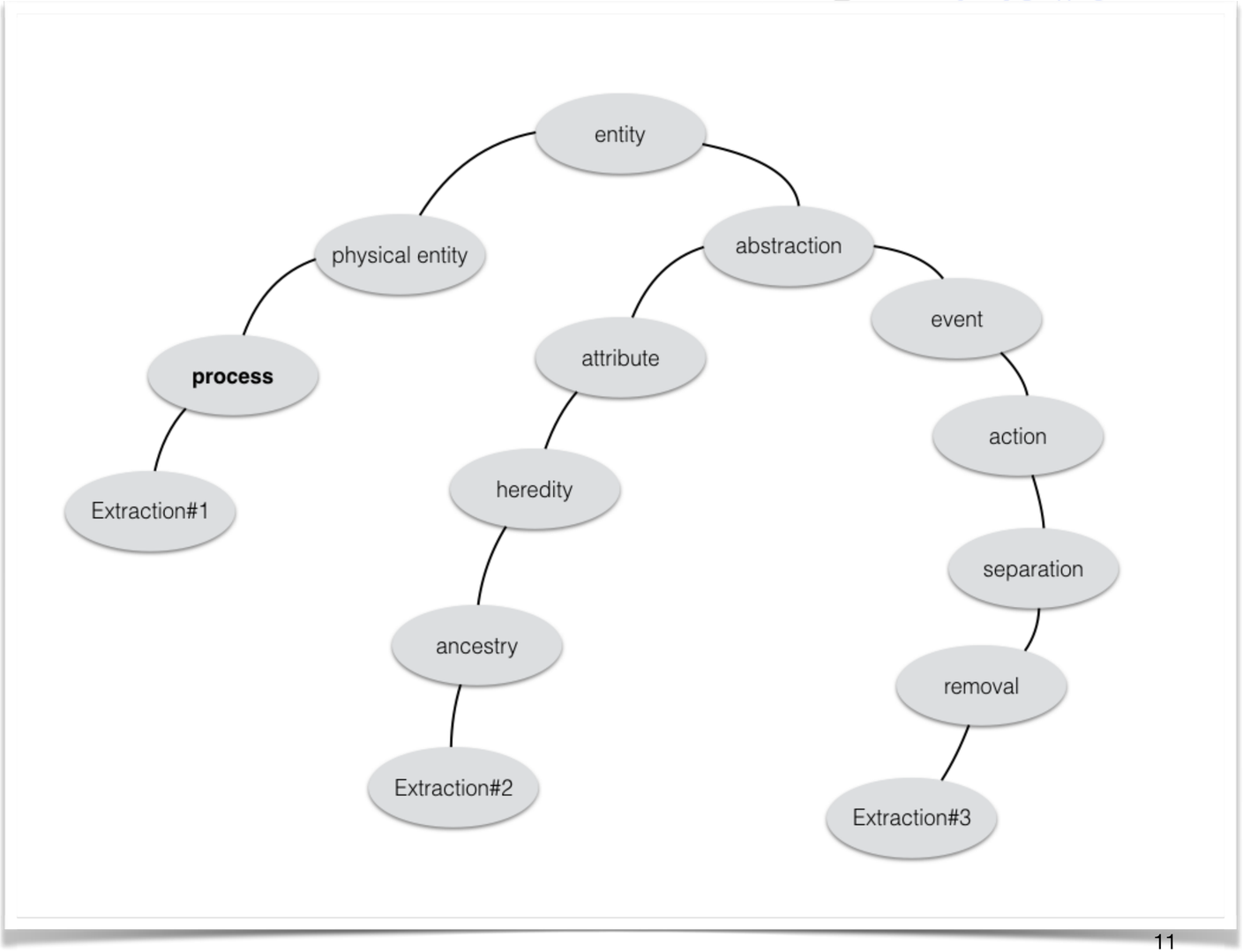
Méthode base pour le défi B

- Classificateur SVM
- Caractéristiques de base:
  - Prefixes et suffixes de taille 3, 4, 5 (vus dans le training) de la séquence (*keyphrase*) candidate
  - Capitalisation de la keyphrase
  - Nombre de chiffres dans la keyphrase
  - Nombre de tirets dans la keyphrase
  - Nombre de mots dans la keyphrase

- Exemple: Information Extraction
  - inf, info, infor
  - ion, tion, ction
  - Capitalisé
  - Pas de chiffres
  - Pas de tirets
  - 2 mots

Caractéristiques extraites à partir de WordNet

- Synpath de WordNet
- Par exemple: synpaths pour *Extraction*
- Le synset ***process*** apparaît dans un des chemins



Caractéristiques extraites à partir de WordNet

- Trouver pour chaque catégorie C de Science IE les 20 synsets qui maximisent la probabilité $p(s|C)$ (calculé sur le training set) en minimisant au même temps la probabilité $p(s|\neg C)$

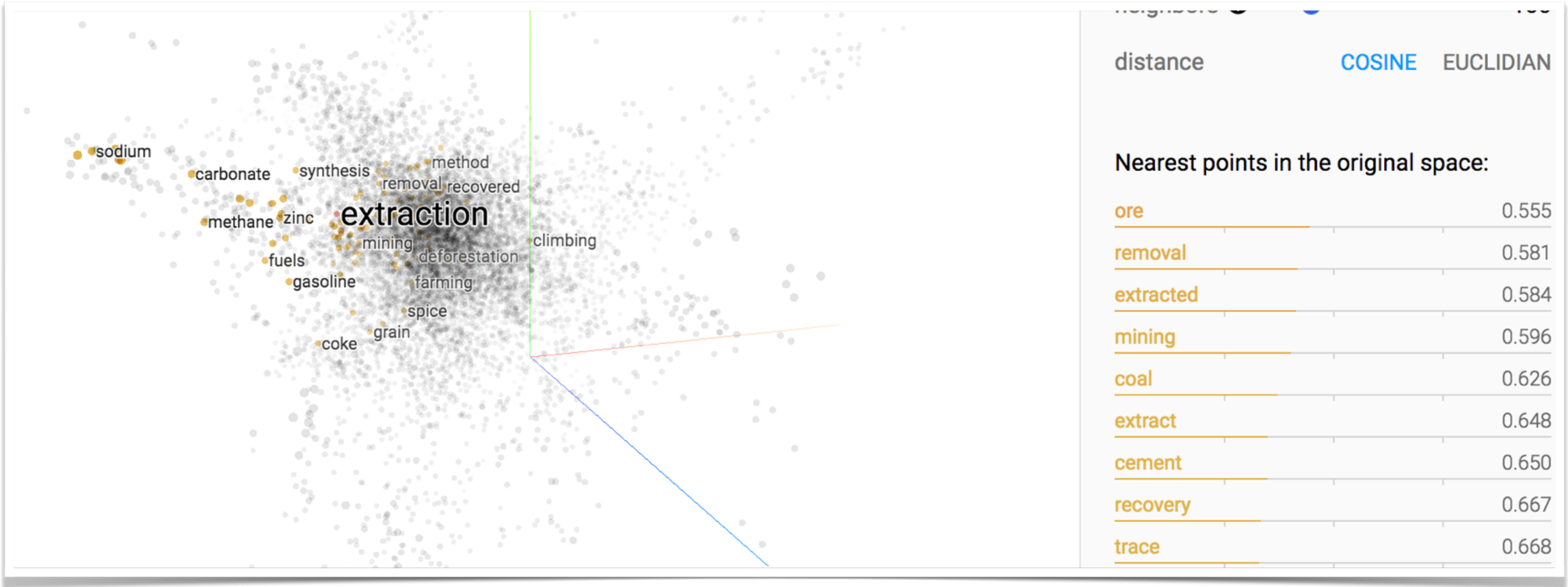
- -> 20 caractéristiques binaires

- Les 5 synsets les plus représentatifs de chaque classe:

PROCESS	*MATERIAL*	*TASK*
psychological_feature.n.01	*physical_entity.n.01*	*science.n.01*
event.n.01	*object.n.01*	*possession.n.02*
abstraction.n.06	*whole.n.02*	*natural_science.n.01*
act.n.02	*artifact.n.01*	*question.n.02*
cognition.n.01	*matter.n.03*	*subject.n.01*

Plongements de mots

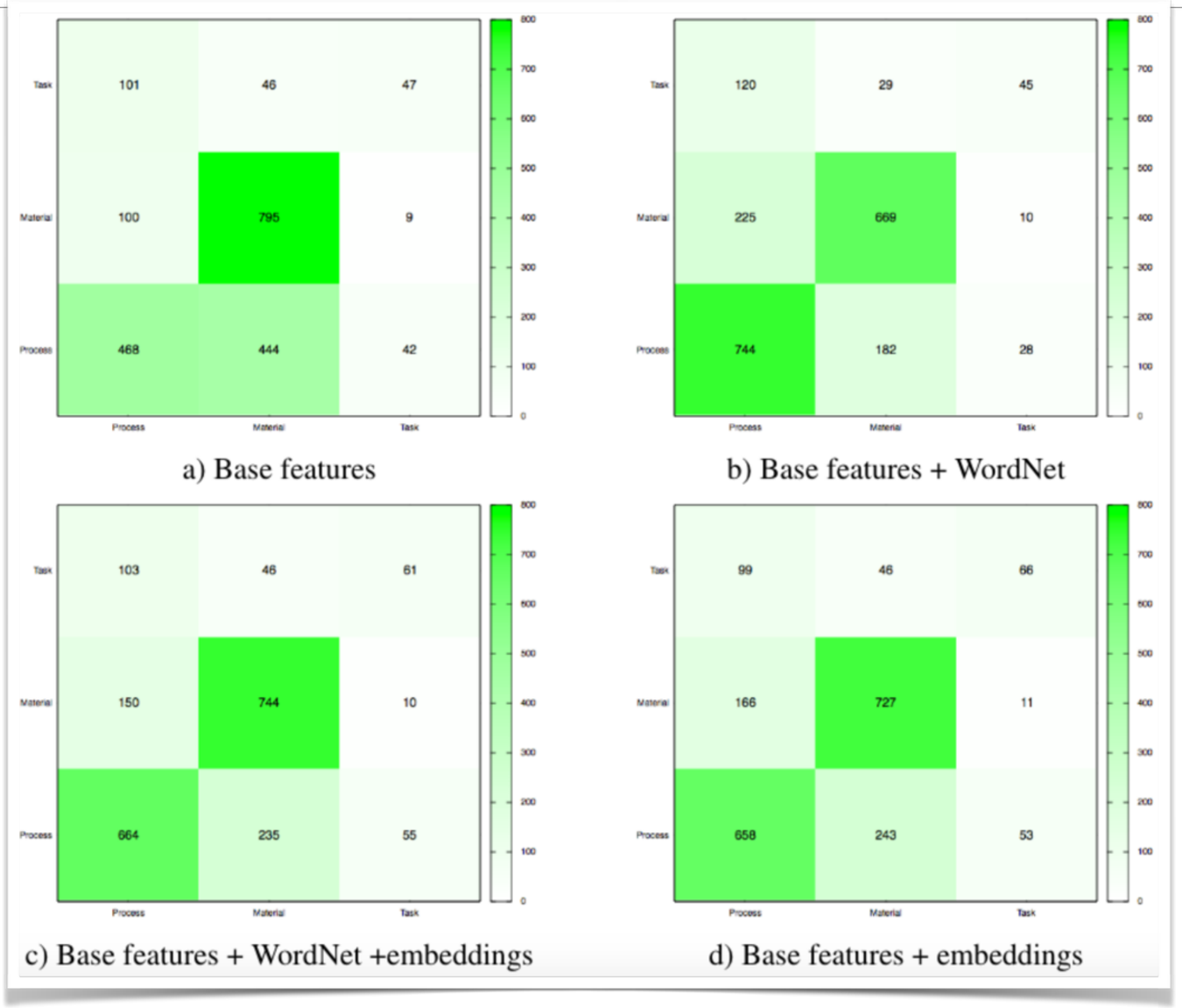
- On utilise des vecteurs de taille 300 (Google news pre-trained)
- Si plus d'un mot, on utilise le max (par colonnes) entre les vecteurs



Résultats

	PROCESS	*MATERIAL*	*TASK*	*all*
Base	.577	.726	.322	.619
Base + WN	**.728**	.750	.325	.700
All features	.710	**.778**	.381	**.716**
Base + Embeddings	.701	.764	**.407**	.701
best@SemEval2017	.660	.760	.280	.670

Résultats



a) Base features

b) Base features + WordNet

c) Base features + WordNet +embeddings

d) Base features + embeddings

Conclusions

- L'intégration de connaissances externes permet d'améliorer le taux de discrimination entre les différentes catégories
- Les résultats obtenus avec des caractéristiques liées à la structure et le signifié de la keyphrase sont supérieurs à ceux obtenus avec des caractéristiques liées au contexte (CRF utilisé dans la participation officielle)
- Toutefois, les experiences montrent aussi que la catégorie TASK est difficile à distinguer de la catégorie PROCESS
  - Problème d'annotation ou catégorie qui n'est pas assez bien définie?
  - Exemple - in the training set :
    - "synthetic assessment method" <- PROCESS
    - "synthetic assessment method based on cloud theory" <- TASK

Quelques pointeurs

- [Chavalarias 2013] David Chavalarias et Jean-Philippe Cointet. “Phylomemetic patterns in science evolution - the rise and fall of scientific fields”. PLOS ONE, 8(2).

- [Gabor 2016a] Gabor K., Zargayouna H., Buscaldi D., Tellier I., Charnois T. Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. In: LREC 2016. Portorose, Slovenia.

- [Gabor 2016b] Gábor K., Zargayouna H., Tellier I., Buscaldi D., Charnois T. Unsupervised Relation Extraction in Specialized Corpora Using Sequence Mining. IDA 2016: 237-248

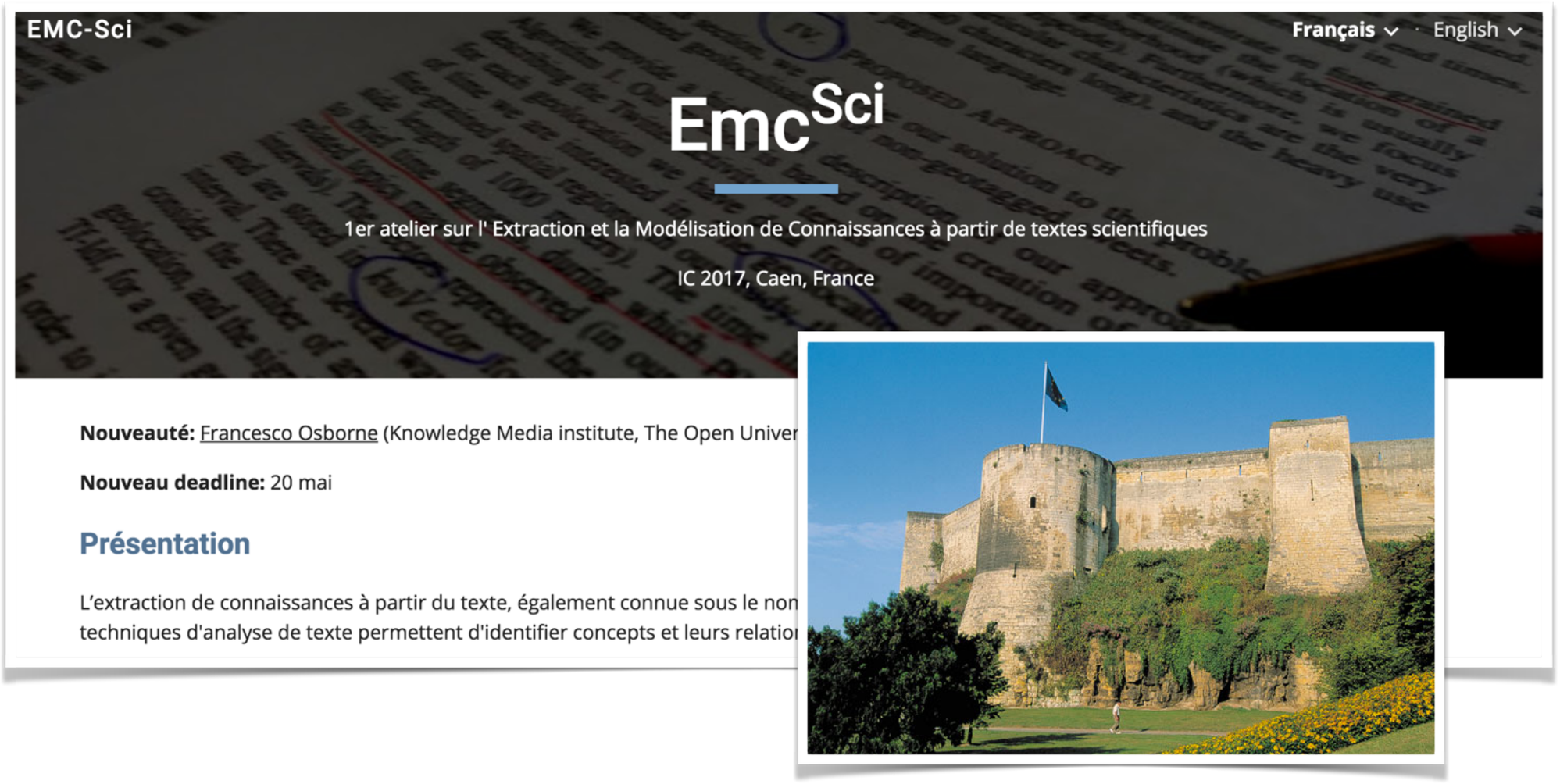
- [Gangemi 2016] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. “Semantic Web Machine Reading with FRED". Semantic Web, Preprint, to appear 2016

Un peu de pub…

- Atelier Emc-Sci @ IC 2017

Un peu de pub…

- Atelier Emc-Sci @ IC 2017

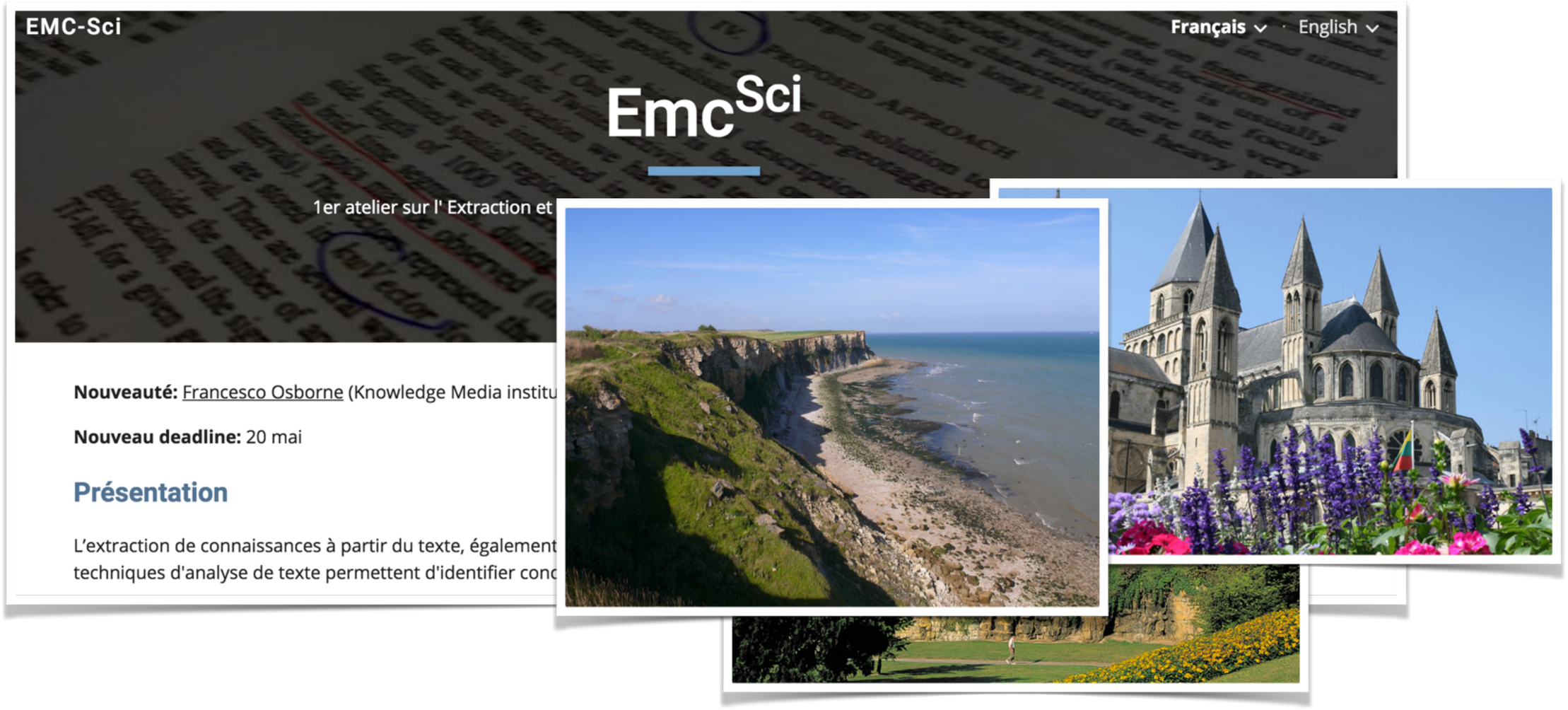


Un peu de pub…

- Atelier Emc-Sci @ IC 2017

Un peu de pub…

- Atelier Emc-Sci @ IC 2017



Un peu de pub…

- Atelier Emc-Sci @ IC 2017

Merci!