

Les web services et la datavisualisation Istex

Des web services dédiés à la fouille de textes

03 octobre 2024

Valérie Bonvallot (valerie.bonvallot@inist.fr)

Justine Revol (justine.revol@inist.fr)

ANF TDM 2024

Exploration documentaire et extraction d'informations

1 TDM : Quelques rappels

2 TDM à l'Inist et démos

3 Présentation des TPs

4 A vous de jouer





Vos attentes ?

<https://docs.google.com/presentation/d/1AbXbkgnjvVm6OKXly-On398HNdoCWF9HZfCibbo5mmY/edit?usp=sharing>

1 **TDM : Quelques rappels**

2 TDM à l'Inist et démos

3 Présentation des TPs

4 A vous de jouer



Environnement



La fouille de texte
en 1 coup d'oeil

<https://www.mauricelargeron.com/analyse-de-texte-et-seo/>



*Practical Text Mining and Statistical Analysis for Non-structured Text
Data Applications (2012)*

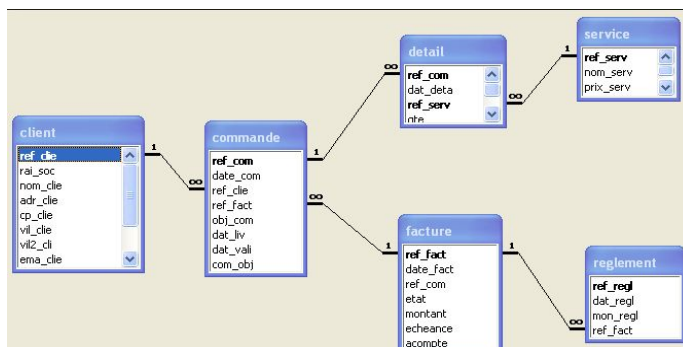
TEXTE

➔ 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	polysémie : 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

LANGUE



Quelques techniques de TAL (structurer les données)

Stanford CoreNLP : <http://corenlp.run/>

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

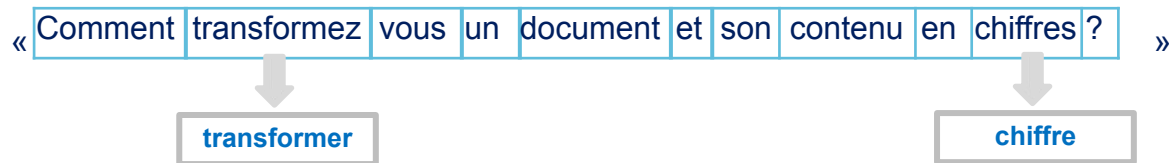
Tokenisation

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

POS tagging (Part Of Speech)



Lemmatisation (forme canonique)



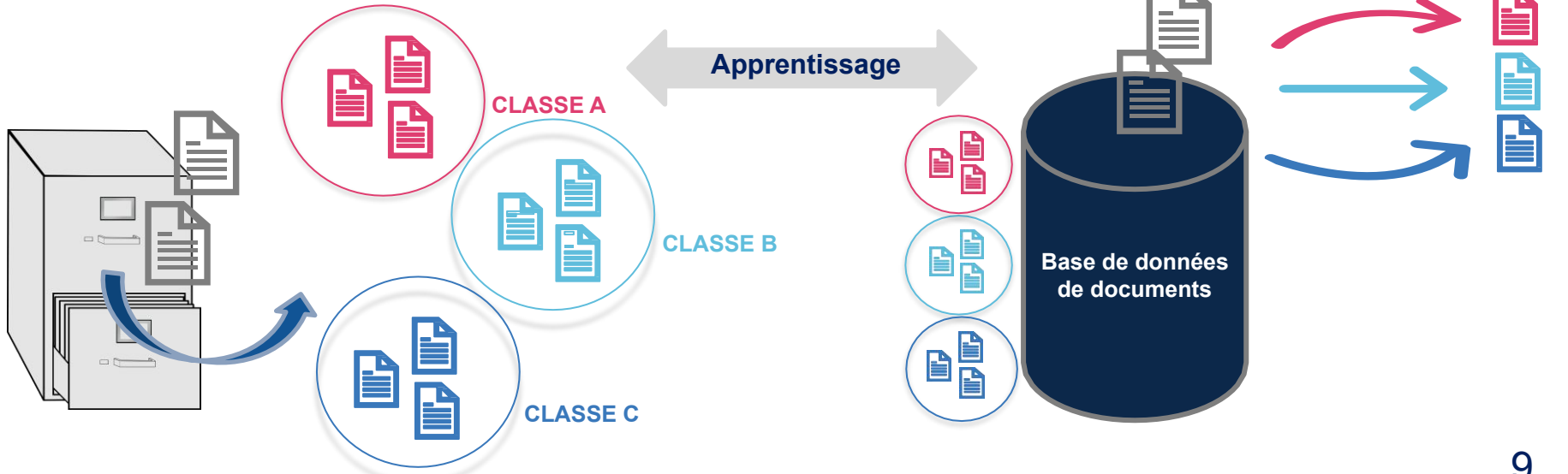


Quelques techniques de TDM

Problématique

Classer les documents selon :

- les thèmes de ces documents
- les zones géographiques considérées
- ...



CLASSIFICATION

CLASSIFICATION SUPERVISÉE

Apprentissage sur données déjà étiquetées

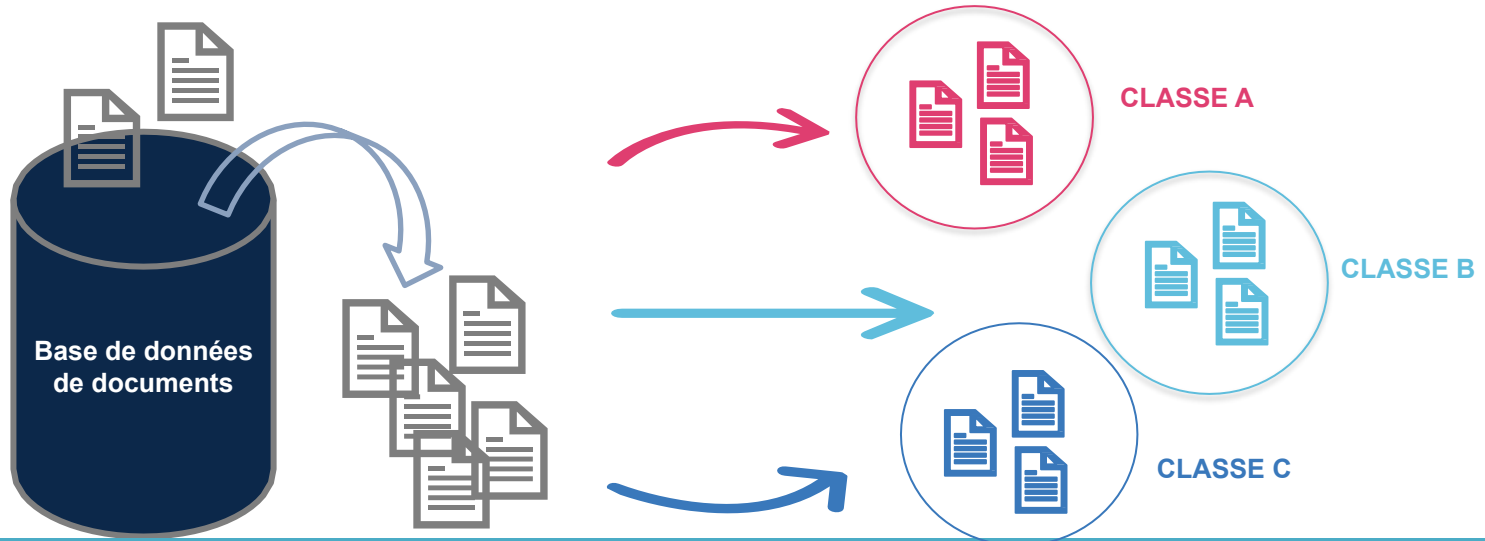


Quelques techniques de TDM

Problématique

Classer les documents selon :

- les thèmes de ces documents
- les zones géographiques considérées
- ...



CLASSIFICATION

CLASSIFICATION NON SUPERVISEE
(clustering-regroupement)

TDM



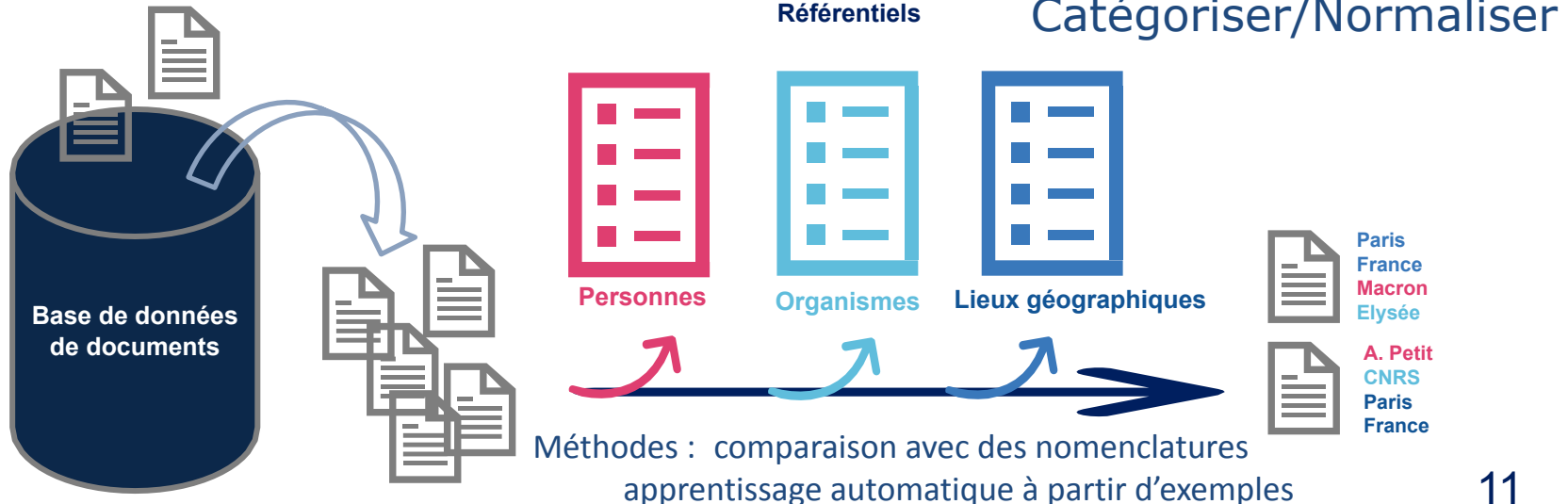
Quelques techniques de TDM

Extraction d'info (reconnaissance de termes)

RECONNAISSANCE D'ENTITES NOMMEES

Problématique

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



TDM

Indexation



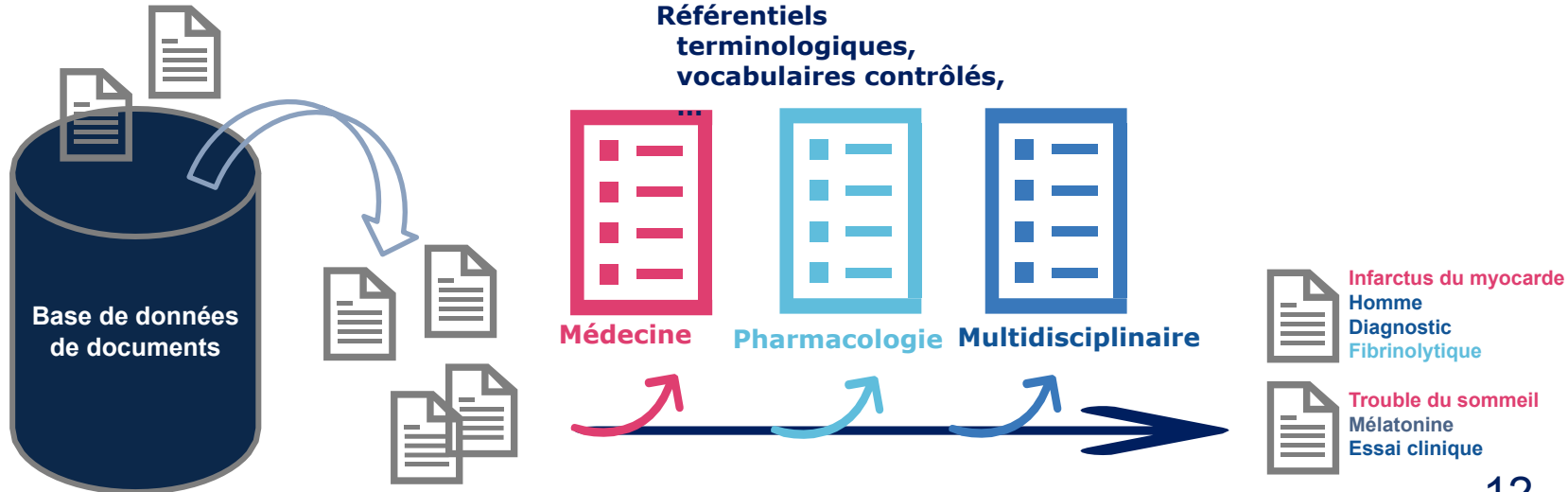
Quelques techniques de TDM

LIBRE ET/OU CONTROLÉE

(contrôlée = par rapport à des référentiels)

Problématique

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



1 TDM : Quelques rappels

2 **TDM à l'Inist et démos**
Outils et services

3 Présentation des TPs

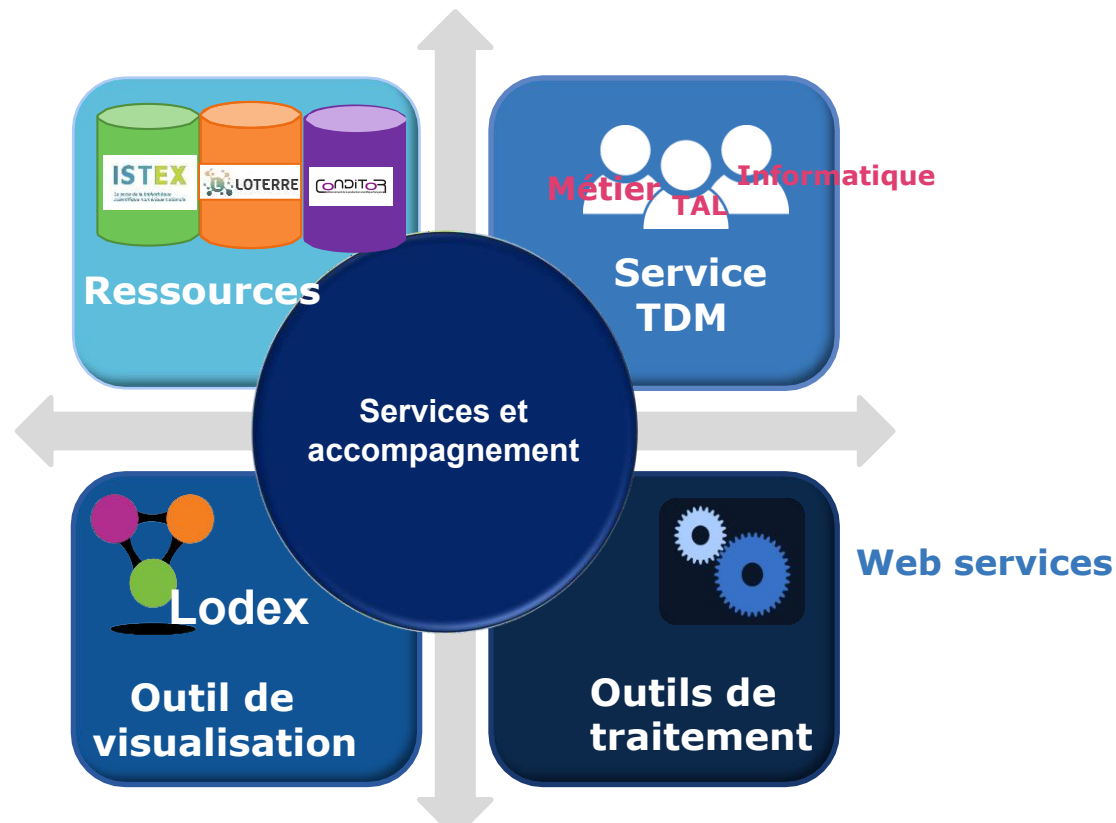
4 A vous de jouer



Quelles opportunités à l'INIST-CNRS ?

Base bibliographique

Base terminologique





Le TDM pour tous grâce à des WEB SERVICES dans LODEX

Pour faciliter l'accès aux techniques de fouille de données, l'Inist développe des web services autour du traitement de l'information scientifique et technique.



VALORISATION
de programmes
et algorithmes TDM

WEB SERVICES DE TEXT MINING

ENRICHISSEMENTS DE DONNÉES

- Homogénéisation
- Indexation / classification
- Attribution d'identifiants

Récupération d'informations de Crossref, Unpaywall...

VISUALISATION et TABLEAUX DE BORD avec LODEX



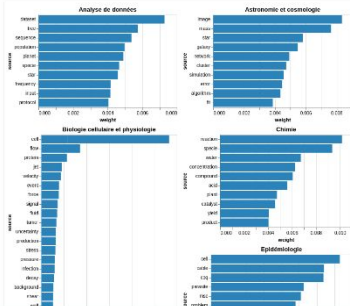
Pour les décideurs, professionnels de l'information et utilisateurs non spécialistes



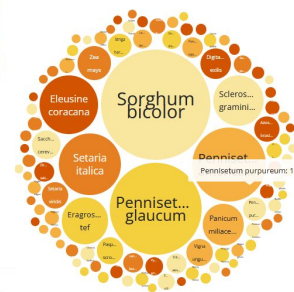
Contact : valerie.bonvallot@inist.fr



Caractérisation des thématiques par méta-clés



Noms d'espèces détectées (IRC3species)

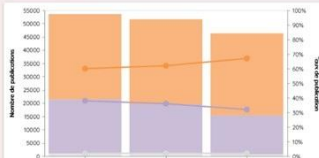


Science ouverte - Poids du libre accès dans la production CNRS

Etude commanditée par la Direction des Données Ouvertes de la Recherche (DDOR) et réalisée par l'Inist-CNRS. Les références, pour les années de publication 2017 à 2019, ont été téléchargées (déchargement en mars 2021) dans la base Core Collection de Clarivate Analytics et ont été enrichies en mars 2021 par des données du SAPPs sur les instituts et des données obtenues à partir du service Unpaywall.

CNRS - Publications
151 431

Années - Accès ouvert

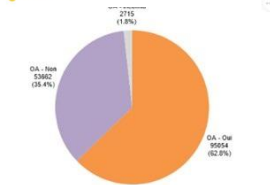
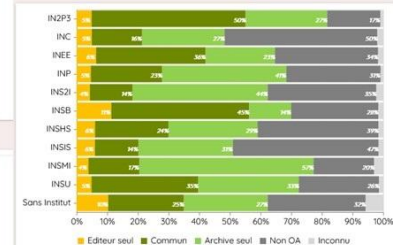


VOIR LES DÉTAILS

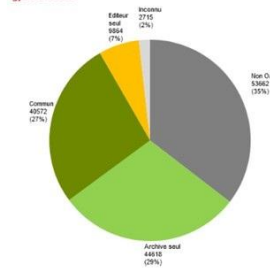
Diffusion des publications



CNRS - Instituts - Types d'accès



Types d'accès



Des outils et services de TDM pour tous

Des web-services pour aider...

ISTEX TDM

Les services Istex pour la fouille de textes

<https://services.istex.fr/>

Web service (WS)

interface et protocole d'échange en ligne de données

1 WS = 1 tâche, un traitement spécifique

Peu de compétences informatiques (transparence du langage, pas d'installation)

Paramétrage minimal

Données issues de différentes sources

Recensement et description

Modalités d'utilisation

Cas d'usage - illustration

Rechercher un web service

Tapez ici votre recherche, p.ex. : Classification

RECHERCHER

Nos derniers web services

Résumés - Texte intégral

entityTag
EXTRACTION D'ENTITÉS NOMMÉES
(PERSONNES, LOCALISATIONS,
ORGANISMES ET AUTRES)

Adresses et affiliations

IdRorDetect
ASSOCIER UN IDENTIFIANT ROR À
UNE ADRESSE D'AFFILIATION

Résumés - Texte intégral

noiseDetect
 DÉTECTION DE BRUIT D'UN CORPUS

Résumés - Texte intégral

sciencematrixClass
CLASSIFICATION EN DOMAINES
SCIENTIFIQUES SCIENCE-METRIX

Trouvez un service web correspondant à vos besoins

Nous développons et mettons à votre disposition des outils de TDM (Text and Data Mining) faciles à mettre en œuvre, couplés à un outil de création de tableaux de bord dynamiques.

Actuellement **38** web services sont disponibles

COMMENT LES UTILISER ?

VOIR LA DOCUMENTATION

Des outils et services de TDM pour tous

Des web-services pour aider...

ISTEX TDM
Les services Istex pour la fouille de textes

Rechercher un web service

Tapez ici votre recherche, p.ex. : Classification

RECHERCHER

Nos derniers web services

Résumés - Texte intégral

entityTag
EXTRACTION D'ENTITÉS NOMMÉES
(PERSONNES, LOCALISATIONS,
ORGANISMES ET AUTRES)

Résumés - Texte intégral

noiseDetect
 DÉTECTION DE BRUIT D'UN CORPUS

Résumés

textClustering
EXTRACTION DE CLUSTERS D'UN
CORPUS

Adresses et affiliations

IdRorDetect
ASSOCIER UN IDENTIFIANT ROR À
UNE ADRESSE D'AFFILIATION

Résumés - Texte intégral

sciencematrixClass
CLASSIFICATION EN DOMAINES
SCIENTIFIQUES SCIENCE-METRIX

Texte intégral

textExtract
EXTRACTION DU TEXTE À PARTIR
D'UN PDF

VOIR TOUS LES SERVICES

ISTEX TDM
Les services Istex pour la fouille de textes

<https://services.istex.fr/>

OBJET TRAITÉ

- Adresses et affiliations (9)
- Auteurs (2)
- Éléments catalographiques (3)
- Résumés (21)
- Texte intégral (21)

LANGUES (3)

- Anglais (31)
- Français (23)
- Autre (13)

TRAITEMENT (7)

- Classification (8)
- Extraction d'entités nommées (9)
- Homogénéisation (7)
- Indexation (6)
- Traitement automatique du langage (3)
- Prétraitement (3)
- Validation (1)

TYPE DE DONNÉES (2)

- Corpus (4)
- Document (30)

entityTag Extraction d'entités nommées (Personnes, Localisations, Organismes et autres)

Ce web service extrait d'un texte diverses entités nommées. Deux variantes existent : la première fonctionne sur des textes indépendamment de la langue et propose 4 types d'entités, la seconde fonctionne sur des textes en anglais et propose davantage de...

noiseDetect Détection de bruit d'un corpus

Ce web service traite non plus du texte mais de corpus de textes en anglais. En effet, le résultat obtenu pour chacun des documents dépend des autres. L'algorithme repère la liste des identifiants des documents considérés comme du bruit dans...

textClustering Extraction de clusters d'un corpus

Ce web service traite non plus du texte mais de corpus de textes en anglais. En effet, le résultat obtenu pour chacun des documents dépend des autres. L'algorithme extrait plusieurs groupes (clusters) d'un corpus afin d'y classer les différents textes...

IdRorDetect Associer un identifiant ROR à une adresse d'affiliation

Ce web service prend en entrée une affiliation pour interroger l'API ROR et renvoie un identifiant.

sciencematrixClass Classification en domaines scientifiques Science-Metrix

Ce web service classe des documents en anglais dans les trois niveaux de la classification Science-Metrix.

textExtract Extraction du texte à partir d'un PDF

Ce web service transforme un PDF en texte en excluant les éléments qui perturberaient le traitement de fouille de texte ultérieur. Il doit pas être un PDF image.

Accueil > Web-services > Classification en domaines scientifiques Science-Metrix

sciencematrixClass - Classification en domaines scientifiques Science-Metrix

Description Utilisation Cas d'usage

Niveau d'utilisation : Débutant
Niveau de validation : Expérimental

Objectif

Ce web service classe des documents en anglais dans les trois niveaux de la classification Science-Metrix.

Méthode

Les trois niveaux de la classification sont renvoyés dans un tableau dans un champ "classif" et le i-ème élément du tableau correspond au i-ème niveau.

Le modèle utilisé est un réseau de neurones, entraîné par apprentissage supervisé en utilisant la bibliothèque fastText. Les labels utilisés pour les dispositions étaient des domaines scientifiques Science-Metrix de revue et non de documents : nous avons appliqué l'algorithme de Science-Metrix en utilisant

Des outils et services de TDM pour tous

Des web-services pour aider...

ISTEX TDM
Les services **Istex** pour la fouille de textes

<https://services.istex.fr/>

Onglet “Description”

- Niveau d'utilisation / Niveau de validation
- Objectif
- Méthode, modèles et ressources
- Métrique et précaution d'utilisation
- Variantes
- Références
- Ces web services qui peuvent vous intéresser

Onglet “Utilisation”

- URL pour Lodex
- Exemple du traitement (entrée □ sortie)
- Pour aller plus loin
- Lien vers OpenApi (pour tester)
- Lien vers le code source (gitbucket-github)

Onglet “Cas d'usage”

- Définition des besoins
- Illustrations

Des outils et services de TDM pour tous

Des web-services pour aider...

Utilisables par des néophytes via

LODEX outil de visualisation

Description

Utilisation

Cas d'usage

URL DU WEB SERVICE À RENSEIGNER DANS LODEX PRÉCALCUL EST :
<https://text-clustering.services.istex.fr/v1/noise-lodex>



Interface IA Factory

ISTEX IA Factory
L'IA appliquée à vos corpus

CHARGEZ VOS CORPUS ET DÉCOUVREZ LES RÉSULTATS DES SERVICES TDM

IA Factory est une interface de chargement de corpus et d'exploitation d'outils de TDM vous permettant d'explorer vos corpus documentés en quelques clics :

- Téléchargez vos données et choisissez le format et le champ à traiter.
- Choisissez le service web de TDM que vous voulez effectuer.
- Remplissez votre adresse mail.

À l'issue du traitement vous recevrez un mail avec un lien de téléchargement du résultat.

1 Téléversement 2 Configuration 3 Vérification 4 Confirmation

Convertisseur
Transformation d'un fichier ISTE X (format tar.gz) en fichier corpus (v1/istex-tar.gz)

Le fichier est transformé en fichier corpus exploitable par un web service asynchrone
(* Nom du champ à exploiter comme identifiant de ligne (par défaut value)

Dans un corpus, retourne la liste des identifiants des documents considérés comme du bruit. (v1/noise)

Des outils et services de TDM pour tous

Des web-services pour aider...

...mais aussi en ligne de commande, appelés dans des programmes



The screenshot shows the Swagger UI for the 'terms-extraction' API. At the top, there is a Swagger logo and a dropdown menu labeled 'Select a definition' with 'terms-extraction - Extraction de termes' selected. Below this, the API title 'terms-extraction - Extraction de termes' is displayed with version indicators '1.0.0' and 'OAS 3.1'. The URL 'https://terms-extraction.services.istex.fr' is shown. A description states: 'Extraction de termes à partir de textes en anglais ou en français.' Below the description are links for 'web services', 'Terms of service', and 'Inist-CNRS - Website'. A 'Servers' section contains a dropdown menu with the value '[scheme]://{hostname} - EZS server'. The 'Computed URL' is 'https://terms-extraction.services.istex.fr/'. Under 'Server variables', there are two input fields: 'scheme' with a dropdown set to 'https' and 'hostname' with the text 'terms-extraction.services.is'.

...et des tests en ligne (swagger).

<https://openapi.services.istex.fr/?urls.primaryName=terms-extraction%20-%20Extraction%20de%20termes#/terms-extraction/post-v1-teeft-en>

Des outils et services de TDM pour tous

Des web-services pour aider...

langDetect Détection de la langue d'un texte

Le web service détecte la langue d'un document
texte.

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."
```

ISTEX TDM

Les services Istex pour la fouille de textes

Après

```
<> "en"
```

Des outils et services de TDM pour tous

Des web-services pour aider...

Teeft Extraction de termes d'un texte via Teeft

Le service web Teeft extrait, par défaut, les 5 termes les plus spécifiques d'un texte en anglais ou en français. Il permet ainsi d'avoir une idée de ce dont il est question dans le texte.

ISTEX TDM

Les services Istex pour la fouille de textes

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...

Classification supervisée

pascalFrancisClass **Classification en domaines** **scientifiques Pascal-Francis**

Le web service classe automatiquement 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 possède un domaine scientifique homogène, dans la mesure où les...

Avant

```
"Rhesus Monkey Model Self Injury effect  
Relocation Stress Behavior Neuroendocrine  
Functionbackground self injurious behavior  
SIB disorder many individual clinical  
nonclinical population state stress arousal  
longitudinal datum relationship increase  
(...)  
significant stressor rhesus macaque stressor  
increase self behavior sleep disturbance  
monkey SIB finding life stress SIB human  
disorder HPA axis result potential role CBG  
long term neuroendocrine response major  
stressor"
```

Après



```
"003": "Sciences humaines et sociales",  
"770": "Psychologie. Psychanalyse. Psychiatrie.",  
"770D": "Psychopathologie. Psychiatrie."
```

ISTEX TDM

Les services **Istex** pour la fouille de textes

Des outils et services de TDM pour tous

Des web-services pour aider...

Détection de code RNSR

rnsrRuleDetect Attribution d'identifiant(s) RNSR à une adresse (Alignements)

Le web service attribue, à l'aide de règles, un ou plusieurs identifiants RNSR à partir d'une adresse d'affiliation d'auteur et d'une année de publication. Quand aucun code RNSR n'est trouvé, le service renvoie un tableau vide.

rnsrLearnDetect Attribution d'identifiant(s) RNSR à une adresse (Apprentissage)

Ce web service attribue un ou plusieurs identifiant(s) RNSR à partir d'une adresse d'affiliation d'auteur en langue française.



Avant

"CNRS UMR AMAP MONTPELLIER FRA"

Après

<> RNSR://200317641S

Des outils et services de TDM pour tous

Des web-services pour aider...

ISTEX IA Factory
L'IA appliquée à vos corpus

<https://ia-factory.services.istex.fr/>

CHARGEZ VOS CORPUS ET DÉCOUVREZ LES RÉSULTATS DES SERVICES TDM

IA Factory est une interface de chargement de corpus et d'exécution d'outils de TDM vous permettant d'exploiter vos propres données en simplement 3 étapes:

- Téléchargez vos données et choisissez le format et le champ à traiter,
- Choisissez le service web de TDM que vous voulez exécuter,
- Remplissez votre adresse mail.

À l'issue du traitement vous recevrez un mail avec un lien de téléchargement du résultat.

1

Téléversement

2


Configuration

3

Vérification

4

Confirmation

 Déposer un fichier

Fichier manquant

SUIVANT

Par exemple :

- chargez vos données au format CSV : VosData.csv (dans le cas de corpus ISTEEX on pourra utiliser le fichier compressé fourni par ISTEEX),
- choisissez le convertisseur adapté au csv (« Transformation d'un fichier ISTEEX en fichier csv ») (Dans le cas de corpus ISTEEX on pourra utiliser « Transformation d'un fichier ISTEEX (format tar.gz) en fichier corpus »),
- dans la liste déroulante, sélectionnez le nom de la colonne que vous voulez traiter,
- puis sélectionnez le web service que vous voulez exécuter,
- et enfin remplissez votre adresse mail, vous recevrez un message au lancement du service et à la fin du traitement avec un lien de téléchargement des résultats.

Des web services en ligne sans passer par Lodex (en cours)

Extrait des thématiques d'un corpus. (/v1/lda)

Applique un algorithme d'alignement avec idRef prévu dans le cadre du projet Rapido (/v1/rapido)

Extraction des termes d'un fichier corpus en anglais (/v1/en)

Extraction des termes d'un fichier corpus en français (/v1/fr)

Valide l'ensemble des références bibliographiques d'un PDF. (/v1/bibcheck-pdf)

Classification en domaines scientifiques Science-Metrix (/v1/sciencematrix-class)

Dans un corpus, retourne la liste des identifiants des documents considérés comme du bruit. (/v1/noise)

Crée différents `clusters` à partir d'un ensemble de textes courts ou d'un ensemble de listes de mots-clés. (/v1/clustering)

Des outils et services de TDM pour tous

... et un catalogue d'outils

ISTEX DATA

Catalogues de données, de ressources et d'outils.

<https://data.istex.fr/instance/tm-tools-explorer>

Le catalogue TM Tools explorer

AbLang

Modèle de langue sur les anticorps.



ABLTagger

Etiqueteur morphosyntaxique pour l'islandais.



ABNER

ABNER est un outil logiciel open source pour l'analyse...



Adaboost

AdaBoost est un méta-algorithme de classification...



AdaGram Python

Implémentation d'Adagram en Python.



AdaGram.jl

Implémentation du skip-gram adaptatif en Julia.



Explorez divers outils de TDM

Cette application Lodex a été créée pour visualiser de manière simple des références d'outils de TDM sélectionnés depuis une liste de trois cents outils spécialisés dans le traitement automatique du langage et l'exploration de texte.

Pour en savoir plus

MÉTHODOLOGIE

Actuellement **548** outils sont référencés dans ce catalogue

Créateur

Repose sur une **ontologie : OntoTM**

Publiée sur le portail terminologique Loterre

Réalisé avec Lodex



Des outils et services de TDM pour tous

... et un catalogue d'outils

- 548 outils recensés
- 10 facettes
- 8 graphes
- 2 langues (bilinguisme)

The screenshot displays the ISTE DATA website interface. At the top, the logo 'ISTEX DATA' is visible with the tagline 'Catalogues de données, de ressources et d'outils.' Below the logo is a search bar with the placeholder text 'Vous pouvez saisir votre recherche ici'. The main content area is divided into a left sidebar with filters and a main grid of tool cards. The filters include: 'Outils de fouille de données', 'Licence', 'Pays de création', 'Tâches des outils', 'Langue(s) traitée(s)', 'Langage de programmation', 'Format(s) d'entrée', 'Interface utilisateur', 'Système(s) d'exploitation', and 'Domaine d'application'. The main grid shows 10 tool cards, each with a title and a brief description. The tools listed are: ClinicalBERT, TopicModels.jl, CodeBERT, Twitmo, Analec, KLUE-BERT, Tropes, CancerBERT, encodeur universel de phrases, and BERTTurk. At the bottom of the grid, there is a link 'VOIR PLUS DE RÉSULTATS (10 / 548)'. The footer contains navigation icons for 'Accueil', 'Graphiques', and 'Recherche', along with a language selector set to 'français' and a 'Voir Plus' menu icon.

Des outils et services de TDM pour tous

... et des news

The screenshot displays the IStex website interface. At the top, the IStex logo is accompanied by the tagline 'Le socle de la bibliothèque scientifique numérique nationale'. Navigation links include 'Base documentaire', 'Constitution de corpus', 'Exploration et enrichissement', 'Offre de formations', and 'Institutions adhérentes'. A central banner highlights the 'service de la recherche française' with statistics: 28,0 M de documents, 9 492 revues, 438 909 ebooks, and publications from 1473 to 2023. Below this, a 'À la une' section features a featured article about a new TDM web service on IStex, dated 26 September 2024. A sidebar on the right lists other news items, with the top one dated 10 September 2024 regarding the presentation of the TDM service, which is highlighted with a red box.

ISTEX
Le socle de la bibliothèque scientifique numérique nationale

Base documentaire Constitution de corpus Exploration et enrichissement Offre de formations Institutions adhérentes

service de la recherche française

28,0 M de documents 9 492 revues 438 909 ebooks publiés de 1473 à 2023

À la une Voir toutes les actualités →

10 septembre 2024
Istex...de la base documentaire au TDM : Présentation en visio
En savoir plus →

6 août 2024
Jean Zay et le festival de Cannes
En savoir plus →

31 juillet 2024
Présentation d'Istex en visio à la demande
En savoir plus →

15 juillet 2024
Il y a 900 ans naissait Aliénor d'Aquitaine reine de France puis d'Angleterre...
En savoir plus →

26 septembre 2024 : Outils associés
Un nouveau web service de TDM sur Istex
Lire l'article →

Des outils et services de TDM pour tous

QRuiz.net



IT'S
QUIZ
TIME!

<https://quizz.net/Q/?NmFB7u>

<https://fr.dreamstime.com/phrase-son-temps-quiz-tableau-vert-sale-bord-image228637570>

1 TDM : Quelques rappels

2 **TDM à l'Inist et démos
Démonstrations**

3 Présentation des TPs

4 A vous de jouer



Des outils et services de TDM pour tous



CORPUS Biodiversité
La conservation des mammifères sur la liste rouge de l'IUCN*

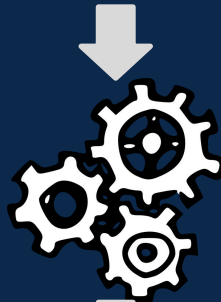
<https://vie-collection.corpus.istex.fr/ark:/67375/8BV-2VJML8C0-C>

Démarche TDM

Extraire un corpus, **Collecter** les données

Nettoyer les données, **Prétraiter** les données

(formats : pdf, txt, jpeg, xml, json, ... encodages de données et les transformer si besoin)



TDM

- Extraction
- Etiquetage
- Vectorisation
- Modélisation

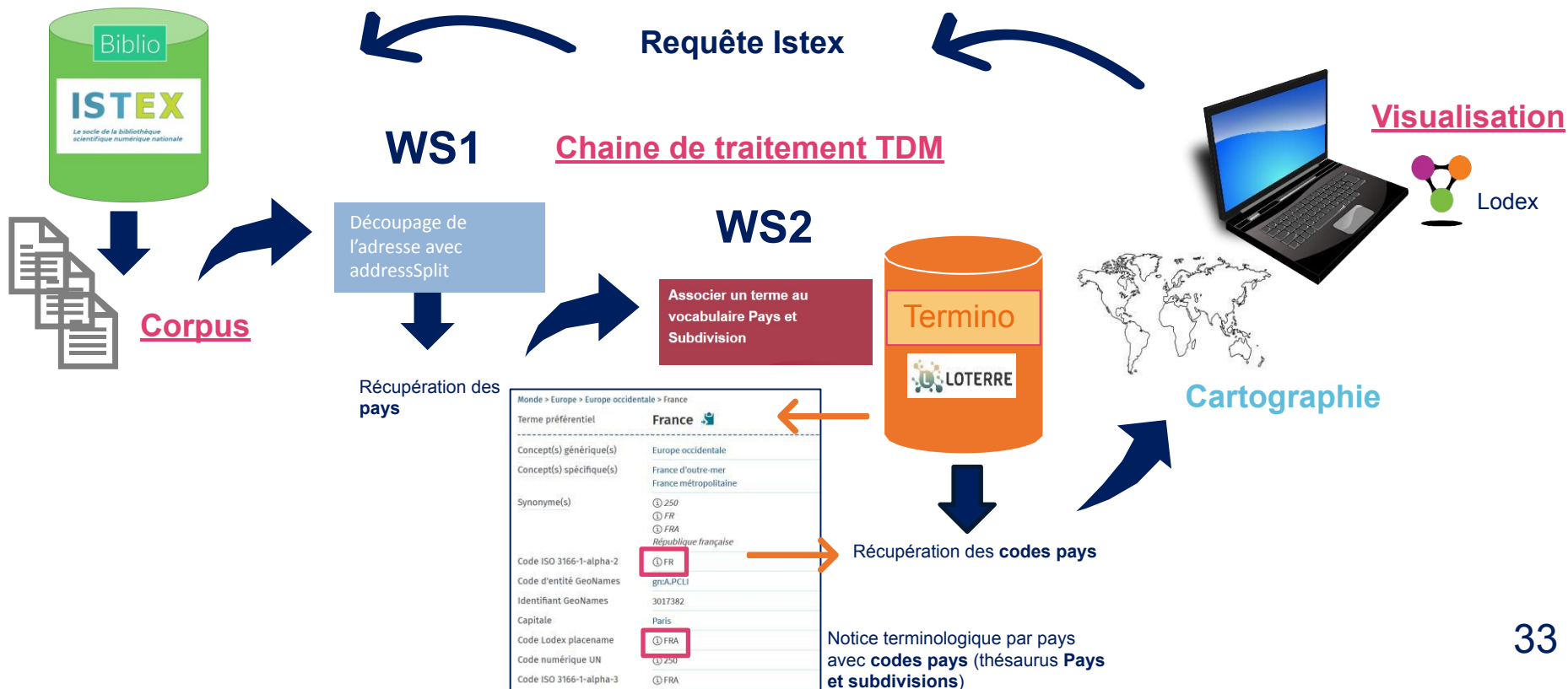


Résultats

Visualiser les résultats dans un outil dédié

Affiner et interpréter les résultats

Démarche TDM dans notre cas d'usage



Les WS en détails

Le web service [addressSplit](#) découpe une adresse au format texte en plusieurs champs.

Inist-CNRS

2, rue Jean Zay

CS 10310

F-54519 Vandœuvre-lès-Nancy

France



```
"house": "inist-cnrs",  
"house_number": "2",  
"road": "rue jean zay cs 10310",  
"postcode": "f-54519",  
"city": "vandœuvre-lès-nancy",  
"country": "france"
```

WS1

Découpage de l'adresse avec
addressSplit

URL DU WEB SERVICE À RENSEIGNER DANS LODEX ENRICHISSEMENT EST :

<https://affiliations-tools.services.istex.fr/v1/adresses/parse>



URL à copier-coller dans Lodox

Les WS en détails

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.

```
{
  "id": 2,
  "value": "Grand-Duché de Luxembourg"
},
```



```
{
  "id": 2,
  "value": {
    "id": "grandduchedeluxembourg",
    "cartographyCode": "LUX",
    "about": "http://data.loterre.fr/ark:/67375/9SD-HXXBRCFQ-F",
    "prefLabel@fr": "Luxembourg",
    "prefLabel@en": "Luxembourg",
    "wikidataURI": "https://www.wikidata.org/wiki/Q1842",
    "geonameURI": "https://www.geonames.org/2960313",
    "countryCode": "LU",
    "latitude": "49.765700074151",
    "longitude": "5.965223432344",
    "localization@en": [
      "Western Europe"
    ],
    "localization@fr": [
      "Europe occidentale"
    ]
  }
},
```

Code pays

WS2

Associer un terme au vocabulaire Pays et Subdivisions

URL DU WEB SERVICE À RENSEIGNER DANS LODEX EST :

<https://loterre-resolvers.services.istex.fr/v1/9SD/identify>



URL à copier-coller dans Lodox



Démarche TDM dans notre cas d'usage

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

CORPUS Biodiversité

La conservation des mammifères sur la liste rouge de l'IUCN*

<https://vie-collection.corpus.istex.fr/ark:/67375/8BV-2VJML8C0-C>

-  Quelles sont les 3 premières espèces concernées en France ?
-  Quels pays sont concernés par le « Lion » ?

Réponse aux questions

visualisation des données dans Lodex...



data.istex.fr

data.istex.fr expose le Triple Store des données ISTEX via son SPARQL Endpoint.

◀ Data ISTEX / Corpus scientifiques / Corpus actualité / Collection Sciences de la Vie

Biodiversité

La conservation des mammifères sur la liste rouge de l'IUCN



- Auteur(s) ▾
- Revue/Monographie ▾
- Editeur ▾
- Année de publication ▾
- Langue de publication ▾
- Type de publication ▾
- Type de document ▾
- Mots-clés d'auteur ▾
- Mot-clés Teef ▾
- Noms d'espèce détectés ▾
- Catégorie INIST ▾
- Catégorie Science Metrix ▾



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.)

Q Vous pouvez saisir votre recherche ici

4487 ressources trouvées sur un total de 4487

Trier Ascendant Exporter

Niche dimensions of New England cottontails in relation to habitat patch size

We examined physical condition, niche dimensions, and survival of New England cottontails (*Sylvilagus transitionalis*) that occupied 21 habitat patches of different sizes during winter. Rabbits on small patches (≤2.5 ha) were predominantly males, and both sexes had lower body mass than individuals on large patches (≥5.0 ha). Niche indices (β , where β ranges from 0 to 1, and values approaching 1 indicate generalized resource use) of habitat use revealed that rabbits on small patches used a greater variety of microhabitats (based on understory stem density, B_s , and proximity to cover: B_c) than...

Ecologia 1993

Influence of fur trade, famine, and forest fires on moose and woodland caribou populations in northwestern Ontario from 1786 to 1911

Hudson's Bay Company records were used to estimate the 1786-1911 annual number of moose (*Alces alces andersonii*) and caribou (*Rangifer tarandus caribou*) involved in trade by northern Ojibwa natives to the company post at Osaburgh House (51°10'N 90°15'W) in northwest Ontario, Canada. The human population for the early 19th century, and the number and severity of human starvations from 1786 to 1911 were estimated. The extent of forest fires in the region around Osaburgh was documented using a "fire-dav" index computed from Hudson's Bay Company journals and usina...

Environmental Management 1993

A review of otters (Carnivora: Mustelidae: Lutrinae) in Malaysia and Singapore

Four species of otters have been recorded from Malaysia and Singapore in the past: Lutra lutra (Common or Eurasian Otter), Lutra sumatrana (Hairy-nosed Otter), Lutrogale perspicillata (Smooth Otter) and Amblyonyx cinereus (Oriental Small-clawed Otter). All four are listed in the Threatened Species Categories of the IUCN Red List of Threatened Animals. L. lutra is designated 'Vulnerable' and the status of the other three Asian species are 'Insufficiently known' due to lack of information. From a review of the available literature and collation of museum records from Malaysia and Sincapore, the past stat...

Hydrobiologia 1994

Seasonal changes in the diets of migrant and non-migrant nectarivorous bats as revealed by carbon stable isotope analysis

Three species of nectar-feeding bats migrate from tropical and subtropical Mexico into the Sonoran and Chihuahuan deserts during the spring and summer months. We examined geographic and seasonal changes in the diet of one migrant species, *Leptonycteris curasoae*, using carbon stable isotope techniques to determine the relative importance of C3 and CAM (Cactaceae, Agavaceae) plants in its diet. We also examined the diet of a non-migratory nectar-feeding bat, *Glossophaga soricina*, from southern Mexico using the same techniques. We found that *L. curasoae* feeds extensively or...

Ecologia 1993

The quandary of local people—Park relations in Nepal's Royal Chitwan National Park

This paper analyzes five major causes of park-people conflicts that have occurred in Nepal's Royal Chitwan National Park. The causes include illegal transactions of forest products from the park, livestock

Réponse aux questions visualisation des données dans Lodex...



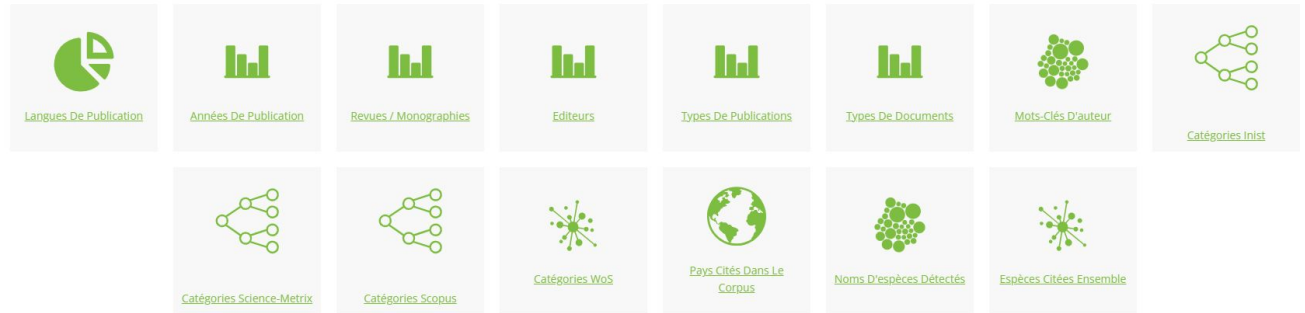
data.istex.fr

data.istex.fr expose le Triple Store des données ISTEEX via son SPARQL Endpoint.

◀ Data ISTEEX / Corpus scientifiques / Corpus actualité / Collection Sciences

Biodiversité

La conservation des mammifères s



Accueil



Graphiques



Recherche



Accueil



Graphiques



Recherche



Voir Plus

Réponse aux questions visualisation des données dans Lodex...



data.istex.fr

data.istex.fr expose le Triple Store des données ISTEX via son SPARQL Endpoint.

◀ Data ISTEX / Corpus scientifiques / Corpus actualité / Collection Sciences de la Vie

Biodiversité

La conservation des mammifères sur la liste rouge de l'IUCN



Auteur(s) ▼

Revue/Monographie ▼

Editeur ▼

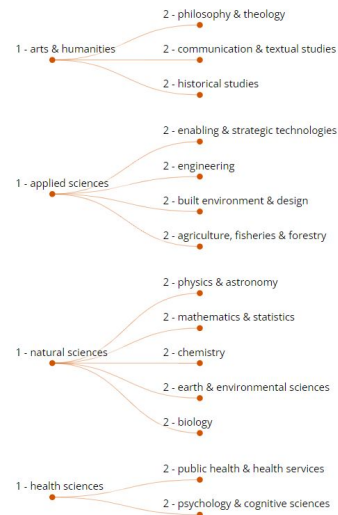
Année de publication ▼

Langue de publication ▼

Type de publication ▼

Type de document ▼

Catégories Science-Metrix



Accueil



Graphiques



Recherche

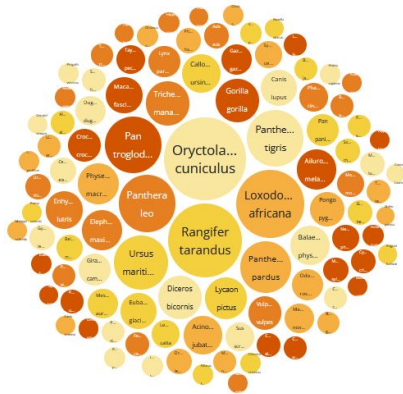


Voir Plus

Réponse aux questions visualisation des données dans Lodex...

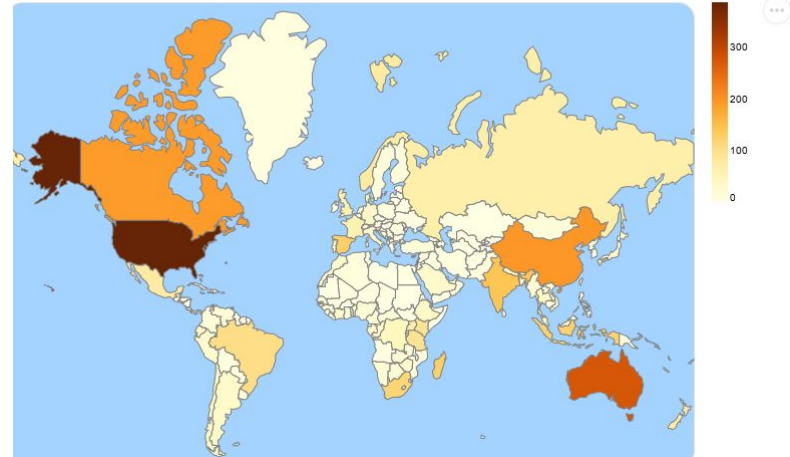
Utilisation des **facettes** pour faire varier les données

Noms d'espèces détectées



Quelles sont les 3 premières espèces concernées en France ?

Pays cités dans le corpus



Quels pays sont concernés par le « Lion » ?

Réponse aux questions visualisation des données dans Lodex...



data.istex.fr

data.istex.fr expose le Triple Store des données ISTEEX via son SPARQL Endpoint.

◀ Data ISTEEX / Corpus scientifiques / Corpus actualité / Collection Sciences de la Vie



Biodiversité



La conservation des mammifères sur la liste rouge de l'IUCN



Accueil



Graphiques



Recherche



Voir Plus



Admin



Déconnexion

Réponse aux questions visualisation des données dans Iodex...

Tableau de données sous forme tabulaire

LODEX

DONNÉES AFFICHAGE

DÉPUBLIER

Données

Enrichissements

Précalculs

Ressources cachées

COLONNES FILTRES DENSITÉ AJOUTER

uri	Titre	Auteur(s)	Affiliation(s)	Revue ou mon...	Auteur(s) monogra...	ISSN	e-ISSN	Volume	Numéro	Page début	Pa
"ark:/67375/WNG-LD"	"Multiscale distributor	"Mahi Puri", "Arjun Sri	"Wildlife Conservatio	"Diversity and Distribu		"1366-9516"	"1472-4642"	"21"	"9"	"1087"	
"ark:/67375/WNG-TPF"	"Direct evidence impli	"Blair Hardman", "Dori		"Ecological Managem		"1442-7001"	"1442-8903"	"17"	"2"	"152"	
"ark:/67375/WNG-WC"	"Drivers of red fox (Vu	"F. Díaz-Ruiz", "J. Can	"Instituto de Investig	"Journal of Zoology"		"0952-8369"	"1469-7998"	"298"	"2"	"128"	
"ark:/67375/WNG-3PE"	"Contemporary niche	"Fernando Martínez-I	"CIBIO/InBio, Centro	"Diversity and Distribu		"1366-9516"	"1472-4642"	"22"	"4"	"432"	
"ark:/67375/WNG-LX"	"Fiddling in biodiversit	"S. M. Durant", "T. Wa	"Institute of Zoology,	"Diversity and Distribu		"1366-9516"	"1472-4642"	"20"	"1"	"114"	
"ark:/67375/WNG-B8"	"Rates of Disturbance	"P. E. Komers", "Z. Stz	"Department of Biosci	"Global Change Biolog		"1354-1013"	"1365-2486"	"19"	"9"	"2916"	
"ark:/67375/WNG-KHK"	"Evolutionary history :	"A. Witting", "R. Patel"	"Leibniz Institute for	"Journal of Zoology"		"0952-8369"	"1469-7998"	"299"	"4"	"239"	
"ark:/67375/WNG-TP"	"Effects of ranger stat	"A. Ghoddousi", "A. Ki	"Workgroup on Endg	"Animal Conservation"		"1367-9430"	"1469-1795"	"19"	"3"	"273"	
"ark:/67375/WNG-SDI"	"Cascading Effects of	"Alison M. Behie", "Su	"School of Archaeol	"Biotropica"		"0006-3606"	"1744-7429"	"46"	"1"	"25"	
"ark:/67375/WNG-2LC"	"Mapping perceptions	"Nicola K. Abram", "Er	"Living Landscape A	"Diversity and Distribu		"1366-9516"	"1472-4642"	"21"	"5"	"487"	
	"The Mediterranean n	"Alexandros A. Karan	"MOM/Hellenic Socie	"Mammal Review"		"0305-1838"	"1365-2907"	"46"	"2"	"92"	
"ark:/67375/WNG-JKC"	"Genetics at the verge	"M. Casas-Marce", "L	"Department of Integ	"Molecular Ecology"		"0962-1083"	"1365-294X"	"22"	"22"	"5503"	
"ark:/67375/WNG-IHT"	"What Ecological and	"Rafael Reyna-Hurtar	"El Colegio de la Fro	"Biotropica"		"0006-3606"	"1744-7429"	"48"	"2"	"246"	

55 colonnes chargées 7 colonnes enrichies

Lignes par page 25 1-25 of 500

Réponse aux questions visualisation des données dans Iodex...

LODEX

DONNÉES

AFFICHAGE



- Données
- Enrichissements**
- Précalculs
- Ressources cachées

COLONNES FILTRES DENSITÉ AJOUTER

DOI	DOI revue/mon...	IodexStamp	Repérage des esp...	AffiliationPretraitem...	enrichissement_ad...	AffiliationsNettoyees	AffiliationsHomog	Code pays	bibCheck
WNG-LD: "10.1111/ddi.12335"	"10.1111/(ISSN)1472-	{"importedDate":"Mon	["Meiurus ursinus"]	["Wildlife Conservator	[{"id":"Wildlife Consen	[{"id":"usa"]	[{"id":"usa"]	[{"cartographyCode":	{"doi":"10.1111/ddi.12
WNG-TPf "10.1111/emr.12210"	"10.1111/(ISSN)1442-	{"importedDate":"Mon	["Felis catus","Lagorcl	[]	[{"id":"","country":	[{"id":"","country":	[{"id":"","country":	[{"cartographyCode":	{"doi":"10.1111/emr.12
WNG-WC "10.1111/jzo.12294"	"10.1111/(ISSN)1469-	{"importedDate":"Mon	["Oryctolagus cunicul	["Instituto de Investiga	[{"id":"Instituto de inve	[{"id":"portugal","un	[{"id":"portugal","un	[{"cartographyCode":	{"doi":"10.1111/jzo.12
WNG-3PE "10.1111/ddi.12406"	"10.1111/(ISSN)1472-	{"importedDate":"Mon	["Giraffa cameloparda	["CIBIO/InBIO, Centro	[{"id":"CIBIO/InBIO, Ce	[{"id":"portugal","un	[{"id":"portugal","un	[{"cartographyCode":	{"doi":"10.1111/ddi.12
WNG-LX "10.1111/ddi.12157"	"10.1111/(ISSN)1472-	{"importedDate":"Mon	["Acinonyx jubatus hei	["Institute of Zoology,	[{"id":"Institute of Zool	[{"id":"uk","usa","un	[{"id":"uk","usa","un	[{"cartographyCode":	{"doi":"10.1111/ddi.12
WNG-BBf "10.1111/gcb.12266"	"10.1111/(ISSN)1365-	{"importedDate":"Mon	["Rangifer tarandus c	["Department of Biosc	[{"id":"Department of E	[{"id":"canada","un	[{"id":"canada","un	[{"cartographyCode":	{"doi":"10.1111/gcb.12
WNG-KHk "10.1111/jzo.12348"	"10.1111/(ISSN)1469-	{"importedDate":"Mon	["Panthera pardus"]	["Leibniz Institute for Z	[{"id":"Leibniz Institute	[{"id":"germany","un	[{"id":"germany","un	[{"cartographyCode":	{"doi":"10.1111/jzo.12
WNG-TPi "10.1111/acv.12240"	"10.1111/(ISSN)1469-	{"importedDate":"Mon	["Ovis vignei","Panthe	["Workgroup on Enda	[{"id":"Workgroup on E	[{"id":"germany","un	[{"id":"germany","un	[{"cartographyCode":	{"doi":"10.1111/acv.12
WNG-SDI "10.1111/btp.12072"	"10.1111/(ISSN)1744-	{"importedDate":"Mon	["Alouatta pigra","Cec	["School of Archaeolo	[{"id":"School of Archæ	[{"id":"australia","un	[{"id":"australia","un	[{"cartographyCode":	{"doi":"10.1111/btp.12
WNG-2LC "10.1111/ddi.12286"	"10.1111/(ISSN)1472-	{"importedDate":"Mon	["Pongo pygmaeus"]	["Living Landscape AI	[{"id":"Living Landscap	[{"id":"","australia	[{"id":"","australia	[{"cartographyCo	{"doi":"10.1111/ddi.12
WNG-FGI "10.1111/mam.12053"	"10.1111/(ISSN)1365-	{"importedDate":"Mon	["Monachus monachu	["MOM/Hellenic Societ	[{"id":"MOM/Hellenic S	[{"id":"greece","un	[{"id":"greece","un	[{"cartographyCode":	{"doi":"10.1111/mam.1
WNG-JKc "10.1111/mec.12498"	"10.1111/(ISSN)1365-	{"importedDate":"Mon	["Lynx pardinus"]	["Department of Integri	[{"id":"Department of I	[{"id":"spain","sweden	[{"id":"spain","sweden	[{"cartographyCode":	{"doi":"10.1111/mec.1
WNG-JHT "10.1111/btp.12269"	"10.1111/(ISSN)1744-	{"importedDate":"Mon	["Tayassu pecari"]	["El Colegio de la Froi	[{"id":"El Colegio de la	[{"id":"mexico","un	[{"id":"mexico","un	[{"about":"http://data	{"doi":"10.1111/btp.12

55 colonnes chargées

7 colonnes enrichies



Lignes par page 25

1-25 of 50

Réponse aux questions

visualisation des données dans Iodex...

The screenshot shows the Iodex interface with a green header bar. The 'DONNÉES' tab is selected. A sidebar on the left contains 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main area displays a table with columns: 'COLONNES', 'FILTRES', 'DENSITÉ', and 'AJOUTER'. The table lists various data columns with their source, path, advanced mode status, and launch status.

Nom	Colonne de la source	Sous-chemin	Mode avancé	Statut	Lancer
Repérage des espèces	Résumé	-	×	Non démarré	▶ LANCER
AffiliationPretraitement	-	-	✓	Non démarré	▶ LANCER
enrichissement_adresse	AffiliationPretraitement	-	×	Non démarré	▶ LANCER
AffiliationsNettoyees	-	-	✓	Non démarré	▶ LANCER
AffiliationsHomog	-	-	✓	Non démarré	▶ LANCER
Code pays	AffiliationsHomog	-	×	Non démarré	▶ LANCER
bibCheck	DOI	-	×	Non démarré	▶ LANCER

Lignes par page : 100 ▼ 1-7 sur 7 < >

Partie administrative

Traitement des données dans Lodex

LODEX ☰ ☰ DONNÉES 🖨️ AFFICHAGE 👁️ DÉPUBLIER

Données
Enrichissements
Précalculs
Ressources cachées

Nom ¹
Repérage des espèces ▶ LANCER ⁵

Statut : Non démarré **Enrichissements** VOIR LES LOGS
traitements document par document
(web service **synchrone**)

Mode avancé
URL du web service
https://irc3-species.services.istex.fr/v1/irc3sp ² 📄

Colonne de la source ³ Résumé ▼ Sous-chemin ▼

SUPPRIMER ANNULER SAUVEGARDER ⁴

Aperçu de la valeur*
Résumé
"During the International Tapir Symposium 16–21 Oct 2011, the conservation of Baird's tapir (Tapir...
"We collected data on habitat use and locomotion of the François' langur (Trachypithecus francoisi...
"Aim: Climate change assessments are largely based on correlative species distribution models..."

LODEX ☰ ☰ DONNÉES 🖨️ AFFICHAGE 👁️ DÉPUBLIER

Données
Enrichissements
Précalculs
Ressources cachées

Nom * ¹
noiseDetect ▶ LANCER ⁵

Statut : Non démarré **Précalculs** VOIR LES LOGS
traitements de l'ensemble des documents. Le résultat obtenu pour chacun des documents dépend des autres (web service **asynchrone**)

URL du web service *
https://text-clustering.services.istex.fr/v1/noise-lodex ² 📄

Colonne(s) source(s) *
Résumé ✕ Colonne ³ source(s) * ▼

SUPPRIMER RETOUR SAUVEGARDER ⁴

Aperçu de la valeur*
Résumé
"During the International Tapir Symposium 16–21 Oct 2011, the conservation of Baird's tapir (Tapir...
"We collected data on habitat use and locomotion of the François' langur (Trachypithecus francoisi...
"Aim: Climate change assessments are largely based on correlative species distribution models..."

Partie administrative

Mise en forme des données dans Lodex : Page d'accueil

LODEX DONNÉES AFFICHAGE DÉPUBLIER

PAGE DONNÉES PUBLIÉES

DATASET - Titre DATASET - Description

+ NOUVEAU CHAMP

[T4CB]
Instance - Titre

GÉNÉRAL AFFICHAGE SÉMANTIQUE

Étiquette

Icones(s) du champ

Nom interne
Instance - Titre

Source de la valeur

VALEUR ARBITRAIRE CHOIX DE LA ROUTINE DONNÉE PRÉCALCULÉE COLONNE(S) EXISTANTE(S) DEPUIS UNE SOUS-RESSOURCE

Valeur arbitraire

```
<h1><a href="http://www.cnrs.fr" target=""></a>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;<span style="font-size: 1.5em, color: #0a3d74"><center> ANF TDM 3-4 octobre 2024 <br> Exploration documentaire et extraction d'informations </span><BR><font color = "#A4BD01"> Atelier <br> Les web services et la datavisualisation Istex</h1></font><br><h2><center>TP à partir des documents du corpus scientifique Istex <br><a href="https://vie-biodiversite.corpus.istex.fr/>Biodiversité : La conservation des mammifères sur la liste rouge de l'UICN (Union Internationale pour la Conservation de la Nature) </a></center></h2></center></center>
```

ISTEX Services
Les technologies et les outils ISTEX pour les projets de recherche.

ANF TDM 3-4 octobre 2024
Exploration documentaire et extraction d'informations
Atelier
Les web services et la datavisualisation Istex

TP à partir des documents du corpus scientifique Istex
Biodiversité : La conservation des mammifères sur la liste rouge de l'UICN (Union Internationale pour la Conservation de la Nature)



Partie administrative

Mise en forme des données dans Lodex : Ressource principale (champs de la notice)

LODEX

☰ DONNÉES

📄 AFFICHAGE

👁️ DÉPUBLIER

🏠 Page d'accueil

📄 Ressource principale

📊 Graphiques

🔍 Recherche et facettes

PAGE

DONNÉES PUBLIÉES

+ NOUVEAU CHAMP

Titre [eBCV]

Identifiant [KJZ]

DOI [HFG]

Source [hNM]

Année de publication [vBM]

Affiliations [xZz]

Pays1 [hX0] addressSplit

Pays2 [Zp7b] pays nettoyés à l'aide d'un enrichissement

Pays3 [qj8n] homogénéisation à l'aide d'une table d'équivalence dans un enrichissement

Code pays [tCBh]

Résumé [anMB]

Mots-clésAuteur [a0o]

ISTEX Services
Les technologies et les outils ISTEX pour les projets de recherche.

Titre
Experimental release of an Iberian lynx (*Lynx pardinus*)

Identifiant
<https://doi.org/10.1007/978-3-319-58423-0>

DOI
10.1007/97800058423

Source
Biodiversity & Conservation

Année de publication
1995

Affiliations

- Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Avda. María Luisa s/n, 41013, Sevilla, Spain
- Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Avda. María Luisa s/n, 41013, Sevilla, Spain
- Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Avda. María Luisa s/n, 41013, Sevilla, Spain

Pays1
• spain

Pays2
• spain

Pays3
• spain

Code pays
• ESP

Partie administrative

Mise en forme des données dans Lodex : Ressource principale (champs de la notice)

LODEX DONNÉES AFFICHAGE DÉPUBLIER

Page d'accueil Ressource principale Graphiques Recherche et facettes

PAGE DONNÉES PUBLIÉES + NOUVEAU CHAMP

Titre [eBCV]

GÉNÉRAL AFFICHAGE SÉMANTIQUE

Étiquette
Titre 1

icône(s) du champ
Nom interne

Source de la valeur 2

VALEUR ARBITRAIRE CHOIX DE LA ROUTINE DONNÉE PRÉCALCULÉE COLONNE(S) EXISTANTE(S) DEPUIS UNE SOUS-RESSOURCE

Colonnes(s) existante(s)
Titre Saisir colonne(s) existante(s)

Opérations de transformation TOUT SUPPRIMER

AJOUTER UNE OPÉRATION

Aperçu de la valeur*

Titre

"Rapid ongoing decline of Baird's tapir in Cusuco National Park,..."

"Habitat use and locomotion of the François' langur (Trachypitecus..."

"Contemporary niche contraction affects climate change predictions ..."

"Fiddling in biodiversity hotspots while deserts burn? Collapse of th..."

"Securing a future for wild Indochinese tigers: Transforming..."

"A preliminary assessment of the effectiveness of the Mesoamerican..."

"Locomotion behavior of cao vit gibbon (Nomascus nasutus) living i..."

"Recolonization of coastal heath by Pseudomys novaehollandiae..."

Partie administrative

Mise en forme des données dans Lodex : Ressource principale (champs de la notice)

LODEX DONNÉES AFFICHAGE DÉPUBLIER

PAGE DONNÉES PUBLIÉES

+ NOUVEAU CHAMP

Titre [eBCV]

LODEX DONNÉES AFFICHAGE SÉMANTIQUE

GÉNÉRAL AFFICHAGE SÉMANTIQUE

Visible

Afficher avec un format

Aucun format appliqué

largeur %

Annoter un autre champ

Champ à annoter

Afficher en tant que champ composé

Champs composant l'affichage

TOUS GRAPHIQUE LODEX TEXTE URL AUTRE

SPARQL - Texte
En choisissant ce format la page à renseigner permet d'écrire des requêtes SPARQL et d'aligner nos données avec par exemple : Wikidata, IDRef, dataBNF, Mirabel...

Texte - Badge pour identifiant
Ce format permet d'afficher un identifiant pérenne (soit un DOI, soit un PMID), et d'y ajouter un lien vers la ressource.

Texte - Balises HTML
Le format HTML permet d'afficher dans l'application un bloc de HTML.

Texte - Chiffre en gras
Ce formatage modifie le texte en gras

Texte - Code
Colore syntaxiquement la valeur du champ en suivant la syntaxe du langage donné en paramètre (parmi XML, JSON, INI, Shell, SQL, Javascript).

Texte - JavaScript/HTML templating (modèle)
Créez un modèle avec EJS (HTML + JavaScript) et ajoutez vos données avec les routines LODEX.

ANNULER

Partie administrative

Mise en forme des données dans Lodox : Graphique

The screenshot displays the LODEX administrative interface. At the top, a green navigation bar contains the 'LODEX' logo, a menu icon, and buttons for 'DONNÉES' and 'AFFICHAGE'. A secondary bar shows 'PAGE' and 'DONNÉES PUBLIÉES' with a 'DÉPUBLIER' button. A left sidebar lists navigation options: 'Page d'accueil', 'Ressource principale', 'Sous-ressources', 'Graphiques' (highlighted with a green box), and 'Recherche et facettes'. The main content area features a '+ NOUVEAU CHAMP' button and a list of data fields, each with a visualization icon and a label:

- Année de publication (plus de 3 publi... [qt94])
- Mots-clésAuteur [pWZZ]
- Espèces (plus de 10 publications) [UCf6]
- Topics [ETqv]
- Topics / Année (plus de 5 publications) [arX1]
- Année de publication (plus de 3 publications)
- Mots-clésAuteur
- Espèces (plus de 10 publications)
- Topics
- Topics / Année (plus de 5 publications)
- Répartition internationale
- Coopérations internationales
- bibCheck

Partie administrative

Mise en forme des données dans Lodox : Graphique

The screenshot shows the Lodox administrative interface. The top navigation bar includes 'LODEX', 'DONNÉES', and 'AFFICHAGE' (highlighted with a blue box). A 'DÉPUBLIER' button is visible on the right. The left sidebar contains navigation options: 'Page d'accueil', 'Ressource principale', 'Sous-ressources', 'Graphiques' (highlighted with a green box), and 'Recherche et facettes'. The main content area displays a configuration page for a field named 'Année de publication (plus de 3 publi... [qt94]'. The 'GÉNÉRAL' tab is selected (highlighted with a green box). The 'Source de la valeur' section shows options: 'VALEUR ARBITRAIRE', 'CHOIX DE LA ROUTINE' (selected), 'DONNÉE PRÉCALCULÉE', 'COLONNE(S) EXISTANTE(S)', and 'DEPUIS UNE SOUS-RESSOURCE'. Below this, a dropdown menu for 'Choisir une routine' is open, showing options like 'distinct-alpha-3-ISO639-from' and 'distinct-by'. The 'distinct-by' option is selected, and its description is shown in a tooltip: 'Compte, pour chaque élément du champ représenté (identifiant), le nombre de fois où cet élément apparaît.' Below the routine selection, there is a 'Champs de la routine' dropdown set to 'Champ N°1'.

Partie administrative

Mise en forme des données dans Lodox : Graphique

The screenshot shows the Lodox administrative interface. At the top, there is a green navigation bar with 'LODEX', 'DONNÉES', and 'AFFICHAGE' (highlighted with a blue box). A 'DÉPUBLIER' button is visible on the right. On the left, a dark sidebar contains navigation options: 'Page d'accueil', 'Ressource principale', 'Sous-ressources', 'Graphiques' (highlighted with a green box), and 'Recherche et facettes'. The main content area is titled 'ANNÉE DE PUBLICATION (plus de 3 publi... [qt94])' and has tabs for 'GÉNÉRAL', 'AFFICHAGE' (highlighted with a green box), and 'SÉMANTIQUE'. Below the 'AFFICHAGE' tab, there is a 'Visible' toggle (checked) and a section for 'Afficher avec un format'. Under this section, 'Aucun format appliqué' is selected. A dropdown menu is open, showing options: 'TOUS', 'GRAPHIQUE' (highlighted with a green box), 'LODEX', 'TEXTE', 'URL', and 'AUTRE'. The 'Graphique - Aster Plot' option is expanded, showing its description: 'Ce format permet de faire un graphique sur le champ multivalué textuel mots-clés, dans la ressource, et non au niveau du jeu de données, pour visualiser les ressources qui ont le plus de similarité entre elles.' Other chart options listed include 'Carte de chaleur', 'Carte proportionnelle', 'Cartographie', 'Cartographie de flux', and 'Coordonnées parallèles'.

Partie administrative

Mise en forme des données dans Lodex : Recherche et facettes

The screenshot shows the administrative interface of the Lodex system. At the top, there is a green navigation bar with the 'LODEX' logo on the left, a 'DONNÉES' menu, and an 'AFFICHAGE' button highlighted with a blue box. On the far right of the bar is a 'DÉPUBLIER' button. A left sidebar contains navigation links: 'Page d'accueil', 'Ressource principale', 'Sous-ressources', and 'Recherche et facettes', with the last one highlighted by a green box. The main content area is divided into several sections: 'Champs de recherche' with a search input field containing 'Identifiant [KJIZ]'; 'Facettes' with a list of search criteria including 'Année de publication [V6bM]', 'bibCheck [nc3r]', 'Code pays [bCBh]', and 'Espèces [xcSp]'; and 'Syndication de recherche' with fields for 'Titre de la ressource', 'Description de la ressource', and three detail fields: 'Titre [eBCV]', 'Source [nbNM]', and 'Année de publication [V6bM]'.

This screenshot displays the search results page. At the top, there is a search bar with the placeholder text 'Vous pouvez saisir votre recherche ici'. Below the search bar, the results are organized into a grid. On the left, there is a vertical sidebar with filter facets: 'Année de publication', 'Pays1', 'Pays2', 'Pays3', 'Code pays', 'Mots-clés/Auteur', and 'Espèces'. The main area shows a grid of search results under the heading '500 ressources trouvées sur un total de 500'. The results are sorted by 'TRIÉRIER ASCENDANT' and there is an 'EXPORTER' button. The results are displayed in a grid of cards, each containing a year, a title, and a source. For example, one result is '1995 Experimental release of ... Biodiversity & Conserva...' from '1992 Limits to predator regul... Oecologia'. At the bottom of the results grid, there is a link that says 'VOIR PLUS DE RÉSULTATS (10 / 500)'.

1 TDM : Quelques rappels

2 TDM à l'Inist et démos

3 **Présentation des TPs**

4 A vous de jouer



C'est à vous...

3 TPs au choix

1. Biodiversité



- Extraction de noms d'espèces
- Découpage d'une adresse + séries de curation
- Détection de bruit dans un corpus
- Extraction de thématiques d'un corpus
- Contrôle de références bibliographiques

2. Risques climatiques



- Détection de genre
- Détection de la langue d'un texte
- Classification ScienceMetrix
- entityTag - Extraction d'entités nommées (Personnes, Localisations, Organismes et autres)
- IdRorDetect

3. Réfugiés



- Détection de langue
- Teeft
- Classification Pascal/Francis
- Classification Hal
- Extraction d'entités nommées

C'est à vous...

3 TPs au choix

- **Une instance par personne ou par groupe de 2 personnes si préférence**
- **La récupération des données (fichier zip)**
 - un corpus Istex tar.gz
 - un modèle tar
- **Le déroulé du TP**
 - Biodiversité :
 - Risques climatiques :
 - Réfugiés :
- **Les solutions (fichier zip)**
 - Les 3 corpus Istex enrichis
 - Les 3 modèles avec les enrichissements

C'est à vous...

3 TPs au choix

- Une instance par personne ou par groupe de 2 personnes si préférence
 - Répartition

1	Robert
2	Cretin Dorine
3	Demir Tugce
4	Desset Sophie
5	Franco Daniele

6	Martin Christelle
7	Otilien Ethson
8	Parkhomenko Ksenia
9	Rabia Aïda
10	Rayane Van-hung

11	Ricaud Robin
12	Sidhom Sahbi
13	Sidorets Anna
14	Trzmielewski Marcin
15	Wibaux Laurine



CONCLUSION

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**



EN PRÉVISION

Nous suivre sur les
réseaux sociaux:

[X \(Ex-twitter\)](#)

[Facebook](#)

[LinkedIn](#)

[YouTube](#)

[Fil d'actualités](#)



De nouveaux **web services**



Des actions de **formation de type atelier**

- **5 nov** Media Normandie
<https://www.crfcb.fr/#/program/6003/12974/>
- **20-21 novembre** Enssib



De nouvelles utilisations de Lodex (exploitation de **données issues de Zotero** – expérimentation en cours)

MERCI !
A VOTRE ÉCOUTE...



valerie.bonvallot@inist.fr

justine.revol@inist.fr



ANNEXES

Annexe 1

Liens utiles

- Nous contacter
 - valerie.bonvallot@inist.fr
 - justine.revol@inist.fr
- Sites utiles
 - <https://www.istex.fr/>
 - <https://www.istex.fr/constitution-de-corpus/>
 - <https://scientific-corpus.data.istex.fr/>
 - <https://services.istex.fr/>
 - <https://ia-factory.services.istex.fr/>
 - <https://data.istex.fr/instance/tm-tools-explorer>
 - <https://www.lodex.fr/>

Annexe 2

Quiz

Qu'est-ce que le TDM ?

- création de textes littéraires
- extraction automatique d'informations à partir de textes ou de données
- organisation de documents
- outil de programmation

Annexe 2

Quiz

Quel terme n'est pas en lien avec le TDM ?

- Entité nommée
- Catalogage
- Classification
- Lemmatisation

Annexe 2

Quiz

Quelle étape du TDM consiste à transformer des phrases en mots individuels ?

- lemmatisation
- tokenisation
- clustering
- agrégation

Annexe 2

Quiz

Répondre aux questions suivantes en s'aidant d'objectif TDM et de l'open API

Sélectionner le web service qui n'existe pas dans ISTEEX TDM

- Détection de la langue d'un texte
- Détection d'entités géographiques
- Classification dans les domaines HAL
- Extraction d'entités nommées de biologie

Annexe 2

Quiz

Combien de web services peuvent être utilisés sur du texte écrit en français ?

- 9
- 1
- 21
- tous

Annexe 2

Quiz

Qu'obtient-on lorsque l'on essaie de détecter la langue de cette expression : "Time flies" ? (en testant le service "détection de langues" via openAPI)

Annexe 2

Quiz

En utilisant le bon web service, donner un mot-clé pertinent de : "Le journal CNRS est un site d'information scientifique."

Annexe 3

TDM Qualité



Quelques notions liées à l'évaluation

Rappel : proportion de documents bien classés par rapport à tous les documents d'une classe.

Recall valeur entre [0-1]

$$\text{silence} = 1 - R$$

True positive / True positive + False negative

un rappel de 1 signifie que chaque élément de la classe C a été étiqueté comme appartenant à la classe C

Précision : proportion de documents correctement classés parmi ceux classés dans une classe.

valeur entre [0-1]

$$\text{bruit} = 1 - P$$

True positive / True positive + False positive

une précision de 1 pour une classe C signifie que chaque élément étiqueté comme appartenant à la classe C appartient bien à la classe C

Fmesure : moyenne harmonique de P et R = $2P \cdot R / (P + R)$

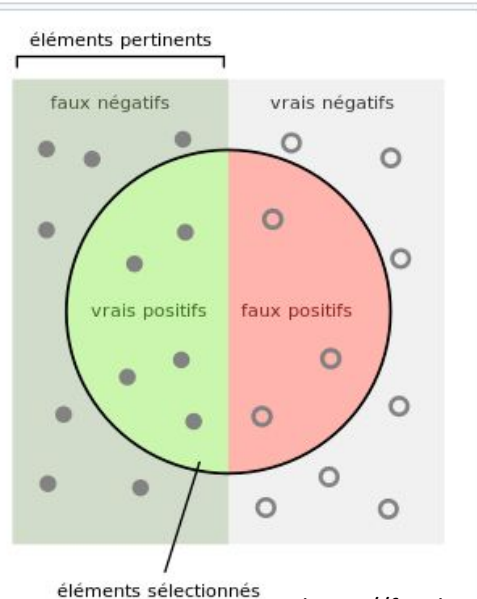
Accuracy : true positive + true negative / nb total

Annexe 3

TDM Qualité



Quelques notions dans l'évaluation



Combien de candidats sélectionnés sont pertinents ?

Précision =



Combien d'éléments pertinents sont sélectionnés ?

Rappel =



Précision et rappel (« recall »). La précision compte la proportion d'items pertinents parmi les items sélectionnés alors que le rappel compte la proportion d'items pertinents sélectionnés parmi tous les items pertinents sélectionnables.

https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel

Confusion matrix		Reality	
		Negative : 0	Positive : 1
Prediction	Negative : 0	True Negative : TN	False Negative : FN
	Positive : 1	False Positive : FP	True Positive : TP

Figure 1. Matrice de confusion

<https://kobias.fr/classification-metrics-matrice-de-confusion/>

Annexe 4



Quelques techniques pour traiter les données : Vectorisation

Embedding
réseaux de neurones

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]