

INRAE

➤ Migale et son offre de service de text-mining

Valentin Loux, Mouhamadou Ba

Journées du PEPI IBIS - 14 & 15 septembre 2023

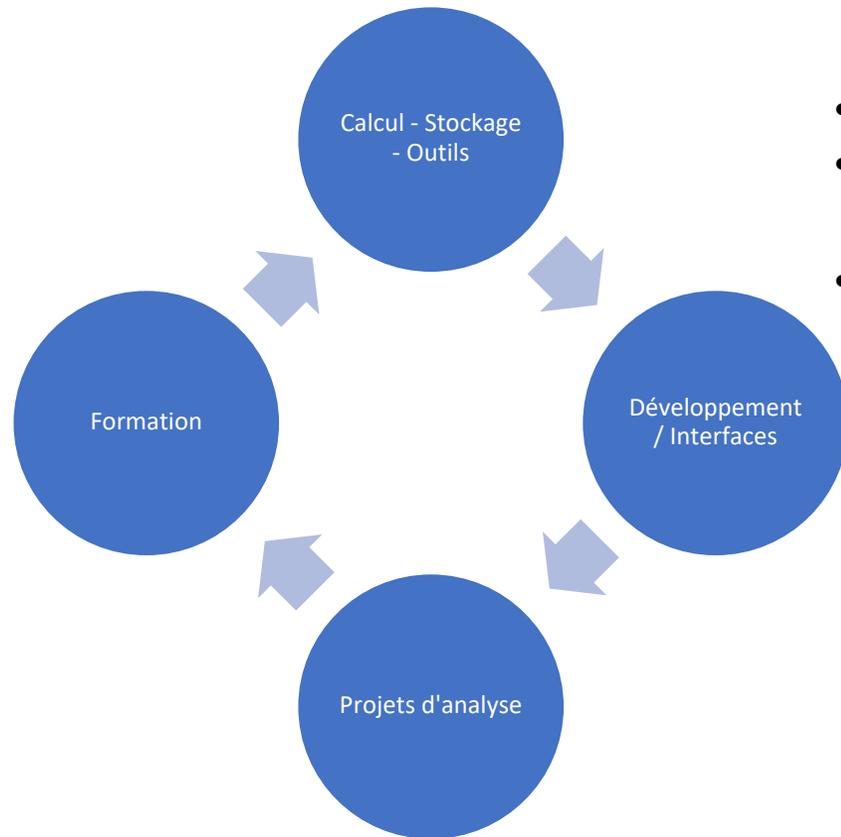
Missions de la plateforme Migale

- 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



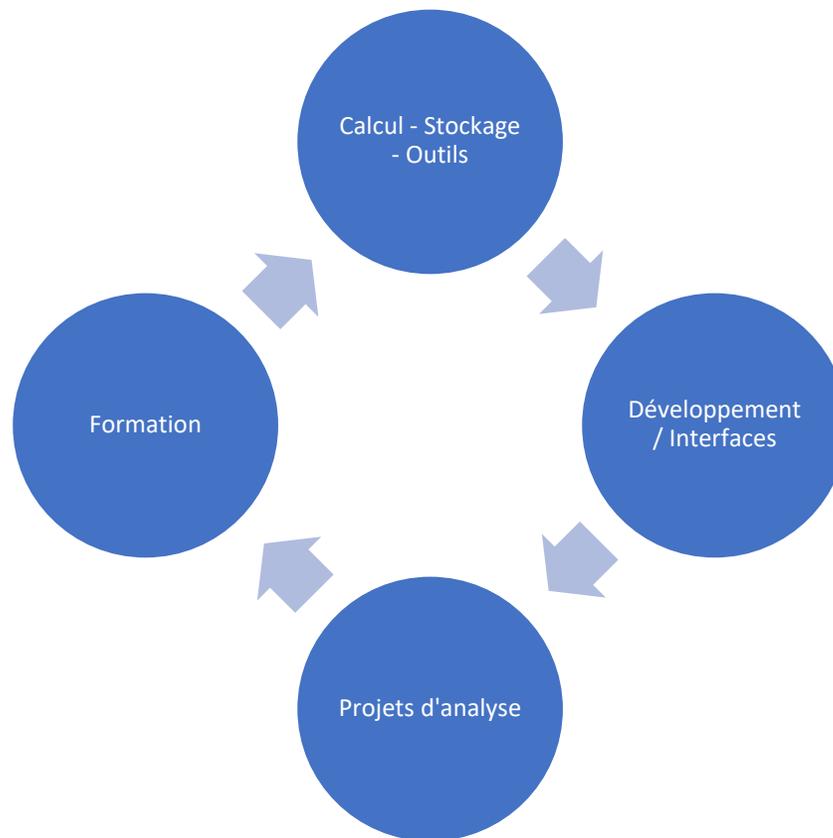
- 2016 : mise en place du service d'analyse de données
- Focus thématique: génomique et métagénomique microbienne
 - Ecosystèmes variés : environnement, aliment...
- Accompagnement global des utilisateurs :
 - Outils
 - Expertise / Analyse
 - Formation

Accent mis sur l'**autonomisation** par la **transparence et la reproductibilité**

- Outils conviviaux
- Tutoriels
- Rapports réutilisables et détaillés (comment / pourquoi)

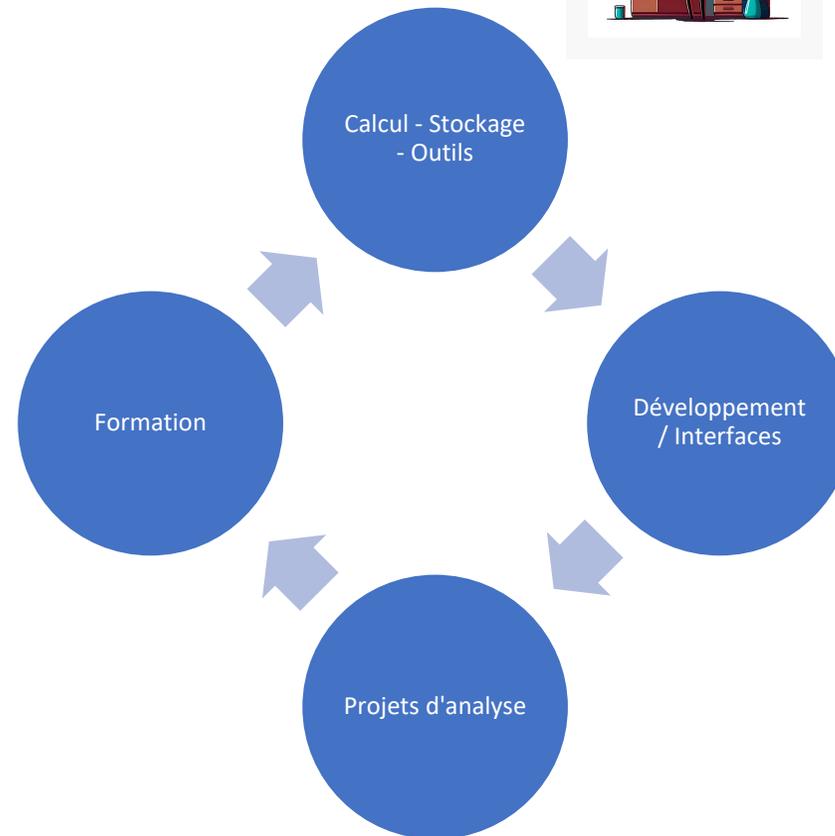
Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse



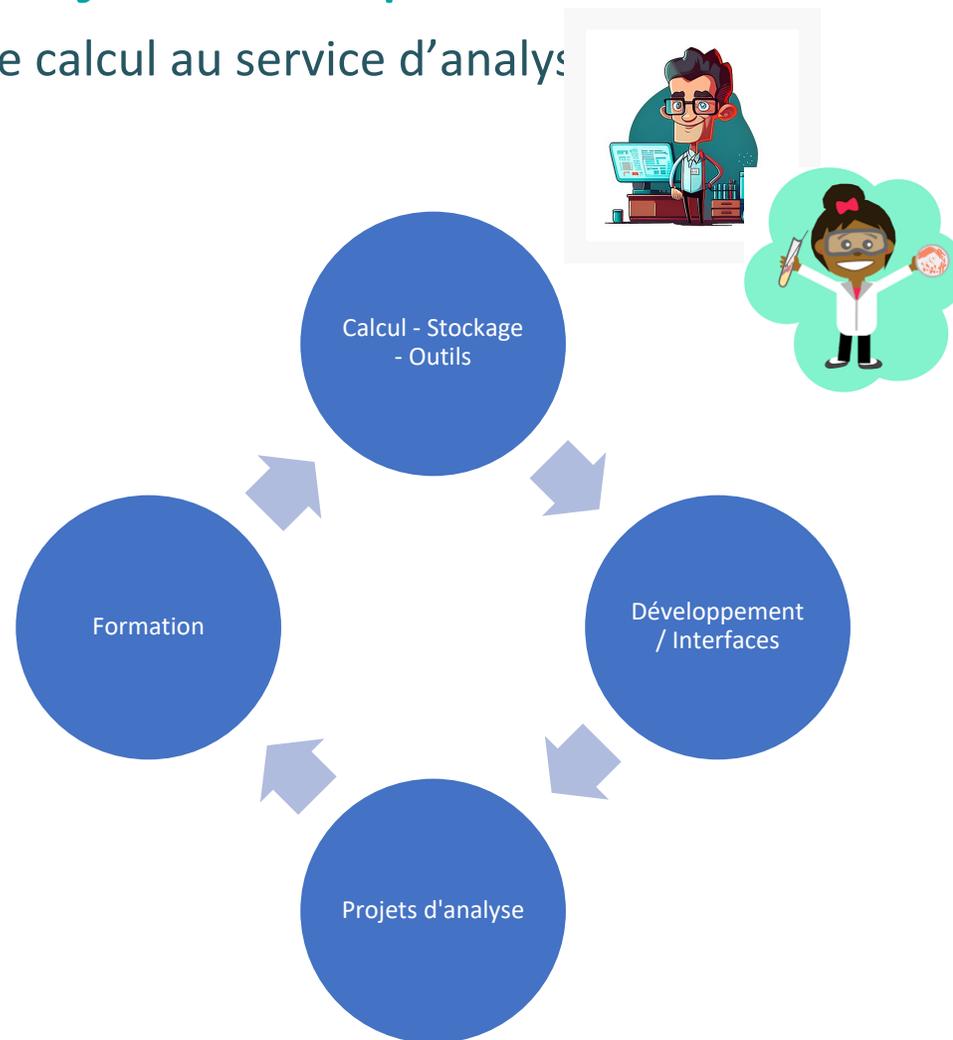
Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse



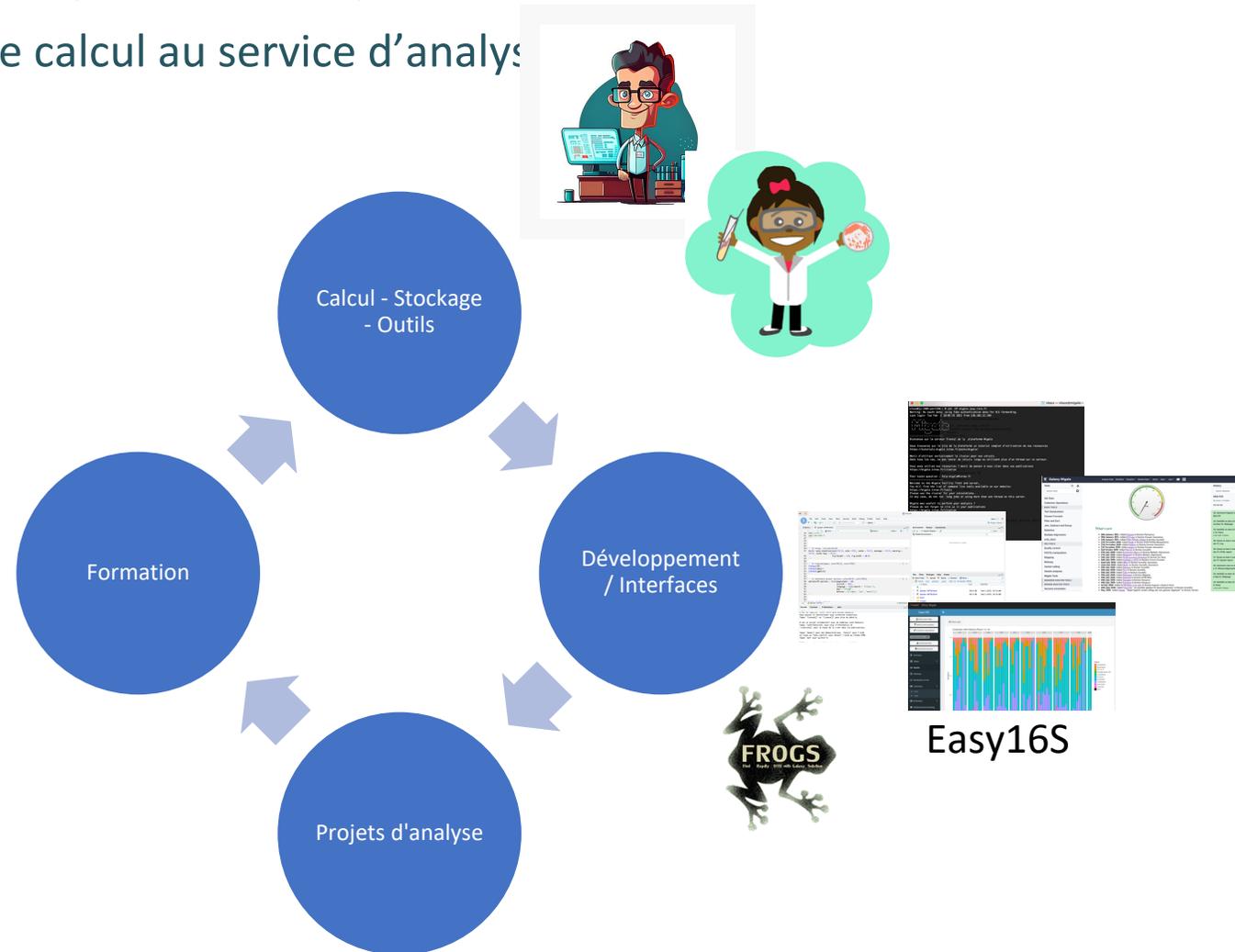
Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse



Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse



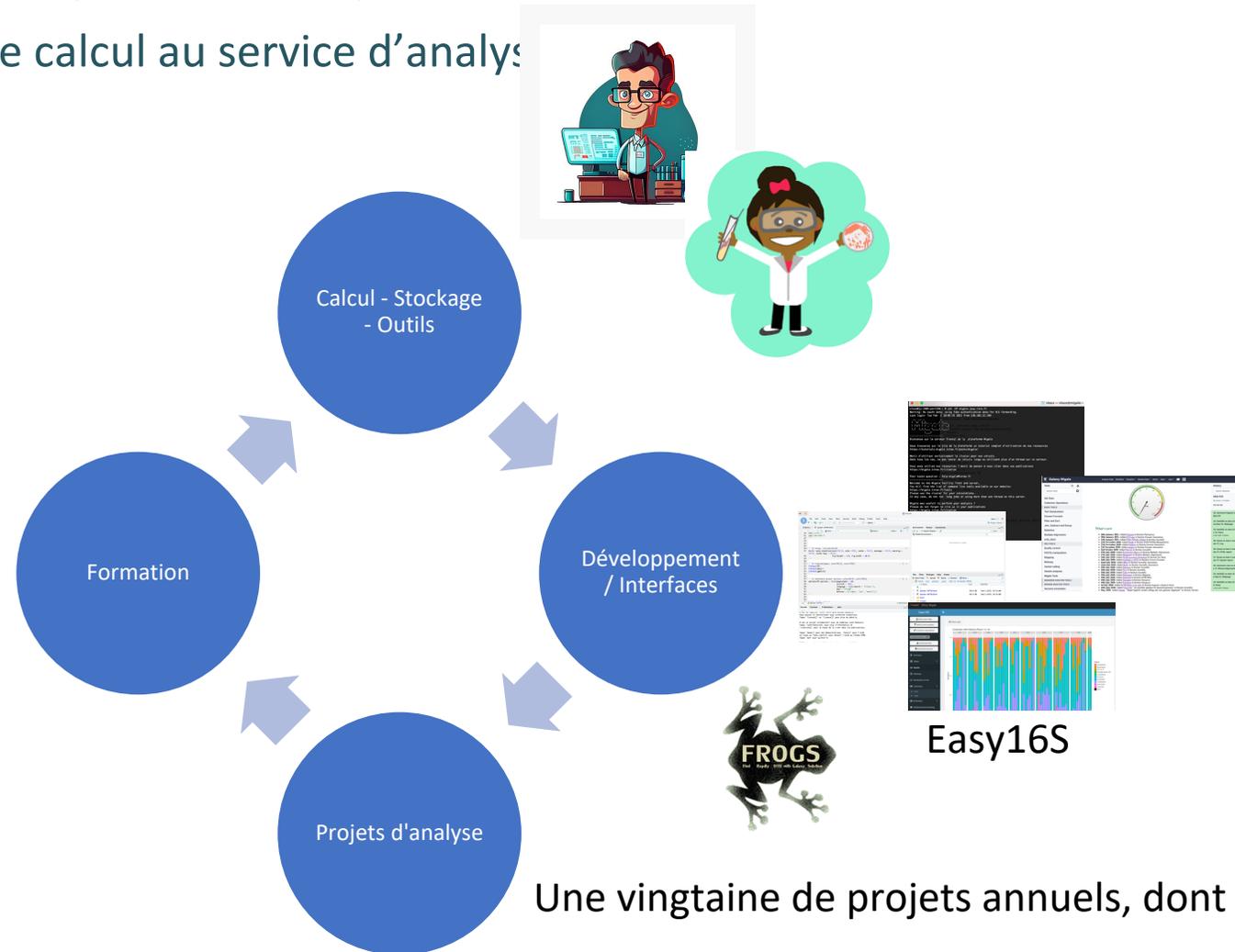
INRAE

Migale et son offre de service text mining

14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse



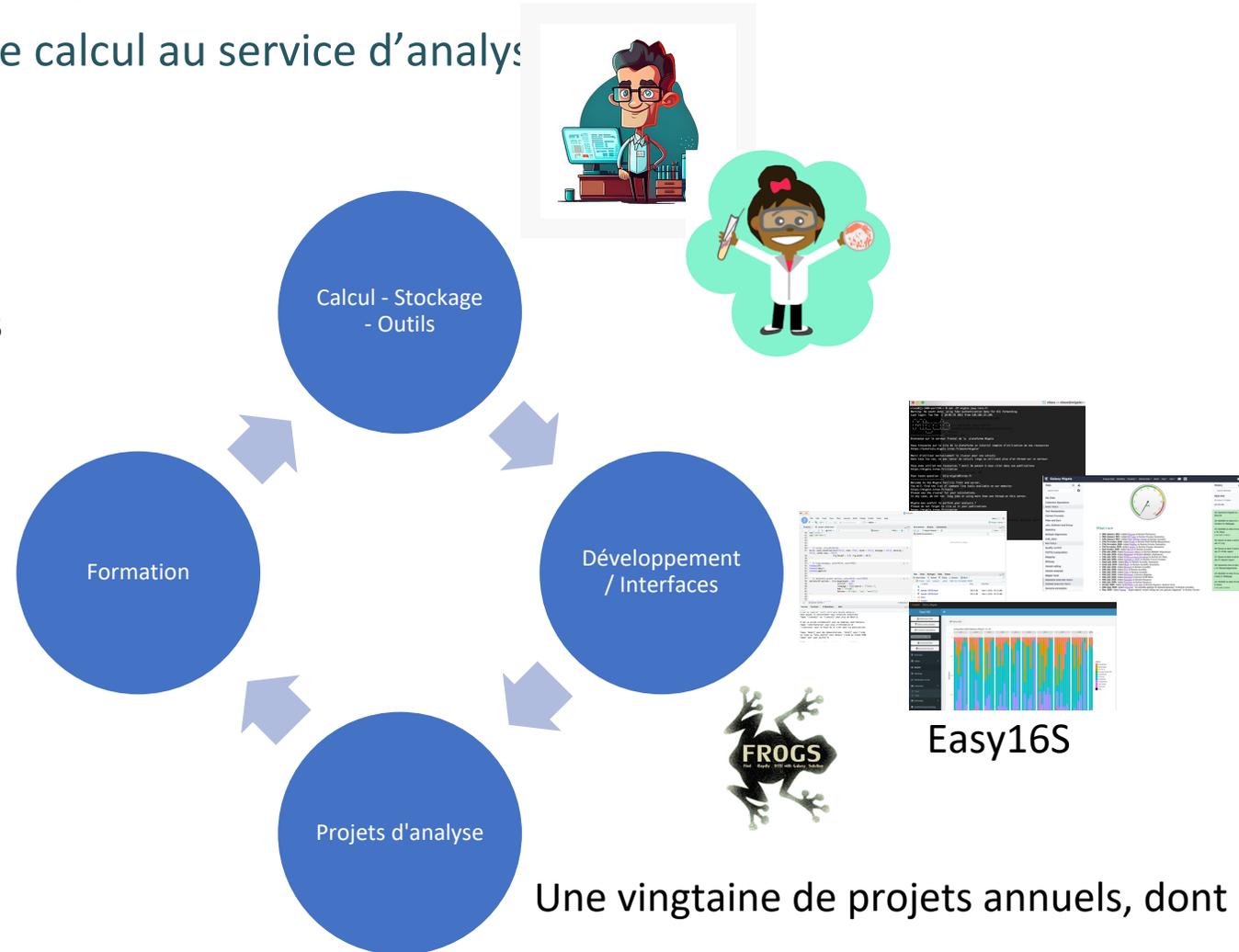
MetaPDO Cheese (Irlinger *et al*, *in prep*)
MetaBar Food (Rué *et al*, 2023)

Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse

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 »



Une vingtaine de projets annuels, dont :

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



INRAE

Migale et son offre de service text mining

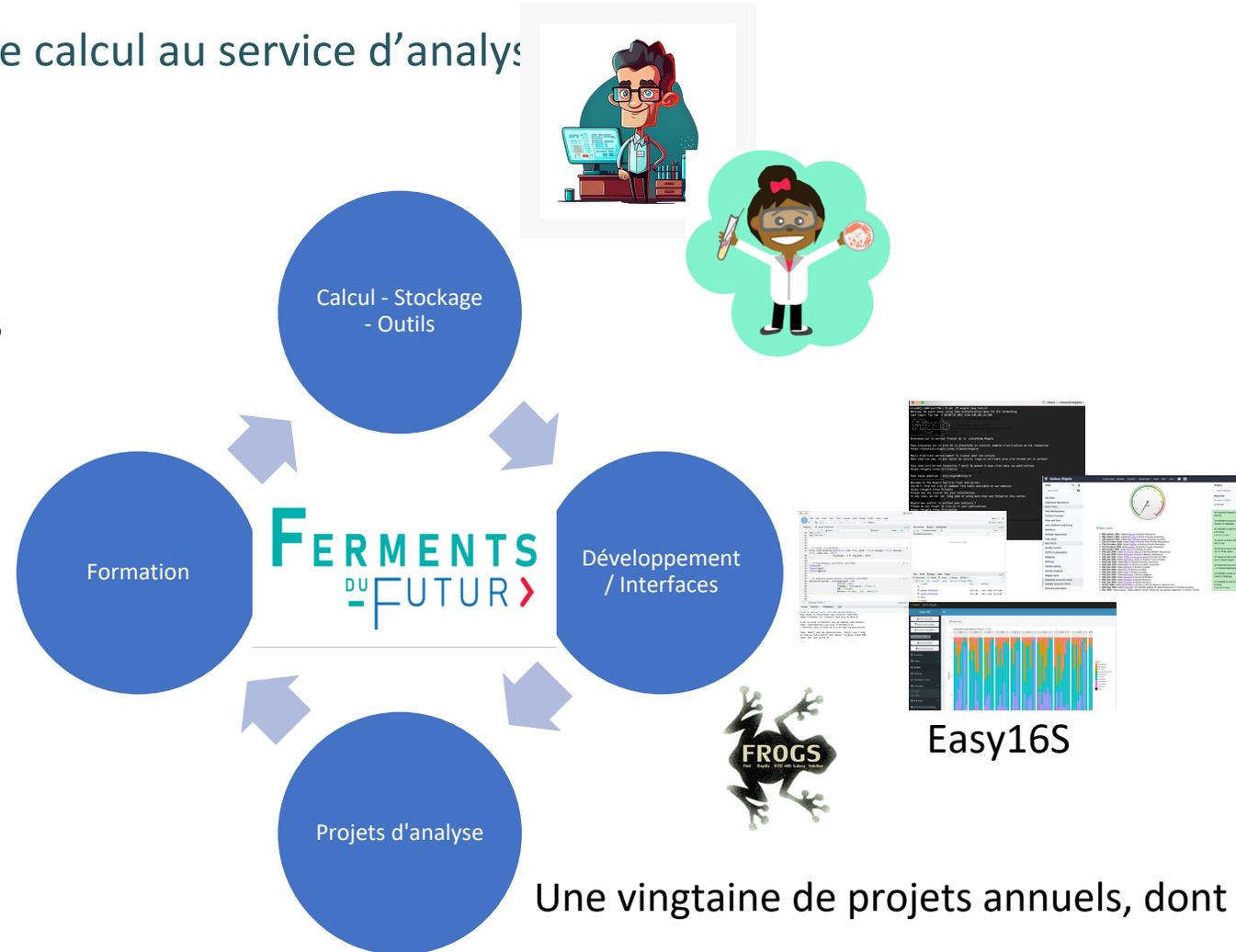
14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

Des résultats et projets complémentaires

De l'environnement de calcul au service d'analyse

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 »



Une vingtaine de projets annuels, dont :

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



INRAE

Migale et son offre de service text mining

14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

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.



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, spécialisée en text-mining

Données et Outils

Analyse de Données

Formation

Environnements et Calcul

Conception et Développement



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'assistance, corpus	pour utilisateurs autonomes		Science de la vie
DKPro Core (TU DARMSTADT) https://dkpro.github.io	Composants de traitement			général

Migale et son offre de service text mining

14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

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>

Mon projet (Locomotion Primates) | Changer de projet | Mon compte | Traitement | Déconnexion

Structurer les classes sémantiques

Classes Associées | Termes liés

locomotor behavior

Racine

body mass

grasping tail

locomotor behavior

arboreal locomotor behavior

gait

gait speed

Voir le terme

Forme de surface : locomotor behavior

Lemme : locomotor behavior

POS Tagging : JJ NN

Formes remplacées :

Mon avis

L'avis des autres utilisateurs

Autres propriétés

Recherche de termes

Forme... Lemme Pos taggi Producte Tête Expansion Type... Langue Source Valid... Vidette

Nb de mots: (min max)	Nb de docs: (min max)	Nb d'occurrences: (min max)	Label d...	Vidette
tree crown periphery	tree crown periphery	NN NN NN	en	tree crown (Variation)
adjacent tree crown	adjacent tree crown	JJ NN NN	en	(adjacent tree crown)
crown of adjacent tree	crown of adjacent tree	NN NN JJ	en	tree crown (Variation)
communicative	communicative	JJ	en	tree crown (Variation)
three-dimensional	three-dimensional	JJ	en	(three-dimensional)
hominoid evolution	hominoid evolution	JJ NN	en	large hominoid evolution

Terms par page 20 | 1-20 de 136027

Voir le terme dans son contexte

Contexte

American Journal of Primatology 87:105-118 (2006) RESEARCH ARTICLES Hand and Body Position During **Locomotor Behavior** in the Aye-Aye (*Daubentonia madagascariensis*) ELISSA KRAKAUER, PIERRE LEMELIN, AND DANIEL SCHMITT Department of Biological Anthropology and Anatomy, Duke University, Durham, North Carolina Aye-ayes (*Daubentonia madagascariensis*) have unique hands among primates, with extraordinarily long fingers in relation to body size. These long digits may be vulnerable to damage from forces during locomotion, particularly during head-first descent. To test this hypothesis, we examined hand and body position in three captive adult aye-ayes while they walked frequently. Previous behavioral studies of aye-aye locomotion reported that *Daubentonia* must curl its fingers during horizontal quadrupedalism and/or descent to reduce potential stresses on its long fingers. To test this hypothesis, we examined hand and body position in three captive adult aye-ayes while they walked quadrupedally on horizontal and oblique branches. Substantial variation in hand position was observed among individuals for each substrate orientation. While hand

Rows per page: 10 | 1-10 of 165

Alvis Search Engine

psychrobacter lives in aquatic

Microorganisms

Next value	Prev	All
Psychrobacter	65	22
Bacteria	25	16
bacterium	15	6
Psychrobacter spp	6	4
Colwellia	3	3
Psychrobacter urea	5	3
Psychrobacter glab	7	3
Polaribacter	3	3
Escherichia coli	5	2

Habitats

Next value	Prev	All
marine environment	47	19
living organism	17	9
soil	19	7
water	10	6
human	8	6
sea salt	12	5
sea ice	23	5

Structure of the O-polysaccharide chain of **Psychrobacter muricicola 2p(7)** isolated from overcooked water brines within permafrost.

Authors: Anna N Kondakova, Kseniya A Novotbilaya-Vasova, Marina S Drobotova, Galina N Senchenkova, Victoria A Shcherbakova, Alexander S Shastkov, David A Gilchinsky, Sergei A Nedospasov, Yury A Kozlov

2012 Carbohydrate research

Abstract: Psychrotrophic bacteria of the genus **Psychrobacter** have not been studied in respect to lipopolysaccharide structure. In this work, we determined the structure of the O-specific polysaccharide of the lipopolysaccharide of **Psychrobacter muricicola 2p(7)** isolated from overcooked (-9°C) **water brines** within permafrost. The polysaccharide was found to be acidic due to the presence of an amide of 2-acetamido-2-deoxy-1-galacturonic acid with glycine ([-GlcNAc6S]), which has not been hitherto found in nature. The following structure of the disaccharide repeating unit of the polysaccharide was established using composition analysis along with 1D and 2D (1H and 13C) NMR spectroscopy: -[4)-α-1-GalNAc6S(1-3)-β-1-GlcNAc(1-3-

Psychrobacter oceanii sp. nov., isolated from **marine** sediment.

Authors: Miketaka Matsuyama, Yuhiko Kitayama, Taroichi Sakai, Hirokazu Kashiwara, Akane Matsunobu, Tokumitsu Okada, Kiyoko Hirota, Daisuke Numoto

2015 International journal of systematic and evolutionary microbiology

Copyright IBSA, 2015

annotation - 1190042491 Evaluation of antibacterial activity of

Evaluation of antibacterial activity of synthetic aliphatic and aromatic monoacylglycerols.

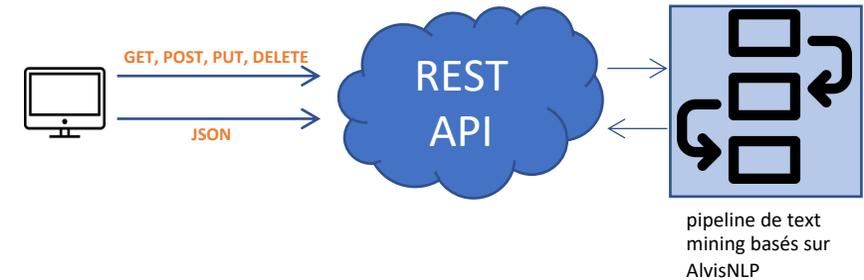
The antibacterial activity of synthetic aliphatic and aromatic monoacylglycerols (MAGs) was studied against two human pathogens: **Staphylococcus aureus** and **Escherichia coli**. The active compounds inhibited selectively **S. aureus**. The most active compounds amongst them were those with medium size aliphatic chain and aromatic MAGs with a electron withdrawing substituents at the aryl ring. The introduction of one or two-carbon spacer between the aryl ring and the carboxylic function did not influence antibacterial effectiveness.

Id	Annotation Set	K Type	Details
10015...	Imported: imported from review by armand ba chere d'ak in campaign 10	Bacteria	Staphylococcus aureus
10047...	Imported: imported from review by armand ba chere d'ak in campaign 10	Habitat	S. human

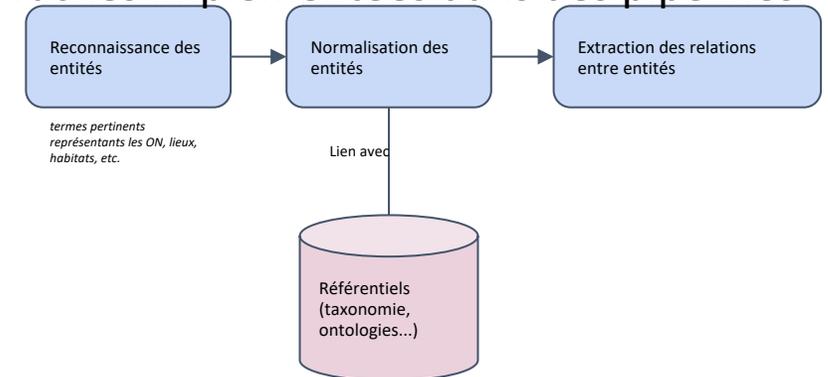
Données et Outils

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



Tâches implémentées dans des pipelines



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 une application web / projet TYDI++

ALVIS WRAPPER API -- Classification ^{0.0.1}
[Base URL: /]
Swagger JSON

API to do classification analysis for the TIERS ESV project

awapi/classifications classification tasks

- GET /awapi/classifications Get the list of all tasks
- DELETE /awapi/classifications remove all the tasks [FOR TESTING ONLY, WILL BE HIDDEN FOR USERS]
- POST /awapi/classifications Create and Run a new task
- GET /awapi/classifications/info Get the api meta info
- GET /awapi/classifications/{id} Fetch a given task and its status
- PUT /awapi/classifications/{id} Re-run a task given its identifier
- DELETE /awapi/classifications/{id} Delete a task given its identifier [FOR TESTING ONLY, WILL BE HIDDEN FOR USERS]
- GET /awapi/classifications/{id}/input Get a task input given its identifier
- GET /awapi/classifications/{id}/log Get a task output given its identifier
- GET /awapi/classifications/{id}/output Get a task output given its identifier

Models

ALVIS WRAPPER API -- Term Extraction ^{0.0.1}
[Base URL: /]
Swagger JSON

API to launch text mining analysis tasks within the Tydi+ project

core main actions on tasks

- GET /core/tasks/extraction/{id} Fetch a given task and its status
- PUT /core/tasks/extraction/{id} Run a task given its identifier
- DELETE /core/tasks/extraction/{id} Delete a task given its identifier
- GET /core/tasks/extractions Get the list of all tasks
- DELETE /core/tasks/extractions remove all the tasks [FOR TESTING ONLY, WILL BE HIDDEN FOR USERS]
- GET /core/tasks/extractions/{id}/input Get a task input given its identifier
- GET /core/tasks/extractions/{id}/log Get a task output given its identifier
- GET /core/tasks/extractions/{id}/output Get a task output given its identifier
- POST /core/tasks/{Lang}/extract Create a new task to be run

Models



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

We can collaborate or help you to analyze your data. The different modalities and procedures are explained [here](#). This form helps us to evaluate the feasibility and dedicated to your needs.

You will be contacted as soon as possible to give you an answer to your request and establish the requirements and delays.

You ask for *
 collaboration support

Category *
 bioinformatics biostatistics
 text_mining

————— Description of your project —————

Project name *
[Text input field]

Aim of the project *
[Text area]

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, classifieurs
Exemples	Corpus "pathway" (PubMed)	Extraction "gènes" et "métabolites"	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

INRAE

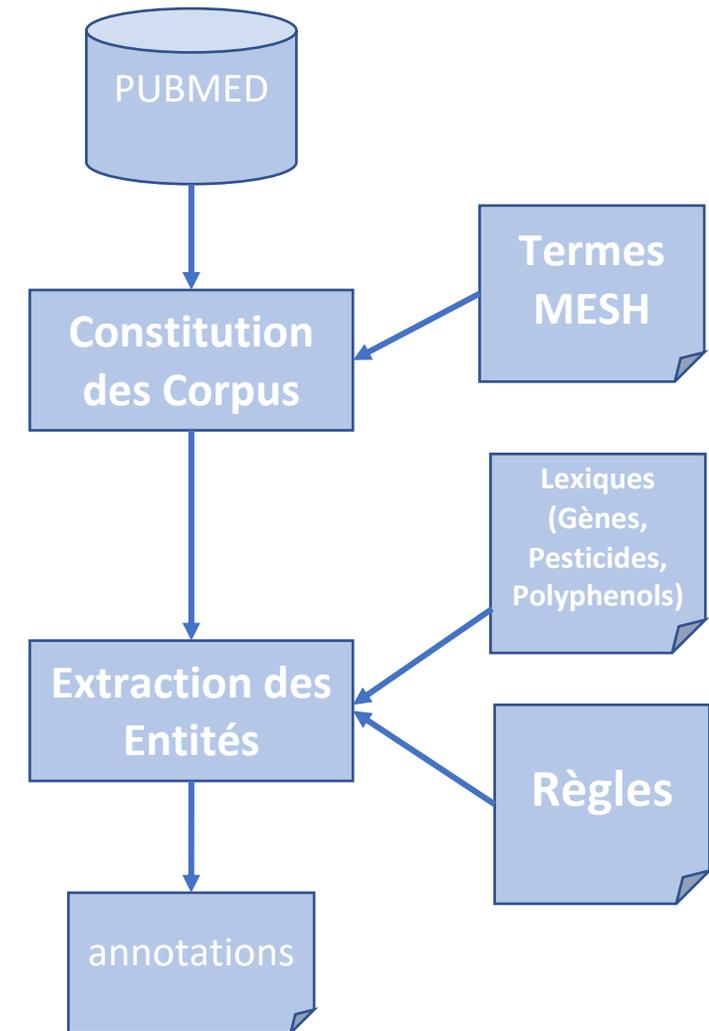
Migale et son offre de service text mining

14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

Analyse de données

Extraction d'informations sur la fertilité et le métabolisme animal

- **Entités à extraire**
 - Gènes, pesticides, polyphénols, espèces, tissus
- **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



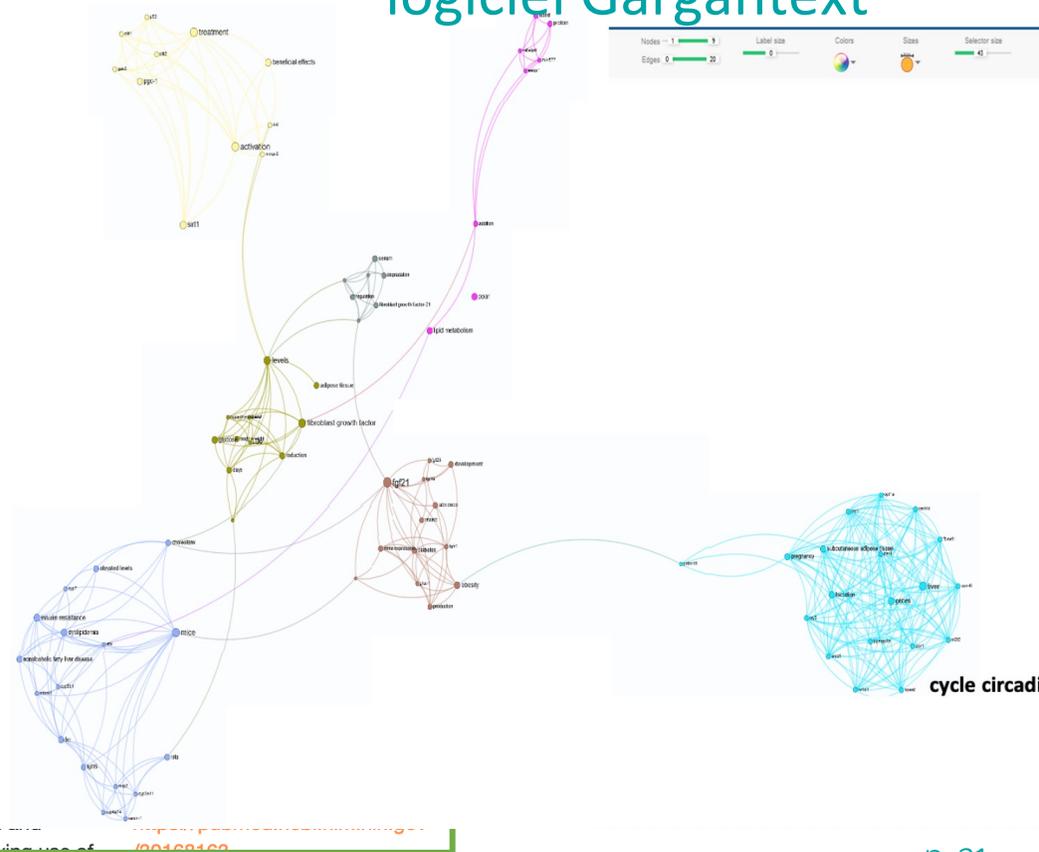
Analyse de données

Extraction d'informations sur la fertilité et le métabolisme animal

Résultats

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 containing hepatocytes in fetal rats.	https://pubmed.ncbi.nlm.nih.gov/6169057
				Effects of engraftment of CD34(-) and CD34(+) from mobilized blood differs from that of CD34(-) and CD34(+) cells from bone marrow.	https://pubmed.ncbi.nlm.nih.gov/11008020
				Gene expression analysis of hypothalamic and pituitary components of the growth hormone axis in fasted and streptozotocin-treated mice. Neuropeptide Y (NPY)-intact (NPY+/+) and NPY-knockout (NPY-/-) mice.	https://pubmed.ncbi.nlm.nih.gov/16244497

Exploitation sous le logiciel Gargantext



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

PMID	PESTICIDE	SCORE	SECTION	TITLE	PMID	GENE	SCORE	SECTION	TITLE
30502743	endosulfan	20	title+abstract	Different effects of α-endosulfan, β-endosulfan, and endosulfan on sex hormone levels, metabolic parameters, and oxidative stress in adult mice treated with endosulfan.	25625345	Luteolin	11	title+abstract	Isolation of Luteolin and glucoside from Dendranthema morifolium Ramat Tzvel and its Pharmacokinetics in Rat.
7806131	casein	20	title+abstract	Temporal effects of prolactin on tyrosine kinase activity, casein synthesis, and casein mRNA accumulation in mammary gland explants.	27334554	Resveratrol	10	title+abstract	Differential responses of Resveratrol on proliferation of progenitor cells and age-related hippocampal neurogenesis.
19539752	ethanol	20	title+abstract	Simultaneous prenatal ethanol and nicotine exposure affect ethanol consumption, ethanol preference, and oxytocin receptor binding in adolescent and adult rats.	28822243	Chrysin	10	abstract+title	Chrysin-induced sperm motility and fatty acid profile changes in reproductive performance of male rats.
15805107	sucrose	20	title+abstract	Phloem-localized, proton-coupled sucrose carrier ZmSUT1 mediates sucrose transport under the control of the sucrose signaling pathway in Zea mays.	30168163	Naringenin	10	abstract+title	Testicular microanatomic and hormonal alterations following use of naringenin in adult rats.

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

migale

<https://migale.inrae.fr>

Collaborateurs dans MalAGE

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



INSTITUT FRANÇAIS DE BIOINFORMATIQUE



FRANCE
GENOMIQUE

INRAE

Migale et son offre de service text mining

14.09.2023/ Journées du PEPI IBIS / V. Loux & M. Ba

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>