Introduction à la fouille de textes et de données (TDM)

et présentation du pôle Text-mining

Mouhamadou Ba – Equipe Migale

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

migale

https://migale.inrae.fr

Introduction

research world produce
2,5 millions articles per year

One article published
every 12 secondes

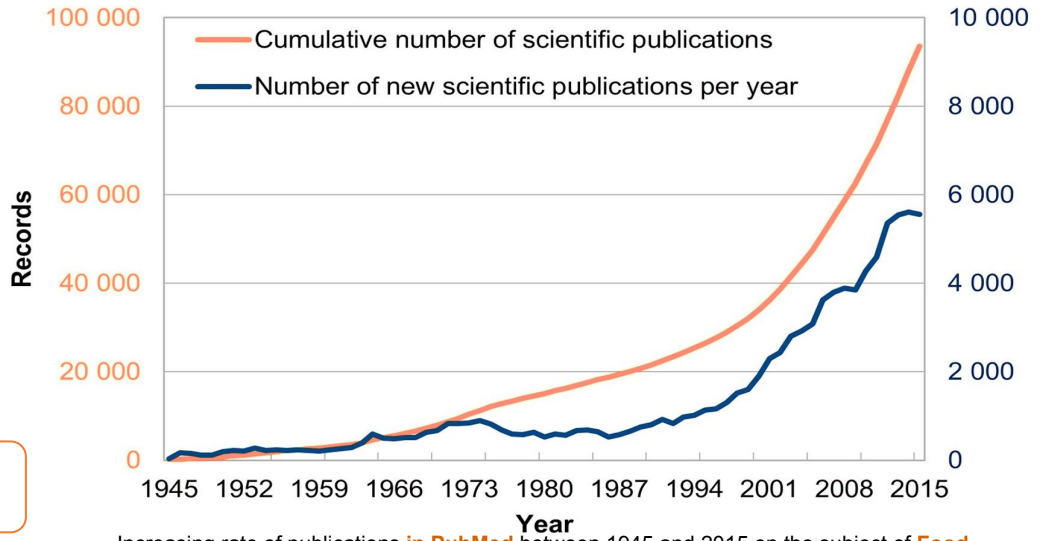
15 zettabytes of useful data
in 2020 (16 trillion of Gb)

1,8 billions of web sites

74,200,000 pages
(Facebook)

350,000 tweets
per minute

Orduña-Malea et al., Scientometrics 2014

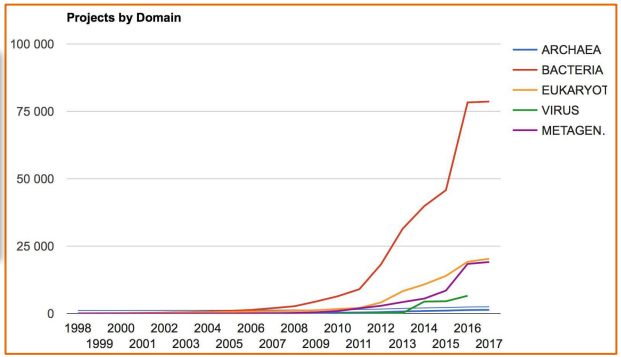


Increasing rate of publications **in PubMed** between 1945 and 2015 on the subject of **Food Microbiology** (Chaix et *al.* 2018).

Croissance exponentielle

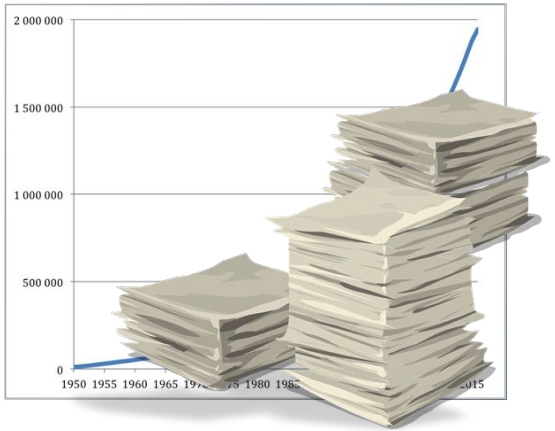
de l'information

génétique



Fouille de texte, un enjeu scientifique majeur

- 50% des articles ne sont jamais lus
- 90% des articles ne sont pas cités
- 80% des articles cités ne sont pas lus



Lokman I. Meho, the rise and rise of citation analysis, 2007.

Simkin & Roychowdhury. Read before you cite!, 2002

Exploiter les données de la recherche

Donner du sens aux données textuelles

Transformer une donnée non structurée en donnée structurée, manipulable par un ordinateur

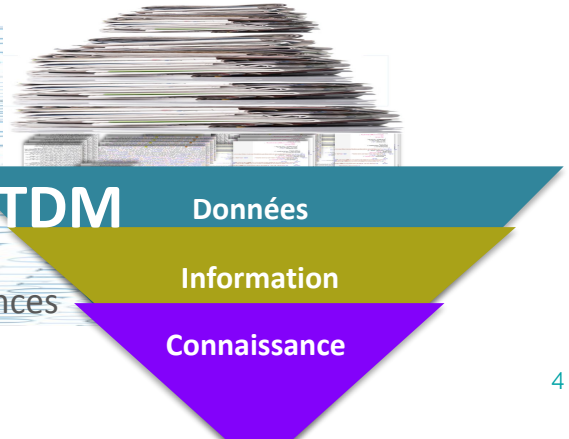
Intégrer le TDM scientifique

au cœur de l'activité du chercheur non spécialiste

Text mining, fouille de texte. Une définition

L'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**,
du Traitement Automatique des Langues, de la Représentation des Connaissances et des Statistiques.



De nouvelles voies pour la découverte de connaissances
Hargreaves Report (Digital Opportunity, 2011)

Exemple

Information d'intérêt

- Microbes
- Habitats
- Lieux géographiques

- Extraire ces informations des articles de PubMeD

- Le titre d'un article (PMID:19329624)

Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand

Exemple

Information d'intérêt

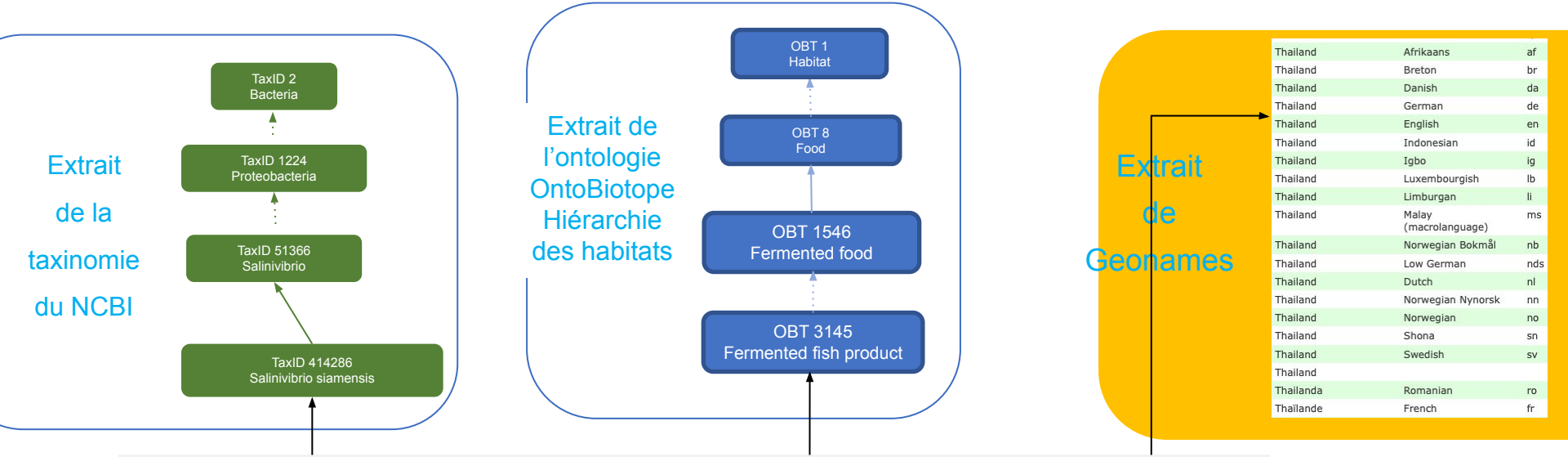
- Microbes
- Habitats
- Lieux géographiques

Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand

Exemple

Information d'intérêt

- Microbes
- Habitats
- Lieux géographiques



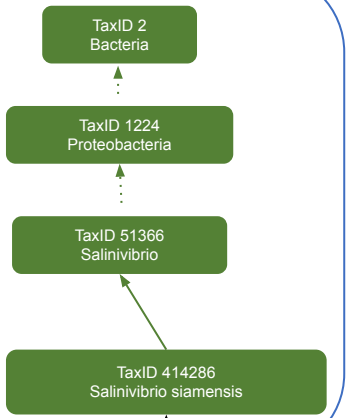
Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand

Exemple

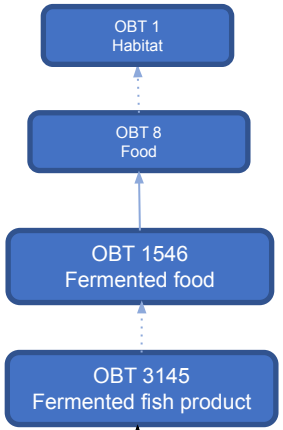
Information d'intérêt

- Microbes
- Habitats
- Lieux géographiques

Extrait de la taxinomie du NCBI



Extrait de l'ontologie OntoBiotope Hiérarchie des habitats



Extrait de Geonames

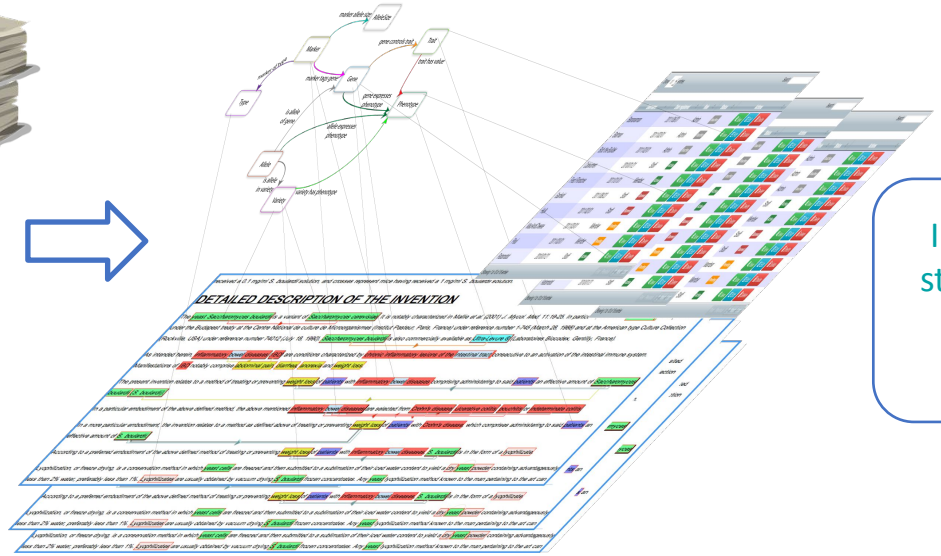
Thailand	Afrikaans	af
Thailand	Breton	br
Thailand	Danish	da
Thailand	German	de
Thailand	English	en
Thailand	Indonesian	id
Thailand	Igbo	ig
Thailand	Luxembourgish	lb
Thailand	Limburgan	li
Thailand	Malay (macrolanguage)	ms
Thailand	Norwegian Bokmål	nb
Thailand	Low German	nds
Thailand	Dutch	nl
Thailand	Norwegian Nynorsk	nn
Thailand	Norwegian	no
Thailand	Shona	sn
Thailand	Swedish	sv
Thailand		
Thailanda	Romanian	ro
Thaïlande	French	fr

Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand

Vit dans

Localisé à

Exemple

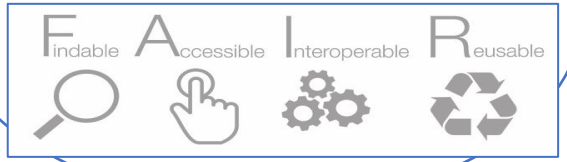


Intégrer, centraliser, structurer et standardiser l'information pour en faciliter l'accès et l'utilisation

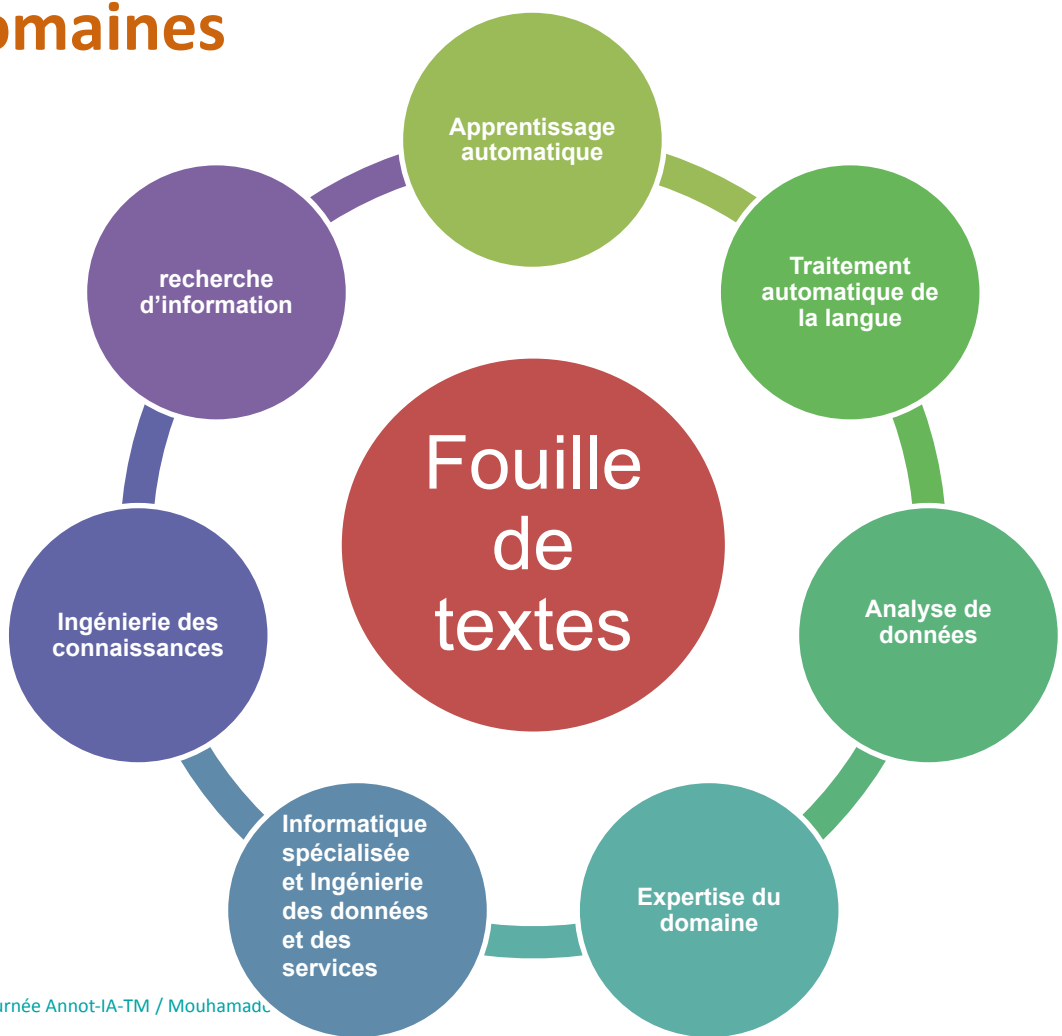
Base de données

Omnicrobe

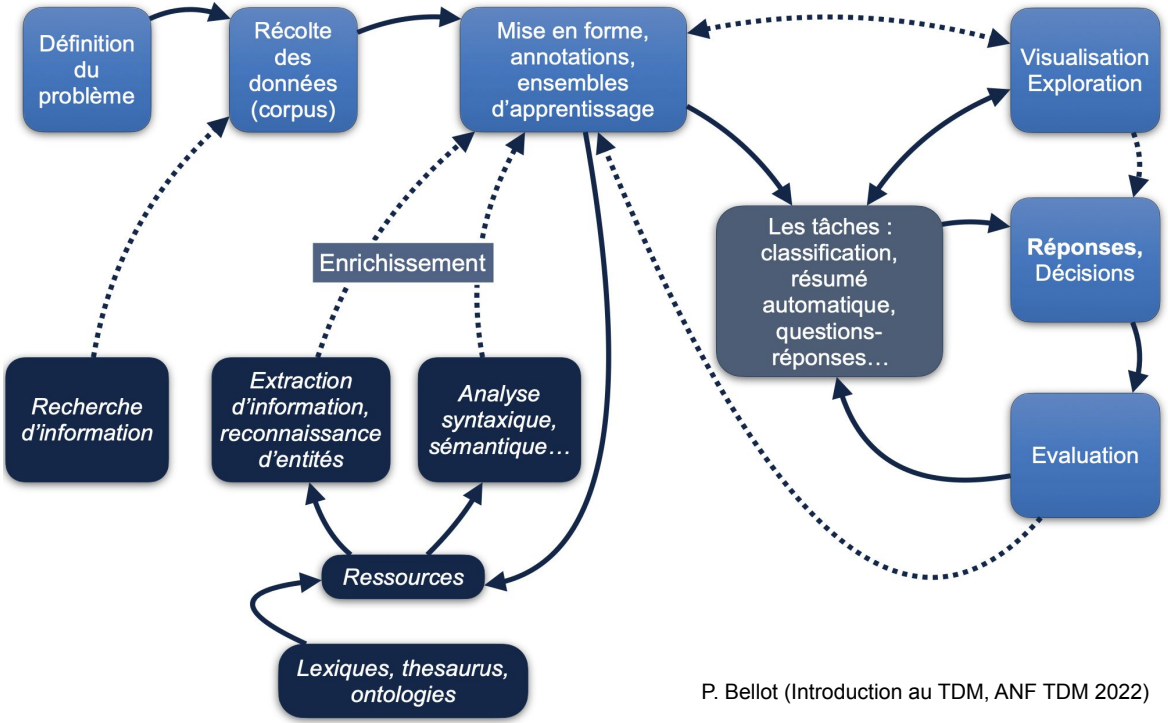
L'extraction d'information contribue aux principes



Plusieurs domaines



Processus de fouille de textes

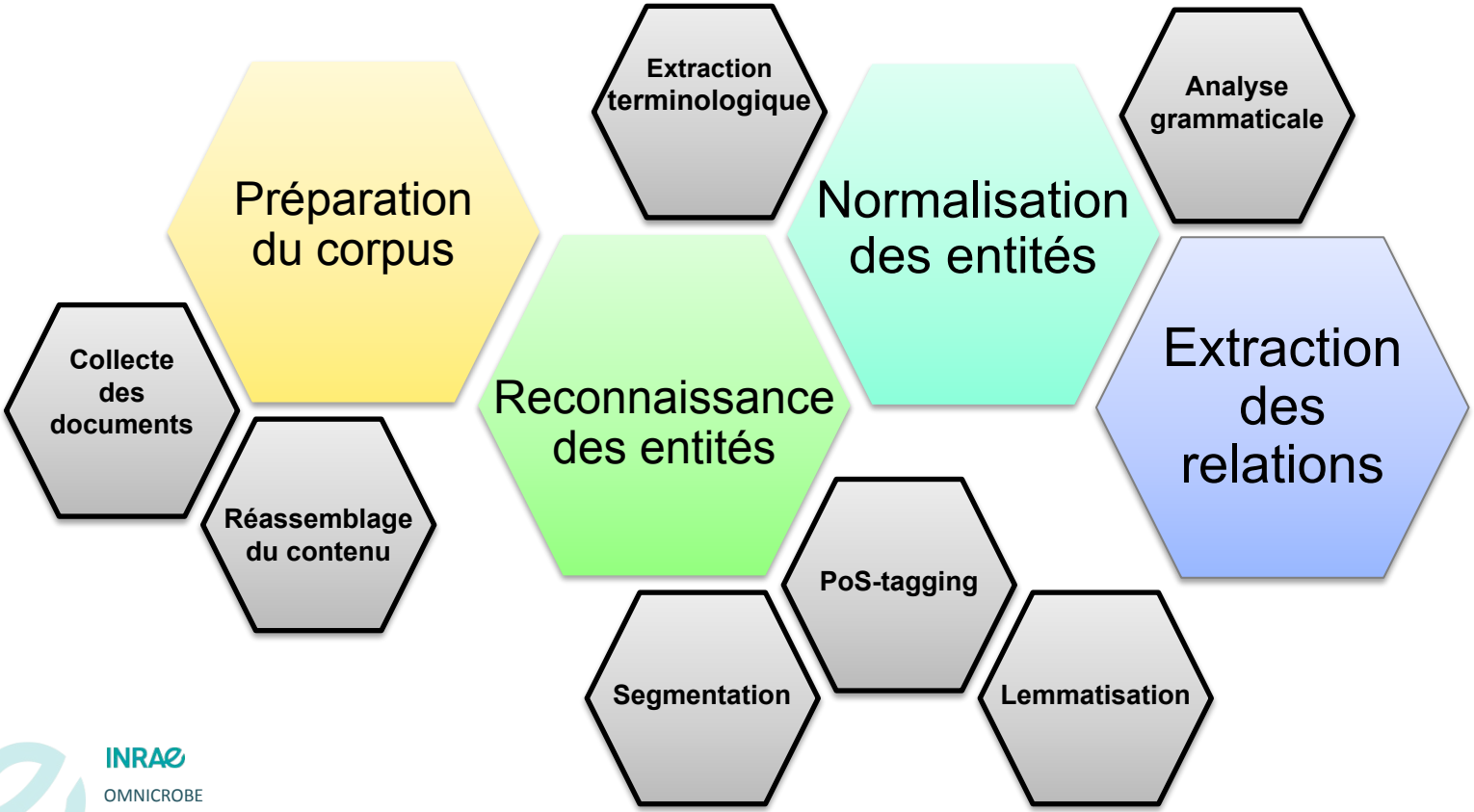


P. Bellot (Introduction au TDM, ANF TDM 2022)

- Recherche d'information
- Classification
- Annotation
- Extraction d'information

I. Tellier (Introduction à la fouille de textes, Université de Paris 3 - Sorbonne Nouvelle)

De multiples traitements intermédiaires



Quelques caractéristiques

Nature des données et des documents

Séquences de caractères

Données non structurées

Genres

Littérature scientifique

Presse technique

Rapports techniques

Champs libres de bases de données

Presse d'information

Réseaux sociaux

Pages web

Brevets

Langues

Formats

Langages de balisage

Langages de visualisation et d'impression

Langages de traitement de texte

Approches

- À base de règles
- Apprentissage machine (supervisé)

- Statistique
- Sémantique
- Descriptive
- Prédictive
- Numérique
- Symbolique

Règles

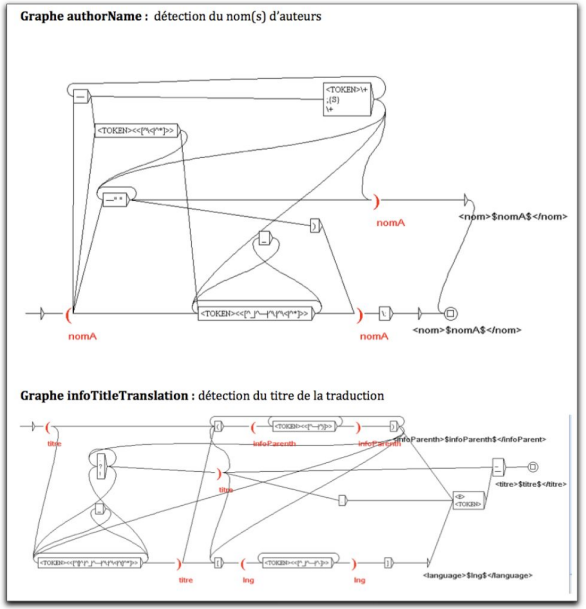
Des patrons et des automates qui formalisent les traitements

The ArbF amino acid sequence shares 55% identity with that of the E. coli BglF permease.

Chilly Gonzales (born Jason Charles Beck ; 20 March 1972) is a Canadian musician who resided in Paris, France [...]

```
[A-Za-z][a-z]{2}[A-Z]
```

```
resid.* in \w+
```



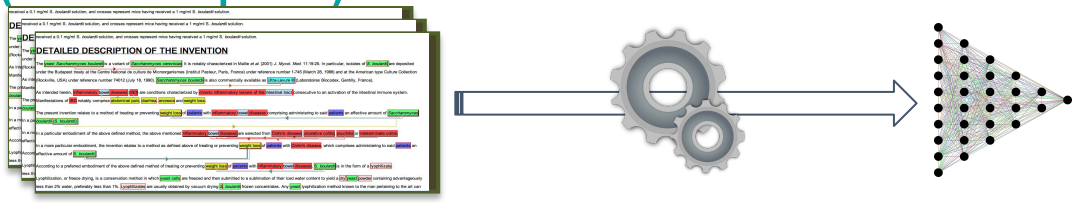
Etude pour l'UNESCO (Univ. Avignon, Open Edition 2011)

Limites

- La performance des règles dépend des données (types d'entités ou de relations, des domaines, et du genre des documents)
- La qualité dépend de la variabilité d'expression dans les documents.
- La maintenance d'un système basé sur des règles écrites à la main peut être coûteuse à maintenir.

Apprentissage supervisé

Des algos qui induisent des modèles à partir de données annotées (d'exemples).



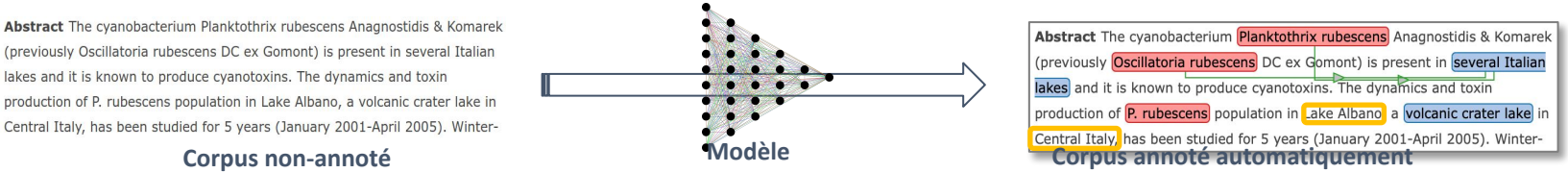
Corpus d'apprentissage annoté | Algorithme d'apprentissage | Modèle

Phase d'apprentissage

- Lors de la phase d'apprentissage, l'algorithme induit un modèle à partir d'exemples annotés par des experts.
- La nature du modèle dépend de l'algorithme utilisé. Ce peuvent être des droites de régression, des arbres de décision, etc.

Apprentissage supervisé

Des algos qui induisent des modèles à partir de données annotées (d'exemples).



Phase de production

- Lors de la phase d'étiquetage, le modèle sert à annoter automatiquement de nouveaux documents.
- Les erreurs peuvent être quantifiées en appliquant le modèle sur le corpus d'apprentissage.

À savoir

- L'élaboration du corpus d'apprentissage représente un effort initial considérable.
- Il existe des corpus d'apprentissage et même des modèles pré-appris pour un certain nombre de types d'annotation.

Corpus d'apprentissage

"Gold Standard"

- Échantillon de documents annotés par des experts afin de représenter formellement le sens du texte.
- Quantité suffisante pour que l'algorithme d'apprentissage opère l'induction et produise des modèles stables.
- La qualité dépend de la conformité de l'annotation au besoin exprimé.

Valorisation

- Un corpus annoté est un jeu de données (recherche et industrie) car il permet de mettre au point les systèmes automatiques.
- Communication académique :
  - organisation d'un **"challenge"** porté par une conférence ou un workshop (ACL, EMNLP, CLEF, CoNLL, BioNLP),
  - Data paper (Scientific Data, Pensoft, BMC Research Notes),
  - archives ouvertes (CodaLab, Papers With Code, LREC),
  - dépôts de code (GitHub, GitLab).

Ressources

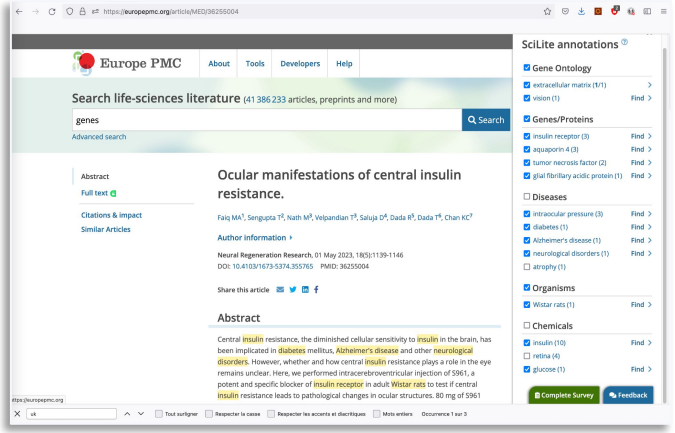
- Contenus
- Outils logiciels
- Services

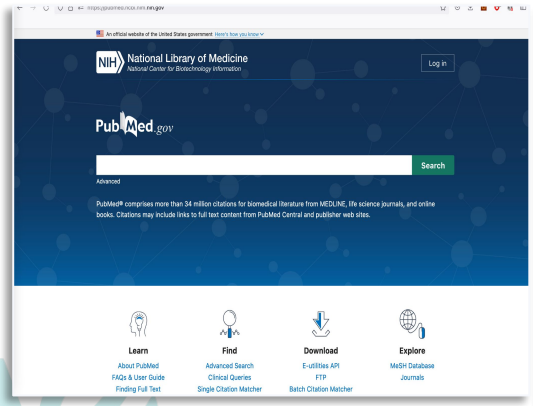
E-infrastructures / Plateformes

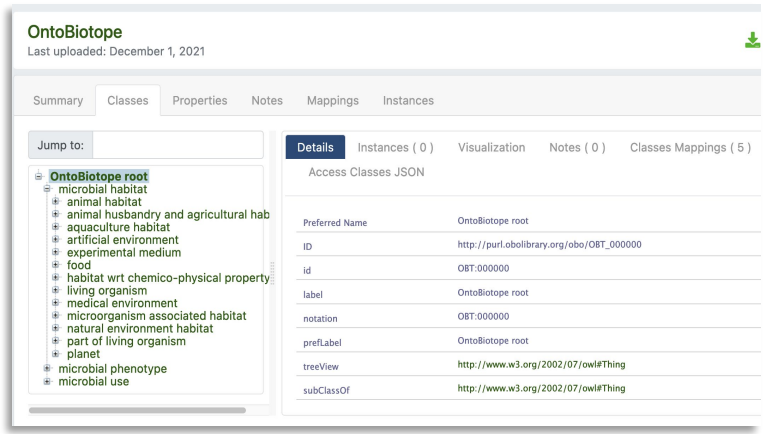
Projets

Contenus

- Données textuelles (articles, rapports, champs de base données,)
  - PubMeD, EPMC, ISTEX, WoS, Frontier, Wiley,
- Dictionnaires, lexiques, thésaurus, taxonomies, ontologies,…
  - Ontobiotope, Taxinomie NCBI, AGROVOC, ChEBI, MeSH, OMIM, GBIF
- Données d'apprentissage, modèles, règles…
  - BioNLP-ST datasets, Wikipedia embeddings
- Banque de données génétiques et Centre de Ressources Biologiques
  - GenBank, GOLD, BioSample, CIRMs, DSMZ

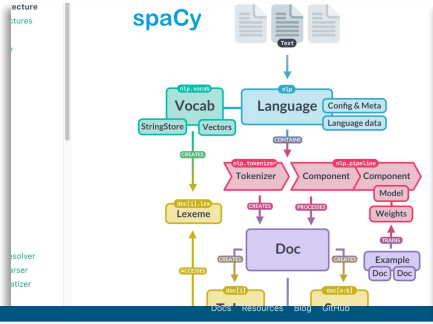


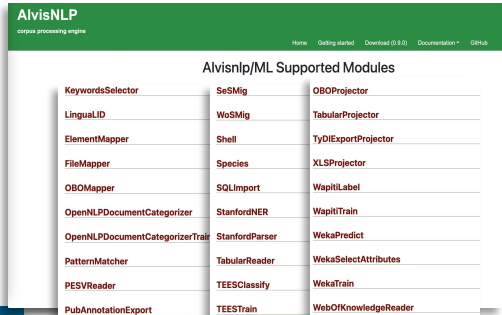


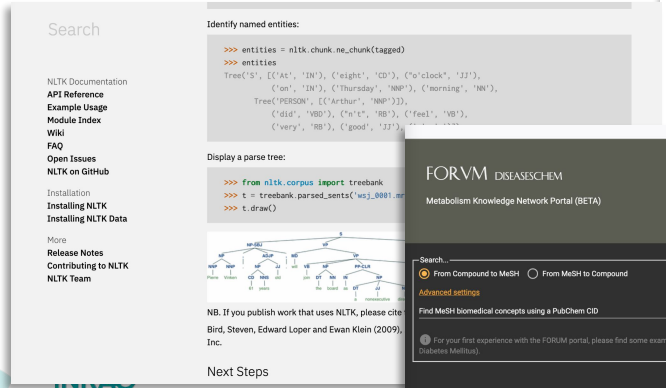


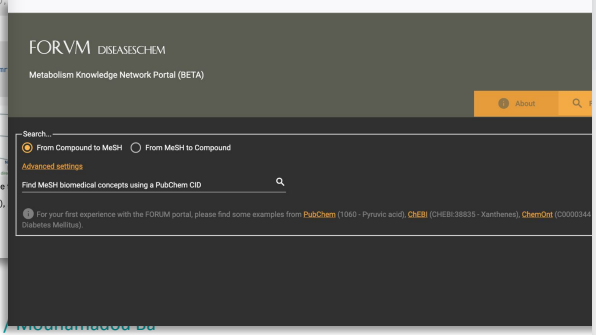
Outils logiciels

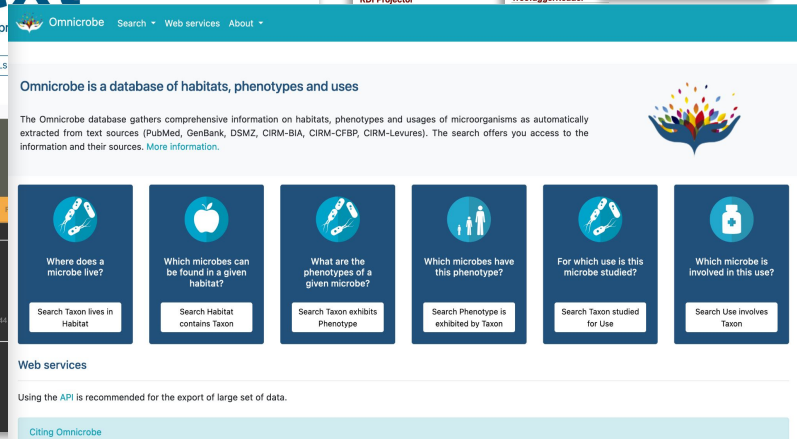
- Un continuum de solutions implémentant des traitements de text-mining (modules spécialisées, APIs, pipelines, micoservices,)
  - AlvisNLP, TermSuite, FastText, Spacy, NLTK, WeKa,…
- Des applications spécialisées, pour l'annotation de textes, confection de terminologies, annotation, visualisation, recherche, traduction,…
  - Omnicrobe, FORUM, AlvisIR, AlvisAE, INCEpTION, Tydi, Gargntext, IRaMuTeQ, CorText,…





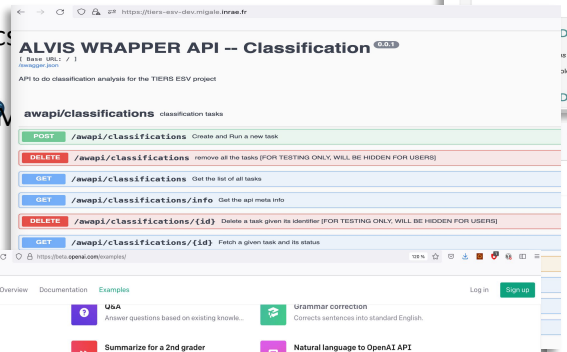


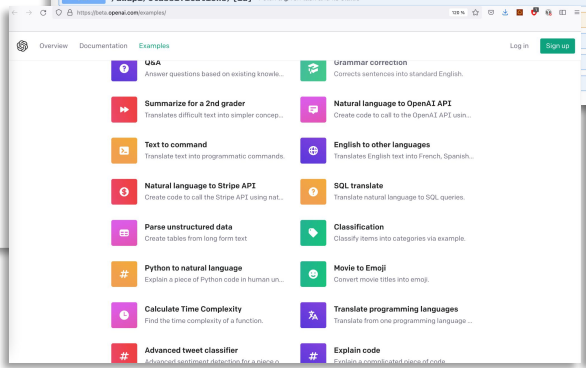


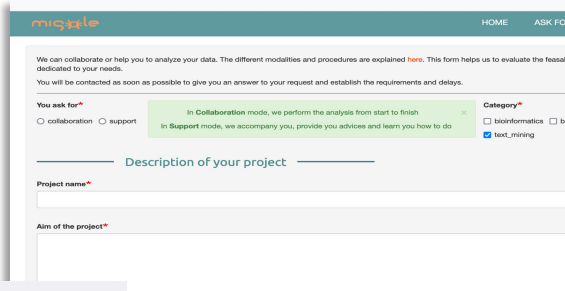


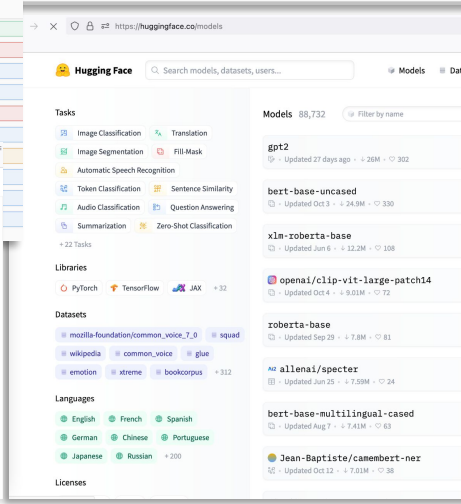
Offre de services

- Mise en disposition de contenu, d'outils, d'environnements de travail
  - Offre de services web dans Migale et à l'INIST
  - Service web de europePMC, NCBI
- Analyse des données textuelles
  - Service d'analyse de Migale
  - Constitution de corpus à l'INIST
- Formation et accompagnement
  - Introduction au text-mining avec AlvisNLP, "Bioinformatics
  - ANF (Action National de Formation) TDM, CNRS et INRAE
  - « Text-mining with the SimText toolset », Galaxy training
- Partage de ressources
  - Partage modèles & algo, HuggingFace, OpenAI









E-infrastructures / plateformes

- Plusieurs plateformes offrent des services de text-mining
  - Centrées sur les contenus o centrées sur outils
  - Spécialisées dans un domaine ou généralistes

- Plateformes
  - Migale (MaIAGE/INRAE)
  - ISTEX (INIST/CNRS)
  - ESFRI (European Strategy Forum on Research Infrastructures)
  - RDA (Research Data Alliance)
  - EOSC (European Open Science Cloud)
  - TGIR (Très Grandes Infrastructures de Recherche)

- D'autres plateformes utilisant Galaxy
  - LAP (Language Analysis Portal)/CLARINO (Université d'OSLO)
  - DBCLS Galaxy (Japon)
  - Alveo (niv. of WESTERN SYDNEY)
  - LAPPS Grid (VASSAR COLLEGE, BRANDEIS Univ., CARNEGIE-MELLON Univ., Univ. of PENNSYLVANIA)

Quelques projets

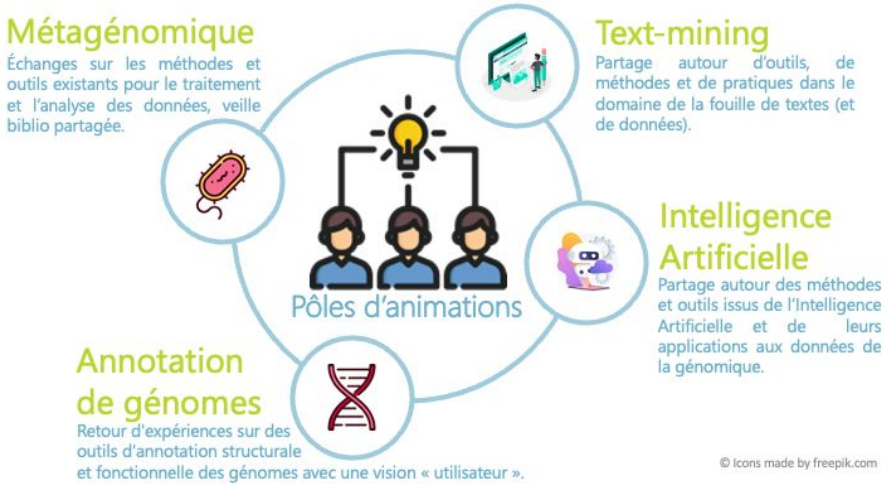
- HoloOligo
  - Extraction d'information concernées les oligosaccharides
  - Partenaires : GenPhySE, PEGASE, MICALIS, GABI, LEMM-CEA, IE PECTOUL, UE3P
- Beyond (/ TIERS-ESV)
  - Les solutions de text-mining comme leviers pour la veille sanitaire animale et végétale
  - Partenaires : Patho, BioSP, Ecodev, BFP, SAVE, TSCF, MaIAGE, PHIM, TETIS, Umr-Pvbmt
- D2KAB
  - Un cadre pour transformer les données de l'agronomie et de la biodiversité en connaissances
  - Partenaires: LIRMM (Université de Montpellier & CNRS), Wimmics (Université Côte d'Azur, Inria & CNRS), DipSO, TSCF, URGI, URFM, MaIAGE, IATE, …
- VisaTM
  - Une réflexion sur une infrastructure de recherche française qui offrent du contenu et des services en text mining
  - Partenaires : INIST (CNRS), MaIAGE (INRAE), LIRMM (Université de Montpellier)
- OpenMinTed (projet Européen H2020)
  - Construire une infrastructure de recherche qui offrent du contenu et des services en text mining
  - Partenaires : Athena Research and Innovation Center (ARC), Univ. of Manchester (NaCTeM), Technische Universität Darmstadt (UKP Lab), EMBL-EBI, INRA (MaIAGE)

Conclusion

- Données textuelles, une source d'information précieuse
- Le Text-Mining, un levier pour accéder aux connaissances

- Renforcer l'utilisation du text-mining pour la bioinformatique
  - Former et accompagner les acteurs
  - Produire des bases de connaissances intégrées
  - Intégration avec les données biologiques
  - Développer des outils innovants pour les acteurs de la recherche

PEPI IBIS : Pôle Text-mining



RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

PEPI IBIS : Pôle Text-mining

Pôle thématique autour de la fouille de textes dans le PEPI IBIS

- rassembler des agents de différents horizons autour de la thématique de text-mining
- échanger sur des outils, des méthodes et des pratiques dans le domaine de l'extraction d'informations à partir de textes

- Rencontre en visio pour échanger
  - autour d'une présentation
  - tous les deux mois
- Une liste pour partager des informations
  - pepi-bioinfostats-textmining@inrae.fr
- Participants
  - Mouhamadou Ba, MaIAGE (animateur)
  - Mathieu Charles, GABI
  - Rémy Decoupes, UMR TETIS
  - Sandra Dérozier, MaIAGE (animatrice)
  - Clément Frainay, Toxalim
  - Raphaël-Gauthier Flores, URGI

PEPI IBIS : Pôle Text-mining

3 réunions d'échange organisées en 2022

Exemples de présentation

- Offre de services de text-mining de Migale, Mouhamadou Ba
- Application de méthodes d'analyses sur la littérature scientifique pour la qualification d'associations composé chimique - concept biomédical, Marion Liotier (stage avec Clément Frainay)

Pôle ouvert à tous!

Vous travaillez ou souhaitez travailler autour des thématiques de la fouille de textes, ou bien utiliser les solutions développées dans ce domaine, rejoignez-nous !

Remerciements

MaIAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement

Migale

- Sophie SCHBATH
- Valentin LOUX
- Mouhamadou Ba
- Hélène CHIAPELLO
- Mahendra MARIADASSOU
- Véronique MARTIN
- Cédric MIDOUX
- Olivier RUÉ
- Valérie VIDAL

Bibliome

- Claire Nédellec
- Robert Bossy
- Louise Deléger
- Arnaud Ferre

StatInfOmics

- Sandra Dérozier

migale
https://migale.inrae.fr

ifb

FRANCE GENOMIQUE

Offres de stage

Deux offres pour Master 2 en développement bioinformatique

- Stage 1 :
  - « Enrichissement et automatisation des sources utilisées de l'application Omnicrobe »
  - Compétences souhaitées
    - Python
    - Snakemake
    - Connaissances sur les banques de données bioinformatiques
  - Contacts
    - Mouhamadou Ba, mouhamadou.ba@inrae.fr
    - Robert Bossy, robert.bossy@inrae.fr
- Stage 2
  - Sujet : « Evolution de l'interface web afin d'ordonner les informations en fonction de leur qualité »
  - Compétences souhaitées :
    - Javascript / Python
    - PostgreSQL
  - Contacts
    - Louise Deléger, louise.deleger@inrae.fr
    - Sandra Dérozier, sandra.derozier@inrae.fr