Webinaire Découverte du TDM

Fabienne KETTANI-SCHMITTHEISLER
Service Text et Data Mining

ANALYSER & FOUILLER L'INFORMATION SCIENTIFIQUE

1. Généralités sur le TDM
2. Cas d'usage
3. Perspectives INIST

cnrs Inist

1 Généralités sur le TDM

Qu’est-ce que le TDM ?

Les évolutions du TDM

Les enjeux du TDM

L’Inist-CNRS et le TDM

Qu’est-ce que le TDM ?

La fouille de textes

Ensemble des méthodes et des traitements informatiques qui consistent à analyser le sens des textes en langage naturel pour en donner une représentation utilisable par les humains et les ordinateurs.

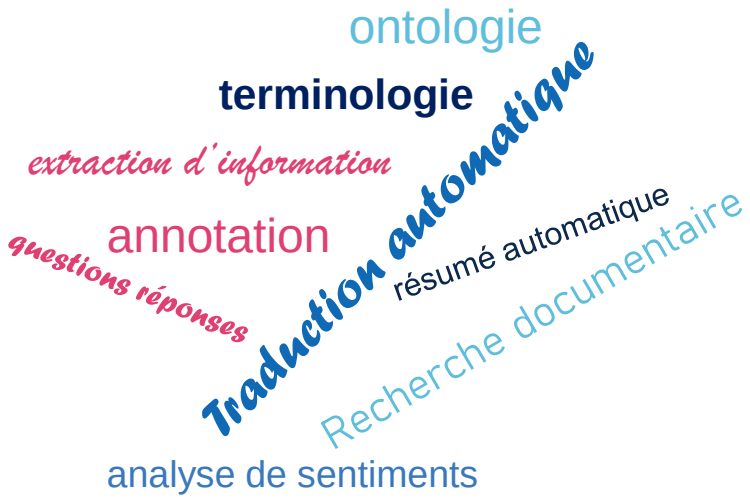
Données → Connaissances

C'est une spécialisation de la fouille de données (data mining) qui fait appel aux méthodes de l'**Intelligence Artificielle** [1], du **Traitement Automatique des Langues** et des **Statistiques**

[1] L'**apprentissage profond** ou **apprentissage en profondeur** (en anglais : ***deep learning***, ***deep structured learning***, ***hierarchical learning***) est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur, du traitement automatisé du langage

La fouille de textes: des technologies qui nous accompagnent déjà largement au quotidien…

- Filtrage de spam
- Recommandations
- Assistant personnel
- Service client, agent conversationnel
- Intelligence économique
- Intelligence stratégique
- Sécurité
- Gestion documentaire
- Assistance au diagnostic médical
- Recherche scientifique
- etc.



Les évolutions du TDM

CONTEXTE


Révolution numérique
Maturité des technologies
Prise de conscience politique
Evolution juridique

Nous ne sommes plus en capacité d'absorber la quantité d'information disponible…

Révolution numérique

De l'**infobésité** galopante (il y a peu) au **déluge** d'informations (aujourd'hui)

Ere du **Big Data**: les 3V : Volume, Vélocité et Variété

Le phénomène Big Data s'amplifie si vite que l'on n'arrive plus à suivre l'évolution des nouvelles unités de mesure : les **exaoctets** (10^{18}octets), les **zettaoctets** (10^{21}), les **yottaoctets** (10^{24})....

> 180 zettaoctets en 2025

Publications scientifiques

50% *des articles ne sont* ***jamais lus***
90% *des articles ne sont* ***pas cités***

ANF TDM 2020/ R. Bossy & C. Nédellec

Image par mcmurryjulie de Pixabay

…mais nous disposons de technologies de plus en plus performantes

Maturité des technologies

30 ans d'expérience en **TAL et IA** (cf ChatGPT…), en partie majorée par l'**implication d'industriels** qui y trouvent un intérêt majeur (analyse de sentiments, de tendances, détection de buzz etc.)

Augmentation très importante de la **puissance de calcul et de stockage** en 40 ans

Evolution majeure des algorithmes: **statistiques versus apprentissage profond**

Le TDM va s'inscrire dans la politique de **Science ouverte**…

On cherche à s'affranchir de la mainmise des éditeurs scientifiques sur les publications et les données de la science et à permettre une meilleure reproductibilité de la recherche.

Prise de conscience politique

2001 

Budapest Open Initiative: problématique du libre accès aux **publications scientifiques** et incitation à l'utilisation des archives ouvertes ou des revues en libre accès, prise de conscience des besoins en licences adaptées

2003 

Déclaration de Berlin: extension de l'ouverture aux données de recherche

(. . .)

Rapport Villani sur l'IA *« Favoriser sans attendre les pratiques de fouille de données (TDM) » (page 35)*

2018 

1er plan national pour la Science Ouverte - **Frédérique Vidal - MESRI**: *« La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens sans entrave, sans délai, sans payement »*

5 M€ /an

2019 

Le Grand Débat: le TDM devient une « réalité publique » https://iscpif.fr/chavalarias/?p=1495

Feuille de route pour la Science Ouverte du CNRS

Engagement des universités: politiques et interlocuteurs désignés pour la science ouverte

2021 

2e plan national pour la Science Ouverte (2021-2024) *« Transformer les pratiques pour faire de la science ouverte le principe par défaut ».* **Objectif:** 100% de publications en accès ouvert en 2030

15 M€ /an

2022 

Plateforme Recherche Data Gouv

… et bénéficier des dispositions légales qui sont prises

Evolution juridique

2016

Loi pour une République numérique:

L'article 38 : **Exceptions au code de la propriété intellectuelle**

Conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche publique."

Introduction d'une **exception au droit d'auteur** ainsi qu'une **exception au droit *sui generis* des producteurs de bases de données**

2019

Directive européenne sur le droit d'auteur et les droits voisins dans le marché unique du numérique ou Directive « Copyright »:

Les **articles 3 et 4 de la directive**, portent sur la "fouille de textes et de données à des fins de recherche scientifique" ; la pratique du TDM (Text and Data Mining). Ces articles prévoient une exception au droit d'auteur "pour les reproductions et les extractions effectuées par des organismes de recherche et des institutions du patrimoine culturel, en vue de procéder, à des fins de recherche scientifique, à une fouille de textes et de données sur des œuvres ou autres objets protégés auxquels ils ont **accès de manière licite**"

2021

Ordonnance de transposition en droit français de la Directive européenne sur le droit d'auteur:

https://www.vie-publique.fr/loi/282569- _ordonnance-completant-transposition-directive- _droits-dauteur

" L'ordonnance consacre ou adapte tout d'abord des **exceptions au droit d'auteur et aux droits voisins** afin de favoriser la **fouille de textes et de données**, l'utilisation d'extraits d'œuvres à des fins **d'illustration dans le cadre de l'enseignement** et la reproduction des œuvres dans un souci de conservation du patrimoine culturel."

2022

Décret n°2022-928 du 23 juin 2022:

https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000045960058

Ce décret fait suite à l'ordonnance du 24 novembre 2021 ci-dessus. Il introduit des modifications du code de la propriété intellectuelle et formalise les **modalités d'application de l'exception en vue de la fouille de textes** et de données (conditions de détention des copies numériques nécessaires à la fouille de textes entre autres)

Quand le droit et la politique s'allient…

Budapest Open Initiative

Déclaration de Berlin

Rapport Villani sur l'IA

1er plan national pour la Science

Le Grand Débat

Feuille de route pour la Science Ouverte du CNRS

5 M€ /an

Engagement des universités

Loi pour une République numérique

Directive européenne sur le droit d'auteur

2e plan national pour la Science Ouverte (2021-2024)

15 M€ /an

Ordonnance de transposition en droit français de la DE

Plateforme Recherche Data Gouv

2001 2003 (. . .) 2016 2018 2019 2021 2022

Décret n°2022-928 du 23 juin 2022

Les enjeux du TDM

SCIENTIFIQUES

Le TDM permet une exploitation et une **réutilisation des produits de la recherche** (publications, ressources sémantiques…)

En tant que tel il contribue à l'**accélération de l'innovation**

… et à **répondre à des questions de recherche** (confirmer des hypothèses posées)

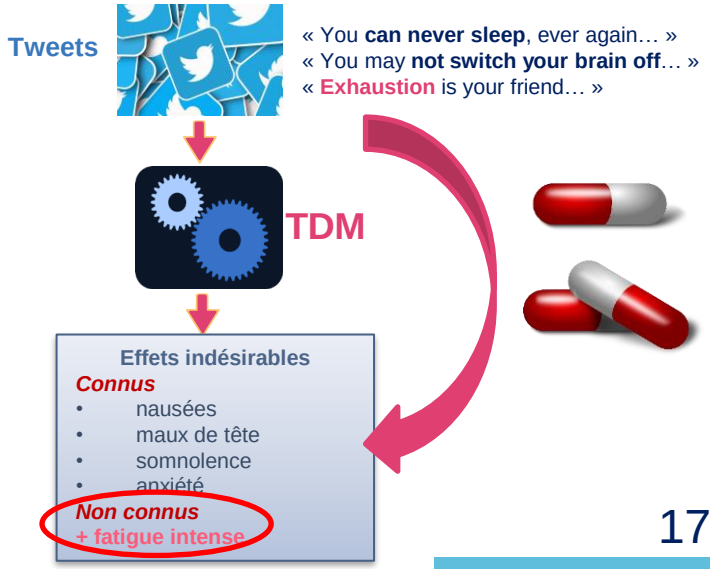
Un challenge: amener le TDM au cœur de l'activité du chercheur

SCIENTIFIQUES

EXEMPLE D'APPLICATION

PHARMACOVIGILANCE

par analyse automatique de tweets pour extraire de la connaissance sur les effets secondaires de l'utilisation des médicaments



O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G.
Pharmacovigilance on twitter? Mining tweets for adverse drug reactions.
AMIA Annu Symp Proc. 2014 Nov 14;2014:924-33. PMID: 25954400; PMCID: PMC4419871.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419871/

METIERS ET COMPETENCES

Le TDM fait appel à de **nouvelles compétences et de nouveaux métiers** (développeurs informatiques spécialisés, ingénieurs de la connaissance…)

Et nécessite de mettre en place de nouveaux parcours de **formation**, à compétences multiples.

Notion d'**accès licite** aux documents / ***Directive européenne***

Principes FAIR (Findable Acessible Interoperable Reusable)

Attention aux biais !! : choix des données, interprétation, etc.

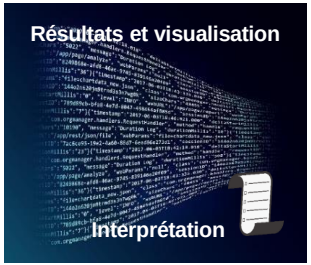
ETHIQUES

TDM

Transparence

Fiabilité

Reproductibilité



Protection des droits (droits d'accès, données personnelles et vie privée…)
Fournisseur: attention aux
Conflits d'intérêt
Exhaustivité (bruit et silence – ce qui n'est pas traité est autant un biais que ce qui est inutile
Fiabilité
Sécurité (stockage)

DES DIFFICULTES ET DES SOLUTIONS

Le TDM repose sur:

- l'exploitation de **texte**
- des traitements automatiques du **langage naturel**
- des traitements informatiques basés sur des outils d'**intelligence artificielle**

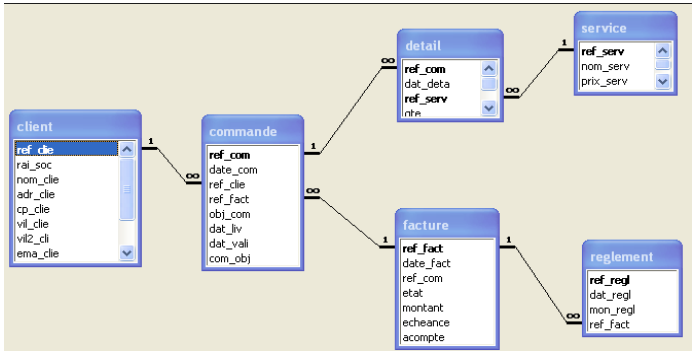
Le texte est une donnée mais avec des caractéristiques spécifiques…

Le texte est une **donnée non structurée**

Un ordinateur interprète de la **donnée structurée**

*« Vous trouverez par la présente le courrier de M. Durand qui **honore le règlement** de sa commande du 22 mai 2019 au sujet de l'achat d'une caisse de 12 bouteilles de Bourgogne »*

QUESTION: la facture de M. Durand est-elle payée ?



… et la langue est complexe

Pour interpréter et comprendre…

Paris	capitale de la France, ville US
ne… pas…	négation
Orange	couleur, fruit, société, ville
Labrador	hyperonymie (chien)
Boire un verre	métonymie

… s'appuyer sur le traitement de la langue

Multilinguisme

Alphabet : latin, cyrillique, grec, arabe, …

Le **découpage** des mots, des phrases, des paragraphes

La **graphie** des mots, leur genre et leur(s) catégorie(s) syntaxique(s)

La **syntaxe** : comment sont construites les phrases

La **sémantique** des mots: désambiguïsation

Stanford CoreNLP : http://corenlp.run/

Quelques techniques de TAL

« Comment transformez vous un document et son contenu en chiffres ? »

Tokenisation

« | Comment | transformez | vous | un | document | et | son | contenu | en | chiffres | ? | »

POS tagging
(Part Of Speech)

Lemmatisation
(forme canonique)

« Comment transformez vous un document et son contenu en chiffres ? »

transformer — chiffre

Stemming
(racinisation)

« Comment transformez vous un document et son contenu en chiffres ? »

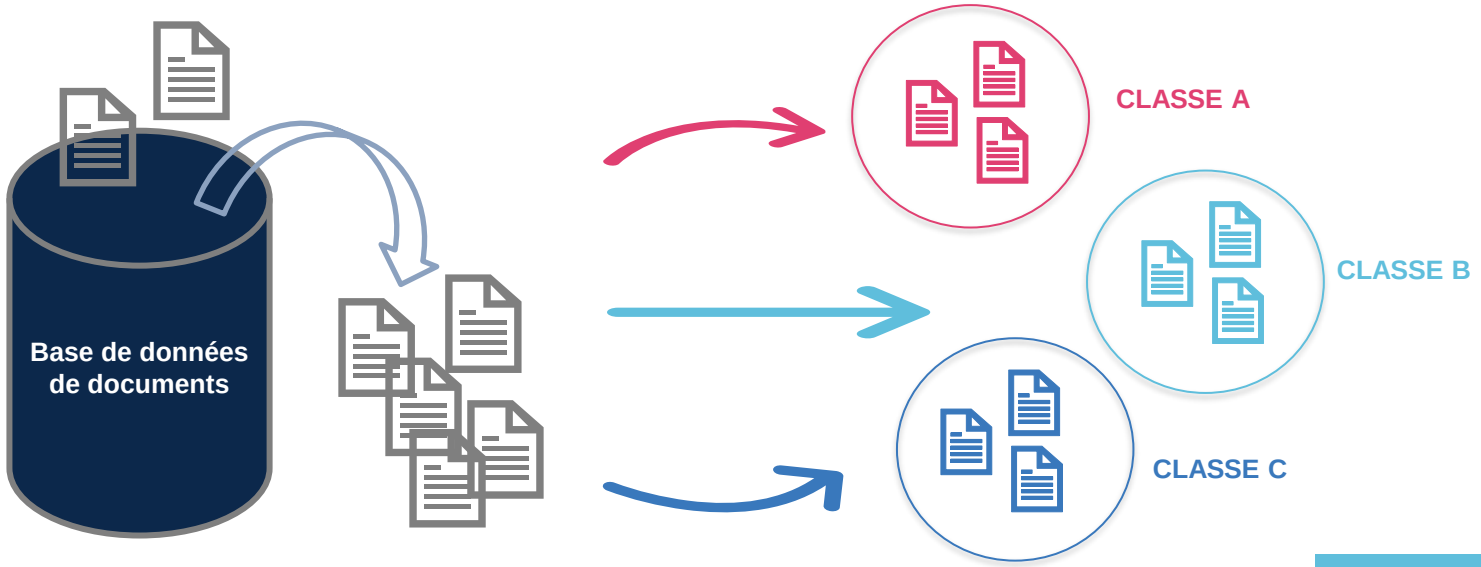
transform — docu — chiffr

Quelques techniques de TDM

CLASSIFICATION NON SUPERVISEE

Problématique

Classer les documents selon (par exemple):
- les thèmes de ces documents
- les zones géographiques considérées…



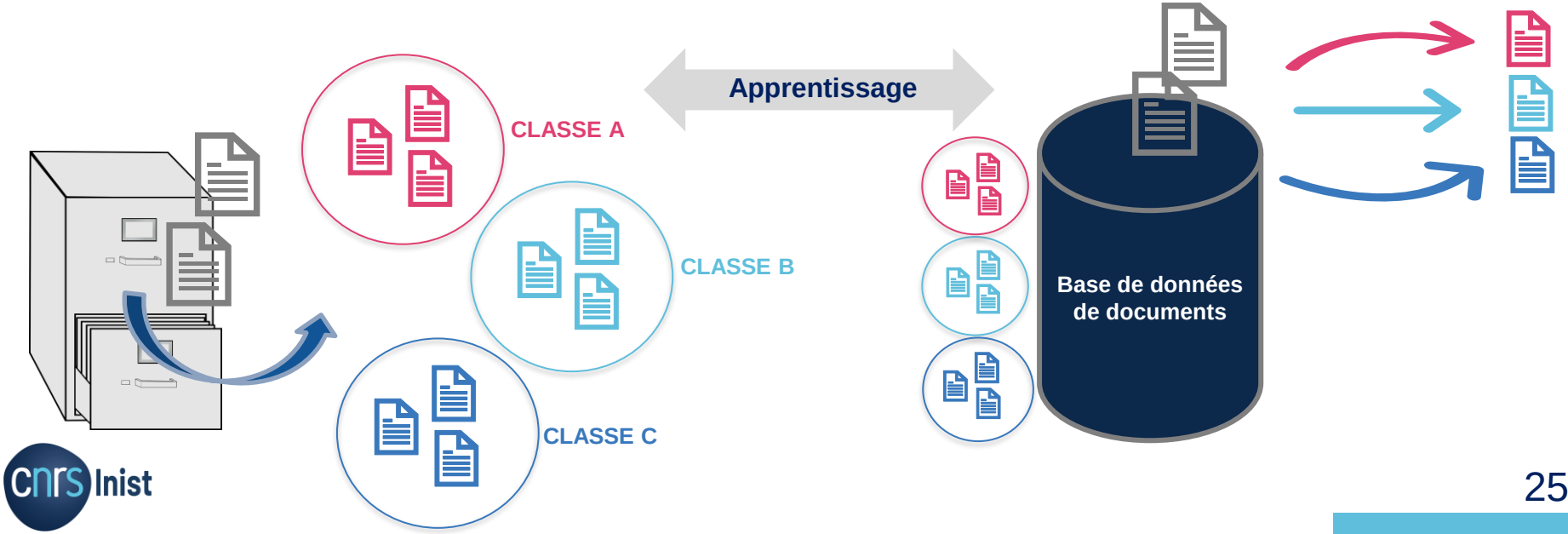
Quelques techniques de TDM

Problématique

Classer les documents selon (par exemple):
- les thèmes de ces documents
- les zones géographiques considérées…

CLASSIFICATION SUPERVISEE

Apprentissage sur données déjà labellisées

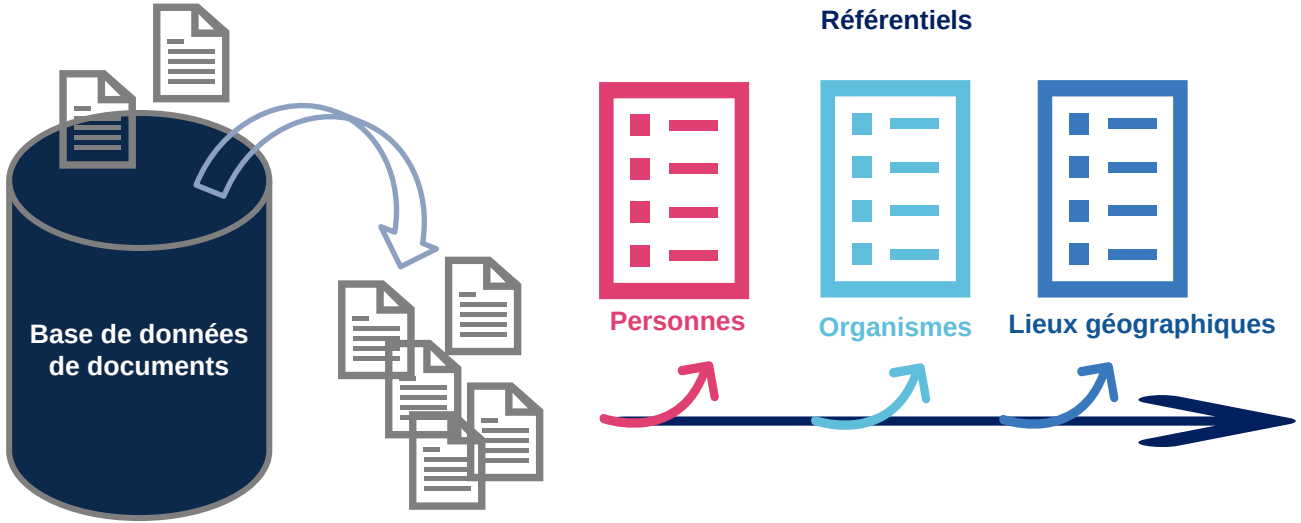


Quelques techniques de TDM

RECONNAISSANCE D'ENTITES NOMMEES

Problématique

Repérage de :
Personnes, lieux géographiques, institutions, sociétés, microorganismes…

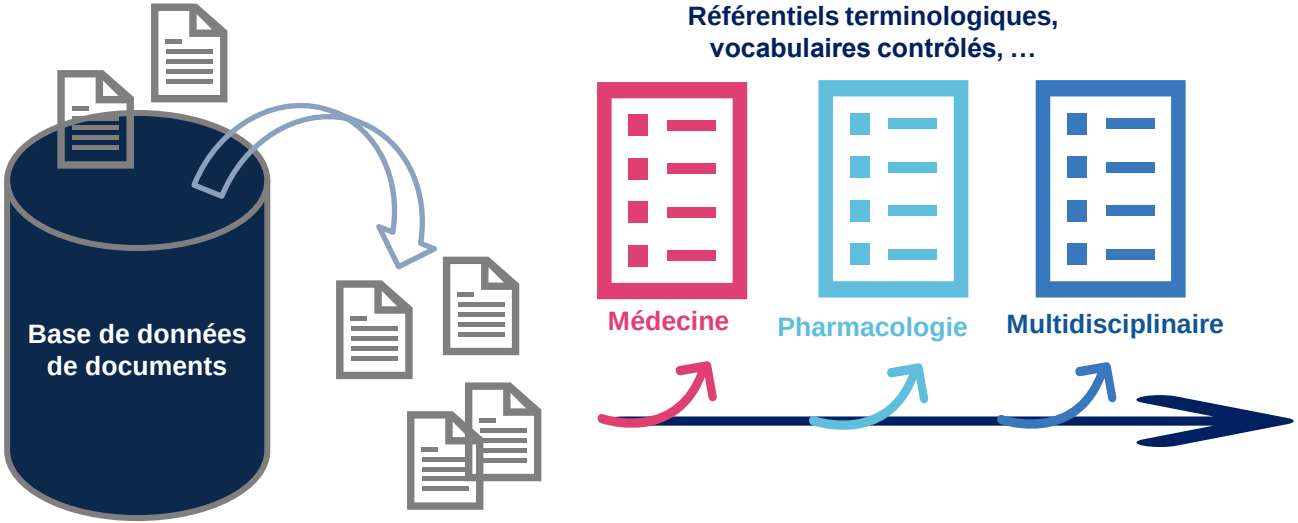


Quelques techniques de TDM

LIBRE ET/OU CONTROLEE

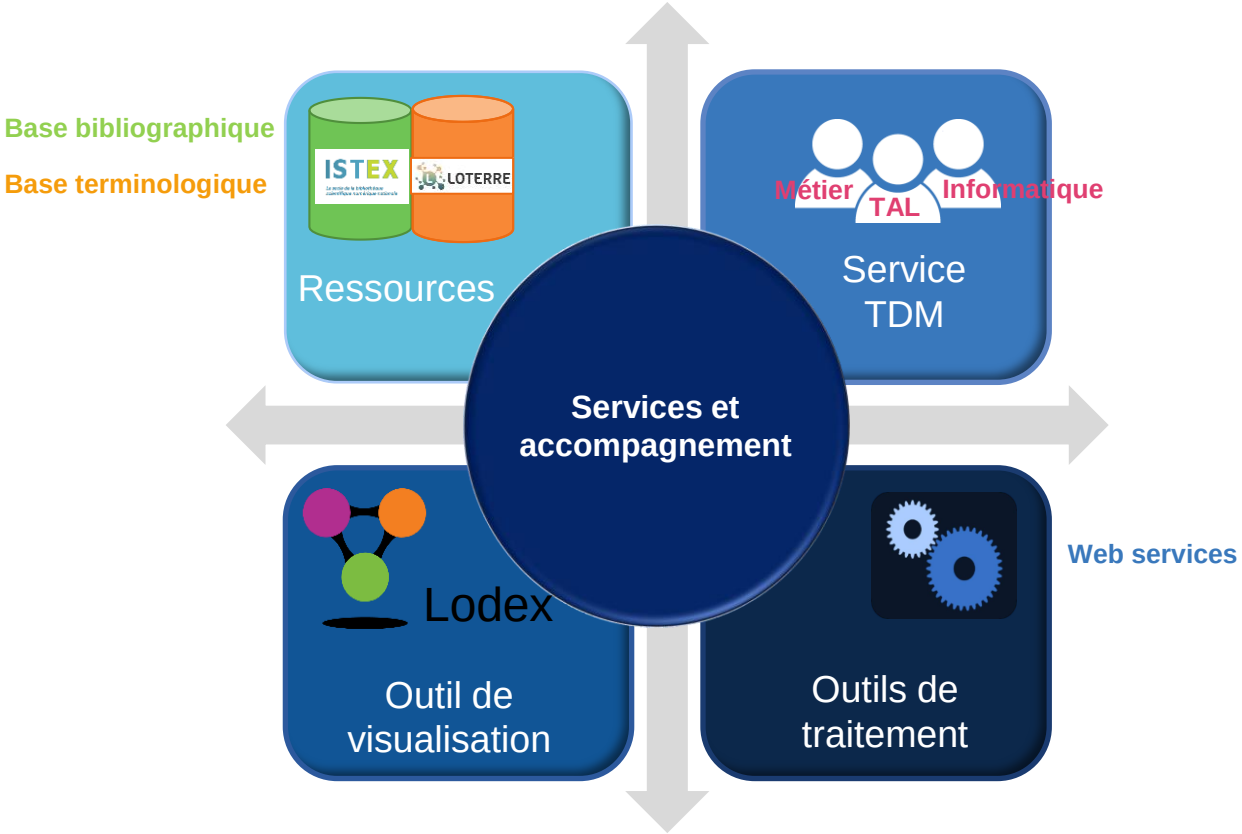
(contrôlée = par rapport à des référentiels)

Problématique Repérage de termes **caractérisant le document** et permettant de le retrouver ensuite au sein d'un corpus



L’ Inist- CNRS et le TDM

Quelles opportunités à l'INIST-CNRS ?



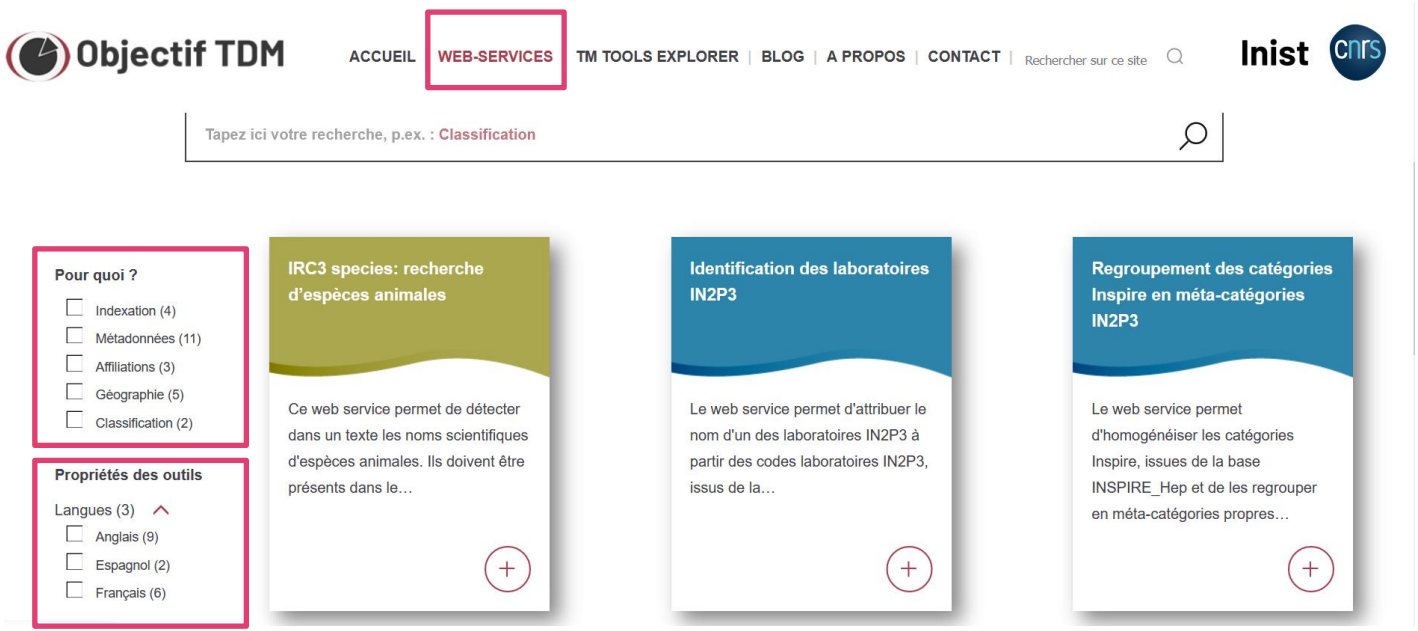
Des outils et services de TDM pour tous

Un site en ligne: OBJECTIF TDM

Des outils et services de TDM pour tous

Des web-services pour aider…

Web service (WS): interface et protocole d'échange en ligne de données
1 WS = 1 tâche



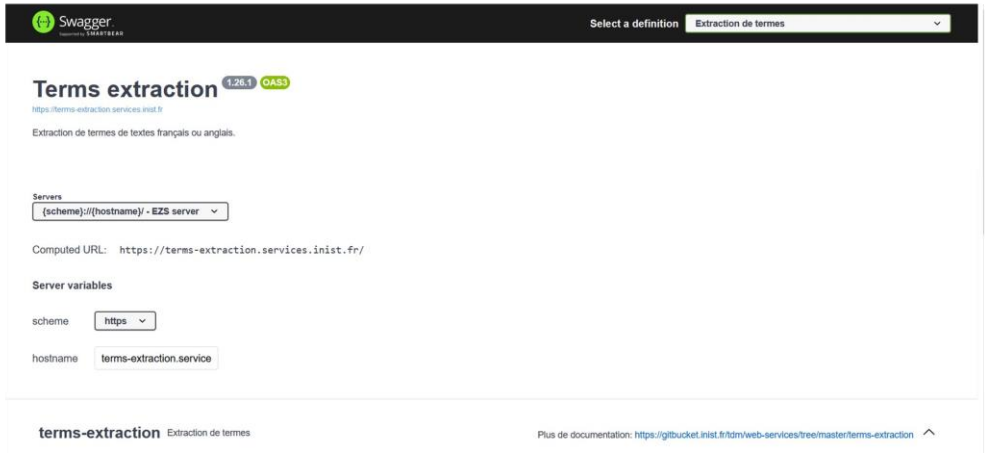
Des outils et services de TDM pour tous

Des web-services pour aider…

Utilisables par des néophytes via LODEX

URL DU WEB SERVICE à renseigner dans LODEX est :
https://terms-extraction.services.inist.fr/v1/teeft/en

…mais aussi en ligne de commande, appelés dans des programmes ou via des interfaces.



Des outils et services de TDM pour tous

Des web-services pour aider…

Extraction de termes d'un texte via Teeft

Le service web teeft extrait les termes les plus pertinents d'un texte en anglais ou en français. Il permet d'avoir…

Avant

"The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 20 January 2020, and later a pandemic on 11 March 2020. As of 2 April 2021, more than 129 million cases have been confirmed, with more than 2.82 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history."

Après

```
"severe acute respiratory syndrome coronavirus2",
"international concern",
"ongoing global pandemic",
"coronavirus disease",
"covid-19",
"december",
"wuhan",
"coronavirus pandemic",
"deadly pandemic",
"covid-19 pandemic"
```

Des outils et services de TDM pour tous

Des web-services pour aider…

Détection de la langue d'un texte

Le web-service detect-lang détecte la langue d'un document texte et renvoie le code langue et la probabilité correspondante. Dans le…

Avant

"User experience design (UXD, UED, or XD) is the process of supporting user behavior[1] through usability, usefulness, and desirability provided in the interaction with a product.[2] User experience design encompasses traditional human-computer interaction (HCI) design and extends it by addressing all aspects of a product or service as perceived by users. Experience design (XD) is the practice of designing products, processes, services, events, omnichannel journeys, and environments with a focus placed on the quality of the user experience and culturally relevant solutions."

Après

"en"

Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils

cnrs Inist

Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils

TM TOOLS EXPLORER

Inist cnrs

Aujourd'hui

Version beta

Outils libres

A venir

V0 prochainement avec rajout de:

- Nouveaux outils
- Modèles de langues
- Algorithmes
- Compétences mises en jeu

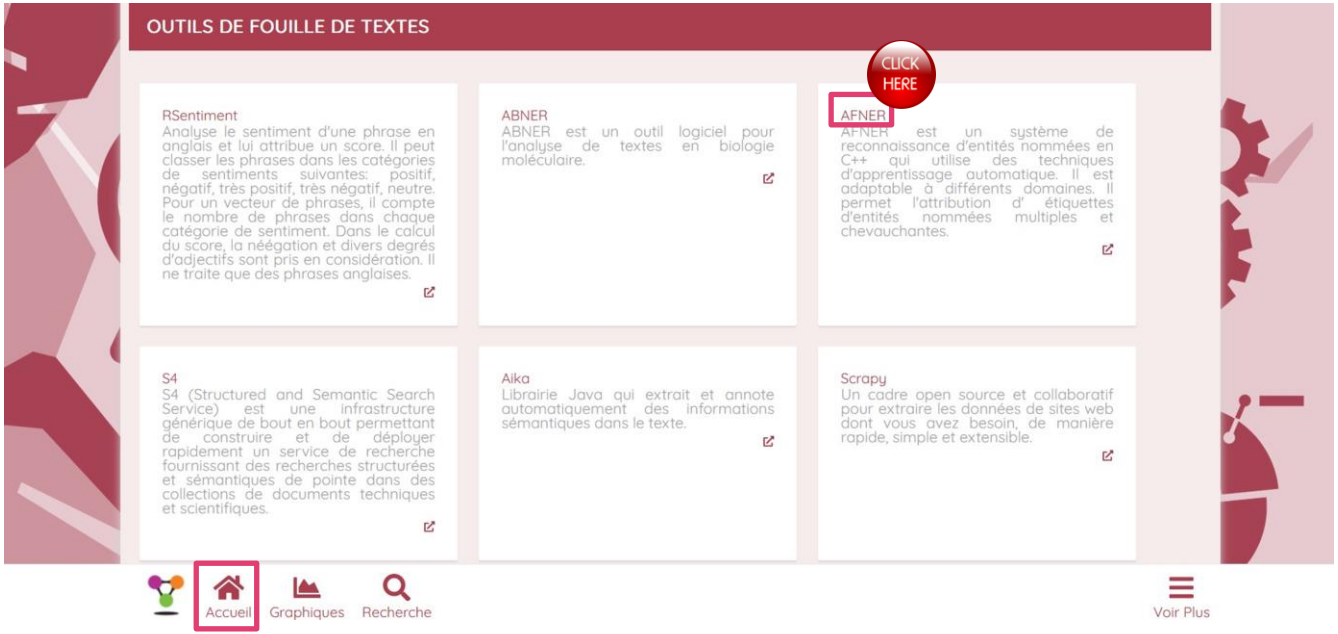
Version beta sur des **outils commerciaux**

Repose sur une **ontologie: OntoTM** qui sera publiée prochainement sur le portail terminologique Loterre

Réalisé avec Lodex

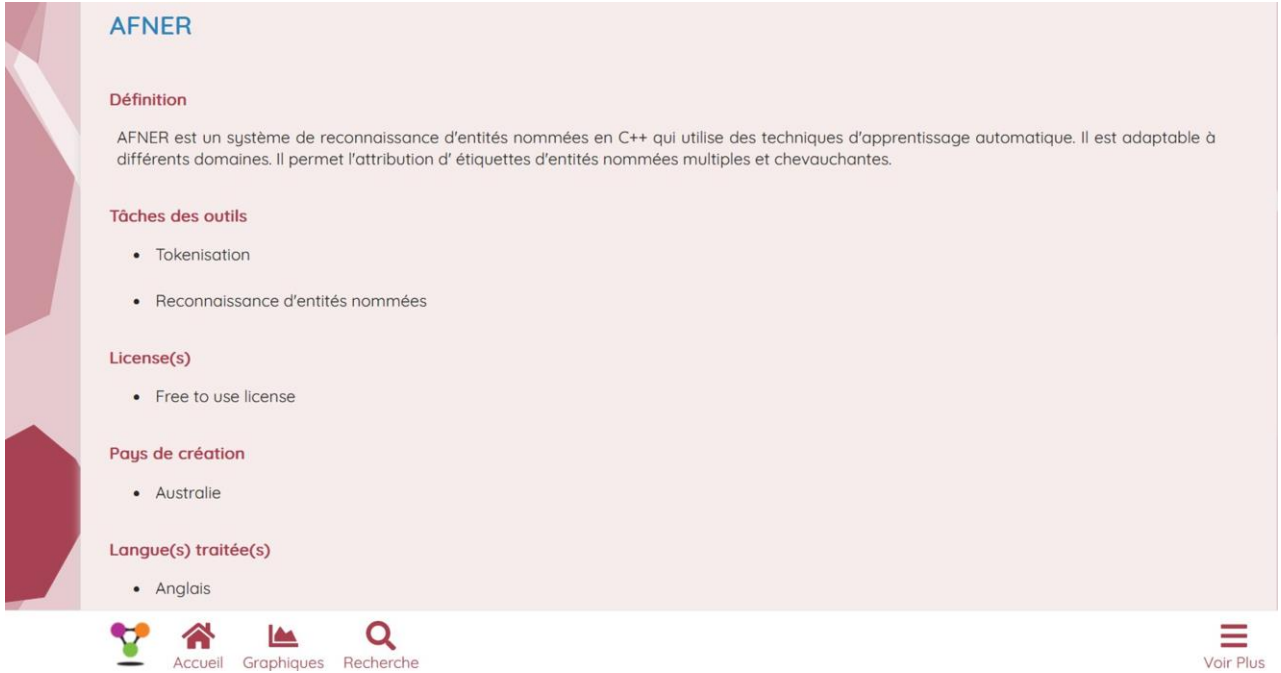
Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils



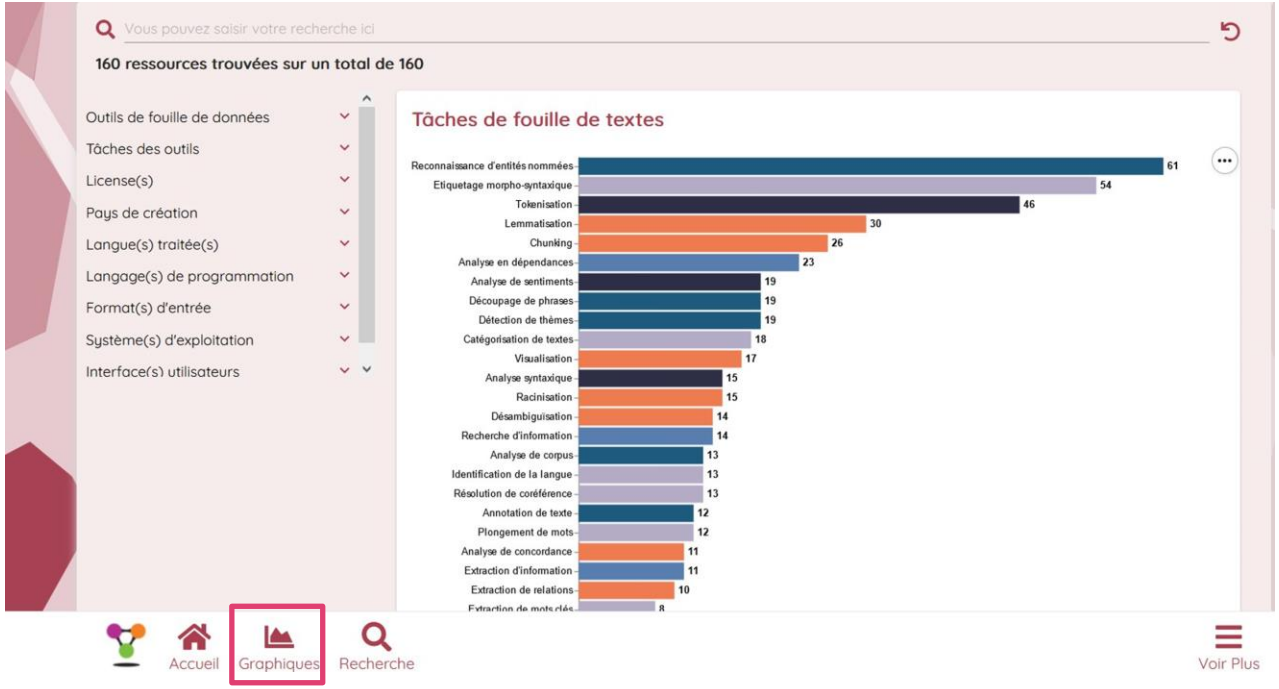
Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils



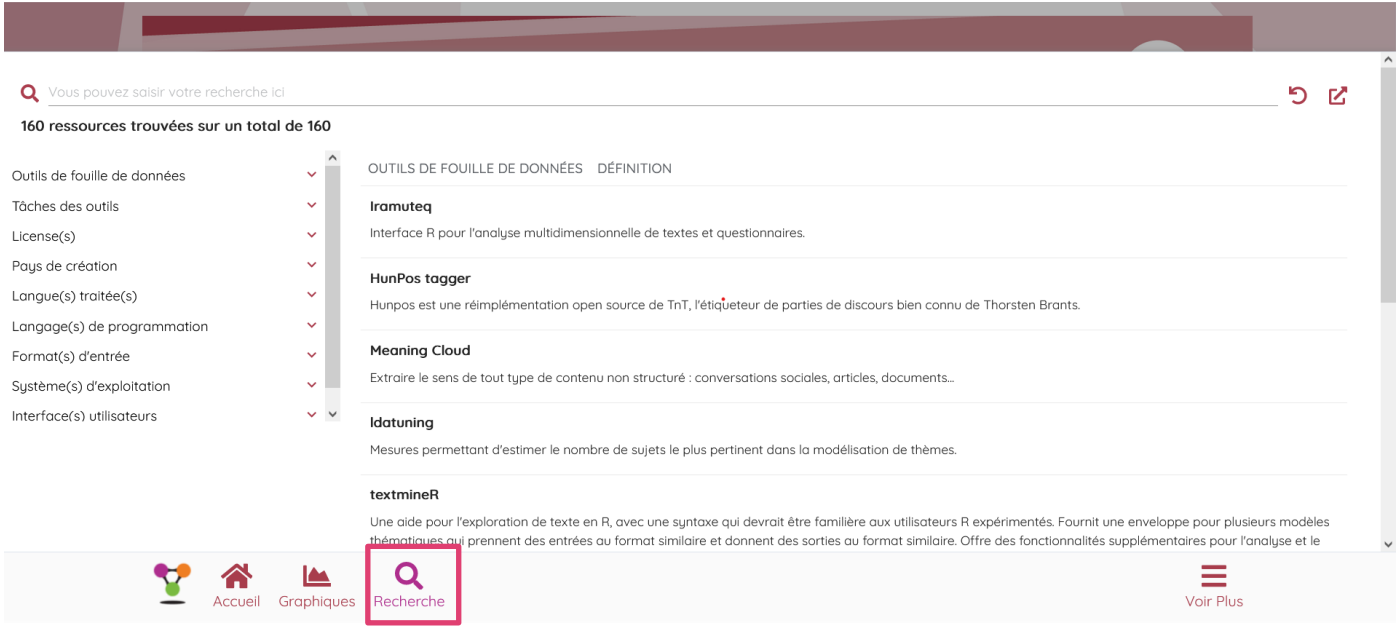
Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils



Des outils et services de TDM pour tous

… mais aussi un catalogue d'outils



Des outils et services de TDM pour tous

… et des news

2

Cas d’usage

CORPUS REFUGIES ISTEX

https://shs-collection.corpus.istex.fr/ark:/67375/QCT-32H73MLV-8

Répondre à 2 questions à partir d'un corpus de notices

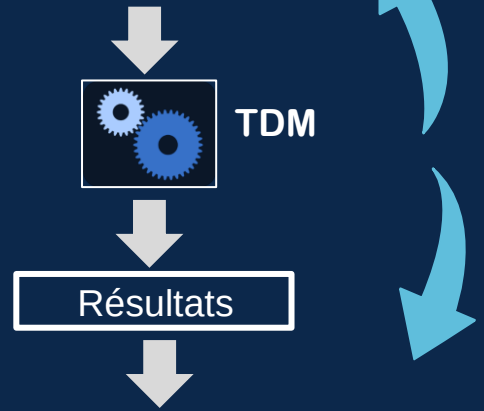
- Quelle part de réfugiés climatiques par rapport aux réfugiés politiques ?
- Quelle répartition géographique ? (visualisation par cartographie)

DEMARCHE TDM

Extraire un corpus

Nettoyer les données

(tenir compte des formats: pdf, txt, jpeg, xml, json, … et des encodages de données et les transformer si besoin)



Visualiser les résultats dans un outil dédié

Affiner et interpréter les résultats

DEMARCHE TDM DANS NOTRE CAS D'USAGE

Biblio ISTEX — Le socle de la bibliothèque scientifique numérique nationale

Requête Istex

Visualisation

Lodex

Chaine de traitement TDM

WS1

Extraction de termes d'un texte via Teeft

WS2

Associer un terme au vocabulaire Pays et Subdivision

Termino — LOTERRE

Corpus

Récupération des **termes géographiques**

Cartographie

Monde > Europe > Europe occidentale > France

Terme préférentiel	France
Concept(s) générique(s)	Europe occidentale
Concept(s) spécifique(s)	France d'outre-mer France métropolitaine
Synonyme(s)	250 FR FRA République française
Code ISO 3166-1-alpha-2	FR
Code d'entité GeoNames	gnA.PCLI
Identifiant GeoNames	3017382
Capitale	Paris
Code Lodex placename	FRA
Code numérique UN	250
Code ISO 3166-1-alpha-3	FRA

Récupération des **codes pays**

Notice terminologique par pays avec **codes pays** (thésaurus **Pays et subdivisions**)

cnrs Inist

LES WS EN DETAILS

WS1

Le web-service **teeft** extrait les termes les plus pertinents d'un texte en anglais ou en français. Il permet d'avoir une idée de ce dont parle le texte. Idéalement, le texte doit contenir plusieurs paragraphes.

Extraction de termes d'un texte via Teeft

```
[{
  "id": "https://fr.wikipedia.org/wiki/Mars_Exploration_Rover",
  "value": "Mars Exploration Rover (MER) est une mission double de la NASA lancée en 2003 et composée de deux robots mobiles ayant
}]
```

```
[
  {
    "id": "https://fr.wikipedia.org/wiki/Mars_Exploration_Rover",
    "value": [
      "deux robots",
      "panneaux solaires",
      "mars exploration rover mer",
      "mission double",
      "deux robots mobiles"
    ]
  }
]
```

URL DU WEB SERVICE à renseigner dans LODEX est :

https://terms-extraction.services.inist.fr/v1/teeft/en

https://terms-extraction.services.inist.fr/v1/teeft/fr?nb=10

Par défaut teeft extrait **5 termes** mais on peut augmenter ce chiffre.

Le vocabulaire **Pays et Subdivision de Loterre** propose pour chaque pays et région française des concepts regroupant informations géographiques, variantes syntaxiques, acronymes, et formes normalisées.

LES WS EN DETAILS

WS2

Associer un terme au vocabulaire Pays et Subdivision

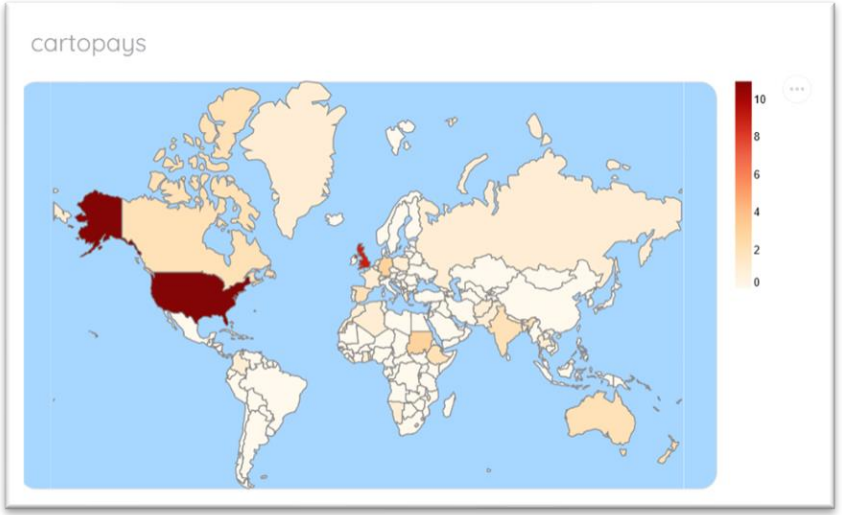
URL DU WEB SERVICE à renseigner dans LODEX est :
https://loterre-resolvers.services.inist.fr/v1/9SD/identify

Code pays

REPONSE AUX QUESTIONS

Visualisation des données dans Lodex

Utilisation des **facettes** pour faire varier les données cartographiques



→ Carto pour les **réfugiés politiques**

→ Carto pour les **réfugiés climatiques**

Lodex est un logiciel open source dédié à la valorisation de données structurées.
L'outil permet de créer des sites web offrant des interfaces pour explorer les données au travers une liste de fiches ou une série de graphiques dynamiques (histogrammes, cartes, diachronies, etc.)

ALLER UN PEU PLUS LOIN…

Comparaison indexation teeft et indexation par mots clés d'auteur

→ Carto pour les **réfugiés politiques**

→ Carto pour les **réfugiés climatiques**

teeft

mots clés d'auteur

A RETENIR

Pour faire du TDM il faut:

Des données

- sur lesquelles on a les droits adéquats
- qui sont « propres » (**GIGO**: Garbage in → Garbage out) et cela suppose toujours un travail conséquent de pré-traitement

Déterminer **un objectif**

Connaître un minimum **les outils et techniques/les ressources** pour utiliser les plus adaptés à l'objectif

Savoir **interpréter les résultats**

Le TDM n'est jamais qu'une **aide**

3 Perspectives INIST

EN PREVISION

- De nouveaux **web services**
- Des actions de **formation de type atelier**
- De nouvelles utilisations de Lodex (exploitation de **données issues de Zotero** – expérimentation en cours)

Nous suivre sur les réseaux sociaux:
Twitter
Facebook
LinkedIn
You tube
Fil d'actualités

Merci !

A votre écoute…

ANALYSER & FOUILLER
L'INFORMATION
SCIENTIFIQUE

fabienne.kettani@inist.fr