



Mastère SIO - Octobre 2021 - Projet Fil Rouge

V I.I

Thématique globale du projet: Transformation automatisée de la donnée en connaissance.

Objectifs pédagogiques :

1. Donner à chaque enseignant un support partagé par tous (étudiants, enseignants) pour appuyer ses exemples de cours et éventuellement la validation de son module de cours. L'intérêt de la démarche est que les étudiants sont mis en situation de réfléchir sur plusieurs aspects très différents d'un même problème connu, selon les thématiques et les technologiques abordées dans chaque cours. D'où le nom de « Fil Rouge », un lien qui éclaire la complémentarité des enseignements et fait le lien entre eux ;
2. Mettre en place un cadre pédagogique et des livrables identifiés, tout en permettant une grande liberté de démarche dans l'originalité de la solution apportée (module TLA), sachant qu'il n'existe pas de méthode formelle définitive pour arriver au résultat demandé. La créativité et la part de recherche personnelle fera la différence ;

Objectifs techniques : Étudier et développer l'ensemble d'une chaîne de traitements en Python, de la collecte des données en passant par la validation, la reconnaissance d'entités nommées, la mise en relation, la restitution, la déduction de nouvelles données, l'interrogation, la représentation de la connaissance produite. Être en mesure d'apporter un indicateur de la qualité et de la précision de la chaîne de traitement fait partie des objectifs (cf « Livrables » en diapo 6).

Mots clefs : Python, « Web Scraping », API Rest, Architecture d'application, « Software As A Service », Traitement Langue Naturelle, Ontologies, Graphes de connaissances.



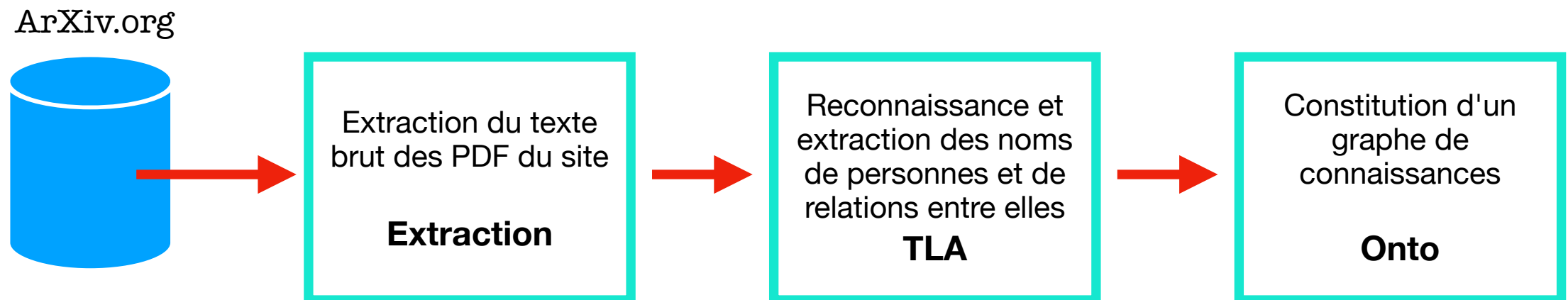
Mastère SIO - Octobre 2021 - Projet Fil Rouge

V 1.1

Objectifs du projet :

Le projet consiste à établir un graphe de connaissances (une sorte de « google scholar ») des auteurs, des documents produits, avec qui, en citant qui, extraits de toutes les publications en « **Computer Science & AI** » du site ArXiv.org. Le graphe de connaissances ainsi construit devra permettre de répondre à des interrogations comme « qui influence qui ? » (interrogation en SPARQL), voire même à l'aide de règles (SWRL) d'établir de nouvelles relations entre les auteurs (Bonus).

Ce projet peut être synthétisé en première approche comme un pipe-line de traitements à trois modules :

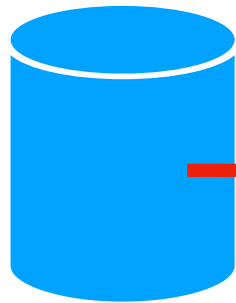




Mastère SIO - Octobre 2021 - Projet Fil Rouge

V 1.1 - Module Extraction

ArXiv.org



Extraction du texte
brut des PDF du site

Extraction

- Le point de départ est l'API décrite ici :

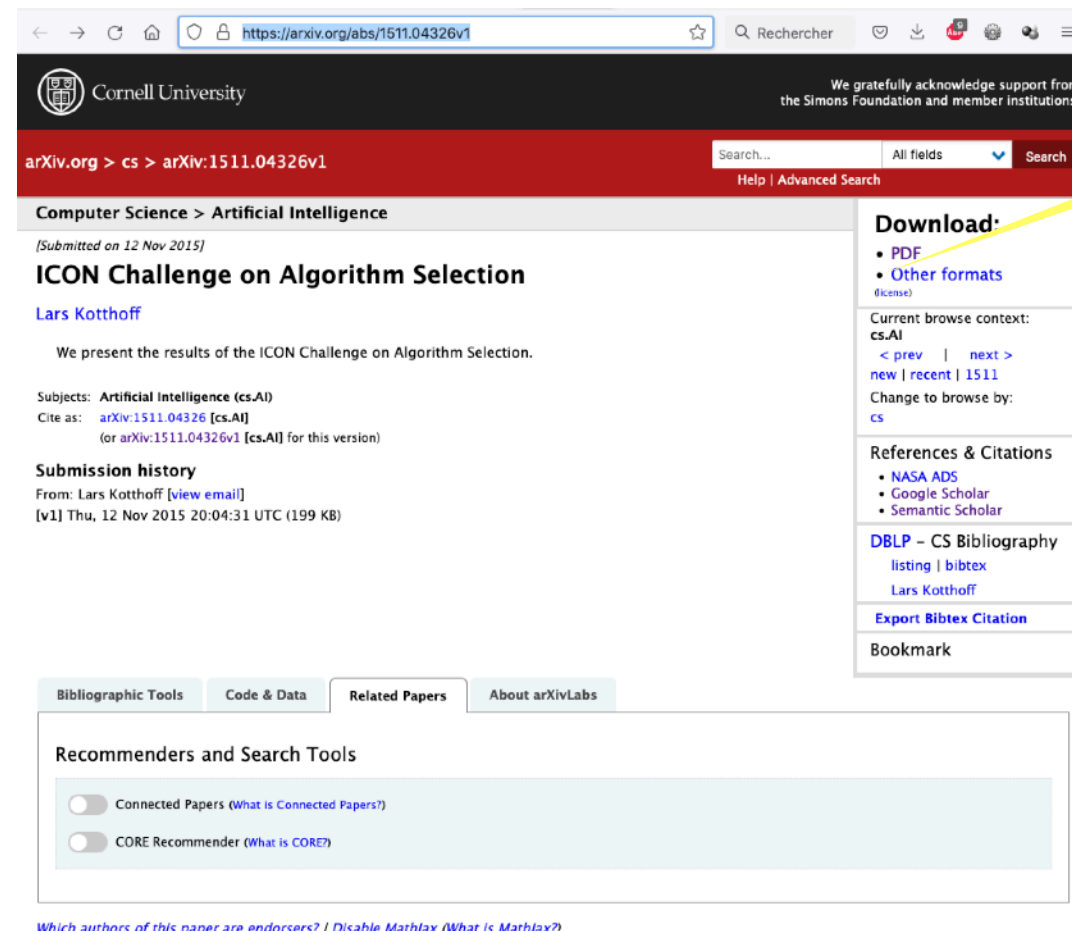
<https://arxiv.org/help/api>

et qui pourrait être traitée, par exemple, avec les modules Python "urllib" (gestion des requêtes) et "feedparser" pour le retour de l'API au format RSS (XML) ;

- La sortie est du texte brut associé à une étiquette qui est le nom de la publication

API →

Liste d'urls du type **<https://arxiv.org/abs/1511.04326v1>**




Le lien qui nous
intéresse est ici

- ➡ Pas de fichier PDF directement accessible, d'où la nécessité de devoir « fouiller » dans l'HTML (« web scraping ») : le module 'BeautifulSoup' est un bon point de départ ;
- ➡ Transformer le PDF en texte : module pdftotext par exemple ;




Mastère SIO - Octobre 2021 - Projet Fil Rouge

V 1.1 - Module TLA



Reconnaissance et
extraction des noms
de personnes et de
relations entre elles



TLA

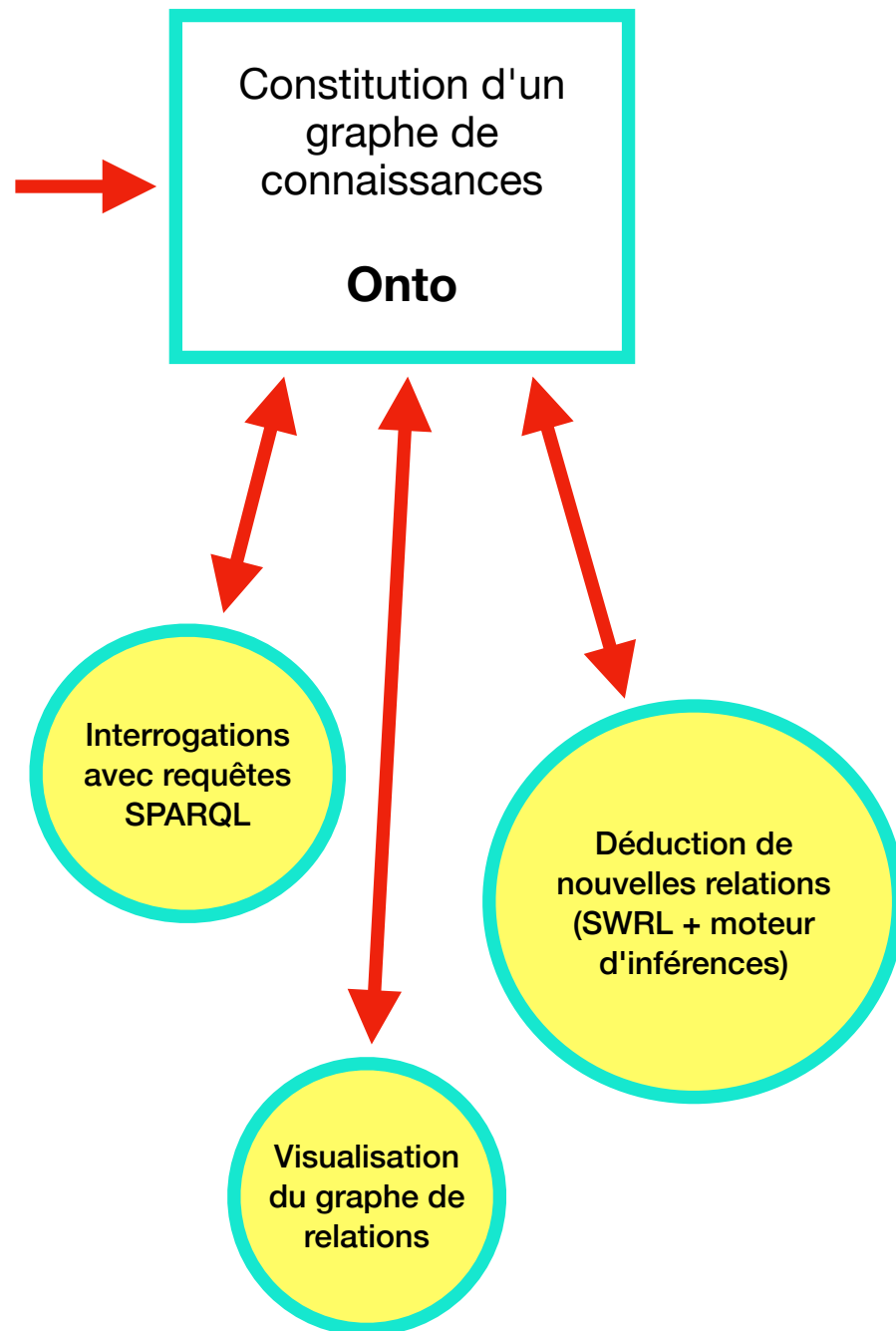
- Les textes en entrée sont en anglais ;
- C'est le domaine du traitement du langage naturel. Il faut identifier dans le flux de mots ceux qui sont relatifs à des noms propres, et c'est la plus grosse difficulté théorique de ce projet car il n'existe pas d'algorithme déterministe pour ce faire avec 100% d'efficacité. Vous allez devoir développer une stratégie, faire des hypothèses, utiliser des heuristiques ;
- L'examen de la documentation du module NLTK est un bon point d'entrée mais ne sera très certainement pas suffisant pour obtenir un résultat satisfaisant ... ;
- Cf présentations de **Badih Ghattas** sur le traitement du langage naturel (sans doute sous forme de conf. vidéo), en marge des cours IML & IDL ;
- À la sortie ce sont des noms propres associés entre eux avec des relations (à deviner, comment ?), à affiner selon ce qui sera retenu des relations voulues dans le graphe (voir le dernier module) ;
- Ce module est potentiellement (très) consommateur de ressources (mémoire, cpu,...) il est à penser « as a service » (hébergé par exemple dans un Cloud), géré via une API qui permette de l'invoquer et de récupérer le résultat de l'extraction.

Question à se poser : comment valider les résultats et avec quel critère ? C'est en effet avec cette caractéristique que vous allez pouvoir défendre votre projet en vantant sa précision. Ce point est capital.

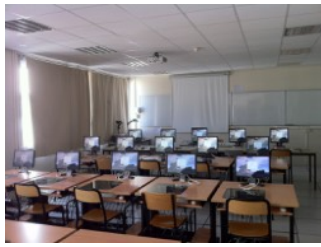


Mastère SIO - Octobre 2021 - Projet Fil Rouge

V 1.1 - **Module ONTO**



- Cours de **Sylvie Després** ;
- Usage de l'outil Protégé ;
- Usage d'un module Python adapté : owlready (cf présentation ultérieure en distanciel de F.Laissus à positionner dans l'EDT quand ce sera pertinent)
- Données nominales extraites au module précédent, assemblées ici en un graphe de connaissances. Dans le cadre de ce projet c'est l'organisation de l'Ontologie FOAF qui est choisie. Toutes les relations ne sont pas utilisées, à affiner en fonction de « l'intelligence » du module d'extraction d'entités nommées et de relations qui précède ;
- **Bonus** : découverte de relations entre les personnes grâce à l'ajout de règles écrites en SWRL et usage d'un résolveur. Par exemple (trivial) si « X est une personne auteur de T un texte et Y une personne auteur de T un texte alors X et Y sont co-auteurs de T » (Pellet conseillé dans ce cas, des précisions lors de la présentation de F.Laissus).



Mastère SIO - Octobre 2021 - Projet Fil Rouge

V I.I

Livrables :

- L'ensemble du projet est à remettre sur Edunao (des précisions suivront) au plus tard le vendredi 8 avril 2022 à 23h59 ;
- Le projet est individuel : chacun remet l'intégralité des livrables sous son nom, par contre les travaux, notamment de réflexion peuvent se faire en groupe ;
- Le livrable est une archive comprimée au format zip qui comprend :
 1. Un rapport technique sur l'ensemble de l'architecture, qui explique les choix et justifie les performances du modèle (question de la diapo 4), la description de l'API qui permet d'accéder au module TLA, un exemple d'interrogation avec le langage SPARQL, quelques mots d'explications sur les éventuelles règles SWRL écrites et les déductions de nouvelles relations trouvées (BONUS). 25 pages maximum.
 2. Le source Python commenté de l'ensemble du « pipe-line » de traitement (hors module TLA), sous forme d'un Notebook (Jupyter vs JupyterLab) ;
 3. Le source Python du module « as a service » (diapo 4) constituant externe du « pipe-line » ;
 4. Un fichier au format XML (par exemple sortie de l'outil Protégé) avec l'ontologie FOAF chargée, des individus et relations ajoutées.