Atelier ANF TDM 2022

Exploration documentaire et extraction d'information

DES WEB SERVICES DEDIES A LA FOUILLE DE TEXTES

04 octobre 2022
Paris

FABIENNE KETTANI INIST-CNRS
JUSTINE REVOL INIST-CNRS

PROGRAMME

9h30- 9h40 Présentations et échanges avec les participants

9h40- 10h Brève introduction au TDM et utilisations possibles.
Services mis en œuvre à l'Inist-CNRS
Exemple du Corpus Mémoire

10h-11h ATELIER partie 1

Niveau simple

Utilisation et applications de 3 types de web services sur un corpus et exploitation dans l'application Lodex:

- Détection de langue
- Géotaggueur
- Classification Pascal/Francis

15 mn

11h15-12h30 ATELIER partie 2

Un peu plus de complexité…

Utilisation d'un web service au choix

La fouille de textes

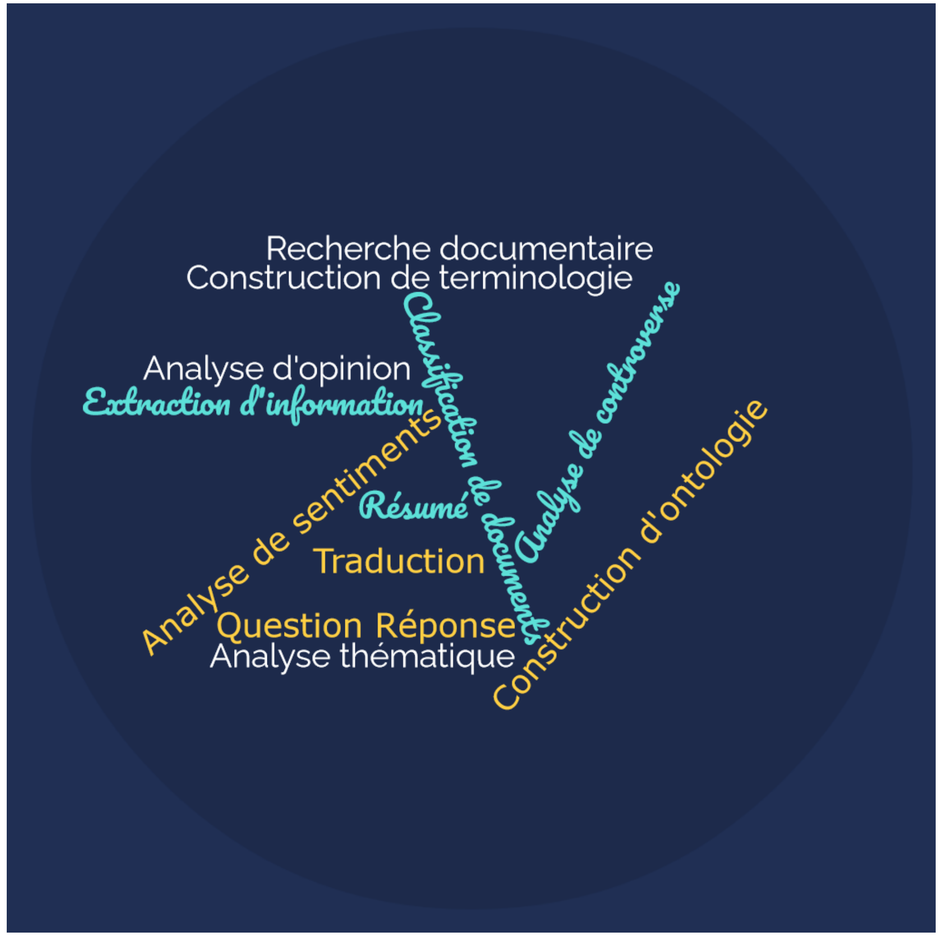
DEFINITION

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.

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**[1] (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.

CONTEXTE

Révolution numérique

Maturité des technologies

Prise de conscience politique

Evolution juridique

Nous ne sommes plus en capacité d'absorber la quantité d'informations qui est disponible…

Révolution numérique

Infobésité galopante il y a 2-3 ans

→ **Déluge** d'informations

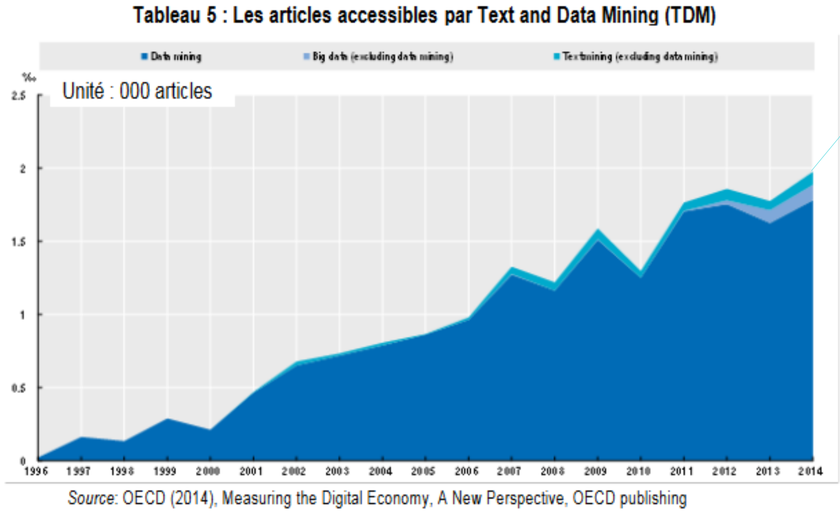
Le phénomène Big Data s'amplifie si vite que l'on n'arrive plus à suivre l'évolution des nouvelles unités de mesure : les **exaoctets** (10^{18}octets), les **zettaoctets** (10^{21}), les **yottaoctets** (10^{24})....

> 180 zettaoctets en 2025

Publications scientifiques

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

TDM



1996 **2014**

Source: OECD (2014), Measuring the Digital Economy, A New Perspective, OECD publishing

Nous disposons de technologies de plus en plus performantes…

Maturité des technologies

- Une expérience et une capitalisation de **30 ans** de travail à la fois sur le **TAL** et l'**IA**
- Une puissance de **calcul** et de **stockage** x1 milliard en 40 ans
- Une évolution majeure des algorithmes : **statistiques vs apprentissage profond**
- Implication forte des **industriels** (analyse de tendance, de sentiment, détection de buzz)

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

Prise de conscience politique

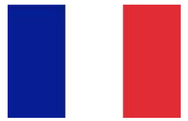
Budapest Open Initiative: problématique du libre accès aux publications scientifiques et incitation à l’utilisation des archives ouvertes ou des revues en libre accès, prise de conscience des besoins en licences adaptées

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)

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.»
https://www.ouvrirlascience.fr/category/science_ouverte/

5 M€ /an



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: politique et interlocuteurs désignés pour la science ouverte

2e Plan national pour la science ouverte (2021-2024): « Transformer les pratiques pour faire de la science ouverte le principe par défaut » → 100% de publications en accès ouvert en 2030

15 M€ /an

Plateforme Recherche Data Gouv

2001 2003 2018 2019 2021 2022

Evolution juridique

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

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**"

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.”

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)

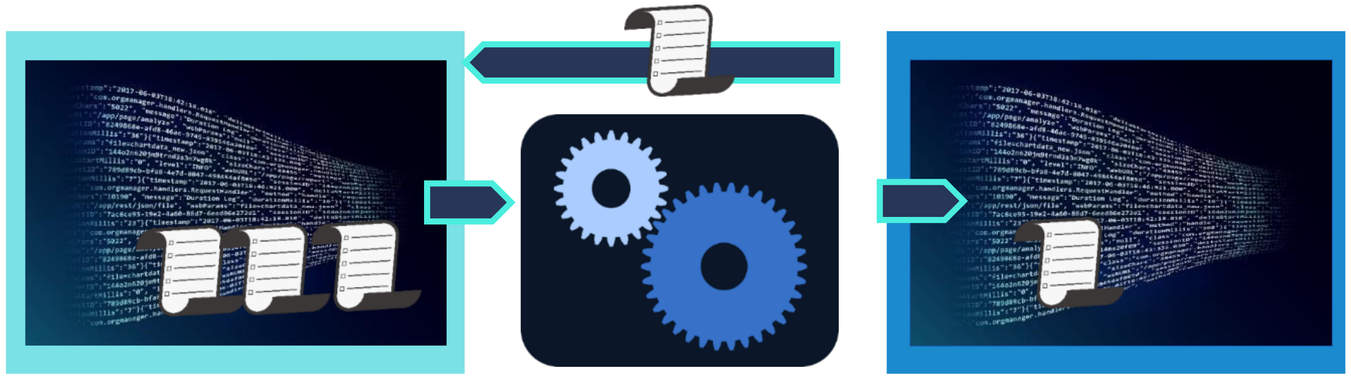
2016 2019 2021 2022

ENJEUX DU TDM

- **Réutiliser les produits de la recherche**: composants de TDM et contenus (publications, ressources sémantiques)
- Le TDM en tant qu'**accélérateur de l'innovation**
- Intégrer le TDM au **coeur de l'activité du chercheur non spécialiste**
- **Nouveaux métiers, nouvelles compétences**: développeurs informatiques spécialisés, ingénieurs de la connaissance,…

ASPECTS ETHIQUES DU TDM

- Transparence
- Fiabilité
- Reproductibilité



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

Constitution du corpus **Algorithmes** **Utilisation des résultats**

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

et FAIR- Findable/Accessible/Interoperable/Reusable

POURQUOI EST-CE COMPLIQUE ?

Le texte, une donnée pas comme les autres…

Données structurées

Données non structurées

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

La facture de M. Schmitt est-elle payée ???

POURQUOI EST-CE COMPLIQUE ?

La complexité de la compréhension des langues…

S’appuyer sur le traitement de la langue…

- **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 phares
- La **sémantique** des mots : désambiguïsation

Pour interpréter et comprendre…

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

DES WEB SERVICES POUR AIDER

Où les trouver ?

https://objectif-tdm.inist.fr/

Mais aussi des outils…

Inist cnrs

TM Tools Explorer

Version beta en évolution…
Repose sur une ontologie – publication à venir
Outils libres pour le moment

European language grid

UN EXEMPLE DE CE QUE L'ON PEUT FAIRE

Corpus mémoire-neurosciences

Lodex Mémoire-Neurosciences