# Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques

Pascal Cuxac, Alain Collignon, Stéphanie Gregorio, François Parmentier

# Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques

**Pascal Cuxac**

Chef de projet Fouilles de textes

INIST-CNRS, Vandœuvre lès Nancy, France

pascal.cuxac@inist.fr

**Alain Collignon**

Ingénieur de recherche

INIST-CNRS, Vandœuvre lès Nancy, France

alain.collignon@inist.fr

**Stéphanie Gregorio**

Ingénieur de recherche

INIST-CNRS, Vandœuvre lès Nancy, France

stephanie.gregorio@inist.fr

**François Parmentier**

Ingénieur de recherche

INIST-CNRS, Vandœuvre lès Nancy, France

francois.parmentier@inist.fr

## Résumé

Dans cet article nous présentons une approche automatique visant à désambiguïser et aligner des entités géographiques de type placeName. Une méthode basée sur des plongements lexicaux permet, à partir d'un apprentissage non supervisé de lever l'ambiguïté face à un terme polysémique. Cela permet alors un alignement automatique avec différents réservoirs (BNF, wikidata…) possédant un triplestore. Nous utilisons alors les technologies du web sémantique, pour à la fois exposer les données de façon différente (data.istex) mais également autoriser des requêtes complexes impossibles à résoudre à partir de moteurs de recherche classiques. Nous aborderons un cas concret basé sur le réservoir ISTEX, et une évaluation qualitative de la méthode sera proposée.

## Mots clés

Web de données , Données ouvertes liées , Alignement automatique , Désambiguïsation , Entités géographiques

## Title

From massive databases to Web of data: disambiguation and alignment of geographical entities in scientific texts


## Abstract

In this paper we present an automatic approach to disambiguate and align geographic entities. A method based on word embeddings allows, from unsupervised learning, to remove ambiguity with polysemic terms. This allows automatic alignment with different databases (BNF, wikidata...) having a triplestore. We then use semantic web technologies, both to expose the data in a different way (data.istex) but also to allow complex queries that cannot be solved from traditional search engines. We will discuss a concrete case based on the ISTEX database, and a qualitative evaluation of the method will be proposed.



## Keywords

Web of Data , Linked Open Data , Automatic alignment , Disambiguation , Geographic entities

## INTRODUCTION

A l'ère du Web, nous assistons au développement de données en libre accès (OpenData, Open Access) et de collections issues de bibliothèques traditionnelles qui sont maintenant accessibles librement : Gallica, Europeana, Digital Public Library of America. De récentes initiatives nationales ont également permis le développement d'importantes archives scientifiques (ISTEX en France, SwissBib en Suisse, GBV en Allemagne, Scholars Portal en Ontario). Le développement des humanités numériques (Bouzidi & Boulesnane, 2017) produit également de vastes volumes de données textuelles dans toutes les disciplines SHS sans exception.

Afin de faciliter l'interrogation de ces réservoirs et aider les utilisateurs dans l'analyse de corpus, il est important de pouvoir enrichir les textes (les annoter) avec différentes données ou métadonnées construites en utilisant des fonctionnalités d'extraction de connaissances.

Parallèlement, le web sémantique est présenté comme étant le web pour lequel les ordinateurs interprètent les métadonnées afin de mieux assister l'utilisateur dans sa recherche de l'information (Berners-Lee, Hendler, & Lassila, 2001). La sémantisation des données se développe rapidement, laissant entrevoir des possibilités d'interrogations « augmentées » en liant des jeux de données, permettant de réinterroger la notion d'information.

Le projet ISTEX [1] (initiative d'excellence en Information Scientifique et Technique) a pour objectif de permettre à la communauté ESR française (Enseignement Supérieur et Recherche) d'accéder en ligne, à une bibliothèque numérique regroupant l'essentiel des publications scientifiques mondiales dans toutes les disciplines scientifiques et en texte intégral. Dans ce cadre, avec plus de 22 millions de textes intégraux, nous bénéficions d'un réservoir de données massives. La mise en ligne de ces informations en texte intégral structuré permet de développer des fonctionnalités d'extraction de connaissances.

Durant la mise en place d'ISTEX, des méthodes d'enrichissement ont été adaptées et appliquées concernant notamment la classification thématique, l'indexation, l'extraction d'entités nommées (Collignon & Cuxac, 2017).

Dans la suite de cet article nous allons nous focaliser sur des entités nommées et plus spécialement sur des entités géographiques de type « placeName » désignant des noms de lieux géopolitiques ou administratifs (ville, région, pays, etc.). Ce type d'annotation revêt une importance grandissante dans des domaines très variés comme en sciences de la vie ou en sciences de la terre (Karl, 2018). Ces entités ont été extraites dans une précédente étape grâce au programme Unitex-Cassys (Cuxac & Thouvenin, 2017), mis en œuvre par le Laboratoire d'Informatique de Tours [2].

Afin d'interroger ISTEX de façon autre qu'une interrogation booléenne, nous avons décidé d'utiliser les technologies du web sémantique, pour à la fois exposer les données de façon différente (data.istex [3]) mais également autoriser des requêtes plus complexes.

## 1 – DÉSAMBIGUÏSATION ET ALIGNEMENT AUTOMATIQUE

### 1.1 État de l'art

La désambiguïsation de termes est une thématique importante quand on fait du traitement automatique de la langue. Comment savoir si "apple" dans un texte, concerne la société informatique ou le fruit ? Est-ce que "rock" s'apparente à une roche ou à un style de musique ?

La chaine de caractère “France” est-elle associée au pays ou à l’écrivain ? Pour un humain, la compréhension du contexte permet de désambiguïser sans se poser de question, pour un ordinateur cela est beaucoup plus complexe.

Si l’on recherche une chaîne de caractères associée à une entité géographique dans la base de données géographiques geonames [4] on peut construire des alignements multiples : par exemple “Xincun” renvoie 1013 réponses différentes. Des filtres de tri sont possibles, connaissant le continent ou le pays ou encore le type d’entité (lieu peuplé, parc, montagne…), mais l'ambiguïté reste difficile à lever.

La recherche d’informations géographiques dans les textes est un sujet qui se développe très rapidement (Sallaberry, 2013), et dans ce cadre il est crucial de lever les ambiguïtés. Il existe différentes méthodes d’extraction d’entités nommées dans les textes, mais quand on a extrait le terme “Paris”, “Athens” ou encore “Georgia”, comment savoir si on a respectivement la capitale de la France ou la ville du Texas, la capitale de la Grèce ou la ville de Géorgie (USA), l’état des Etats-Unis ou la république ?

Ce processus de désambiguïsation est indispensable en amont de toute opération d’alignement : en effet comment aligner correctement une entité si on ne sait pas à quoi elle se réfère ? Comme le signale Bougriou (Bougriou, 2016) “(la) *désambiguïsation des toponymes est devenue une tâche primordiale dans plusieurs domaines tels que le web sémantique, traitement automatique des langages naturels*...”

Pour certains auteurs, extraction et désambiguïsation sont étroitement dépendantes (Habib et Van Keulen, 2011). Certains abordent l’extraction et la désambiguïsation à l’aide de grammaires locales (Martineau et al. 2007), d’autres commencent par une analyse lexicale et contextuelle pour alimenter un modèle d’apprentissage actif (Chihaoui et al., 2018), d’autres appliquent des méthodes de classification non supervisée (Bougriou, 2016).

Ces dernières années ont vu se développer à une vitesse fulgurante les méthodes de “plongement lexical” (“*Word embeddings*”), dans la lignée de ce que les media appellent “Intelligence Artificielle” et les méthodes d’apprentissage profond (“*deep learning*”). Nous reviendrons un peu plus loin sur ces méthodes, sans toutefois entrer dans la théorie sous-jacente. Nombre d’auteurs ont vu là une piste intéressante afin d’apprendre automatiquement le sens des mots (Iacobacci et al., 2016) (Foppiano et Romary, 2018).

## 1.2 Les représentations vectorielles des mots

Les méthodes de plongement lexical ou “word embeddings” se sont démocratisées à la suite des travaux de Mikolov (Mikolov et al, 2016). Sa méthode Word2Vec [5] a rapidement essaimé, on citera par exemple Glove [6], FastText [7], Sense2vec [8] ...

L'intérêt de ces approches par auto-encodeur réside dans la puissance d’un apprentissage non supervisé sur de gros volumes de données afin de calculer un vecteur numérique de dimension raisonnable (souvent de l’ordre de 200 à 300 dimensions) pour chaque terme rencontré, en prenant en compte son contexte d’apparition. Des adaptations de cette méthode peuvent vectoriser des documents, permettant, par rapport aux méthodes classiques, comme TF-IDF, d’avoir un nombre de dimensions réduit et des matrices qui ne sont plus creuses. Les vecteurs étant calculés on peut ensuite, par exemple, évaluer les proximités entre mots (ou documents), ou encore construire des classes de mots (ou documents).

Cependant une même chaîne de caractères avec des sens différents sera toujours vectorisée de façon unique. Le vecteur calculé fera la synthèse des contextes trouvés et ne sera pas capable de distinguer les différents sens du mot.

Adagram (Bartunov et al. 2016) a été développé afin de pallier les problèmes de désambiguïsation de termes pouvant avoir des sens différents suivant leur contexte d'apparition. Cette méthode est dérivée de Word2Vec, et intègre une méthode bayésienne non supervisée afin de produire, si nécessaire, plusieurs prototypes de vecteurs associés à chacun des sens détectés et liés aux contextes d'apparition.

## 1.3 Notre méthodologie avec Geonames

Nous avons mis en place une méthodologie utilisant l'algorithme Adagram, et schématisée sur la figure 1.

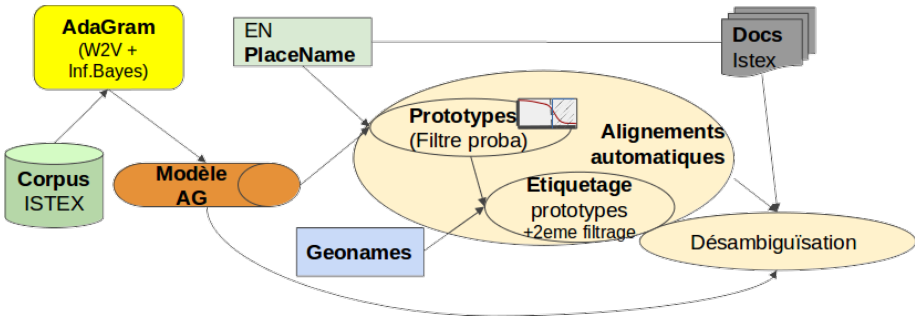


Figure 1 : Méthodologie de désambiguïsation schématisée

Nous partons d'un corpus d'apprentissage en texte intégral dont chaque document contient au moins une entité nommée. Nous appliquons un prétraitement rapide qui consiste à uniformiser la casse des caractères et appliquer un dictionnaire de mots vides. Le résultat est donné en entrée de l'algorithme Adagram, paramétré pour calculer au plus cinq prototypes (ou clusters) différents pour chaque token.

L'approche étant probabiliste, chaque prototype calculé pour une même entité est pondéré par une probabilité que cette même entité soit affectée à ce prototype dans le corpus d'apprentissage. Nous utilisons alors un filtre dynamique qui va repérer une rupture brusque dans l'évolution de ces probabilités pour éliminer les prototypes les moins probables.

Chaque prototype restant est caractérisé par les "n" termes qui le constituent, classé par poids décroissant. Chaque terme est ensuite envoyé à l'API geonames qui nous renvoie des informations telles que le pays, l'état, la province, les coordonnées géographiques... Les données communes aux "n" termes sont utilisées pour étiqueter le prototype : par exemple, l'entité "Paris" peut avoir deux prototypes, le premier constitué de [île de france, saint-denis, montmartre] va être étiqueté "France", le deuxième constitué de [dallas, fort worth, houston] va être étiqueté "Texas, USA". Si un prototype renvoie des données trop hétérogènes ou n'étant pas des données géographiques alors celui-ci sera oublié : par exemple un prototype constitué de [PSG, Neymar, football] ne sera pas pris en compte. Au final de ce processus, à chaque entité nommée correspond un ou plusieurs prototypes (vecteurs) que l'on est capable d'aligner (mettre en correspondance) avec geonames en lui attribuant un identifiant unique. Dans notre exemple "Paris, France" sera "http://www.geonames.org/2988507" et "Paris, USA, Tx" sera "http://www.geonames.org/4717560".

A cette étape nous avons construit un modèle qui va être ensuite utilisé pour désambiguïser les entités dans les documents. Pour chaque entité détectée par Unitex, nous avons les documents correspondants dans le réservoir ISTEX. Nous utilisons alors le modèle construit pour aligner les entités dans les documents en fonction de notre apprentissage.

## 1.4 Les résultats

**SemEval**

Nous avons testé notre approche en utilisant les données de la campagne d'évaluation SemEval 2019 [9] : il s'agit de campagnes d'évaluation annuelle et internationale de systèmes d'analyse sémantique. La tâche 12 intitulée "*Toponym resolution in scientific papers*" [10] comporte une sous-tâche de désambiguïsation des toponymes fournissant des corpus annotés par des experts à des fins d'apprentissage automatique. L'intérêt est d'avoir un corpus annoté, reconnu internationalement, et nous permettant de corriger notre algorithme et de nous évaluer en termes de performance.

Notre approche étant non supervisée, tous les corpus fournis peuvent être utilisés pour notre évaluation. L'évaluation se fait en suivant les préconisations des organisateurs de la tâche, c'est à dire en calculant une précision $P_{ds}$, un rappel $R_{ds}$ et une Fmesure suivant les formules suivantes :

$$P_{ds} = \frac{T_{CD}}{T_{CD} + T_{ID}}$$

$$R_{ds} = \frac{T_{CD}}{T_N}$$

$$F_{ds} = \frac{2 * Pds * Rds}{Pds + Rds}$$

avec :

$T_{CD}$ = Nombre de toponymes correctement désambiguïsés

$T_{ID}$ = Nombre de toponymes incorrectement désambiguïsés

$T_N$ = Nombre total de toponymes dans le corpus

| | $F_{ds}$ |
|---|---|
| Corpus échantillon | 0.85 |
| Corpus d'apprentissage | 0.83 |

Tableau 1 : Évaluation de notre méthode en termes de Fmesure sur les corpus SemEval'19 task 12-2

Les résultats du tableau 1 montre que la méthode a de bonnes performances. Pour avoir une idée de l'ordre de grandeur, le vainqueur de la tâche 12-2 a gagné avec une valeur de $F_{ds}$ = 0.823. Toutefois ce dernier résultat a été obtenu sur un corpus différent. Cependant les ordres de grandeur sont là, puisqu'avec deux corpus différents nous obtenons un score supérieur à 0.8.

Nous nous sommes intéressés alors aux échecs de la méthode, pour comprendre d'où venaient les erreurs et si l'on pouvait apporter une correction. Nous avons détecté plusieurs typologies d'erreurs :

- Erreur due à l'apprentissage : pour ce test nous avons utilisé un corpus d'apprentissage de 400 000 documents, pas forcément adapté aux corpus traités. De ce fait certains toponymes sous-représentés ou même non représentés ne peuvent être détectés par la méthode de façon correcte.
- Dans geonames, les placeName sont identifiés via leur niveau administratif. Ainsi "Nancy" aura un identifiant différent suivant que l'on parle du chef-lieu de canton, de la métropole ou de la préfecture. Cependant les coordonnées géographiques seront bien les mêmes.

Partant de cette constatation, et sachant que notre problématique est bien d'identifier et d'aligner des lieux géographiques, nous avons refait nos évaluations en ne prenant plus en compte l'identifiant geonames comme étiquette, mais en vérifiant les coordonnées géographiques. Si les coordonnées géographiques de l'entité que nous détectons sont les mêmes que celles de l'entité étiquetée par l'expert, alors nous considérons notre résultat comme bon (nous tenons compte d'une marge d'erreur sur les coordonnées).

| | $F_{ds}$ |
|---|---|
| Alignement "administratif" (1) | 0.83 |
| Alignement "géographique" (2) | 0.945 |

Tab. 2 : Évaluation de notre méthode en termes de Fmesure sur les corpus SemEval'19 task 12-2 en prenant en compte le niveau administratif (1) ou les coordonnées géographiques (2)

Le résultat (tableau 2) montre une augmentation significative des performances, puisque la Fmesure dépasse grandement les 0.9

**Application à l'archive ISTEX**

Dans un premier temps nous avons repéré 41 placeNames qui potentiellement peuvent avoir au moins 50 alignements différents avec geonames. Le maximum est pour l'entité "Xincun" qui possède 705 identifiants différents (fig. 2). On notera qu'en français "xincun" peut être traduit comme "nouveau village", cela pouvant expliquer la "prolifération" de ce toponyme.

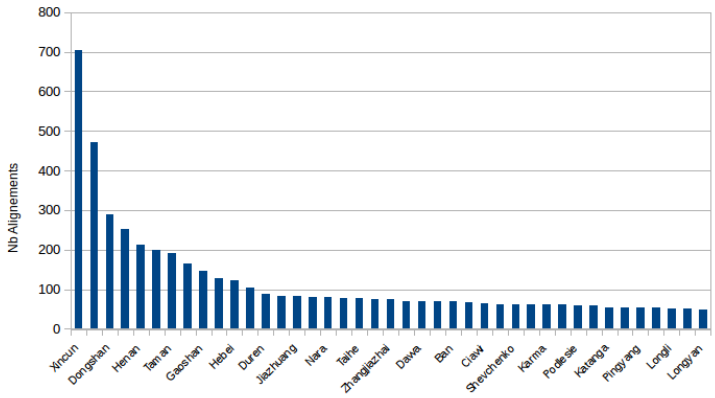


Figure 2 : Répartition des entités de type placeNames ayant plus de 50 alignements différents.

Lorsque l'on applique le modèle calculé sur 318 documents ISTEX contenant l'entité "Xincun", le résultat de la désambiguïsation permet d'aligner ce placeName avec un seul identifiant geonames, en l'occurrence l'uri "http://sws.geonames.org/1789159/ ".

Le résultat global sur les 41 placeNames traités permet de désambiguïser fortement, puisque nous passons de 6069 alignements différents à 152 alignements possibles. La méthode permet également d'éliminer certaines entités extraites qui ne sont pas des toponymes (highway, velocity…).

Dans un second temps la méthode a été appliquée à 4452 toponymes ayant potentiellement entre 2 et 50 alignements dans geonames. Globalement il en résulte un maximum de 4

alignements possibles pour certaines entités, mais dans la plupart des cas un toponyme dans un document est aligné avec une seule uri geonames.

## 2 – Exposition des données - Mise en œuvre

### 2.1 Éléments de contexte

L'ensemble de ces lieux géographiques a permis de réaliser un jeu de données au même titre que d'autres jeux de données constitués à partir d'informations extraites automatiquement (catégories scientifiques) ou récupérés à partir d'informations induites et produites par les documentalistes (types de documents, regroupement des langues, etc.).

L'ensemble de ces données suit un processus permettant la mise en ligne de données liées et ouvertes au travers du site data.istex.fr et interrogeables via un triplestore. Pour cela un workflow respectant le mode opératoire présenté ici (figure 3) a été mis en place afin de rendre les données « sémantiques ».

Les enrichissements issus de l'outil Unitex forment une représentation particulière dénommée facettes au sein de l'api ISTEX [11]. Puis viennent les étapes de curation, d'ajout d'information provenant d'autres réservoirs, de désambiguïsation automatique et d'alignement avec des ressources cœurs.

Ces données propres, documentées et préformatées sont ainsi ingérées par un outil (LODEX) qui permet de sémantiser les différentes informations ainsi que de les publier et de les partager selon les normes du web de données via le site data.istex.fr, et un triplestore [12].

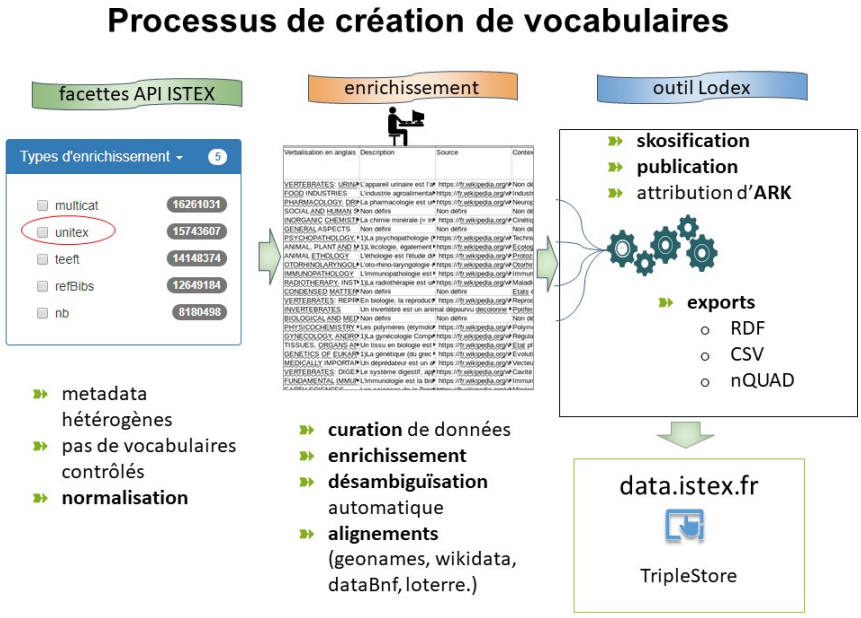


Figure 3 : Processus de création de vocabulaires

### 2.2 - La création de l'ontologie ISTEX

Rendre nos données disponibles sur le Web implique l'utilisation d'une ontologie facilitant leur accès et leur interopérabilité. C'est une notion-clé dans ce qui est appelé aujourd'hui le web de données car il est possible de naviguer et de rebondir sur des informations complémentaires à la ressource initiale. Cette interopérabilité est une des raisons motivant

l'utilisation d'ontologie : d'abord utilisées en intelligence artificielle avant de s'étendre à d'autres champs de l'informatique, les ontologies définissent un vocabulaire commun à un domaine en structurant l'information par des ensembles de concepts (Szabados et Letricot, 2012).

Malgré la multitude d'ontologies disponibles au travers de l'initiative linked open vocabulary (Vandenbussche et al., 2012), afin de pallier nos besoins, nous avons créé une ontologie propre au projet Istex (classes et propriétés). La figure 4, dans un formalisme orienté-objet compatible avec le langage RDF [13], présente les 18 classes permettant de décrire les différents jeux de données ainsi que les différentes *Object Properties* les reliant. Nous avons utilisé neuf ontologies de référence complétées par la création de nos propres classes et propriétés istex, pour obtenir l'ontologie ISTEX [14].

La modélisation du jeu de données « placeNames » appartenant à la classe *istex: PlaceConcept*, classe qui décrit la structure conceptuelle des entités nommées de type lieux, nous constatons que cette classe est reliée d'une part à la classe *bibo:Document* par la propriété ObjectProperty *istex: extractedPlace* et d'autre part à la classe *istex: NamedEntityConcept* (classe qui décrit la structure conceptuelle des entités nommées en général) par la propriété ObjectProperty *istex: dctIsPartOf*. Pour décrire la classe *istex: PlaceConcept*, nous avons utilisé 10 DataProperty issues de 4 vocabulaires de référence (SKOS [15],DBpedia [16], GeoNames [17], DCTERMS [18]) et une DataProperty que nous avons créée permettant de caractériser la requête à réaliser pour accéder aux documents indexés par le concept décrit. En ce qui concerne la classe *istex:NamedEntityConcept*, 8 propriétés ont été utilisées, toutes issues de deux vocabulaires de référence, SKOS et DCTERMS.

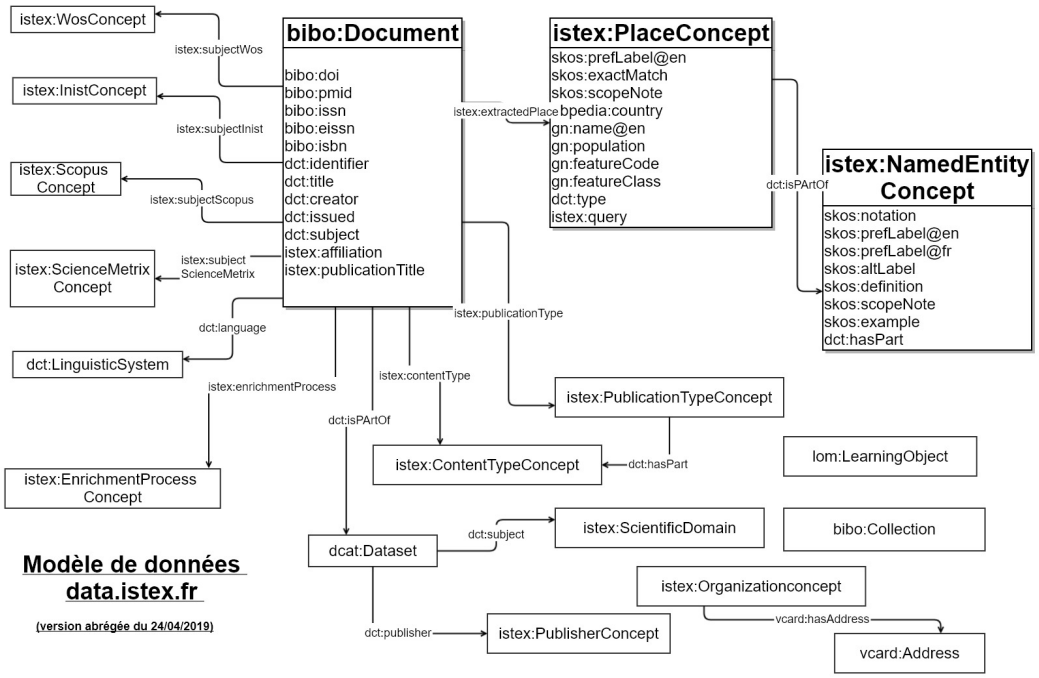


Figure 4 : Ontologie ISTEX - vue abrégée

La mise en place de cette ontologie valorisant les données de l'archive ISTEX a nécessité de modéliser nos données, mais aussi d'utiliser un outil permettant de répondre à ce triple objectif :

- Mettre en œuvre l'ontologie globale pour Istex,
- Créer un site dédié : data.istex.fr regroupant tous les travaux (listes des vocabulaires/jeux de données, accès à l'ontologie Istex,
- Alimenter un triplestore agrégeant toutes les données produites et structurées.

## 2.3 - L'outil LODEX

Dans un environnement professionnel en pleine mutation qui a vu naître de nouvelles activités dans les bibliothèques, la curation, la modélisation, la normalisation, le modèle RDF sont au cœur des préoccupations des data-managers. Ceci a eu pour incidence l'émergence d'outils dédiés comme par exemple LODRefine et Catmandu (Harlow, 2015). Plus près de nos préoccupations, le logiciel CubicWeb est utilisé dans le développement de l'application data.bnf.fr (Le Bœuf, 2013). Ce logiciel présente de nombreuses fonctionnalités, cependant l'usage de ce framework nécessite l'appui technique de la société Logilab.

Nous nous sommes orientés vers le développement d'une solution logicielle libre appelée LODEX (Gregorio et al., 2019). Par rapport aux outils similaires, cet outil se concentre sur trois priorités : "masquer" la complexité des triplets au format RDF, donner envie de structurer son information en augmentant les données (visualisation, interconnexion, etc.) et faciliter la mise à jour ou l'ajout d'information sans refaire un long processus de publication. C'est un logiciel libre dont le code source est accessible sur GitHub [19].

Suite à l'import d'un fichier, l'outil génère automatiquement un URI (Uniform Resource Identifier), identifiant requis pour le web sémantique. Par défaut, l'outil LODEX crée un uid (Unique Identifier). Dans une des étapes de "stylage", l'outil permet de renseigner la propriété ou prédicat des triplets (un triplet RDF est composé de trois parties : sujet - prédicat - objet). La saisie y est facilitée par auto-complétion avec les différentes ontologies présentes dans le Linked Open Vocabularies [20] (LOV). De même, il permet de :

- Annoter un autre champ : par exemple pour préciser la source d'une définition,
- Composer ce champ : au sens du web sémantique, à partir de plusieurs champs. Par exemple, une adresse est composée d'un nom de rue, d'une ville, d'un pays.
- Après curation et sémantisation, le jeu de données est exposé sur le web. Différents exports aux formats du web sémantique (Turtle pour sa lisibilité ; N-Quads et N-Triple pour leur simplicité et JSON pour son application courante dans le web), permettent d'alimenter un triplestore.

Après curation et sémantisation, le jeu de données est exposé sur le web. Différents exports aux formats du web sémantique (Turtle pour sa lisibilité ; N-Quads et N-Triple pour leur simplicité et JSON pour son application courante dans le web), permettent d'alimenter un triplestore.

## 3 – EXPOSITION DES DONNÉES - LES RÉSULTATS

L'outil LODEX a permis d'agréger de manière cohérente toutes les données extraites du fonds ISTEX et de les sémantiser. Les données sont publiées sur le Web, dans le respect des standards du web sémantique, sous une licence ouverte. L'agrégation mène aussi à un SPARQL endpoint [21] contenant un graphe global des données ISTEX. Cette approche par profil se différencie des autres approches, en permettant de créer un graphe progressivement, le résultat permettant de faire des requêtes spécifiques sur les données ISTEX structurées en graphe.

### 3.1 DATA.ISTEX.FR : la publication sur le web

La page d'accueil de data.istex.fr permet d'accéder à l'ensemble des jeux de données et à chaque ressource ou élément constituant un jeu de données.

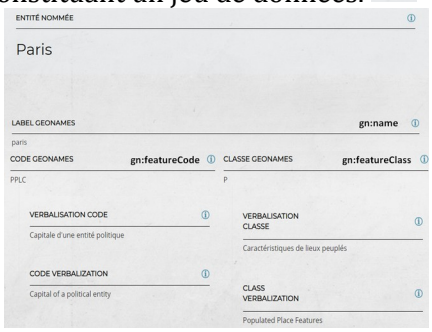


Figure 5 : Mise en évidence des enrichissements geonames

Suite à l'opération de désambiguïsation nous avons utilisé la puissance des requêtes Sparql afin de proposer des enrichissements dynamiques à partir de l'identifiant geonames caractérisant un lieu géographique. Pour ce faire, nous avons généré une requête Sparql permettant d'interroger un triplestore distant et d'afficher dynamiquement les résultats dans data.istex.fr (figure 6).

Après analyse des catalogues concernant les données géographiques accessibles via un triplestore, notre choix s'est arrêté sur :

- Wikidata [22] qui présente la particularité de contenir des informations générales et des liens vers d'autres bases de données
- Data.bnf [23] qui dispose de différents vocabulaires qui sont eux-mêmes alignés afin de faciliter leur réutilisation sur le Web de données.
- Loterre [24] (Linked Open TERminology REsources) est une plateforme développée par l'Inist, destinée à l'exposition, l'interrogation et le téléchargement de ressources terminologiques dans divers domaines comme la géographie.

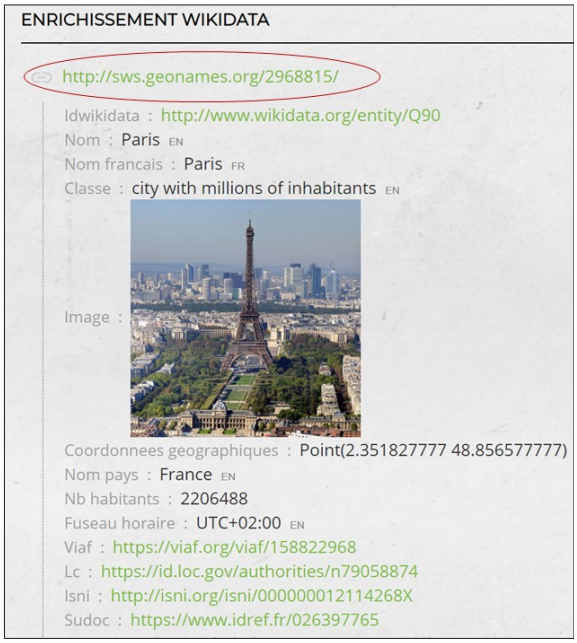


Figure 6 : Illustration d'enrichissement dynamique à partir de wikidata

Par conséquent, grâce à la certitude des identifiants geonames issus des travaux de désambiguïsation, nous pouvons proposer aux usagers de data.istex des enrichissements à la volée pour chaque lieu géographique.

## 3.2 -Pourquoi un triplestore

L'intérêt des triplestores est que nous avons un modèle connu, public et publié, de représentation de l'information, ce qui permet d'interroger des triplestores différents avec des procédures identiques.

L'ontologie istex a permis de construire un graphe global qui est accessible par un sparql endpoint agrégeant toutes les données produites et structurées. Les notices se trouvant dans le triplestore sont constituées par l'apport de différentes sources d'informations. Tout d'abord des informations de niveaux bibliographiques en provenance du réservoir ISTEX puis par des informations de niveau enrichissement. A ce jour, plus de 490 millions de triplets sont contenus dans le triplestore.

L'usage d'un sparql endpoint permet de :

- Bénéficier de la puissance de requêtage sparql en allant plus loin que la simple recherche booléenne,
- Agréger toutes les données produites avec d'autres réservoirs telles que celles contenues dans le triplestore de l'abes, data.bnf et d'autres ...,
- Rendre les données interopérables (alignements, visibilité).

Par ailleurs, on peut aussi interroger les triplestores via des Endpoint Sparql. Cela permet par exemple, de déterminer quels sont les documents ISTEX qui parlent d'un lieu (placeName) en Meurthe et Moselle ? Cette interrogation est impossible à réaliser directement dans l'API ISTEX car le lieu géographique "Meurthe et Moselle" n'est pas forcément présent dans les documents, néanmoins grâce aux différents alignements et à la structuration de l'information en RDF, cette requête en langage Sparql est rendue possible :

```
PREFIX wdt:http://www.wikidata.org/prop/direct/
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?placename ?geonames?doc
WHERE {
SERVICE <https://query.wikidata.org/sparql>
{
        SELECT ?geonames
    WHERE {
    ?wdUri wdt:P1566 ?geonamesID.
    ?wdUri wdt:P131 wd:Q502784.
    BIND(iri(concat("http://sws.geonames.org/",(?geonamesID),"/")) as ?geonames).
}
}
        ?placename skos:exactMatch ?geonames.
    ?doc ixtex:extractedPlace ?placename
}
```

Figure 7 : Exemple de requête

Ce simple exemple nous démontre que les informations structurées en RDF et stockées dans notre triplestore nous permettent d'aller au-delà du simple accès au document de l'archive ISTEX.

## CONCLUSION

Nous avons exposé, dans cet article, une méthode de désambiguïsation et d'alignement automatique de toponymes avec la ressource geonames. Cette méthode par apprentissage non supervisé faisant intervenir des méthodes d'apprentissage profond, a l'avantage de ne pas requérir de ressources d'étiquetage. Produire des corpus étiquetés manuellement, de qualité et de volumétrie suffisante est très coûteux mais également très complexe quand on se trouve dans des cas très généralistes comme ici.

Nous avons, dans nos expérimentations, utilisé le texte intégral, mais nous avons expérimenté la méthode avec succès dans le cas de paragraphes ou même de phrases. Ainsi, si dans un texte une même entité a plusieurs sens, nous pouvons tout à fait les détecter.

Cette méthodologie pourrait être appliquée à d'autres types de données, comme par exemple des noms de personnes.

Grâce à cette méthode de désambiguïsation et d'alignement automatique de toponymes, le double objectif est donc de pouvoir générer de nouveaux enrichissements à partir du fonds ISTEX et de développer une méthodologie didactique afin de les valoriser via le web de données ou Linked Open Data.

L'originalité de nos travaux est de rendre les données extraites d'ISTEX visibles, de les valoriser en respectant les normes du web sémantique, ouvrant ainsi ce fonds documentaire au web. Ce réseau sémantique a pour but de centrer les utilisateurs non plus sur le document mais la donnée elle-même.

Nous envisageons maintenant d'apprendre sur de gros volumes de données, par exemple les 15 millions de documents en anglais du réservoir ISTEX.

## NOTES

[1] http://www.istex.fr/

[2] http://tln.li.univ-tours.fr/Tln_Istex.html

[3] http://data.istex.fr

[4] https://www.geonames.org/

[5] https://github.com/tmikolov/word2vec

[6] https://github.com/stanfordnlp/GloVe

[7] https://fasttext.cc/

[8] https://github.com/explosion/sense2vec

[9] http://alt.qcri.org/semeval2019/

[10] https://competitions.codalab.org/competitions/19948

[11] https://demo.istex.fr

[12] https://data.istex.fr/sparql/

[13] https://www.w3.org/RDF/

[14] https://data.istex.fr/ontology/istex/

[15] https://www.w3.org/2009/08/skos-reference/skos.html

[16] http://dbpedia.org/ontology/

[17] http://www.geonames.org/ontology/documentation.html

[18] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[19] https://github.com/Inist-CNRS/lodex

[20] https://lov.linkeddata.es/dataset/lov

[21] https://data.istex.fr/triplestore/sparql

[22] https://query.wikidata.org/

[23] https://data.bnf.fr/

[24] https://www.loterre.fr/

## REFERENCES BIBLIOGRAPHIQUES


Bartunov S., Kondrashkin D., Osokin A., & Vetrov D. (2016). Breaking Sticks and Ambiguities with Adaptive Skip-gram. *19th International Conference on Artificial Intelligence and Statistics, Mai 2016,* p. 130–138, Cadiz, Spain.

Berners-Lee T., Hendler J., & Lassila O. (2001). The semantic web. *Scientific american*, 2001, vol. 284, n° 5, p. 34–43.

Bougriou N. & Soumia M. (2016), *Désambiguïsation des toponymes guidée par des ressources géographiques*, thèse, Université Abdelhamid Mehri Constantine 2, 2016, 86 p.

Bouzidi L. & Boulesnane S. (2017). Les humanités numériques. *Les Cahiers du numérique*, vol.13, n° 3, p. 19-38.

Chihaoui A., Bouhafs A., Roche M., & Teisseire M., Désambiguïsation des entités spatiales par apprentissage actif, *Revue Internationale de Géomatique*, vol.28, n° 2, p. 163–189

Collignon A. & Cuxac P. (2017). ISTEX : des enrichissements au web de données. *I2D Information, données documents*, vol. 54, n° 4, p. 8-15.

Cuxac P. & Thouvenin N. (2017). Archives numériques et fouille de textes : le projet ISTEX. *Atelier TextMine, Conférence EGC'17*, Grenoble.

Foppiano L. & Romary L., (2018) Entity-Fishing: A DARIAH Entity Recognition and Disambiguation Service, *Digital Scholarship in the Humanities* , Septembre 2018, Tokyo, Japan.

Gregorio S., Collignon A., Parmentier F. & Thouvenin N. (2019). LODEX : des données structurées au web sémantique. *Atelier Web des Données, Conférence EGC'19*, Janvier 2019, Metz.

Harlow C. (2015). Data Munging tools in preparation for RDF: Catmandu and LODRefine. *Code {4}lib journal*, 2015, vol. 30, p. 1-12.

Iacobacci I., Pilehvar M.T., & Navigli R., (2016) Embeddings for Word Sense Disambiguation: An Evaluation Study, *54th Annual Meeting of the Association for Computational Linguistics*, 2016, vol. 1, p. 897–907.

Karl J. W. (2018). Mining location information from life- and earth-sciences studies to facilitate knowledge discovery. *Journal of Librarianship and Information Science, Février 2018, vol.51, n° 4, p. 1007-1021*

Le Bœuf P. (2013). Customized OPACs on semantic Web : the OpenCat prototype. *IFLA Satellite Meeting 2013*, Singapore.

Martineau C., Tolone E. & Voyatzi S., (2007) Les Entités Nommées : Usage et Degrés de Précision et de Désambiguïsation, *26ème Colloque International Sur Le Lexique et La Grammaire* , Bonifacio, France, 2007, p. 105-112

Mena B. Habib & Van Keulen M.,(2011) Named Entity Extraction and Disambiguation: The Reinforcement Effect, *Proc. of MUD*, Seattle, USA, 2011, p. 9-16

Mikolov T., Chen K., Corrado G. & Dean J., (2013) Efficient Estimation of Word Representations in Vector Space, ArXiv:1301.3781, disponible sur : http://arxiv.org/abs/1301.3781 (18/09/2019)

Sallaberry, C. (2013). *Geographical Information Retrieval in Textual Corpora*. John Wiley & Sons, 2013, 133p.

Szabados A.-V. Letricot, R. (2012) L'ontologie CIDOC CRM appliquée aux objets du patrimoine antique. *3e Journées d'Informatique et Archéologie de Paris*, Juin 2012, Paris, France.

Vandenbussche P.-Y., Vatant B. & Charlet J. (2012). Linked Open Vocabularies, un écosystème encore fragile. *Atelier Qualité & Robustesse, 23èmes journées francophones d'ingénierie des connaissances*. 2012, Paris, p. *1-13.*