Introduction au TDM

Formation Cycl@doc

26 mars 2024

cnrs Inist | Institut de l'information scientifique et technique

Qu'est ce que le TDM

L'évolution du TDM

Les enjeux du TDM

Comment faire du TDM

L'INIST-CNRS et le TDM

Photo de Codioful (Formerly Gradienta) sur Unsplash

Qu’est ce que le TDM ?

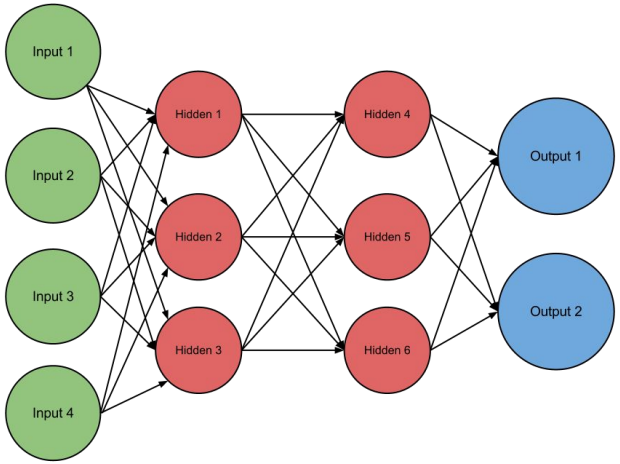
La fouille de textes

DEFINITION

Ensemble des méthodes et des traitements informatiques qui consistent à **analyser le sens de 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**, du **Traitement Automatique des Langues** et des **Statistiques**.

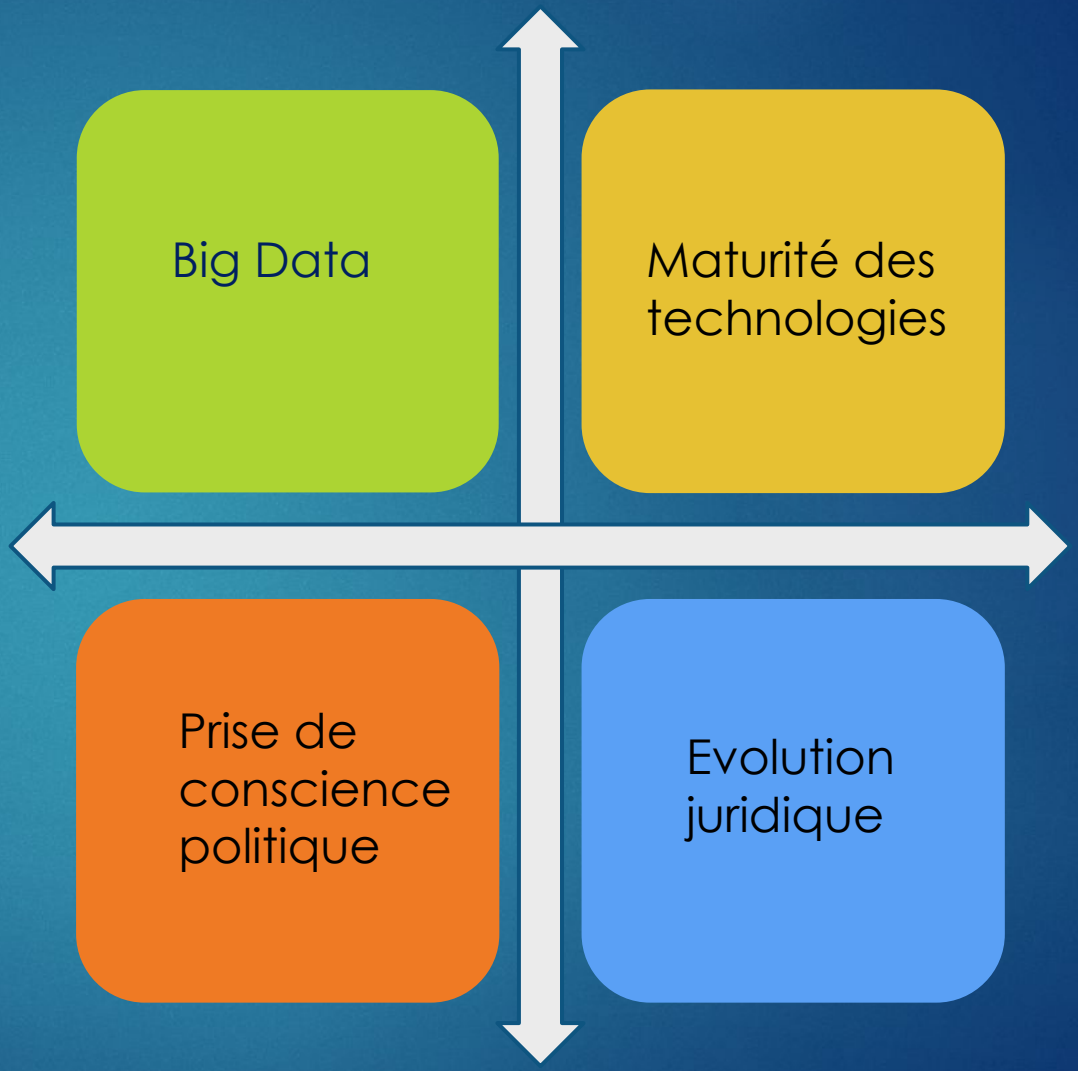


La fouille de textes : des technologies qui nous accompagnent déjà largement au quotidien…

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

L’évolution du TDM

CONTEXTE


Big Data
Maturité des technologies
Prise de conscience politique
Evolution juridique

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

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

Ere du **Big Data ;** les 3V : **V**olume, **V**élocité et **V**ariété

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 générée par bing

… mais nous disposons de technologies de plus en plus performantes

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

Évolution 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.

10

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 la recherche**

(…)

Rapport Villani sur l'I.A *« Favoriser sans attendre les pratiques de fouille de texte et de données (TDM) » (page 35)*

2018 

1er Plan national pour la Science ouverte - Frédérique VIDAL- MESRI **5 M € /an**

« 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.»

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): 15 M € /an

« Transformer les pratiques pour faire de la science ouverte le principe par défaut » 100% de publications en accès ouvert en 2030

2022 

Plateforme Recherche Data Gouv

… et bénéficier des dispositions légales qui sont prises

11

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

Loi pour une République numérique

Rapport Villani sur l'I.A

1er Plan national pour la Science ouverte

5 M € /an

Le Grand Débat

Feuille de route pour la Science ouverte du CNRS

Engagement des universités

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

Décret n° 2022 du 23 juin 2022

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

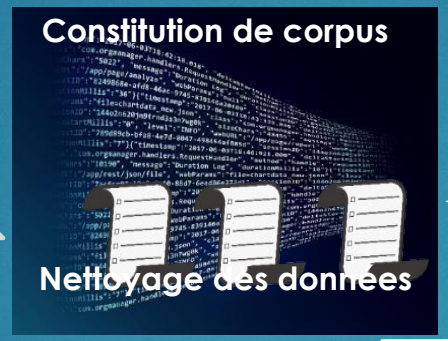
Les enjeux du TDM

DIFFICULTES ETHIQUES ET SOLUTIONS

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

Principes FAIR (Findable Acessible Interoperable Reusable)

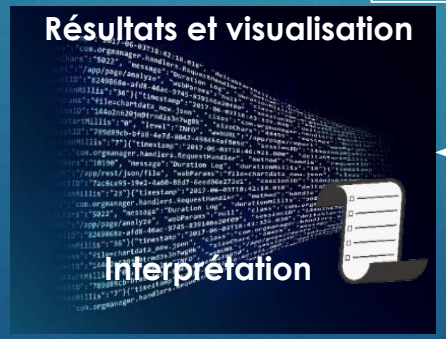
Attention aux biais !! : choix des données, interprétation, etc.



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)

DIFFICULTES TECHNIQUES ET SOLUTIONS

Le TDM repose sur:

1. l'exploitation de **texte**
2. des traitements automatiques du **langage naturel**
3. des traitements informatiques basés sur des outils d'**intelligence artificielle**

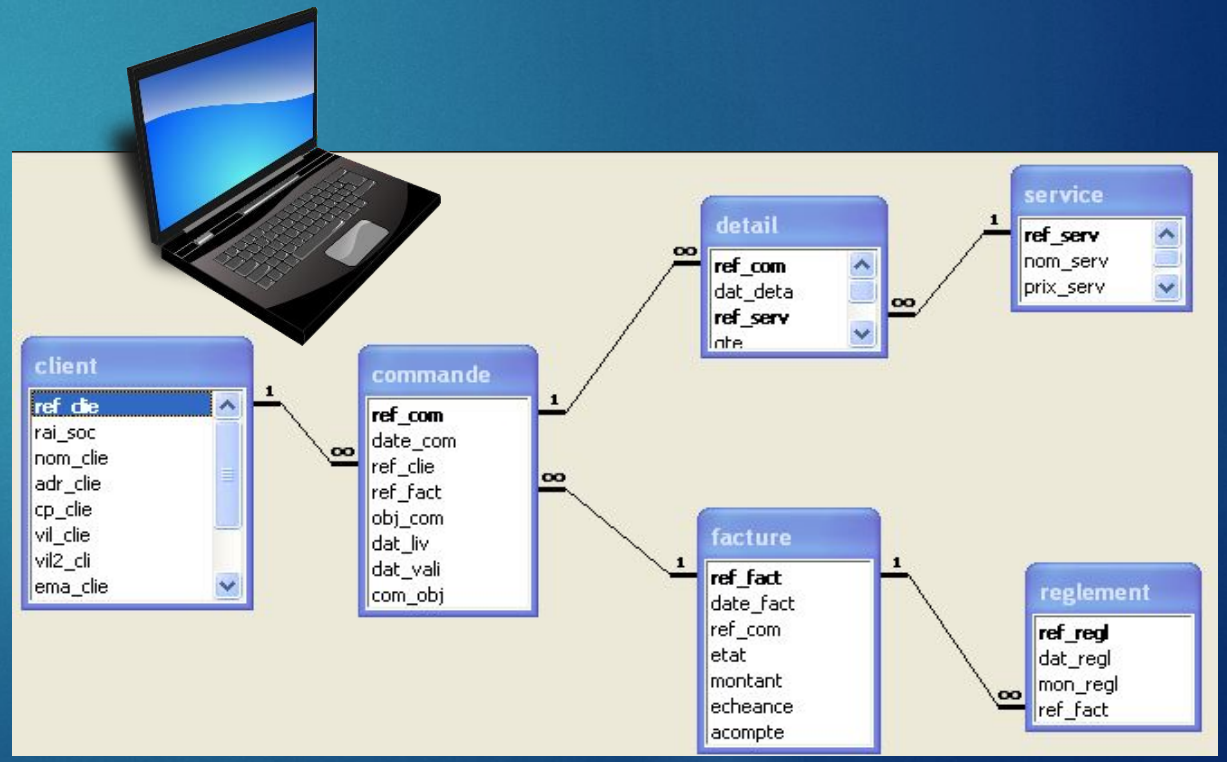
1/ Le texte est une donnée mais avec des caractéristiques spécifiques…

Le texte est une **donnée non structurée**

Un ordinateur interprète de la **donnée structurée**

*« Vous trouverez par la présente le courrier de M. Dupont 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. Dupont est-elle payée ?



2/ 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

3/ Quelques techniques de TDM

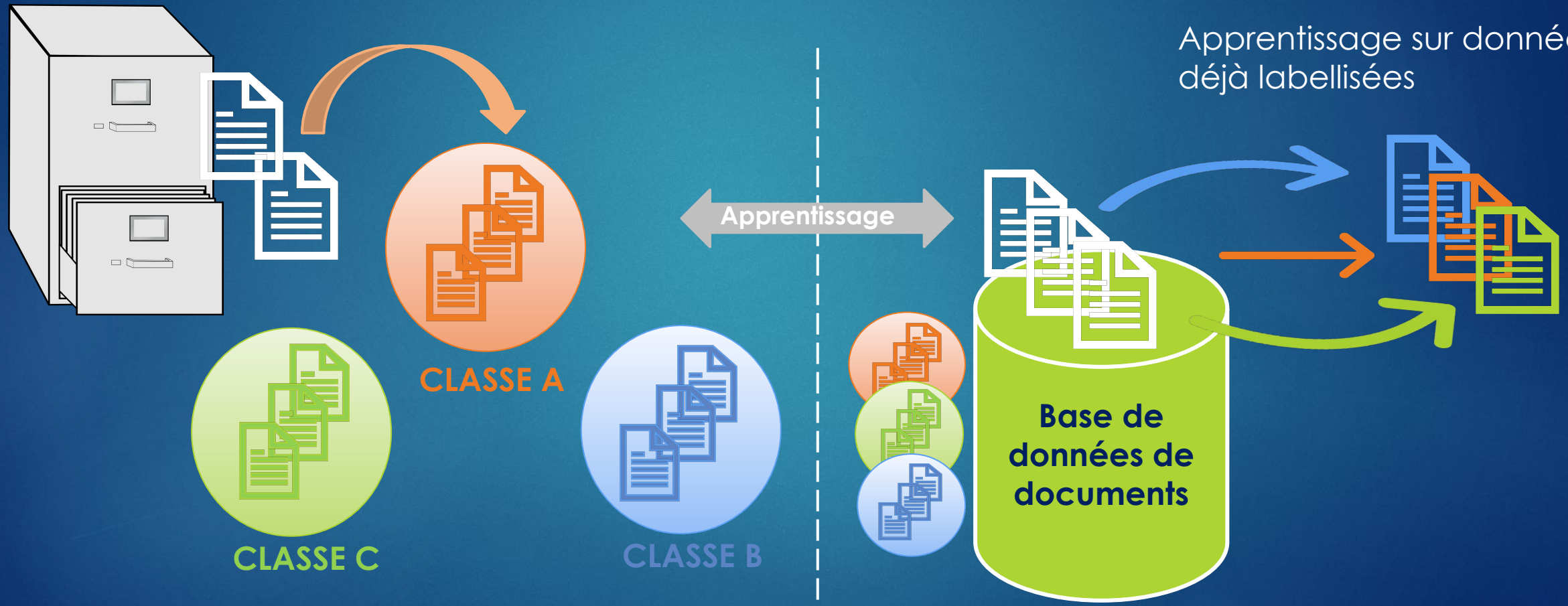
CLASSIFICATION

CLASSIFICATION SUPERVISEE

Apprentissage sur données déjà labellisées

Problématique

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



3/ Quelques techniques de TDM

EXTRACTION D'INFORMATION

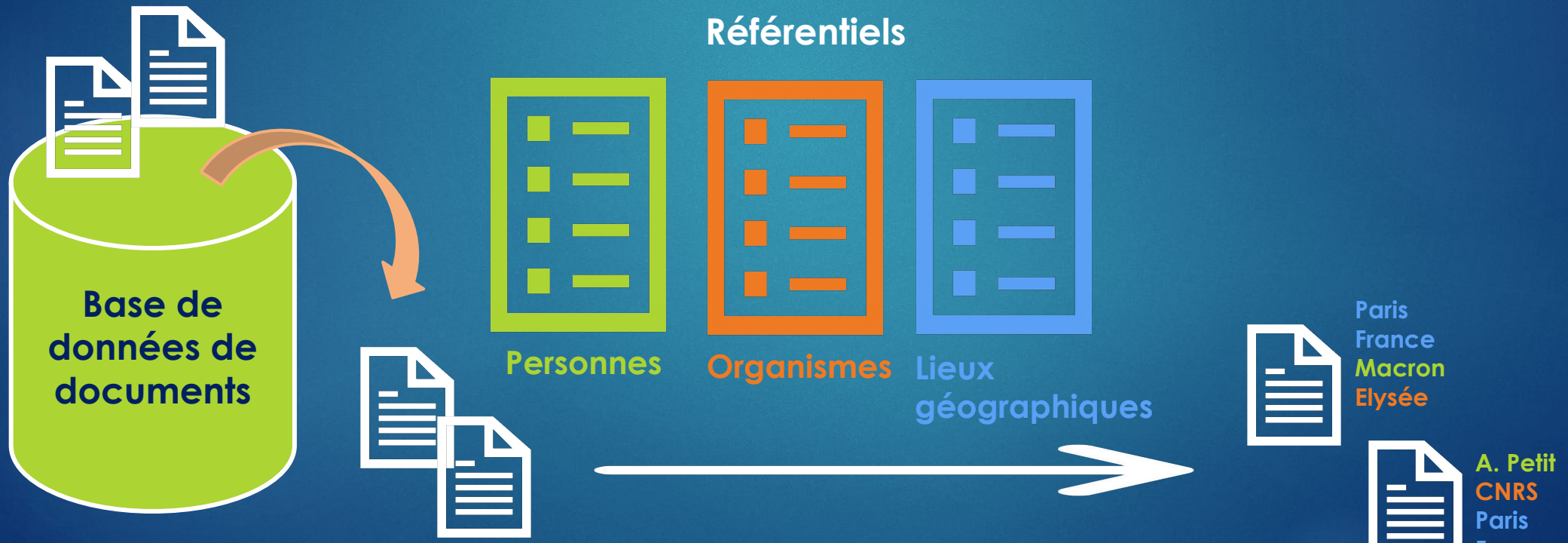
RECONNAISSANCE D'ENTITES NOMMEES

Problématique

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

Réponses possibles

Diverses méthodes possibles : statistiques (fréquence des mots dans le document, après avoir pris soin de supprimer les termes « vides »), vectorielles, utilisation de référentiels,…



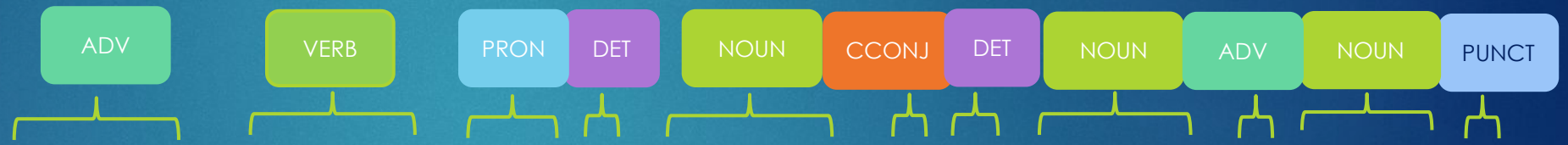
Quelques outils de TDM

1/ Structuration des données

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

Tokenisation

POS tagging
(Part Of Speech)



Lemmatisation

2/ Vectorisation des données

CountVectorizer

Corpus :

1. « Comment transformer le contenu d'un document en chiffres ? »
2. « Je l'ai transformé en chiffres ! »

doc	comment	transformer	le	contenu	de	un	document	en	chiffre	je	avoir
1.	1	1	1	1	1	1	1	1	1	0	0
2.	0	1	1	0	0	0	0	1	1	1	1

CountVectorizer

Corpus :

1. « Comment transformer le contenu d'un document en chiffres ? »
2. « Je l'ai transformé en chiffres ! »

doc	comment	transformer	le	contenu	de	un	document	en	chiffre	je	avoir
1.	1	1	1	1	1	1	1	1	1	0	0
2.	0	1	1	0	0	0	0	1	1	1	1

Après suppression des mots vides

doc		transformer		contenu			document		chiffre		avoir
1.		1		1			1		1		0
2.		1		0			0		1		1

2/ Vectorisation des données

embedding

Phrase : "Comment transformez-vous une phrase en chiffres ?"

Comment ⟶ **[0.01, 0.8, -0.1 , … , 0.2 , -1.4]**

transformez ⟶ **[-0.8, 0.2, … , -1.4]**

…

2/ Vectorisation des données

embedding

Phrase : "Comment transformez-vous une phrase en chiffres ?"

Comment ⟶ **[0.01, 0.8, -0.1 , … , 0.2 , -1.4]**

transformez ⟶ **[-0.8, 0.2, … , -1.4]**

…

Opération (simple ou moins simple) sur l'ensemble des vecteurs.

"Comment transformez-vous une phrase en chiffres ?" ⟶ **[-0.79, 1, … , -2.8]**

L'INIST-CNRS et le TDM

Quelles opportunités à l'INIST-CNRS ?

Qu'est ce qu'un web service ?

Web service (WS): 1 outil = 1 tâche
Utilisable par des non connaisseurs via LODEX, mais aussi en ligne de commande, appelé dans des programmes ou via des interfaces.

Qu'est ce qu'un web service ?

Web service (WS): 1 outil = 1 tâche
Utilisable par des non connaisseurs via LODEX, mais aussi en ligne de commande, appelé dans des programmes ou via des interfaces.

La complexité de nos WS dépend de plusieurs facteurs :
- Type de tâches (classification, extraction, indexation …)
- Types de données (résumé, métadonnées, texte intégral …)

Qu'est ce qu'un web service ?

Web service (WS): 1 outil = 1 tâche
Utilisable par des non connaisseurs via LODEX, mais aussi en ligne de commande, appelé dans des programmes ou via des interfaces.

La complexité de nos WS dépend de plusieurs facteurs :
- Type de tâches (classification, extraction, indexation…)
- Types de données (résumé, métadonnées, texte plein…)

- IA Factory : interface web de l'INIST pour utiliser un WS simplement
- Lodex : outil de data visualization (nos web services peuvent y être utilisés)

Web services : de la simplicité à la complexité !

Quelques exemples de web services :

- **Extraction de termes avec Teeft**
- Détection de genre d'un auteur
- Classification en domaines scientifiques Pascal-Francis

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 une idée de ce dont parle le texte. Idéalement, le texte doit contenir plusieurs paragraphes. Par défaut teeft extrait 5...

AVANT	APRES

[{
"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 pour objectif d'étudier la géologie de la planète Mars, en particulier le rôle joué par l'eau dans l'histoire de la planète. Les deux astromobiles ont été lancés au début de l'été 2003 et se sont posés en janvier 2004 sur deux sites martiens susceptibles d'avoir conservé des traces de l'action de l'eau dans leur sol. Chaque astromobile, piloté par un opérateur depuis la Terre, a alors entamé un périple en utilisant une batterie d'instruments embarqués pour analyser les roches les plus intéressantes (...)"
}]
	[
	{
	"id":
	"https://fr.wikipedia.org/wiki/Mars_Exploration_Rover",
	"value": [
	"deux robots",
	"panneaux solaires",
	"mars exploration rover mer",
	"mission double",
	"deux robots mobiles"
	]
	}
	]

Web services : de la simplicité à la complexité !

Quelques exemples de web services :
- Extraction de termes avec Teeft
- **Détection de genre d'un auteur**
- Classification en domaines scientifiques Pascal-Francis

AVANT	APRES
[{"id": "1", "value": "Jean Christophe, Dupont"}, {"id": "2", "value": "Amke"}, {"id": "3", "value": "Seong-Eun Park"}, {"id": "4", "value": "James A."}]	[{"id": "1", "value": "masculin"}, {"id": "2", "value": "mixte_feminin"}, {"id": "3", "value": "feminin"}, {"id": "4", "value": "masculin"}]

Détection de genre

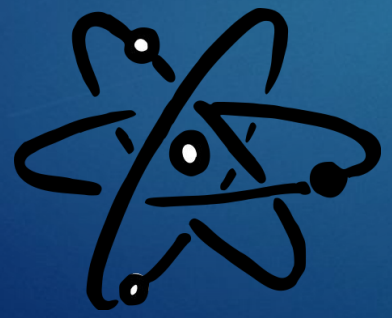
Ce web service permet de détecter le genre à partir d'une liste de prénoms genrés. Cette liste est un mélange entre les données issues de la librairie python gender-guesser et des données issues de la plateforme Kaggle. Elles ont été...

Web services : de la simplicité à la complexité !

Quelques exemples de web services :
- Extraction de termes avec Teeft
- Détection de genre d'un auteur
- **Classification en domaines scientifiques Pascal-Francis**

Classification en domaines scientifiques

Le web service de classification automatique permet de classer des documents scientifiques en anglais dans le plan de classement Pascal (Sciences, Techniques et Médecine) ou Francis (Sciences Humaines et Sociales). Après traitement, chaque document possédera un domaine scientifique homogène, dans...



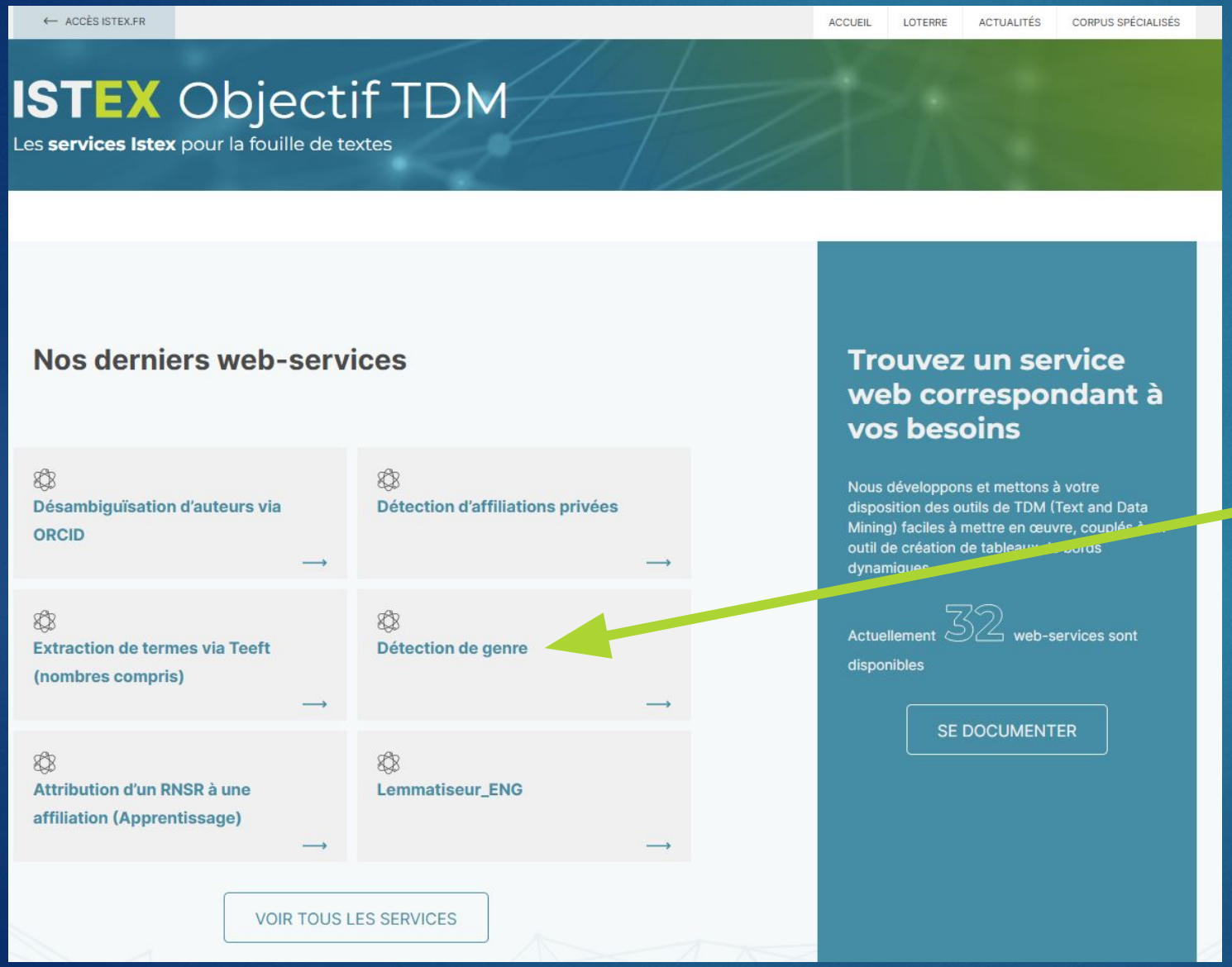
AVANT

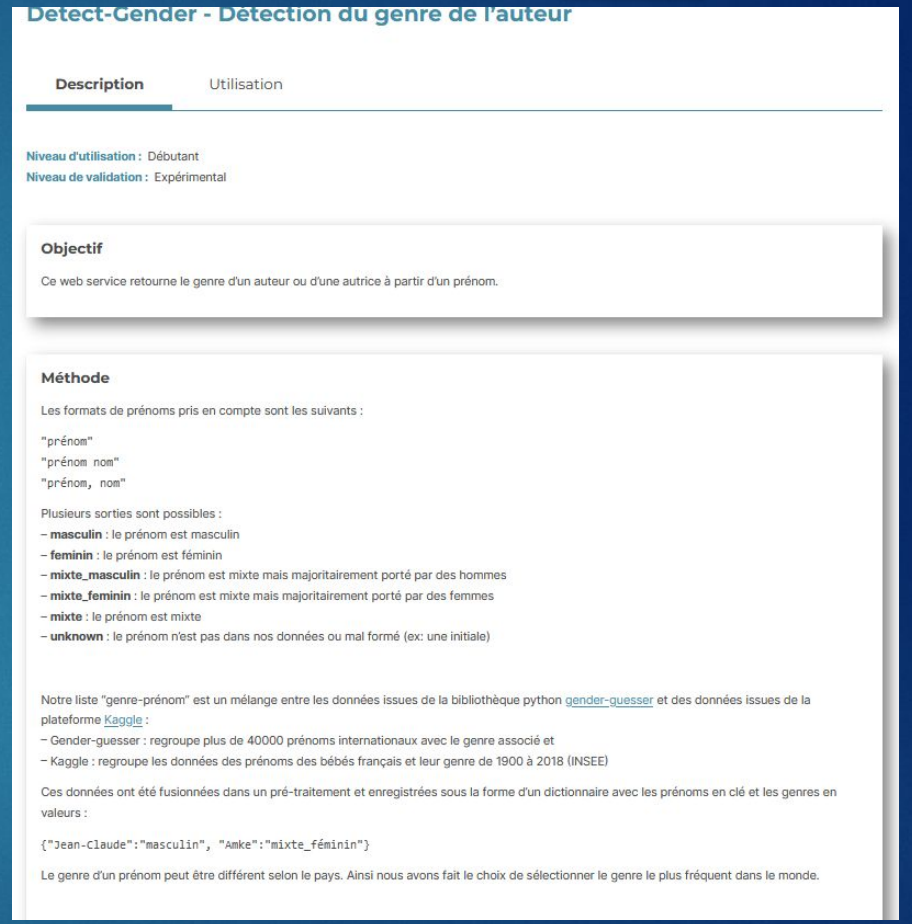

```
[{"idt": "08-040289","value": "Planck 2015 results. XIII. Cosmological parameters.We present results based on full-mission Planck observations of temperature and polarization anisotropies of the CMB. These data are consistent with the six-parameter inflationary LCDM cosmology.[…] However the amplitude of the fluctuations is found to be higher than inferred from rich cluster counts and weak gravitational lensing. Apart from these tensions, the base LCDM cosmology provides an excellent description of the Planck CMB observations and many other astrophysical data sets."}]
```


APRES

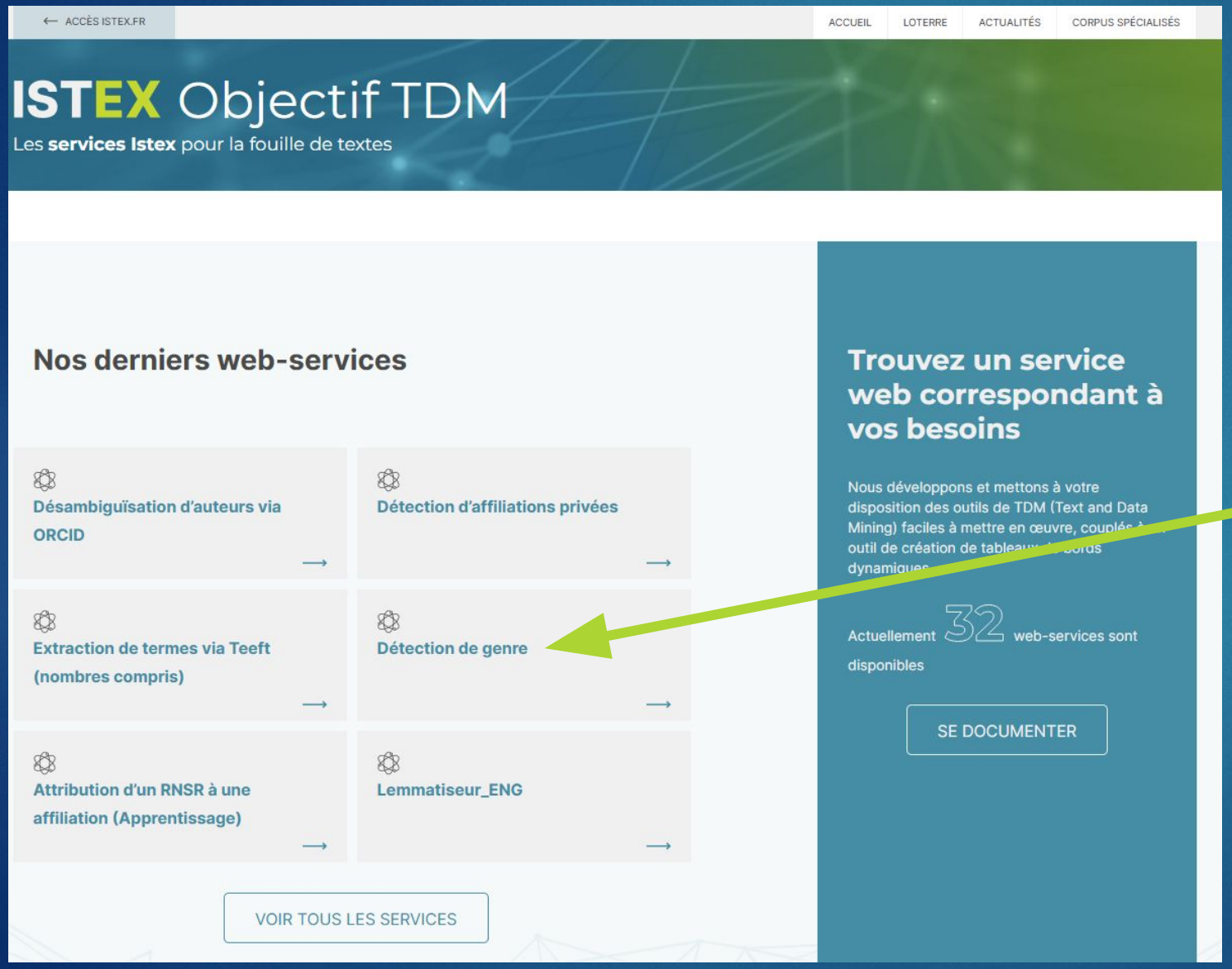

```
[{
    "idt": "08-040289",
    "value": [
        {
            "code": {
                "id": "001",
                "value": "Sciences exactes et technologie."
            },
            "confidence": 1.0000057220458984,
            "rang": 1
        },
        {
            "code": {
                "id": "001E",
                "value": "Terre, océan, espace."
            },
            "confidence": 0.9999549388885498,
            "rang": 2
        },
        {
            "code": {
                "id": "001E03",
                "value": "Astronomie."
            },
            "confidence": 1.0000100135803223,
            "rang": 3
}]
```

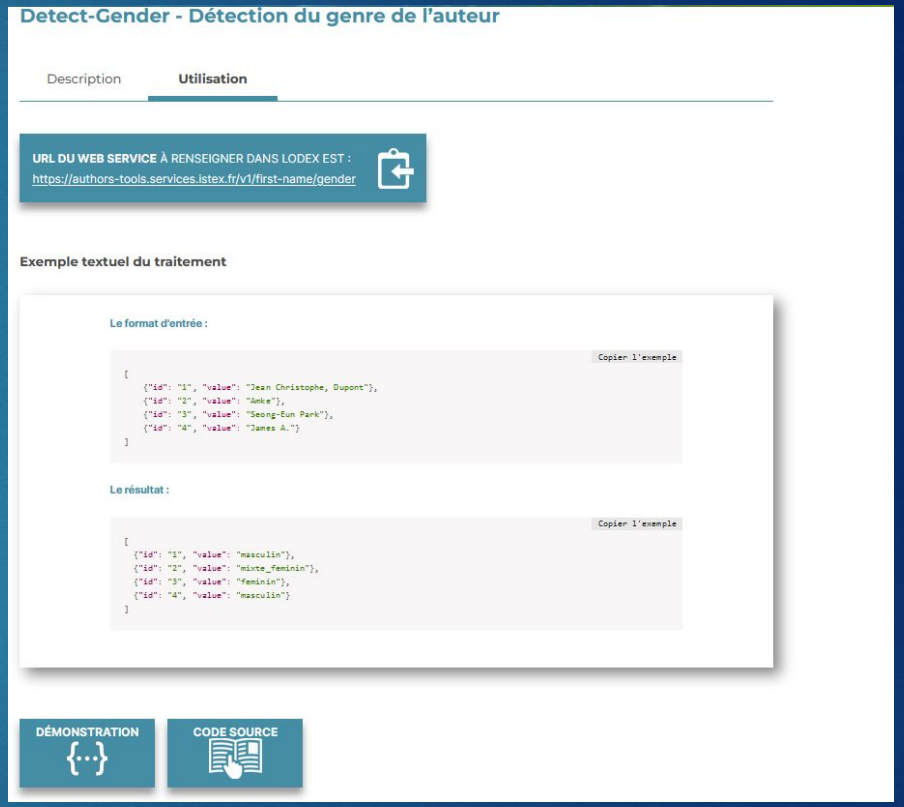
Objectif TDM





Objectif TDM

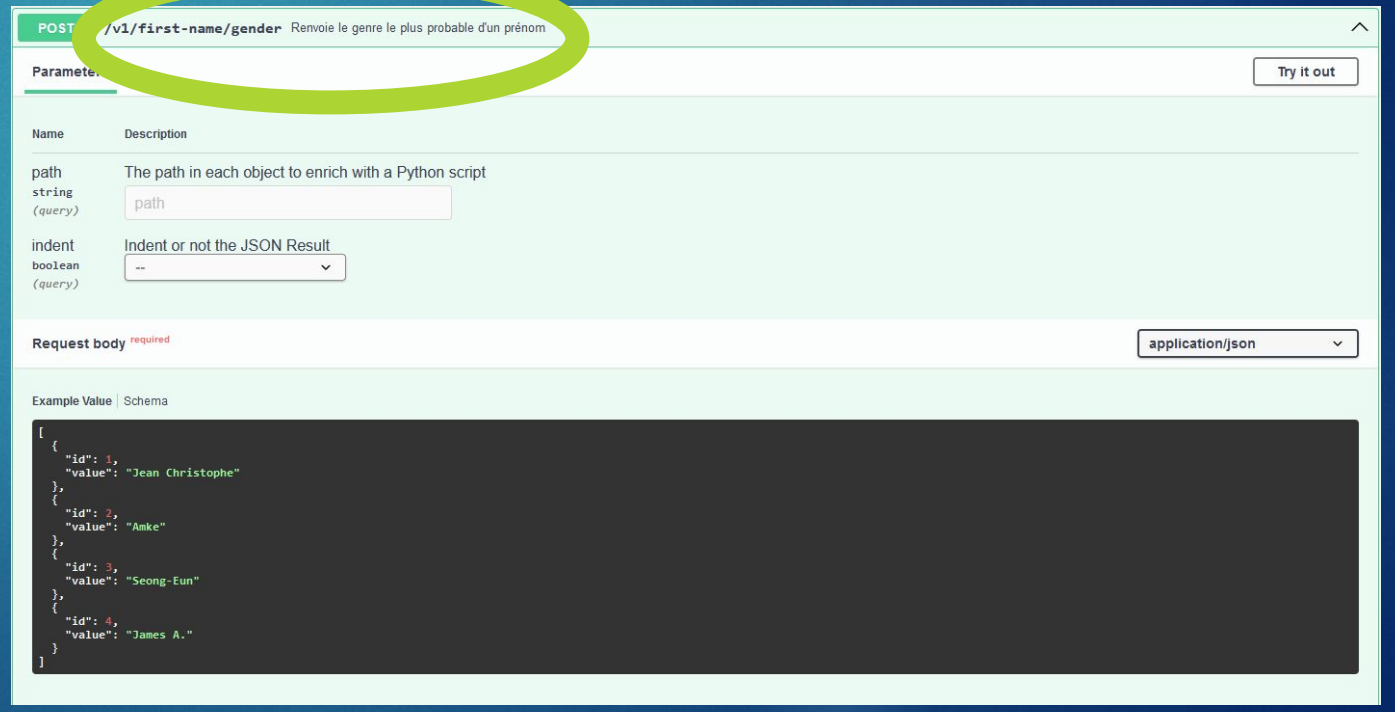




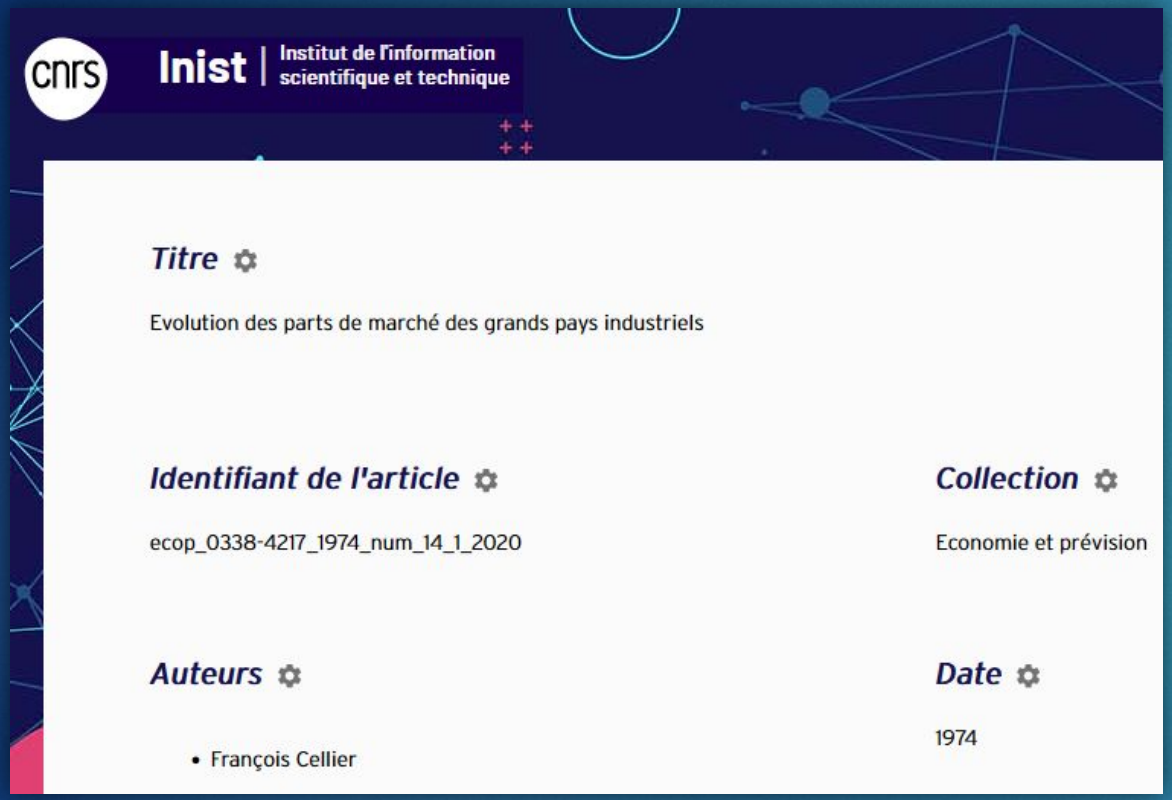
Les Web-Services

https://openapi.services.inist.fr/

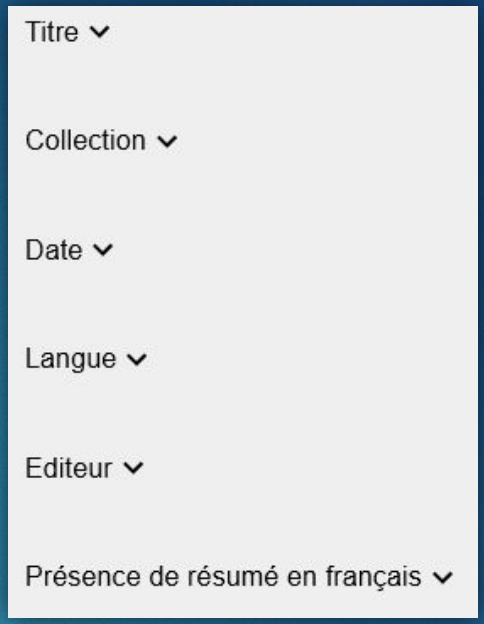
- permet de tester simplement un WS en ligne avec son propre exemple.



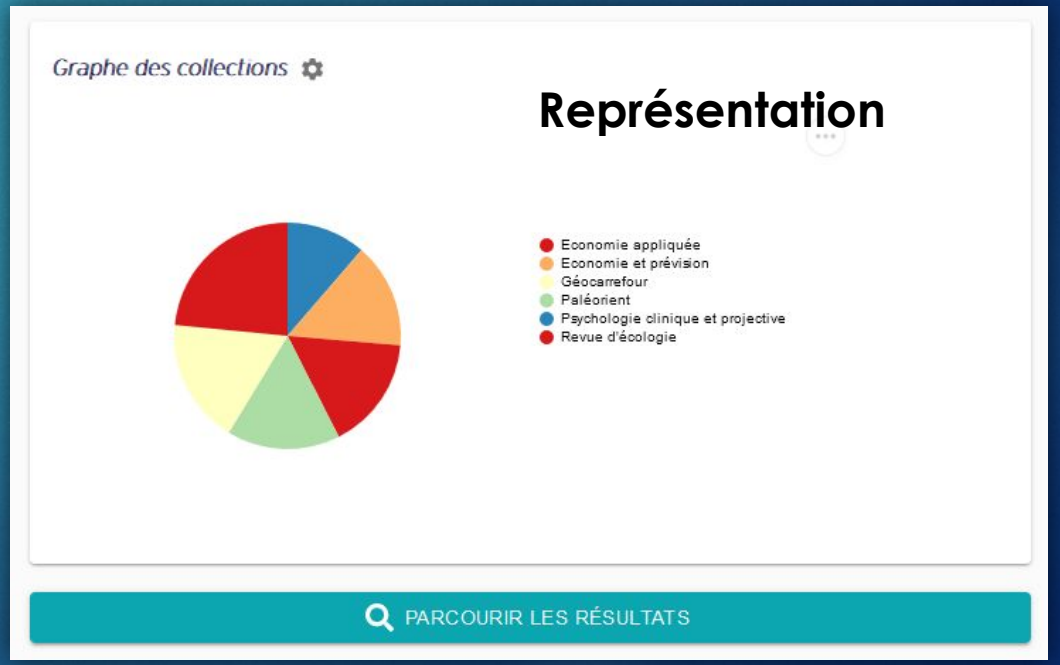
Lodex – Data visualization



Description bibliographique



Filtres/Facettes



Ressources :

- Lien vers objectif TDM : https://services.istex.fr/
- Tutoriel : Lodex pas à pas : https://callisto-formation.fr/course/view.php?id=194
- Site Lodex : https://www.lodex.fr/
- Documentation Lodex : https://www.lodex.fr/docs/documentation/principales-fonctionnalites-disponibles/