Migale et son offre de service de text-mining

Valentin Loux, Mouhamadou Ba

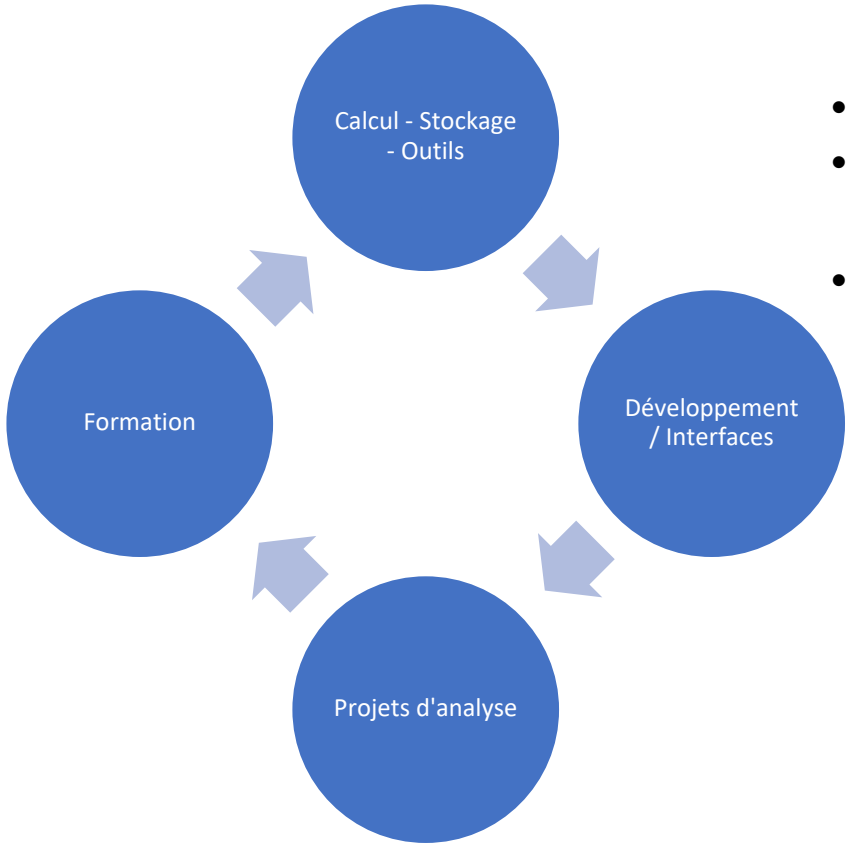
Journées du PEPI IBIS - 14 & 15 septembre 2023

Missions de la plateforme Migale

- Mettre à disposition une infrastructure de calcul scientifique pour la génomique
  - Calcul / Stockage / Outils / Données
  - Diffuser un savoir-faire en bioinformatique et biostatistique
  - Formations / Assistance / Conseil
- Concevoir et développer des applications
  - Interfaces innovantes / Développement et mise à disposition d'outils
- Analyser des données génomiques
  - Métagénomique / Métatranscriptomique / Service / Accompagnement

Des services qui s’enrichissent

De l’environnement de calcul au service d’analyse

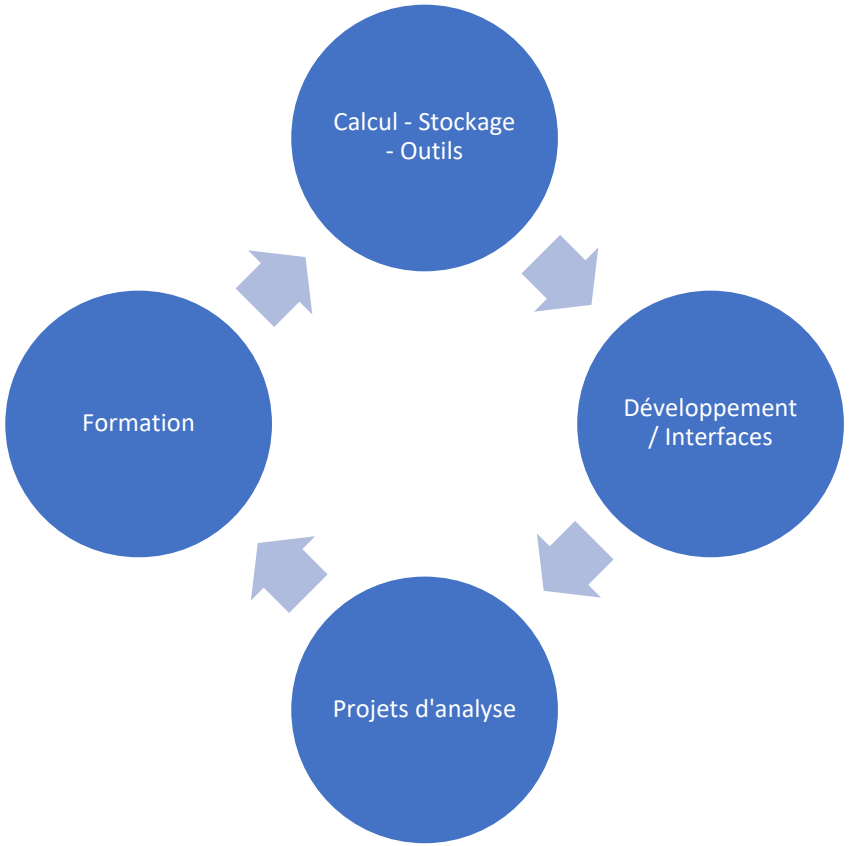


- 2016 : mise en place du service d’analyse de données
- Focus thématique: génomique et métagénomique microbienne
  - Ecosystèmes variés : environnement, aliment…
- Accompagnement global des utilisateurs :
  - Outils
  - Expertise / Analyse
  - Formation

  Accent mis sur l’**autonomisation** par la **transparence et la reproductibilité**
    - Outils conviviaux
    - Tutoriels
    - Rapports réutilisables et détaillés (comment / pourquoi)

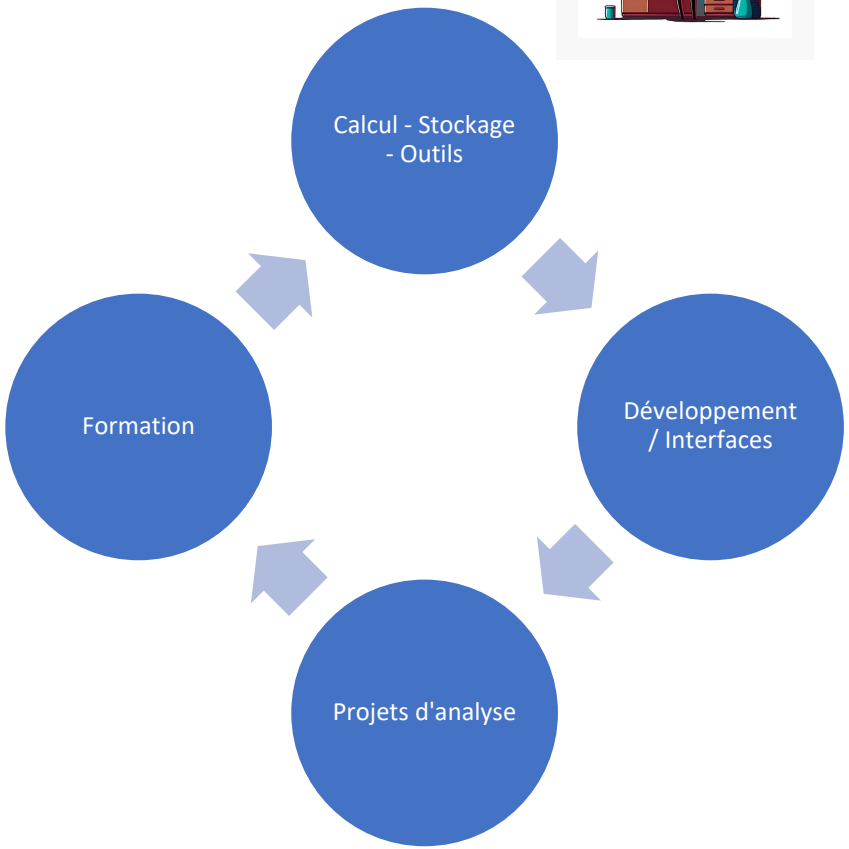
Des résultats et projets complémentaires

De l’environnement de calcul au service d’analyse



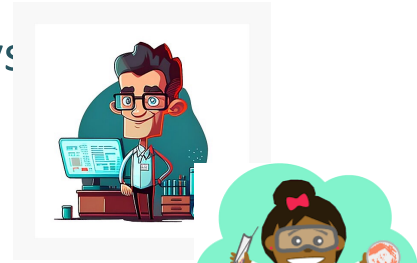
Des résultats et projets complémentaires

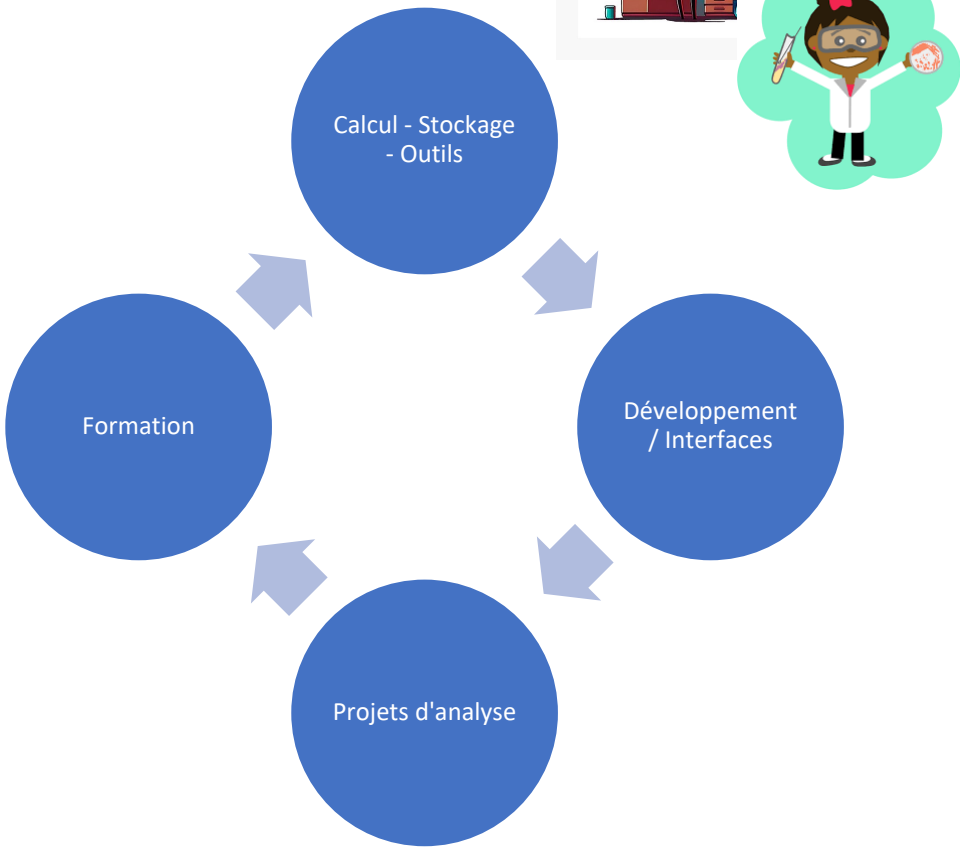
De l’environnement de calcul au service d’analys



Des résultats et projets complémentaires

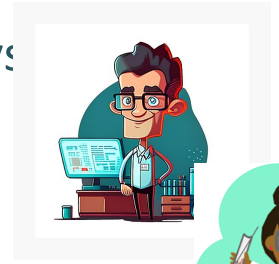
De l’environnement de calcul au service d’analys

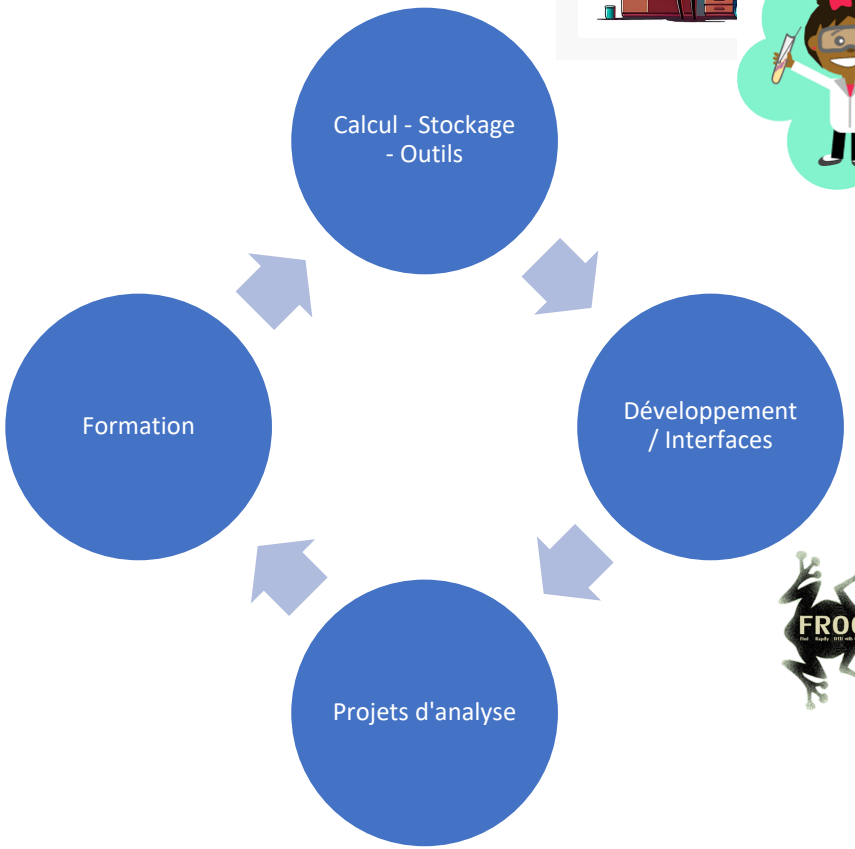


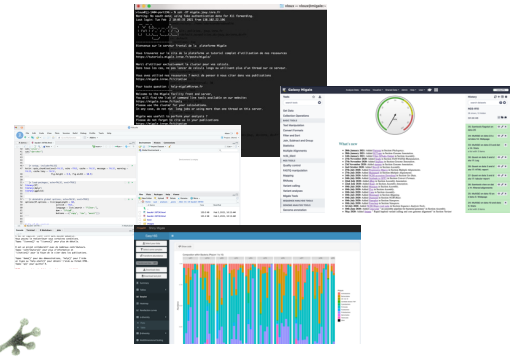


Des résultats et projets complémentaires

De l’environnement de calcul au service d’analys



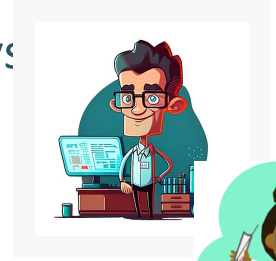


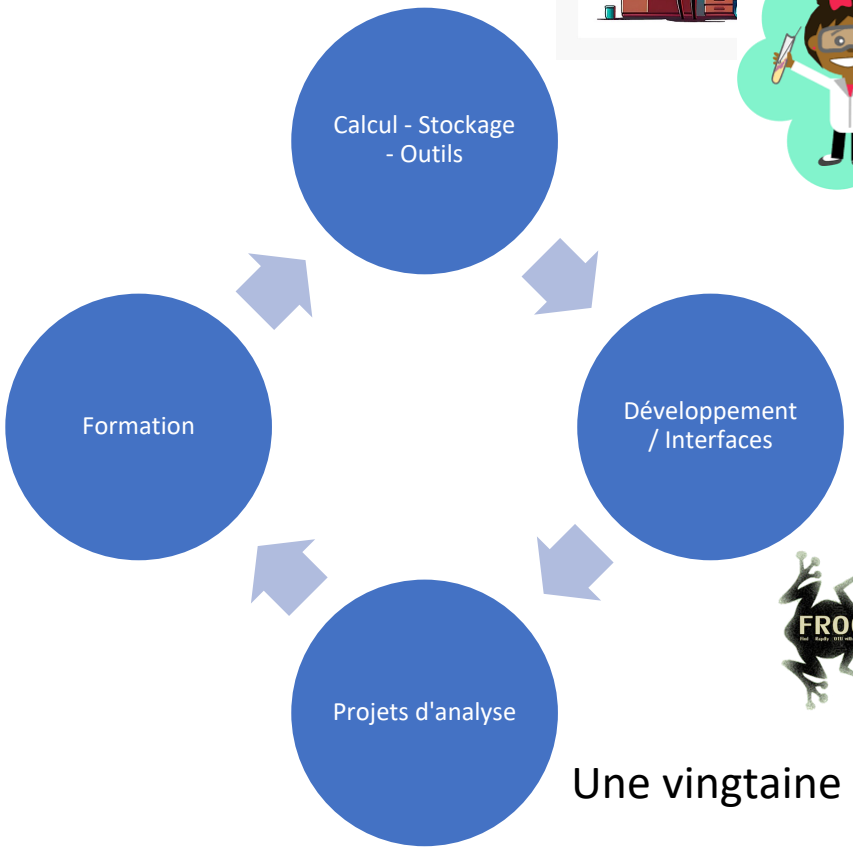


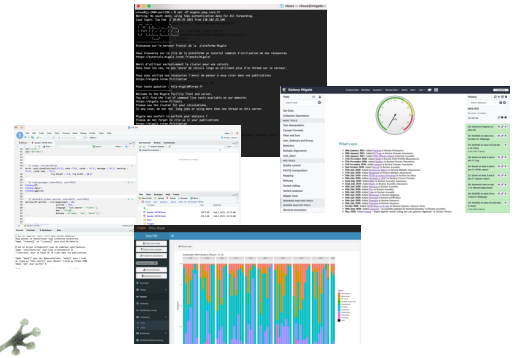
Easy16S

Des résultats et projets complémentaires

De l’environnement de calcul au service d’analys







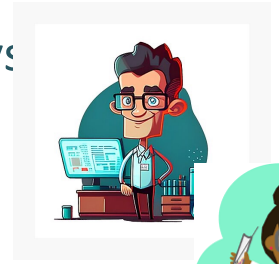
Easy16S

Une vingtaine de projets annuels, dont :

MetaPDO Cheese (Irlinger *et al*, *in prep*)
MetaBar Food (Rué *et al*, 2023)

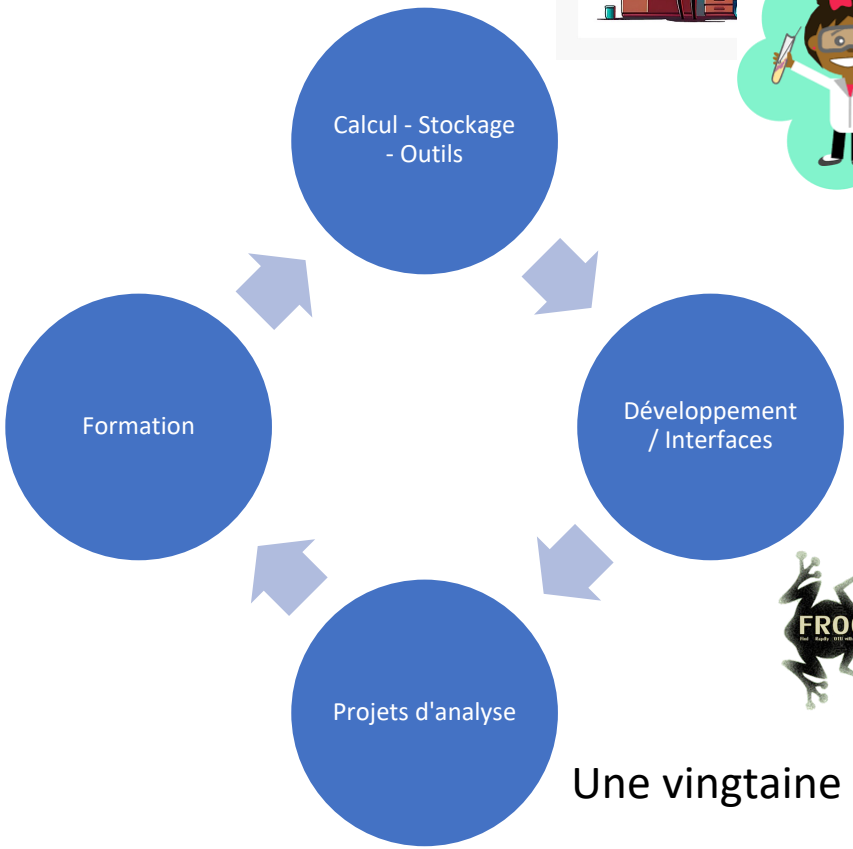
Des résultats et projets complémentaires

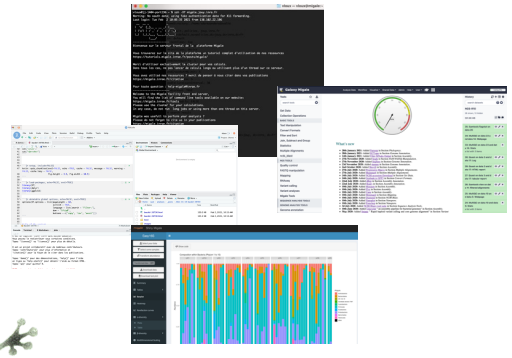
De l'environnement de calcul au service d'analys



15 modules de formations annuels dont:

- Annotation automatique de génomes bactériens
- Comparaison de génomes microbiens
- Analyse de données de metabarcoding
- Analyse de données métagénomiques « shotgun »





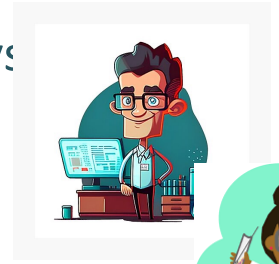
Easy16S

Une vingtaine de projets annuels, dont :

MetaPDO Cheese (Irlinger *et al*, *in prep*)
MetaBar Food (Rué *et al*, 2023)

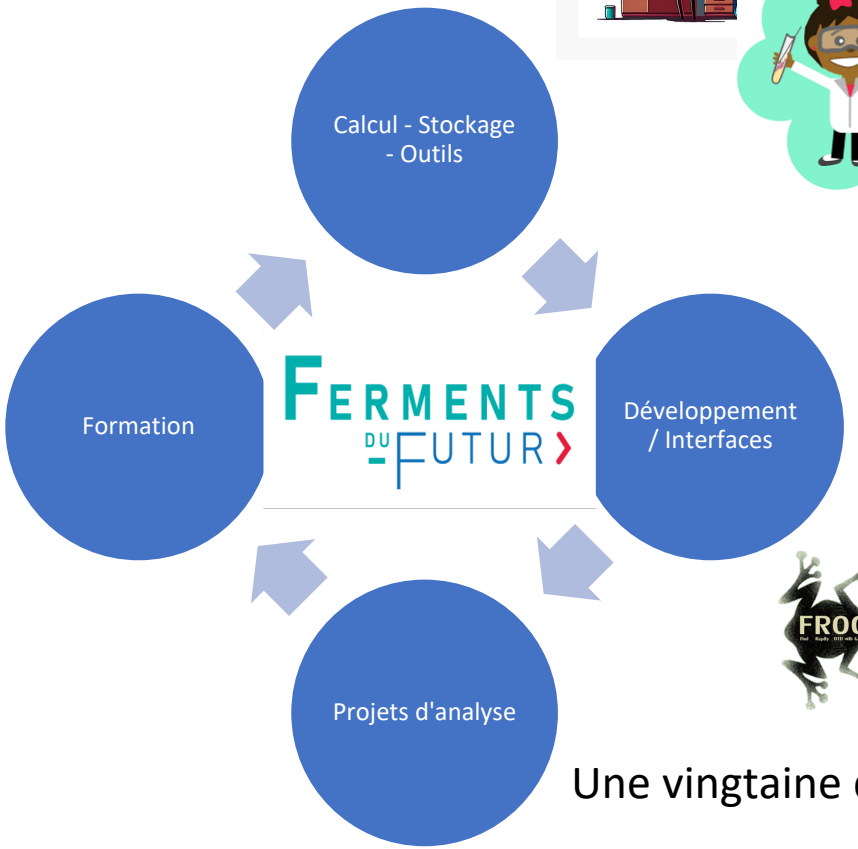
Des résultats et projets complémentaires

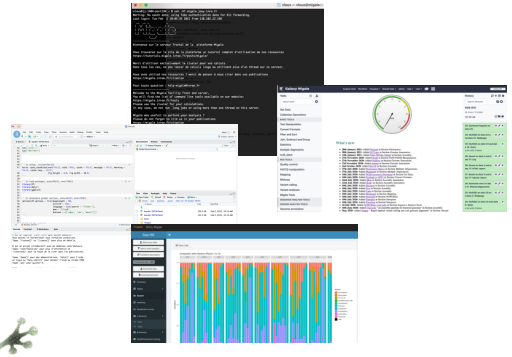
De l'environnement de calcul au service d'analys



15 modules de formations annuels dont:

- Annotation automatique de génomes bactériens
- Comparaison de génomes microbiens
- Analyse de données de metabarcoding
- Analyse de données métagénomiques « shotgun »





Easy16S

Une vingtaine de projets annuels, dont :

MetaPDO Cheese (Irlinger *et al*, *in prep*)
MetaBar Food (Rué *et al*, 2023)

Grand Défi Ferments du Futur

Coordination des développements informatiques de la base de données Ferments du Futur :

- Centralisation des données sur les micro-organismes pouvant intervenir dans la fermentation des aliments
- Base de données = prérequis pour le développement de modèles prédictifs (appels d'offre pré-compétitif annuels)
- Données omiques, phénotypiques, capacités métaboliques, etc. sur les souches et sur les consortias criblés par la plateforme centrale
- Interfaces d'accès pour différents types d'utilisateurs (bioinformaticiens/ informaticiens, biologistes académiques, privés…)
- **4 ingénieurs** en CDI de mission dès 2023 :
  - 2 développeurs (informatiques)
  - 1 bioinformaticien
  - 1 data steward

Opérateur des équipements informatiques (calcul et stockage) nécessaires pour l'entrepôt

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.

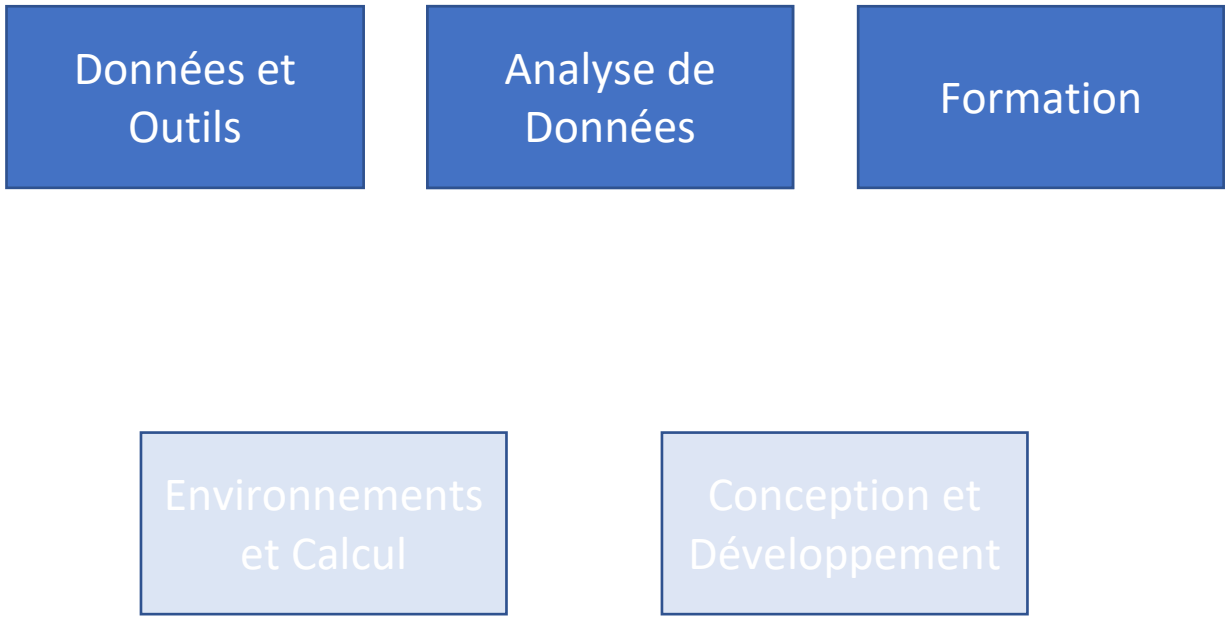
C'est une spécialisation de la fouille de données (*data mining*)
qui fait appel aux méthodes de **l'Intelligence Artificielle**,
du Traitement Automatique des Langues, de la Représentation des Connaissances et des Statistiques.

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

Offre de Service de text-mining de Migale

- Offre de Service : proposition de ressources (données, outils, applications, calcul, formation, assistance, collaboration) à destination des utilisateurs

- Thématique text-mining : ressources (données, outils, applications, formation, assistance, collaboration)
  - Exploitation des données textuelles (articles scientifiques, BDs, bulletins de santé, etc.)
  - Intégration avec les analyses en bioinformatique
  - Pour les chercheurs et ingénieurs en bioinformatique, pas forcément familiers avec le text-mining
  - Collaboration avec l'équipe de recherche Bibliome, specialisée en text-mining



Offres de service similaires

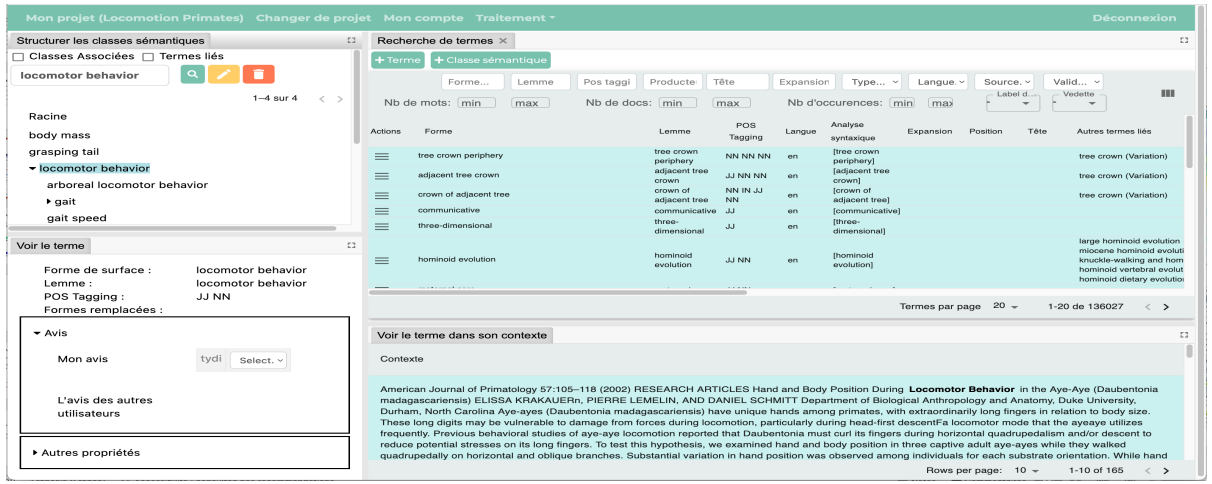
	Outils	**Analyse de données**	**Formation**	**Domaines d'application**
Galaxy Galaxyproject.org/	Outils de traitement		tutoriels	biomedical
ISTEX (CNRS-INIST) https://www.istex.fr/	Outils pour la construction de corpus	internes	ISTEX tour	domaines scientifiques
Clarin (Europe, Pays-bas) https://www.clarin.eu	Données linguistiques numériques, Outils, services	pour utilisateurs autonomes		?
LAPPS Grid (VASSAR COLLEGE, BRANDEIS Univ., CARNEGIE-MELLON Univ., Univ. of PENNSYLVANIA) https://www.lappsgrid.org/	Outils de traitement	pour utilisateurs autonomes		général
Alveo (Univ. of WESTERN SYDNEY) https://www.alveo.edu.au/	Collection de documents, outils de traitement	pour utilisateurs autonomes	tutoriels	général
DBCLS (Japon) https://dbcls.rois.ac.jp	Outils d'assitance, corpus	pour utilisateurs autonomes		Science de la vie
DKPro Core (TU DARMSTADT) https://dkpro.github.io	Composants de traitement			général

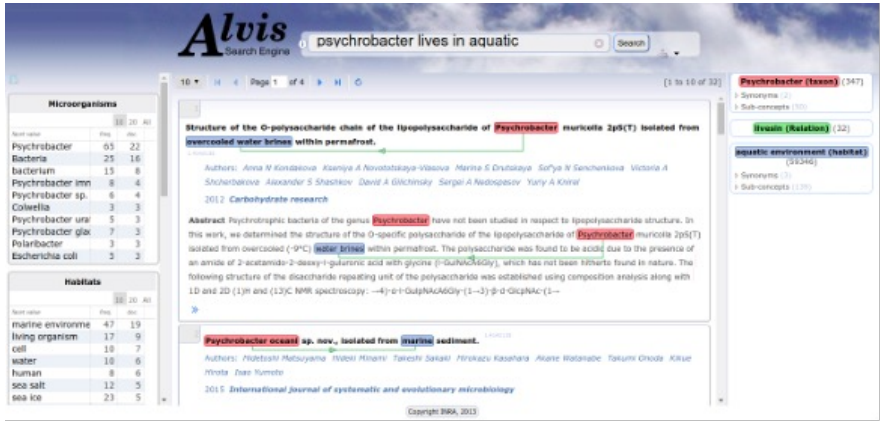
Données et Outils

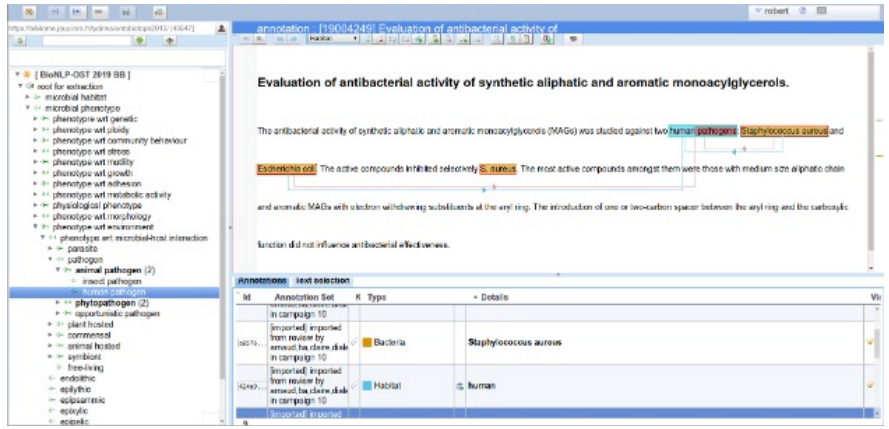
Applications

- Déploiement & Gestion sur demande d'applications d'assistance pour le text-mining pour les collaborateurs
- Plusieurs applications déployées et gérées par Migale dans le cadre des activités en collaboration avec l'équipe Bibliome
  - TYDI et TYDI+ : construction de terminologies
  - ALVISIR : moteur de recherche générique
  - ALVISAE : éditeur d'annotations

- Collaboration autour de la base Omnicrobe hébergées dans Migale, développées et gérées par Sandra dans le cadre du projet Florilege
  - informations sur les habitats, les phénotypes et les usages des micro-organismes extraits automatiquement de sources textuelles (PubMed, GenBank, DSMZ, CIRM-BIA, CIRM-CFBP, CIRM-Levures)
  - https://omnicrobe.migale.inrae.fr



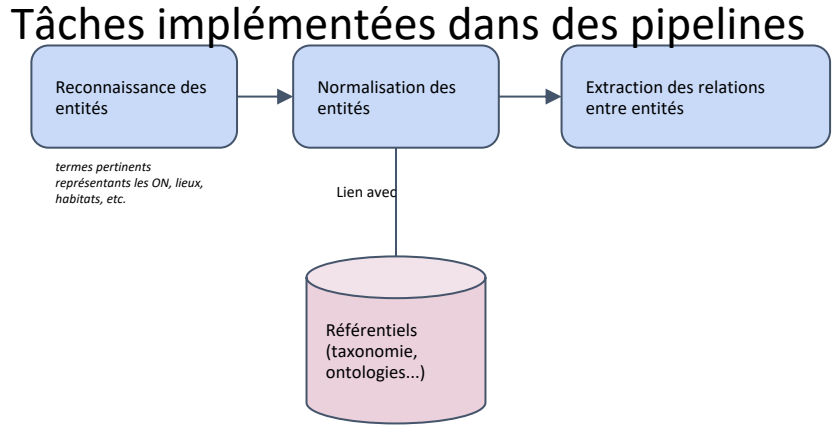




Données et Outils

Services web

- Mise à disposition de services web encapsulant des traitements de text-mining
- Services web basés sur des pipelines d'extraction d'informations spécialisés sur des cas d'usage et développés dans le cadre des projets
- Services web destinées aux partenaires qui souhaitent intégrer des résultats de text-mining dans leurs SI



RestX

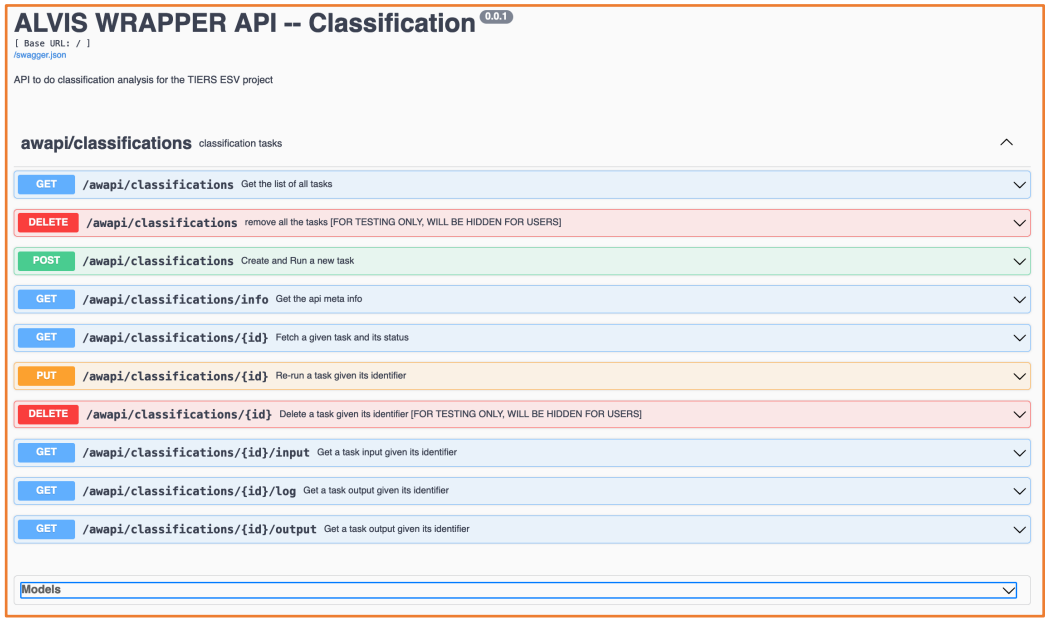
snakemake

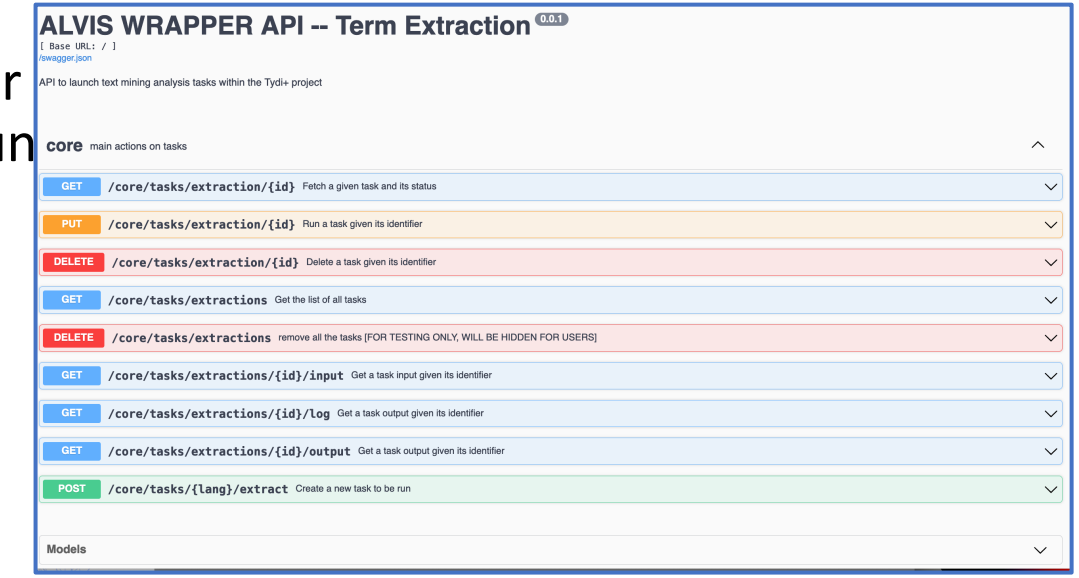
AlvisNLP/ML
corpus processing engine

Données et Outils

Services web

- Services web développés et déployés sur Migale (traitements lancés sur le cluster)
  - Classification de documents pour l'assistance à la veille sanitaire végétale / projet TIERS-ESV
  - Extraction de termes à partir de documents textuels via un application web / projet TYDI++



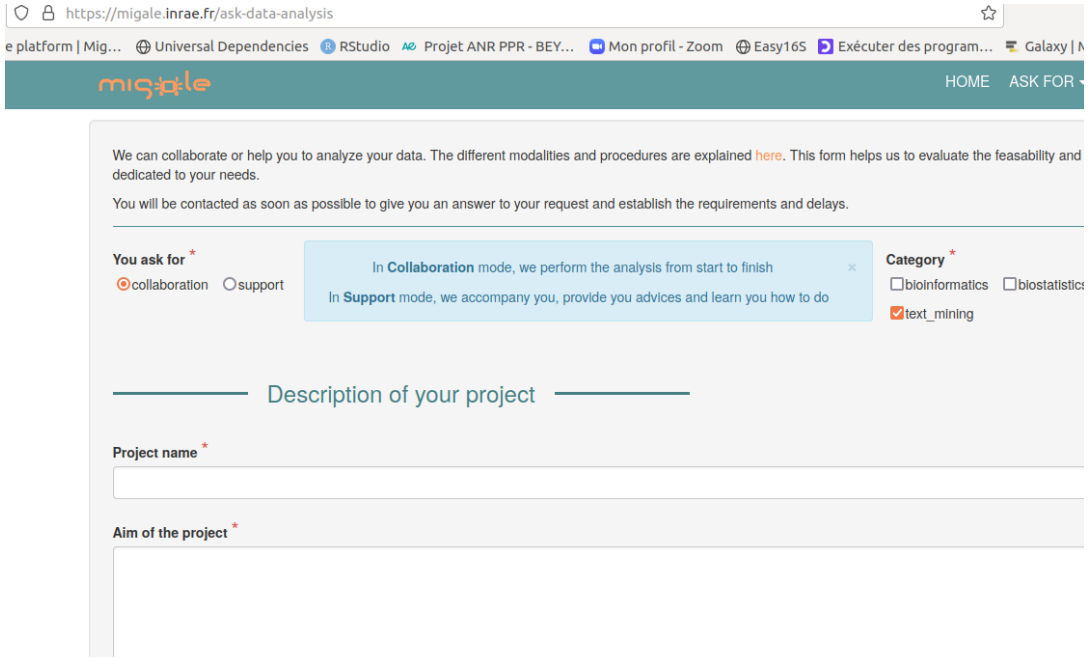


Analyse de données

Données textuelles

Collaboration sur l'analyse des données textuelles en utilisant les solutions offertes par le text-mining

- Mini Projet
  - Demandeurs apportent les besoins et les données
  - Migale s'occupe de les analyser
- Public cible
  - Bio-informaticiens
  - Biologistes, …
- Projets cadrés
  - Calendrier, livrables, suivi, rapports d'analyse, enquêtes de satisfaction



Constitution de corpus thématiques

Extraction d'entités nommées (REN)

Classification de textes

Analyse de données

Données textuelles

Tâches	Constitution de corpus	Extraction d'entités nommées	Classification de textes
Thématiques	Biologie	Microbiologie	Ouvert
Méthodes	Ouvert	Lexiques (*) Patrons apprentissage	apprentissage
Entrées	Textes accessibles (sources en biologie), requêtes	Textes bruts, dictionnaires, macros, textes annotés	Textes bruts, textes annotés
Sorties	Corpus thématiques en biologie	entités biologiques mentionnées dans les textes	elements de textes classés, classifiers
Exemples	Corpus "pathway" (PubMed)	Extraction "gènes" et "metabolites"	Classification phrases avec co-occurrences de gènes/métabolites

AlvisNLP/ML
corpus processing engine

fastText
Library for efficient text classification and representation learning

snakemake

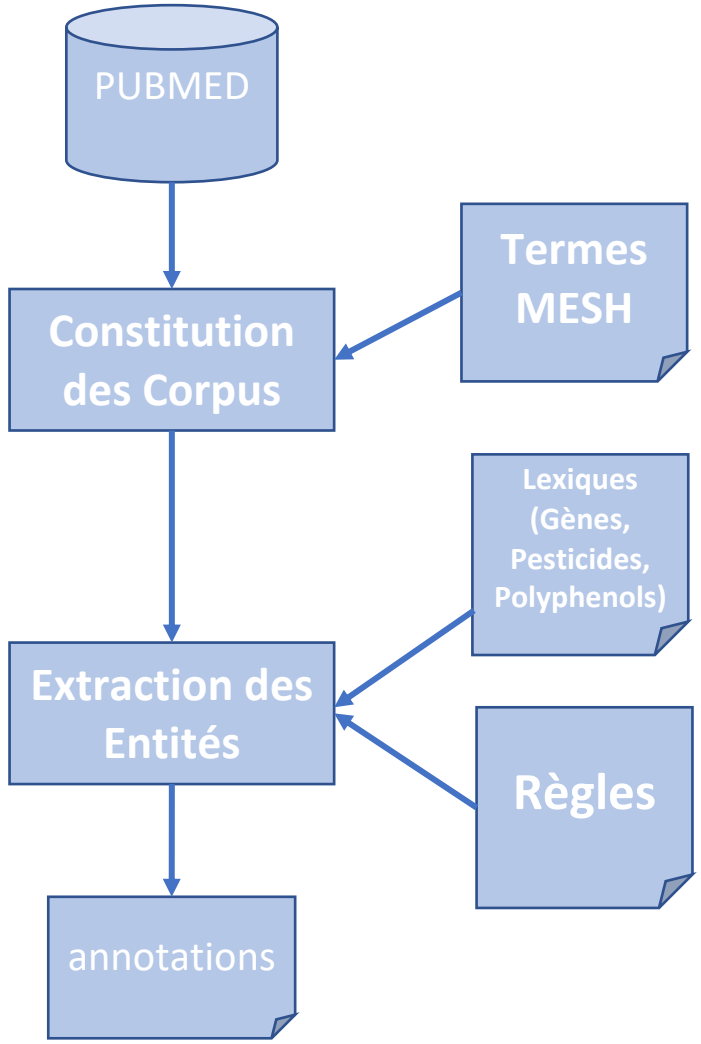
Alvis
Search Engine

spaCy

Analyse de données

Extraction d'informations sur la fertilité et le métabolisme animal

- Entités à extraire
  - Gènes, pesticides, polyphénols, espèces, tissues
- Demandeurs :
  - Biologistes (Laboratoire de Physiologie de la Reproduction & des Comportements, UMR 7247 INRAE/CNRS/Université de Tours/IFCE)
- Utilisation prévue
  - Analyse de tendances
- Projets
  - Nov. 2021 -- Avril 2022
  - Fév. 2023 -- Avril 2023



Analyse de données

Extraction d'informations sur la fertilité et le métabolisme animal

- Résultats

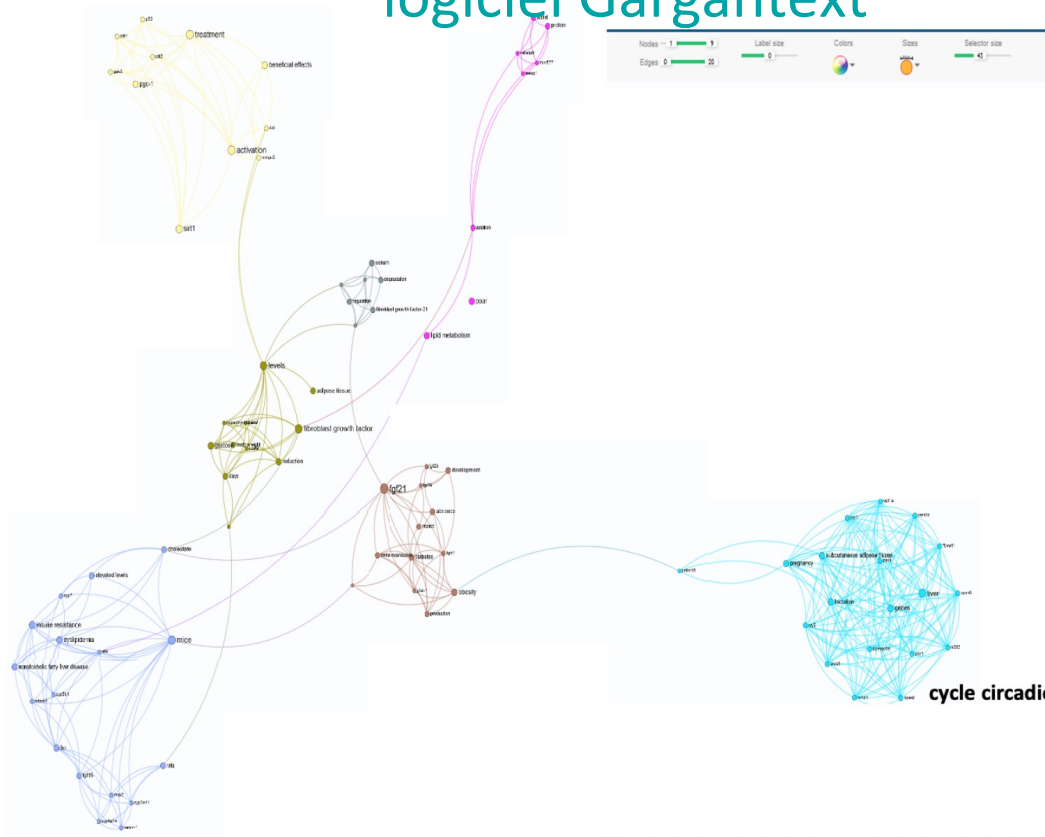
theme	nb.abtracts
ovary	2696858
testis	2807832
metabolism	5486954
ovary AND testis AND metabolism	1061831
ovary OR testis AND metabolism	1319905

PMID	GENE	SCORE	SECTION	TITLE	URL
6169057	AFP	25	title+abstract	Heterogeneity of alpha-fetoprotein(AFP) and albumin containing cells in normal and pathological permissive states for AFP production: AFP containing cells induced in adult rats recapitulate the appearance of AFP ...aining hepatocytes in fetal rats.	https://pubmed.ncbi.nlm.nih.gov/6169057
				...tics of engraftment of CD34(-) and CD34(+) ... from mobilized blood differs from that of ...4(-) and CD34(+) cells from bone marrow.	https://pubmed.ncb/11008020
				...ession analysis of hypothalamic and ...tary components of the growth hormone ...in fasted and streptozotocin-treated ...opeptide Y (NPY)-intact (NPY+/+) and NPY-...kout (NPY-/-) mice.	https://pubmed.ncb/16244497

PMID	PESTICIDE	SCORE	SECTION	TITLE
30502743	endosulfan	20	title+abstract	Different effects of α-endosulfa... β-endosulfan, and endosulfan ... sex hormone levels, metabolic... oxidative stress in adult mice t...
7806131	casein	20	title+abstract	Temporal effects of prolactin o... kinase activity, casein synthesi... casein mRNA accumulation in ... mammary gland explants.
19539752	ethanol	20	title+abstract	Simultaneous prenatal ethanol ... nicotine exposure affect ethan... consumption, ethanol preferen... oxytocin receptor binding in a... and adult rats.
15805107	sucrose	20	title+abstract	Phloem-localized, proton-coup... carrier ZmSUT1 mediates sucr... under the control of the sucros...

...HENOL	SCORE	SECTION	TITLE
...atrol	12	title+abstract	Resveratrol, 4' Acetoxy R... R-equol, Racemic Equol ... Cosmeceuticals to Impro... Health.
25625345 Luteolin	11	title+abstract	Isolation of Luteolin and ... glucoside from Dendrant... morifolium Ramat Tzvel a... Pharmacokinetics in Rats
27334554 Resveratrol	10	title+abstract	Differential responses of ... Resveratrol on proliferati... progenitor cells and agec... hippocampal neurogenes...
28822243 Chrysin	10	abstract+title	Chrysin-induced sperm p... and fatty acid profile cha... reproductive performanc...
30168163 Naringenin	10	abstract+title	Testicular microanatomic... hormonal alterations following use of...

- Exploitation sous le logiciel Gargantext



Formation

"Initiation en Text-mining avec AlvisNLP"

- Objectifs
  - introduire des notions de base en text-mining et Extraction d'Information
  - Initier sur des compétences opérationnelles en Extraction d'Information
- Formation Proposée dans le cycle Migale "Bioinformatique par la pratique"
  - Cycle annuel
- Animateurs
  - Mouhamadou Ba, Robert Bossy
- Format
  - Cours introductif
  - TP, manipulation de workflows de text-mining avec AlvisNLP
- Public
  - (bio)informaticiens

Formation

"Initiation en Text-mining avec AlvisNLP"

- Notions de Traitement Automatique de la Langue (TAL/NLP)
  - Tokenization
  - POS-tagging
  - Lemmatization
  - Reconnaissance d'entités Nommées (REN)
- Notions et manipulation de techniques de REN
  - Projection de lexiques
  - Création de patrons d'extraction
  - Apprentissage supervise
- Création et perfectionnement de workflows pour la REN
- Évaluation de la performance de la REN

Formation

"Initiation en Text-mining avec AlvisNLP"

- Formation dans cadre du projet D2KAB (2020)
  - Animateurs : R. Bossy, M. Ba
  - 11 participants (informaticiens, bio-informaticiens)
  - 1 journée
- Séminaire dans le cadre du hackathon inter-CATI (2021)
  - Animateurs : R. Bossy, A. Chepaikina, M. Ba
  - 9 participants (bio-informaticiens): BOOM, BARIC, GREP, BIOS4Biol
  - 3 demi-journées
- Formation dans le cadre du cycle de la formation de la plateforme Migale (2022)
  - Animateurs : R. Bossy, M. Ba
  - 2 participants
  - 3 demi-journées

Sophie SCHBATH
Responsable scientifique

Valentin LOUX
Responsable opérationnel

Mouhamadou BA
Service Text Mining

Hélène CHIAPELLO
Formations

Christelle HENNEQUET-ANTIER
Analyses Statistiques

Mahendra MARIADASSOU
Biostats

Véronique MARTIN
Formation, Outils et banques de données

Cédric MIDOUX
Analyse de données

Olivier RUÉ
Analyse de données

Valérie VIDAL
Bases de données, support utilisateurs

https://migale.inrae.fr

Collaborateurs dans MaIAGE

- *Bibliome (Robert BOSSY, Louise DELEGER, Claire NEDELLEC)*
- *StatInfOmics (Sandra DEROZIER)*

Citation

*"We are grateful to the **INRAE MIGALE bioinformatics facility** (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help and/or computing and/or storage resources"*

https://migale.inra.fr/citation