# DOING: Données Inteligentes

## Groupe de travail de DIAMS

Proposition de sous-réseaux DIAMS: Mirian Halfeld Ferrari

LIFO - LLL - LIFAT - LIG et al

Novembre, 2019

# Présentation générale

## DOING : GT DIAMS (proposition sous-réseaux)

Transformation des données en information puis en connaissance.
Domaine : TAL, BD et IA

## Participants

- TAL : **LIFAT ; LLL ; LIFO** ; ERTIM (INALCO)
- BD : **LIFO ; LIG** ; IRISA (Rennes) ; Léonard de Vinci Pôle Universitaire
- IA : LIFO
- Associés (avec participation en certains travaux) : UFPR et UFRN (Brésil)
- **Entreprise** : Ennov (Paris)

# Les participants (direct ou indirectement)

**LLL**

- Anne-Lyse Minard-Forst
- Emmanuel Schang

**LIG**

Genoveva Varga-Solar

**Ennov**

Joshua Amavi

**LIFAT**

- Jean-Yves Antoine
- Agata Savary

**LIFO**

- Sylvie Billot
- Gaëtan Caillaut
- Bich Dao
- Jacques Chabin
- Cedric Eichler
- Mirian-Halfeld Ferrari
- Nicolas Hiot
- Giacomo Kahn
- Anaïs Lefeuvre
- Christel Vrain

**Léonard de Vinci Pôle Universitaire**

Christophe Rodrigues

**ERTIM (INALCO)**

Damien Nouvel

**UFRN**

- Martin Musicante
- Ciro Medeiros

**UFPR**

Carmem Hara

**IRISA (Rennes)**

Laurent D'Orazio

# Historique des activités

Réunions

- *7 février 2019 :* (1/2 journée) *Kick-off* DOING : exposés courts sur les travaux de recherche des participants (TAL et BD) et premières discussions
  Participation : Ennov, LLL, LIFAT, LIFO
  Financement : LIFAT/projets (missions Blois-Orléans)
- *4 février 2019 :* (1 journée)
  Discussion sur exemple préparé (expérience entreprise Ennov - domaine de la santé)
  Retour sur participation journée ANR vers projet Grand Débat (domaines divers)
  Participation : Ennov, LIG, LLL, LIFAT, LIFO
  Financement : ICVL (missions et repas)
- *17 mai 2019 :* (1/2 journée)
  Présentation de travaux/perspectives des collègues parisiens (suite réunion ANR Grand Débat)
  Participation : ERTIM (INALCO), Ennov, Léonard de Vinci Pôle Universitaire, LIG, LLL, LIFAT, LIFO
  Financement : LIFO (missions et repas des invités parisiens) ; LIFAT
- *24 octobre 2019 :* (1/2 journée)
  Présentations de travaux des visiteurs brésiliens LIFO
  Discussions sur actions concrètes à lancer : propositions de stage
  Participation : UFPR, UFRN, Léonard de Vinci Pôle Universitaire, LIG,LLL, LIFAT, LIFO
  Financement : financement DIAMS accordé - mais pas utilisé (missions financés par les projets des participants venant de Blois ou Paris)

# Travaux en connexion (antérieurs, en cours ou débutant)

## Extraction d'information - construction de la base de connaissances

- Détection des entités polyexicales (PARSEME FR) :
  - COST PARSEME (2013-2017) LIFAT-LIFO
  - ANR PARSEME-FR (2016-2020) LIFAT-LIFO
  - 2 shared task PARSEME
- Entités nommées : systèmes mXs et CasEN (LIFAT et ERTIM/INALCO)
- Anonymisation, identification des données sensibles (LLL)
- Annotation manuelle et détection automatique des effets secondaires de médicaments dans des tweets : projet campagne d'évaluation SMM4H 2020 (LLL)
- Extraction d'information : domaine général et biomédical - extraction de relations entre entités médicales, détection d'événements, détection d'expressions temporelles, etc. (LLL et al.)

## Manipulation et maintenance base de connaissances

- Mises à jour (RDF) : avec informations inconnues LIFO-ETIS (Cergy) ; UFPR-LIFO
- Mises à jour (RDF) et grammaires de graphe : ANR SENDUP (LIFO)
- Interrogation : LIFAT-LIFO ; LIFO-UFPR-UFTPR
- Anonymisation : ANR SENDUP (LIFO)-UFRN

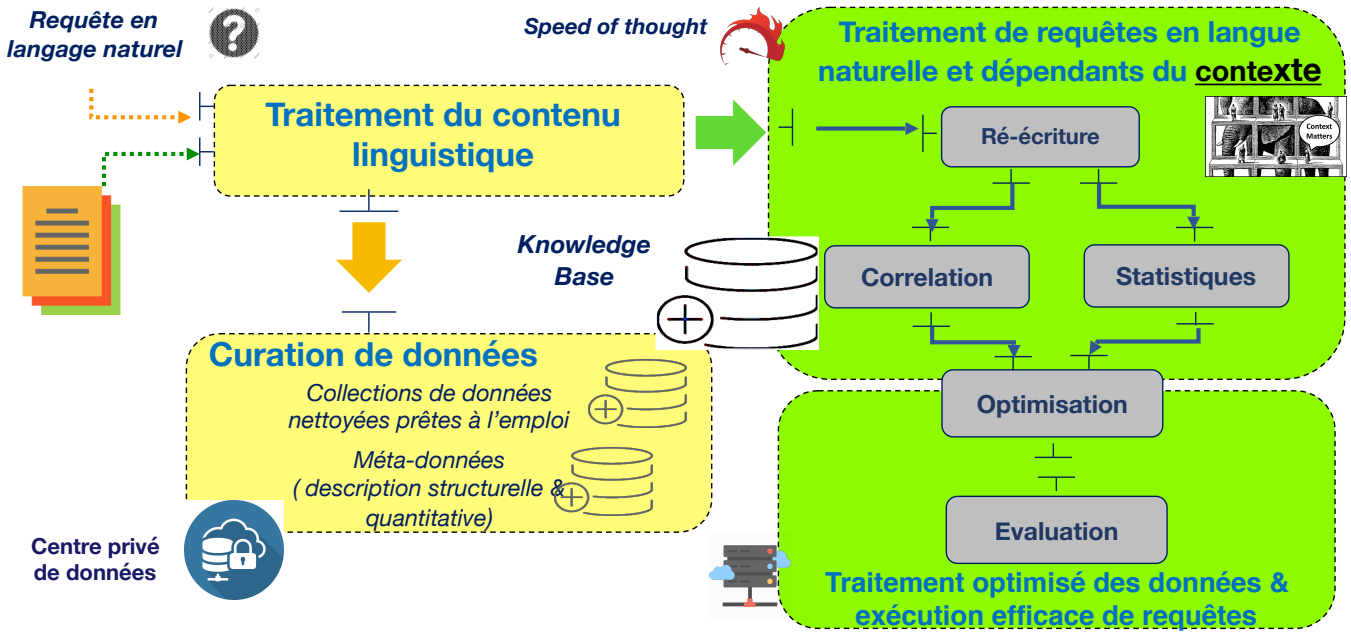
## Stage/Thèse : intersection

Thèse CIFRE (Ennov-LIFO)

# Le contexte scientifique

DOING : Données Intelligentes

# Transformation des données en information puis en connaissance




Extraction d'information: traitement statique

Extraction de connaissance: traitement dynamique (analyse)

# Deux grandes parties

## 1- La transformation des données en information

Extraction de l'information des données textuelles pour peupler une base de connaissances

## 2- La transformation de l'information en connaissance

L'interrogation intelligente et efficace, et la maintenance de bases de connaissances

# 1- La transformation des données en information

## Extraction de l'information des données textuelles pour peupler une base de connaissances

- détection des entités, leur normalisation et l'extraction des relations entre entités ;
- la construction d'un schéma à partir des instances ;
- les mappings entre les schémas ;
- le résumé des données ;
- les problèmes d'anonymisation des données à publier ;
- l'extraction de contraintes d'intégrité à partir des textes ; etc

## Types de données

Données textuelles non structurées (ou semi-structurées) desquelles nous voulons extraire des connaissances

Domaines d'application mis en avant

- santé
- transition énergétique
- textes juridiques concernant les domaines précédents

# Le coeur du problème : un petit exemple

## Entrée/Sortie (version 1)

- Entrée : texte ; version initiale du schéma de la base de connaissance ; terminologie : mapping vers schéma plus au moins initialisé
- Sortie : base de connaissance GRAPHE

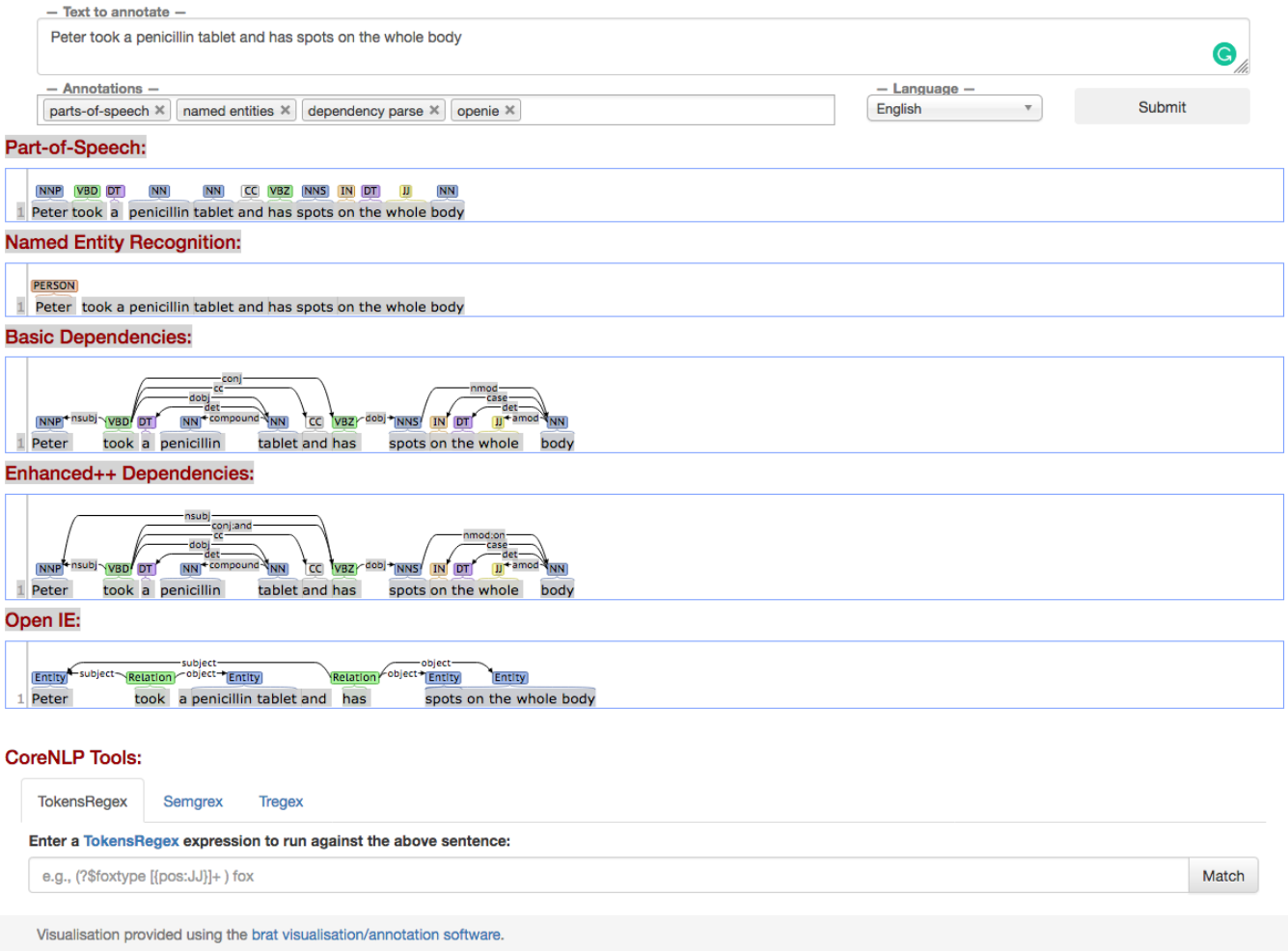
## Quelques questions

- Comment construire une instance à partir du texte ?
- Comment classer les relations encore inconnues (en imaginant les entités connues) ?
- Quelles étapes intermédiaires ?
- L'anonymisation est-elle à traiter à ce niveau ?

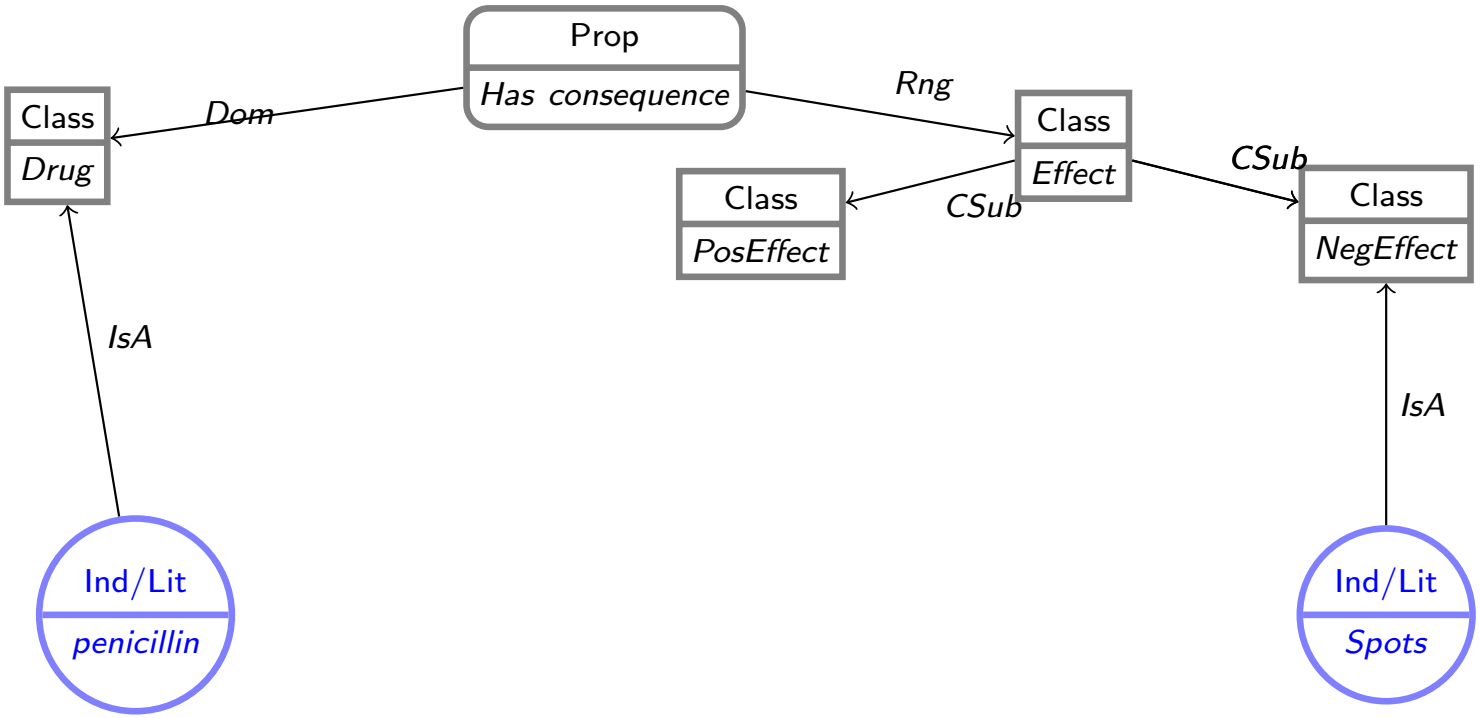
## Texte dans le domaine de la santé/médical

*Peter took a penicillin tablet and has spots on the whole body. Doctor Withsmith, following the protocol, put him under an Antihistaminic treatment*

# Texte : les dépendances



# Graphe à obtenir : extrait du exemple (RDF)



# Des discussions...

## Exemples de difficultés à considérer

- Terminologie selon le domaine de spécialité (utilisation des terminologies disponibles ; techniques pour leur enrichissement)
- Temporalité : l'ordre des événements
- L'existence d'un schéma initial (à évoluer)
- Résumé de textes ou des premières versions des instances
- Étapes diverses : différents *mappings*

# Actions : pour aller vers la concrétisation des idées

## Action concrète dans ce cadre

- Proposition stage Extraction de relations entre entités dans le domaine médical
  - TAL et BD
  - LLL ; LIFAT ; LIFO
  - **Financement ICVL** (Cherche étudiant !)
- Thèse CIFRE (LIFO/Ennov) : interface entre les deux grandes parties de DOING

# 2- La transformation de l'information en connaissance

## Intelligence dans la manipulation d'une base de connaissance

- requêtes en langue naturelle, évaluation-analyse dynamique
- offrir des mécanismes d'analyse efficaces, flexibles, faciles à utiliser et adaptés à l'utilisateur
- nouvelles formes de requêtes : requêtes 'data science', définissant des *pipelines/workflows* capables de rendre des résultat analytiques sur les données de la base.
- réponses fiables prenant en compte le profil de l'utilisateur, le respect des contraintes, du contexte, de la vie privée...
- les questions liées à l'anonymisation en préservant une utilité spécifique (éventuellement selon un contexte) de la base.

## Bases de connaissances

Questions à placer dans le cadre d'une base de données graphe.

# Les requêtes 'data science'

## Thème prospectif : petit exemple

*Déterminer les caractéristiques le plus courantes des personnes ayant le diabète, classer les personnes selon ces caractéristiques et, ensuite, pour chaque catégorie, donner l'âge et le revenu moyen.*

## Les étapes ('pipeline' de tâches)

- utiliser des algorithmes de classification pour trouver les caractéristiques plus marquantes des diabétiques,
- réorganiser les données selon les critères déterminés
- appliquer des fonctions d'agrégation et regroupement sur ces données re-organisées.

# Actions : pour aller vers la concrétisation des idées

## Actions concrète dans ce cadre

- Proposition de stage Vers des requêtes *data science*
  - BD, IA
  - LIG et LIFO
  - Stage prospectif
  - **Financement DIAMS** (cherche étudiant !)
- Collaboration LIFO-UFRN ; intersection ANR-SENDUP : thèse sur questions de l'anonymisation de la base

# Conclusions et perspectives

## Analyse du travail

- Groupe de travail : bonne dynamique, réunions animées, discussions riches et intéressantes ; collaboration nationale et internationale
- Points à renforcer : concrétiser les interactions entre les domaines pour approfondir les échanges.

## Actions lancées

- Stages master : 2 stages proposés ; financement acceptés (ICVL et DIAMS)
- Candidature DOING-MADICS
  (porteurs : Halfeld-Ferrari (LIFO) ; Minard-Forst (LLL) ; Vargas-Solar (LIG))
- Thèses liées à des questions DOING :
  - Ennov-LIFO (CIFRE)
  - Collaboration LIFO-UFRN ; intersection ANR-SENDUP (bourse brésilienne)