Extraction et intégration d'informations dans un contexte FAIR

Claire Nédellec

MaIAGE, INRAE, Université Paris-Saclay

JNE URFIST, 30 sept 2021, Lyon

A chaque question de recherche, son analyse numérique de document

Besoin précis ➔ Extraction d'information, traitement automatique de la langue (TAL)

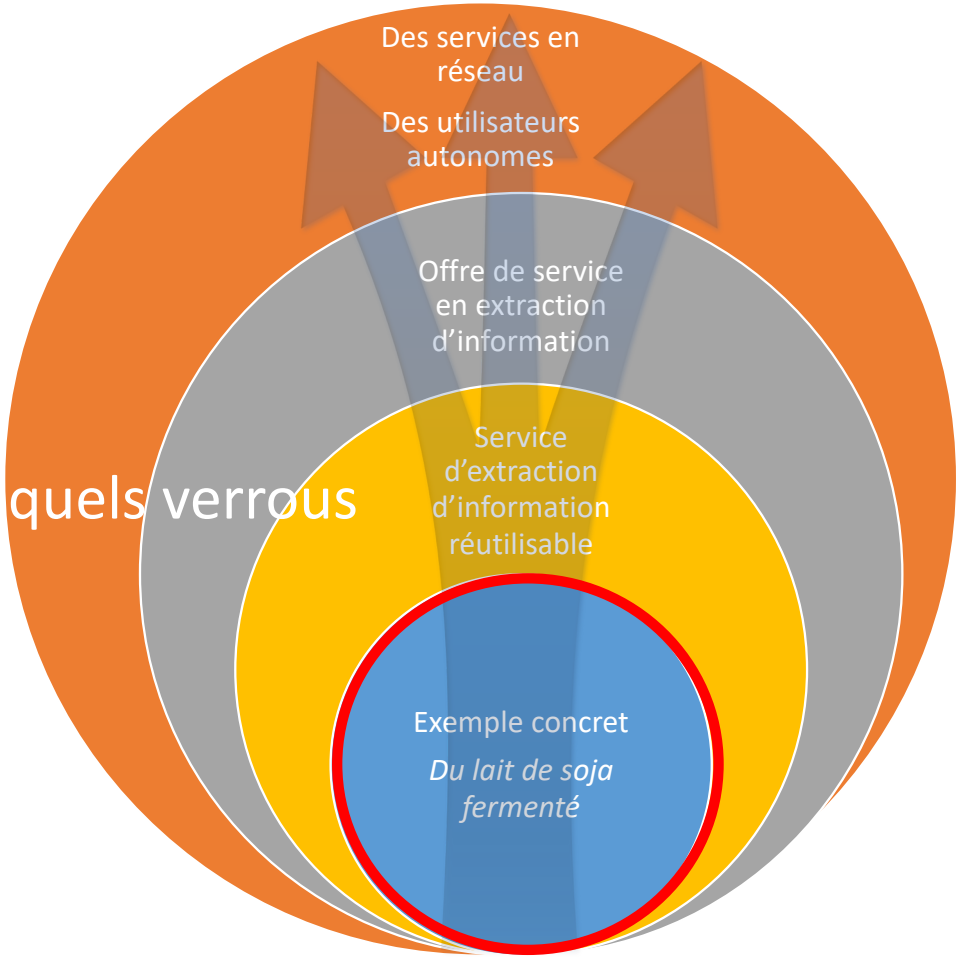
On sait modéliser ce que l'on cherche et on peut cibler l'information à extraire

Besoin exploratoire ➔ Fouille et découverte de connaissances

On ne sait pas caractériser a priori *les connaissances que l'on va dériver du corpus de textes et de données collecté*

Réutilisabilité, Interopérabilité d'outils et de données d'extraction d'information à partir de texte

Quels principes, quels verrous



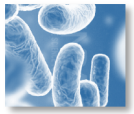
Un besoin, une réponse

Pour concevoir un nouveau produit de type yaourt à partir de jus végétal, extraire l'information de documents scientifiques

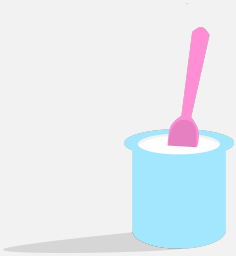


Ingrédients

Quel consortium microbien ?



Fermentation
(propriété d'acidification)

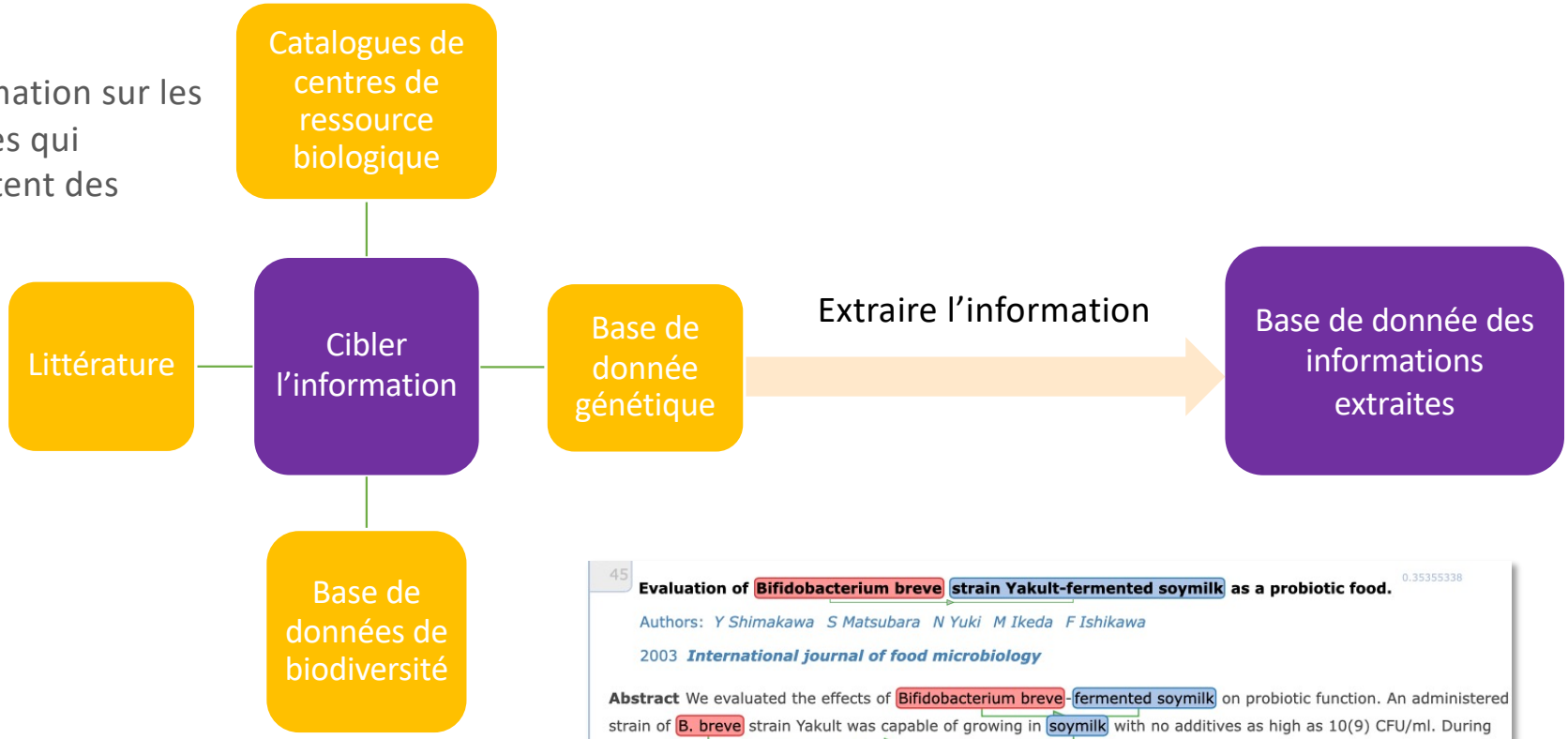


Produit final

- **L'équipe du projet ENovFood**
- Microbiologistes, produits laitiers (STLO)
- Chercheurs en extraction d'information (Bibliome, MaIAGE)
- Développement de service (Plateforme bioinformatique Migale, MaIAGE)

Analyse du besoin : (1) Identifier des souches candidates dans la bibliographie, (2) puis les tester expérimentalement

Sources d'information sur les microbes qui fermentent des plantes



45

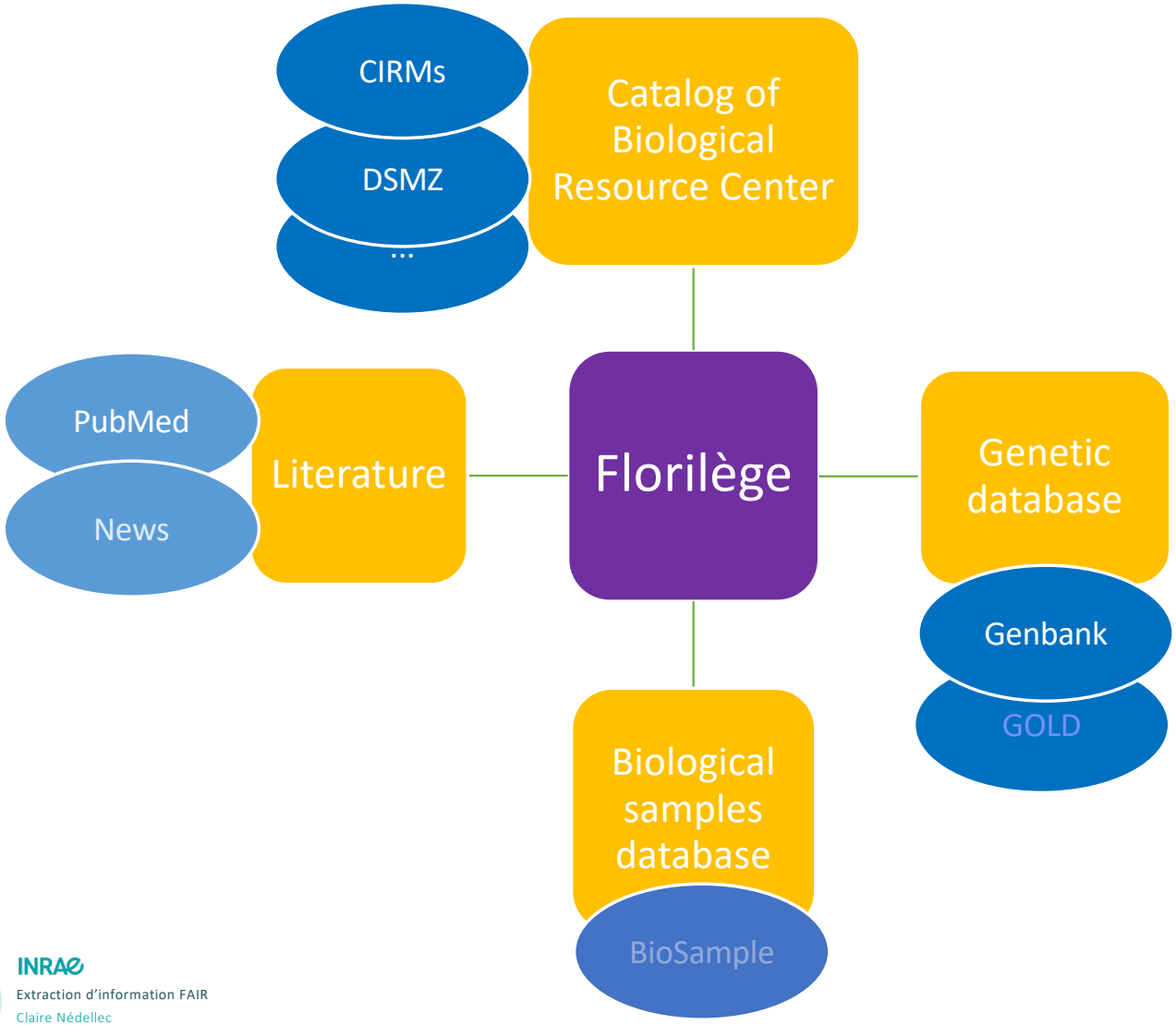
Evaluation of Bifidobacterium breve strain Yakult-fermented soymilk as a probiotic food. 0.35355338

Authors: *Y Shimakawa S Matsubara N Yuki M Ikeda F Ishikawa*

2003 ***International journal of food microbiology***

Abstract We evaluated the effects of Bifidobacterium breve-fermented soymilk on probiotic function. An administered strain of B. breve strain Yakult was capable of growing in soymilk with no additives as high as 10(9) CFU/ml. During storage of the fermented soymilk at 10 degrees C for 20 days, viable counts of the strain did not change. The growth

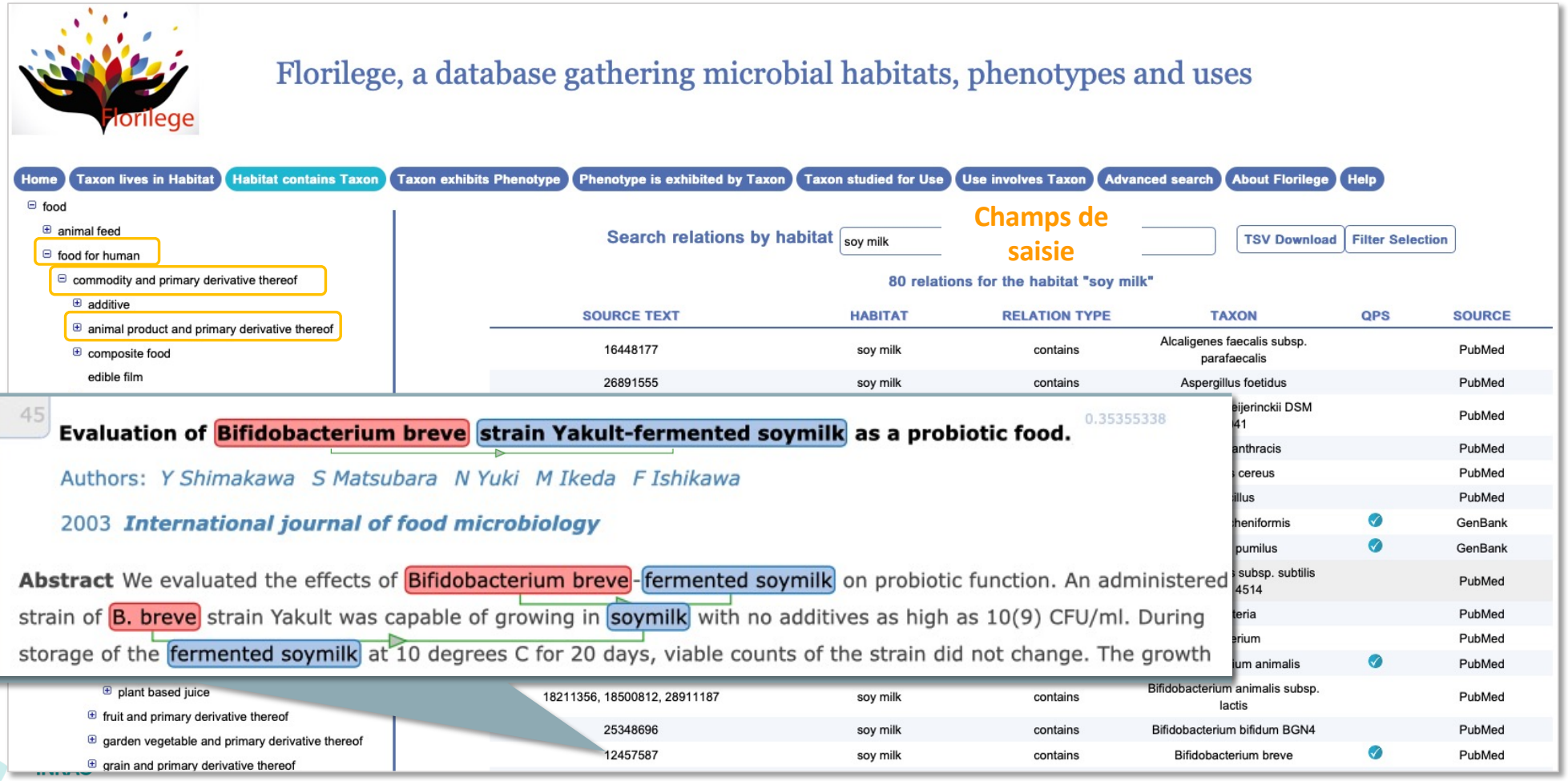
Base Florilege



http://migale.jouy.inra.fr/florilege/

Réalisation : une application d'extraction d'information et son interface

http://migale.jouy.inrae.fr/florilege/



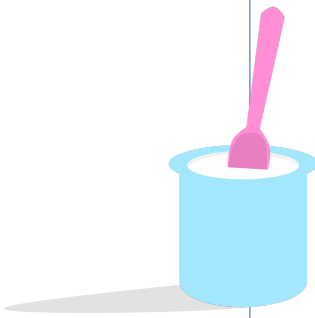
Onto
do

Résultats de la requête

Criblage des souches bactérienne prometteuses par l'extraction automatique d'information

Pré-sélection des espèces par extraction d'information à partir de publications
7 espèces pertinentes
Possédant les propriétés attendues

- Acidification
- Viables dans un mélange de lait et de soja
- Utilisable en alimentation humaine (QPS)
- Disponibles dans un centre de ressources biologique -> 201 souches



Etape expérimentale
Implémentation des souches sélectionnées dans un mélange soja-lait et mesure de l'acidification

Conclusion
Gain signicatif de temps et de moyens économisés en recherche bibliographique et en expérimentation
Ciblage d'un sous-ensemble pertinent d'espèces parmi les milliers d'espèces candidates.
[Harle et al. *Food Microbiol.* 2020.]

Florilege, une ambition plus large que de nouveaux yaourts au soja

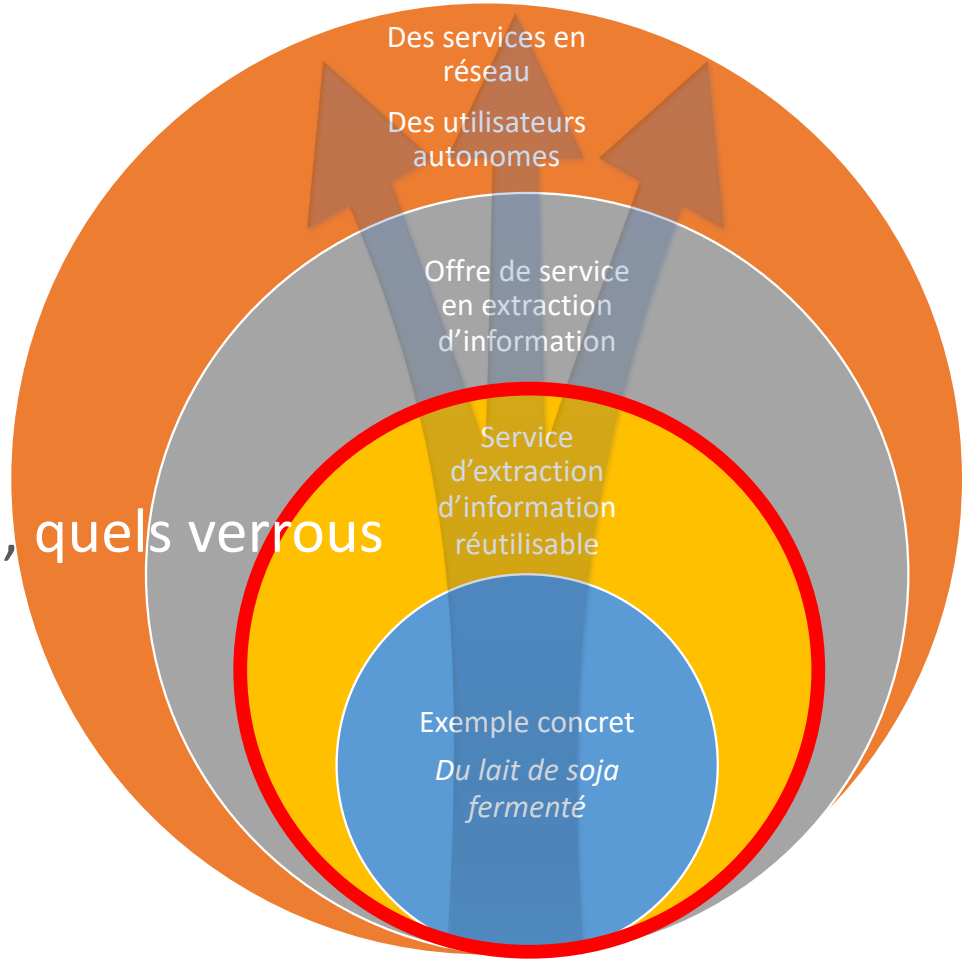
Besoin plus large, issu d'équipes très variées

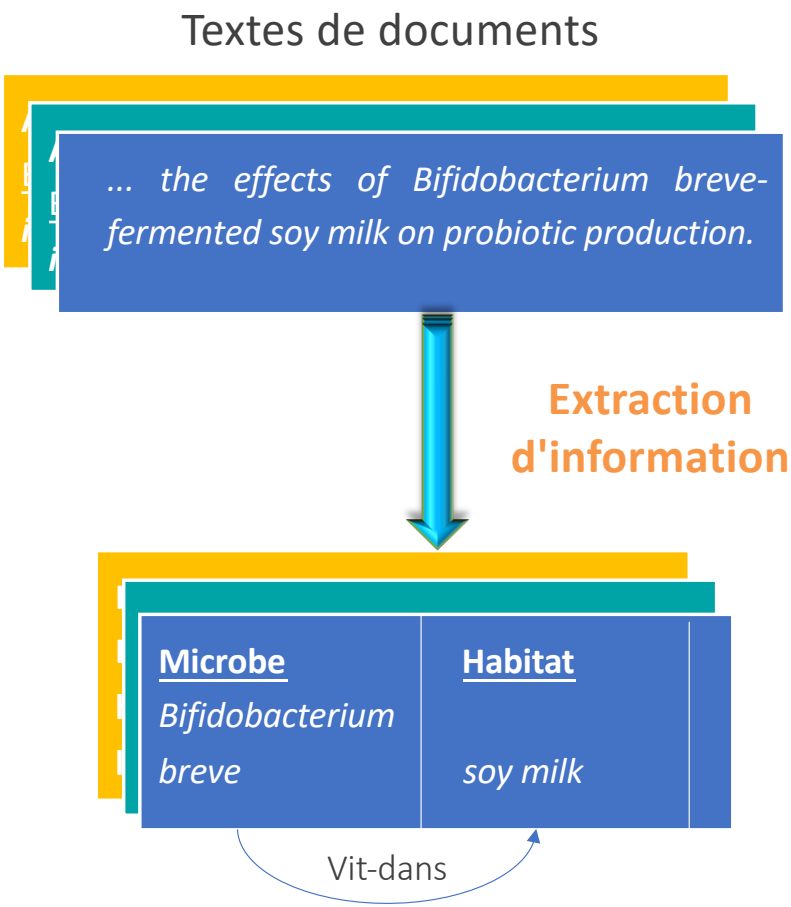
- Quels microbes dans un habitat donné
- Ex : aliments, fonds marins, microbiote intestinal
- D'où provient ce microbe trouvé dans cet aliment
- Quels microbes possèdent un phénotype donné

Construire une base de connaissance

- qui regroupe toutes les informations sur les microbes
- ... que je puisse relier à d'autres informations (ex. génétiques)
- à partir des connaissances contenues dans des publis et BdD

Quels principes, quels verrous



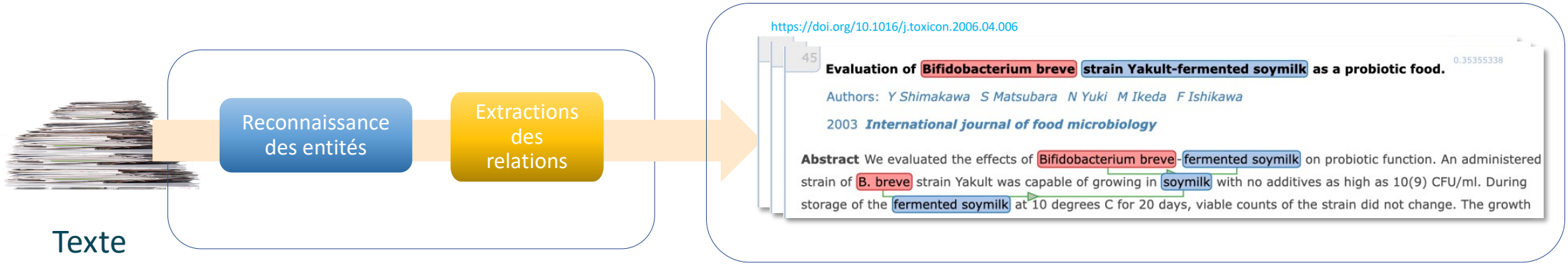


Représentation formelle de l'information

Extraction automatique d'information

Transforme une donnée non structurée, du texte en donnée structurée manipulable par un ordinateur

Extraction automatique d'information



Texte annoté par des entités et des relations

Dans l'exemple

- Les entités de type microbe et habitat
- Reliés par une relation orientée vit-dans

Liage et interopérabilité des données

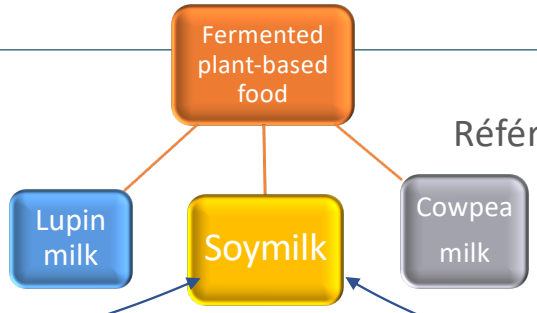
Les données sont comparables, si elles sont représentées dans un même référentiel

Base de données X

Habitat normalisé	Habitat	Taxon
yoghurt	Greek yogurt	L. plantarum ZJ316
soymilk	soybean milk	B. breve ATCC 15700

Base de données Y

Isolement normalisé	Isolement	Souche
yoghurt	traditional yoghurt	Lactobacillus plantarum 2035
soymilk	soyabean milk yoghurt	Lactobacillus plantarum P1201



Référentiel d'habitat microbien

Formaliser l'association entre données du texte et référentiels

Reconnaissance des entités

Normalisation des entités

Extractions des relations

Extraction automatique d'information basée sur des ontologies

Texte

Lives in

Evaluation of Bifidobacterium Breve *-fermented soy milk*

TaxID 414286 Bifodobacterium breve

OBT **3145** Soymilk

L'extraction d'information contribue aux principes

Findable Accessible Interoperable Reusable

Référentiels hiérarchiques

TaxID 2 Bacteria

TaxID 201174 Actinobacteria

TaxID 1678 Bifidobacterium

TaxID 1685 Bifidobacterium breve

Taxinomie du NCBI

Lives in

OBT 1 Habitat

OBT 3061 Legume-based drink

OBT 3366 Soy beverage

OBT 3399 Soymilk

Ontologie OntoBiotope

Méthodes pour la reconnaissance d'entités, l'extraction de relations et la normalisation

Projection de dictionnaires, règles, analyse terminologique

- Limitation : traitement de la polysémie, incomplétude, relations inter-phrases
- Coût de l'adaptation

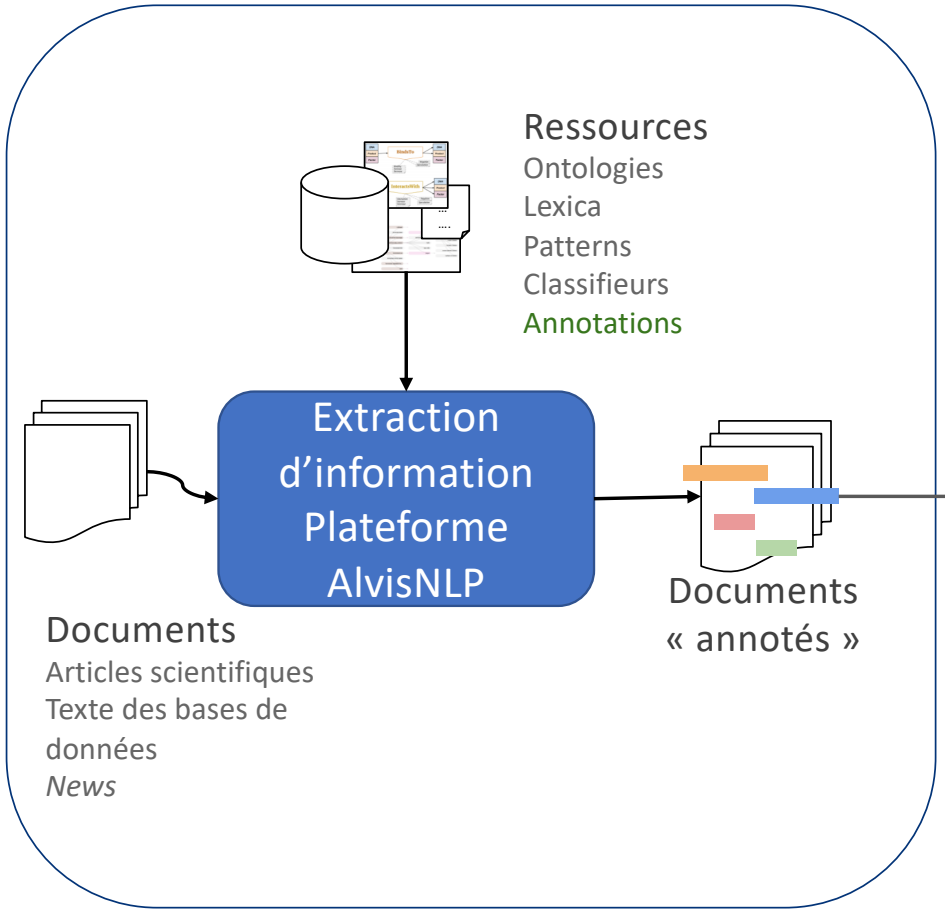
Traitement automatique de la langue, apprentissage automatique

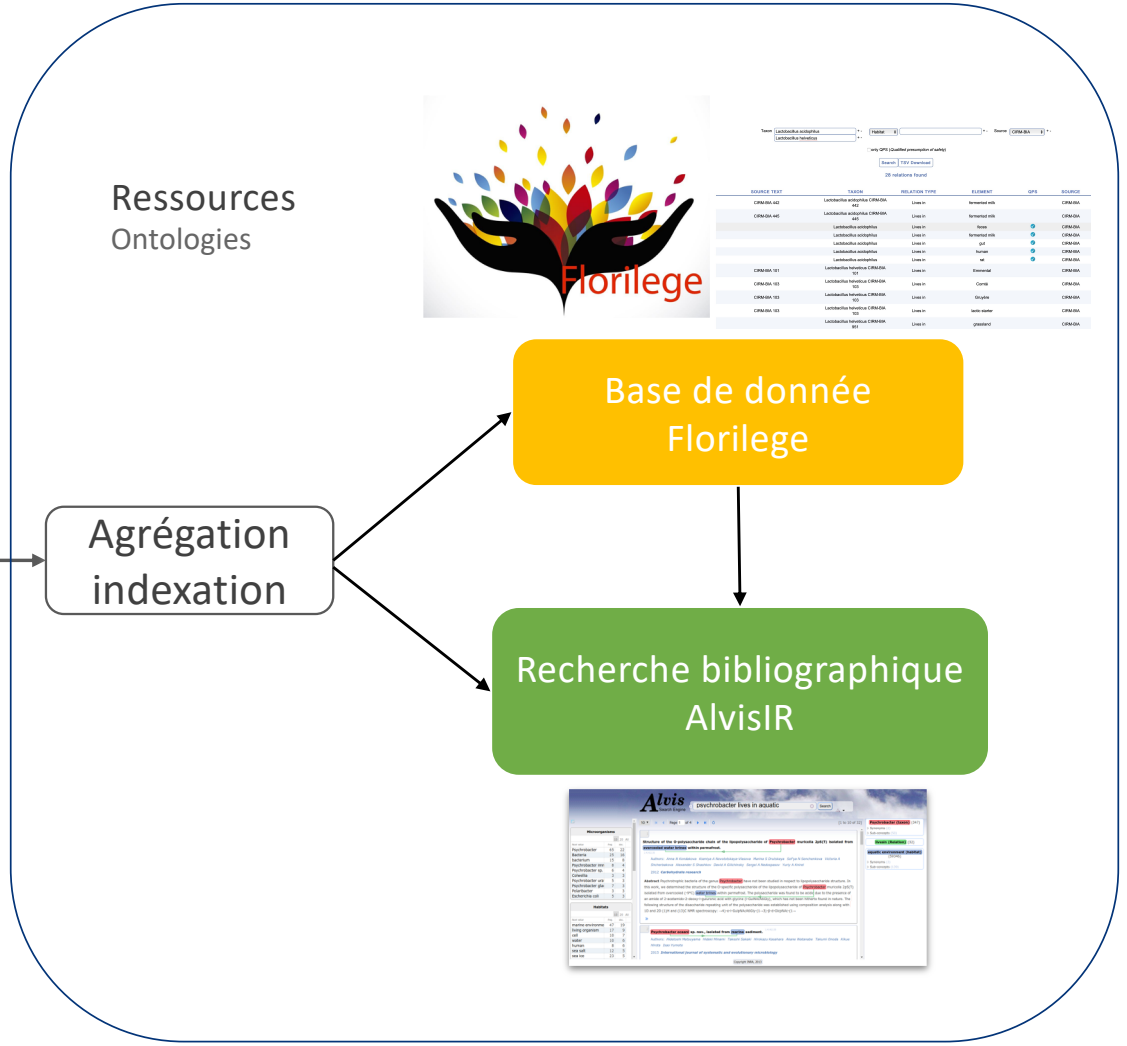
- Progès récents en normalisation par *deep learning*
- Limitation : nombre d'exemples d'entraînement annotés par des experts

Evaluation

- De très nombreux jeux de test (*shared task*) publiés avec les performances des systèmes participants

Implémentation





Méta données des documents

Des entités emboitées, discontinues

PMID- 12457587
DP - 2003 Mar 15
TI - Evaluation of Bifidobacterium breve strain Yakult-fermented soymilk as probiotic food.
PG - 131-6
AB - We evaluated the effects of Bifidobacterium breve-fermented soymilk on probiotic function. An administered strain of B. breve strain Yakult was capable of growing in soymilk with no additives as high as 10(9) CFU/ml. During storage of the fermented soymilk at 10 degrees C for 20 days, viable counts of the strain did not change. The growth inhibition of the strain in a bile-containing medium was lessened by the addition of soy protein. In human feeding experiments, the administered B. breve was recovered at a level of over 10(9) CFU/g faeces, accompanied by an increase in the total number of bifidobacteria. These results indicate that fermented soymilk with B. breve strain Yakult could be a novel type of probiotic food.
FAU - Shimakawa, Y
AU - Shimakawa Y
AD - Yakult Central Institute for Microbiological Research, 1796 Yaho, Kunitachi, Tokyo 186-8650, Japan. yasuhisa-shimakawa@yakult.co.jp
AU - Matsubara S
AU - Yuki N
AU - Ikeda M
AU - Ishikawa F
LA - eng
PT - Journal Article
JT - International journal of food microbiology
MH - Adult
MH - Animals
MH - Beverages/*microbiology
MH - Bifidobacterium/drug effects/*growth & development/isolation & purification
MH - Fermentation
MH - Food Handling/methods
MH - *Food Microbiology
MH - Food Preservation
doi: 10.1016/s0168-1605(02)00224-6.

Index MeSH thématique global au document

Métadonnées locales des informations extraites du document

Entités

Id	Référence	Position	Longueur
1	OBT - Bile containing medium	212	22
2	OBT - Fermented beverage	29	36
3	OBT – Human	326	5
4	OBT - Probiotic food	596	14
5	OBT - Soymilk	53	7
6	NCBI tax - Human	326	5
7	NCBI tax - Bifidobacterium breve	29	21

...

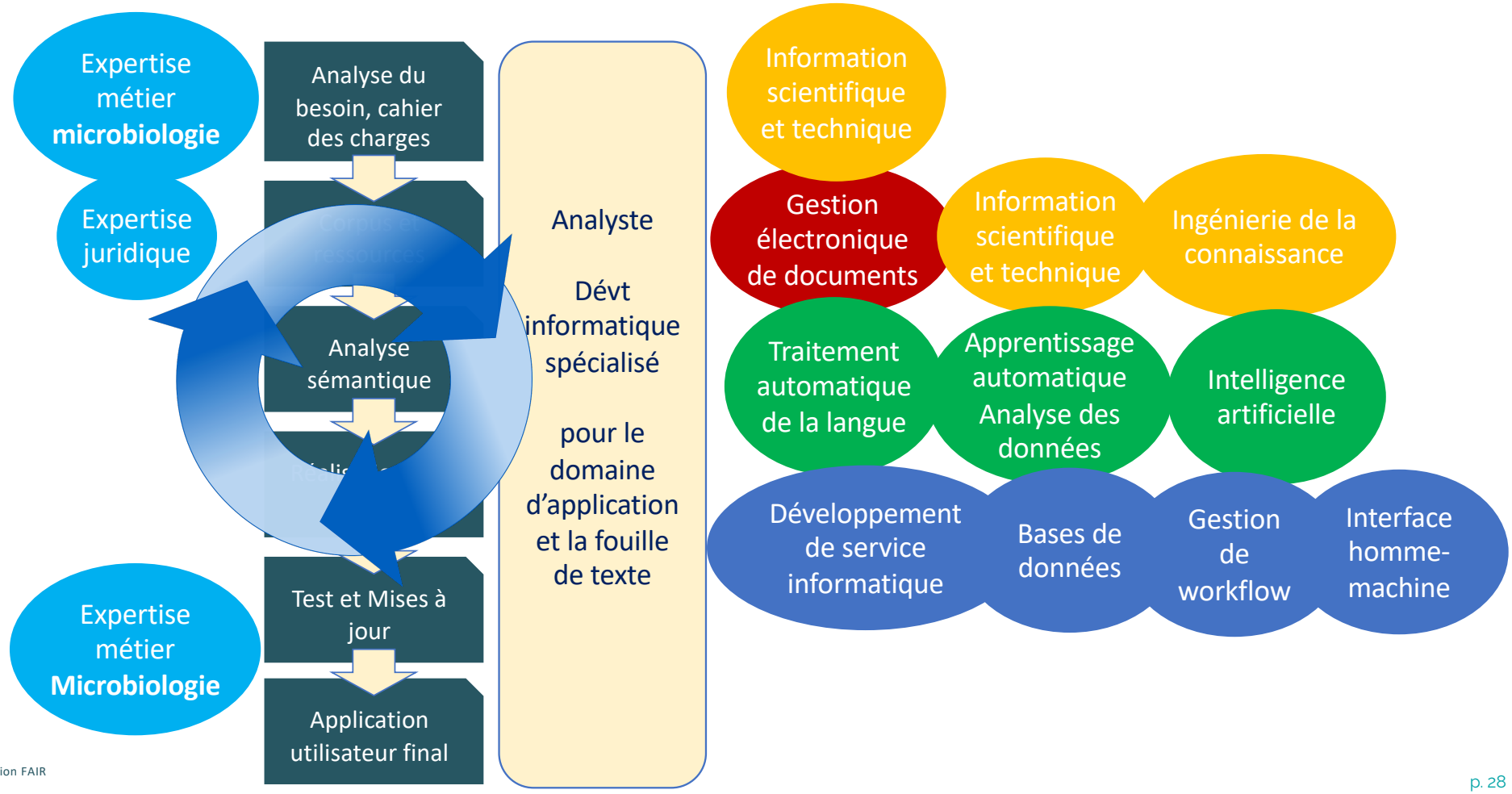
Relation

Type	Argument 1	Argument 2
Lives in	7	5

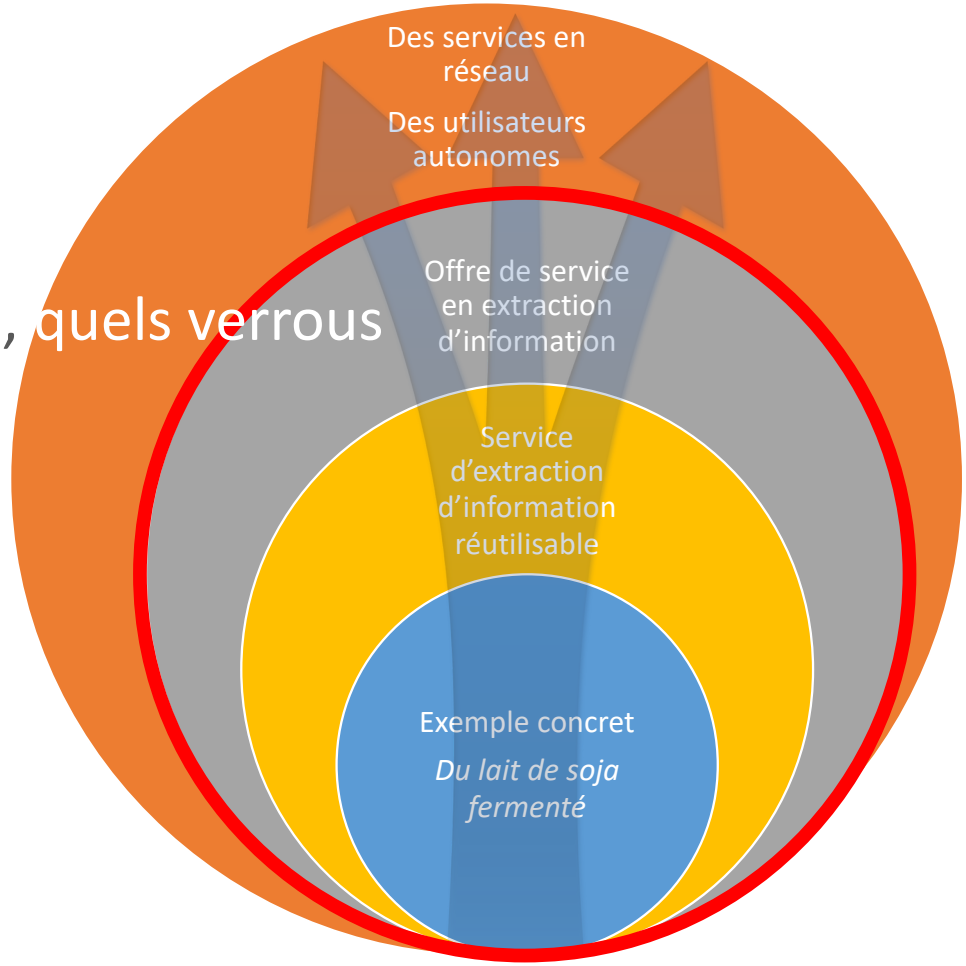
Florilege, un traitement de liage automatique de données textuelles et non textuelles à grande échelle

[Chaix *et al., Food Microbiology*, 2019]

Démarche et compétences pluridisciplinaires



Quels principes, quels verrous



Systématiser la démarche pour offrir des services de TAL sur la plateforme bioinformatique Migale

Florilege : une application *pilote* co-construite de bout en bout sur un exemple

Objectif de la plateforme bioinformatique Migale

- Proposer un nouveau service basé sur le TAL
- Permettant de lier les analyses bioinformatiques classiques (ex. étude de biodiversité par des approches de métagénomiques) à des informations extraites de textes
- Mutualisable avec d'autres plateformes informatiques d'INRAE

Sans être spécialistes du TAL
En s'appuyant sur des équipes de recherche (Bibliome) pour le choix des outils spécialisés

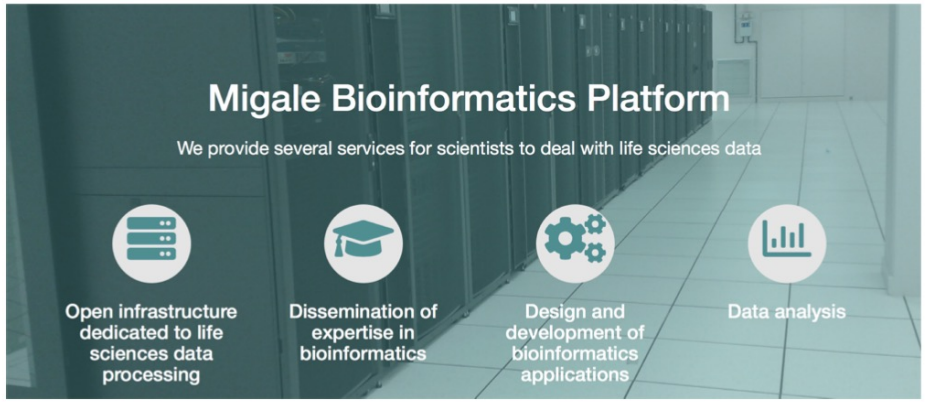
Objectif de l'équipe de recherche Bibliome

- Valoriser les résultats de ses recherches
- Bénéficier d'une plateforme d'expérimentation
- Participer à des projets de recherche et développement en partenariat

La plateforme bioinformatique Migale

Proposer services et formation aux scientifiques pour traiter leurs données en Science de la Vie

Une infrastructure Scientifique Collective d'INRAE, membre de
BioInfOmics : l'IR en bioinformatique d'INRAE
IFB : Institut Français de Bioinformatique
Elixir : le réseau européen pour l'information biologique



Inscrite dans une démarche de Science Ouverte, données, logiciels et services

Traitement Automatique de la Langue et applications

Développer une offre de service en TAL

Dans une application de TAL,
- beaucoup de composants sophistiqués à combiner dans un *workflow*
- Beaucoup de données et ressources très diverses

Comment passer à l'échelle et développer rapidement de nouvelles applications ?

Rationalisation informatique des données et traitements

- Réutilisation, composition, édition et exécution de traitements
- Interopérabilité des composants, données et traitements
- Disposer de moyens automatiques de conception de corpus documentaires spécialisés

AlvisNLP

Bibliome DiPSO
Opscidia

Adaptation des services par

- Des méthodes d'apprentissage automatique et de traitement automatique de la langue
- L'utilisation, la conception et l'enchissement de termino-ontologie spécialisées

Bibliome

Bibliome DiPSO

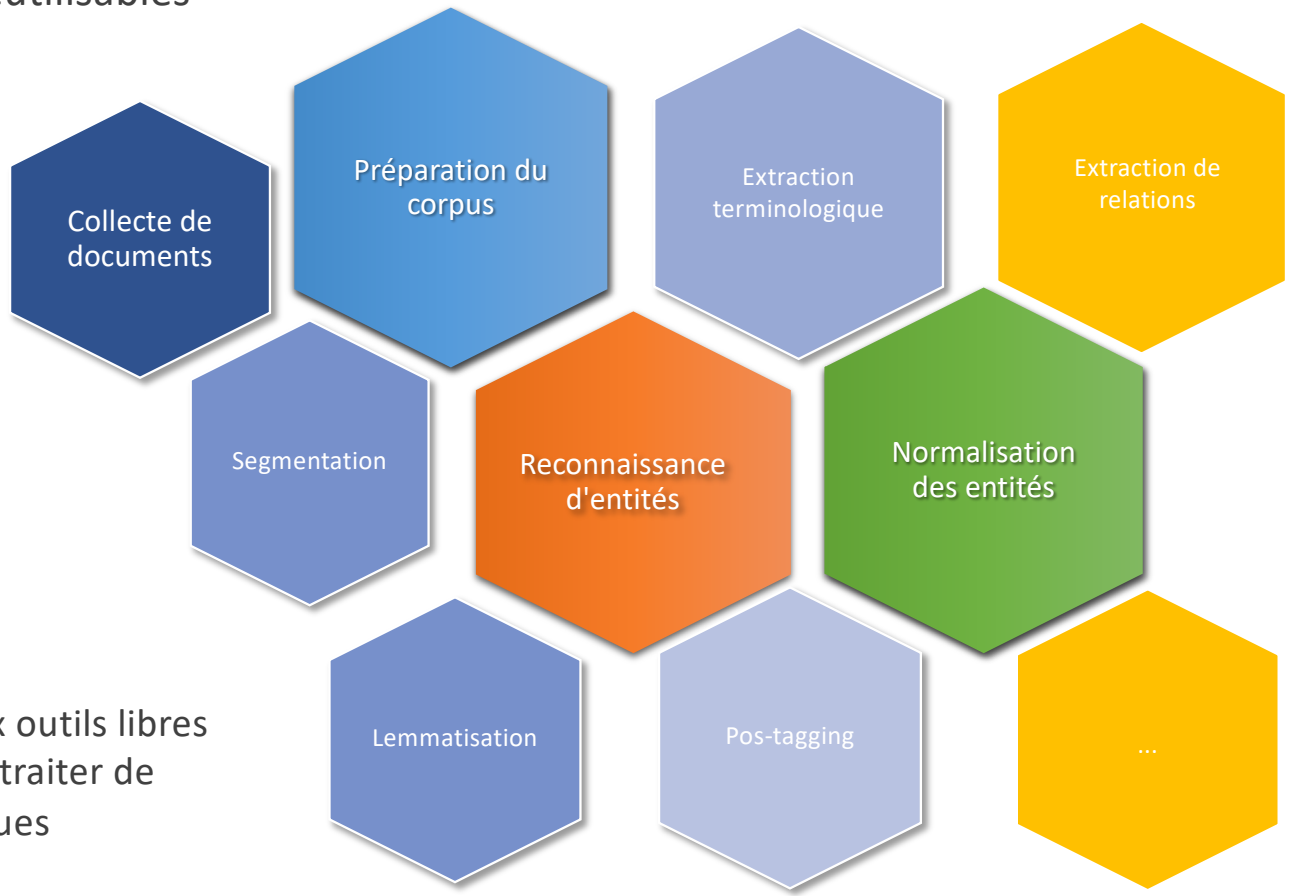
Rassembler des compétences

IST, recherche en TAL, devt en TAL, ingénierie en informatique, compétence SdV

Travailler en réseau : mutualiser, réutiliser

Architecture

Pipeline de modules réutilisables



De très nombreux outils libres à composer pour traiter de nombreuses langues

A chaque question de chercheur, son étude de TAL

L'exploration de textes et de données à des fins de recherche est toujours réalisée en réponse au questionnement d'un scientifique ou d'un groupe de scientifiques.

- Les applications diffèrent par
  - la type et la qualité des résultats attendus
  - La disponibilité de solutions existantes
  - la nature et le volume des sources à explorer
  - la complexité des traitements à mettre en œuvre
  - le degré de connaissances à injecter dans le processus

Pas nécessairement de gros volumes, mais un choix éclairé

Complémentarité des outils, plateformes et infrastructures

Outils/services sur étagère pour des utilisateurs novices

- De plus en plus d'outils libres accessibles
- Des outils clefs en main, facile d'utilisation, mais monolithiques
  - Des outils génériques, mal adaptés à des besoins spécifiques
  - Des outils dédiés spécialisés

Des bibliothèques d'outils et de ressources sémantiques à combiner

- Evaluées, documentées
- Pour des utilisateurs IST, Data Analysts
- A assembler pour concevoir des applications

Ex. CLARIN, SciKitLearn, AgroPortal

Des environnements de composition de traitement automatique de la langue

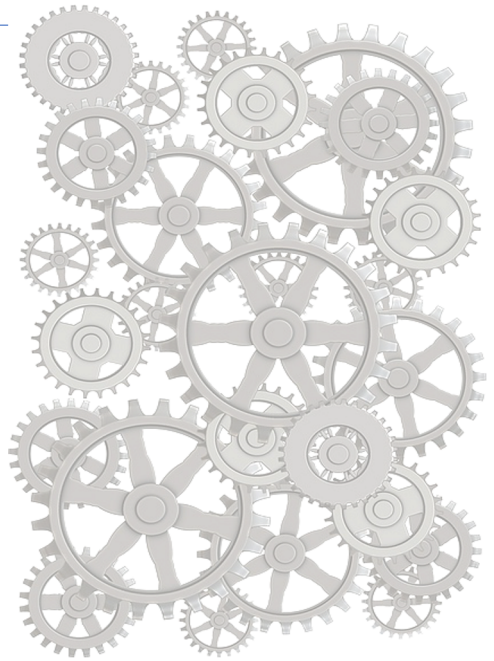
- Des outils clefs en main, analyse syntaxique, reconnaissance d'entités, extraction de relations
- Pour des utilisateurs spécialistes en TAL

Ex. spaCy, StanfordCoreNLP, AlvisNLP

Complémentarité des outils, plateformes et infrastructures

Plateformes techniques, comme Migale

- Pour répondre à la grande diversité des besoins,
- Apportent de la flexibilité : configuration en fonction des besoins
- Pouvoir varier les données
- Mutualiser et réutiliser des parties
- Ajouter de nouveaux services/composants
- Disposent de compétences *data analyst, IST, informaticien*



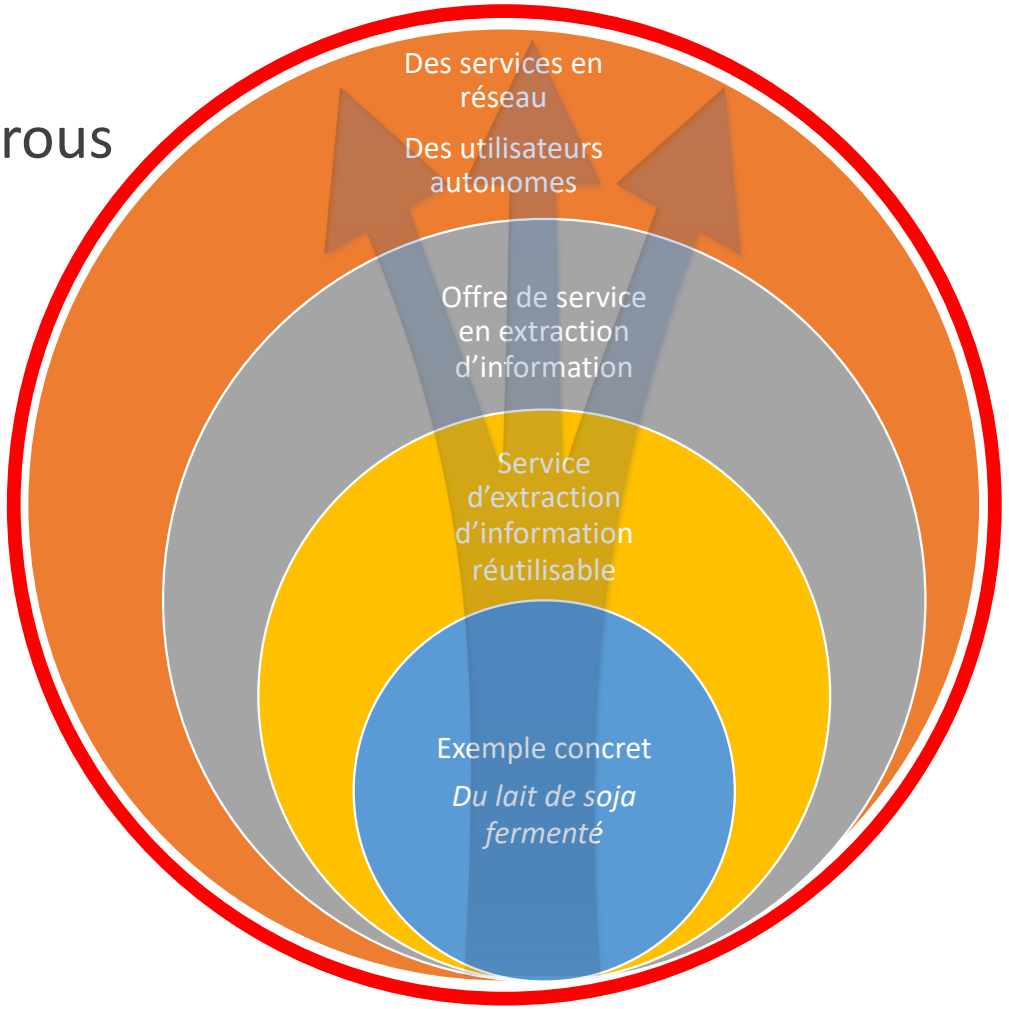
- Pour des utilisateurs novices en fouille de texte
- Des services génériques ou à façon
- Accompagnement humain des utilisateurs

Ex. Bioinformatique : Migale, Humanités numériques : Huma-Num, CoreText, Gargantext

Plateformes souvent isolées, *aux moyens limités*

Haute technicité de la création ou l'adaptation de services

Quels principes, quels verrous

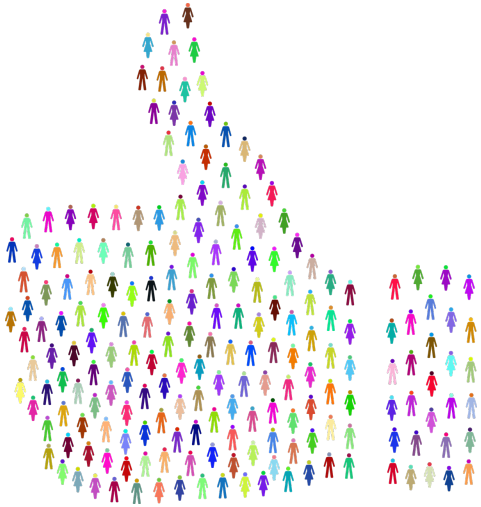


Passer à l'échelle,

de petites plateformes, avec un ou deux ingénieurs
à une offre de service élargie pour des utilisatuers plus autonomes

- Le projet d'infrastructure européenne OpenMinTeD
- Le projet d'étude du CoSO Visa TM

Missions d'un dispositif, d'une e-infrastructure avancée de TAL



Faciliter l'appropriation et l'utilisation des technologies de TAL par les chercheurs et les acteurs de l'appui à la recherche

Expertise, accompagnement, animation

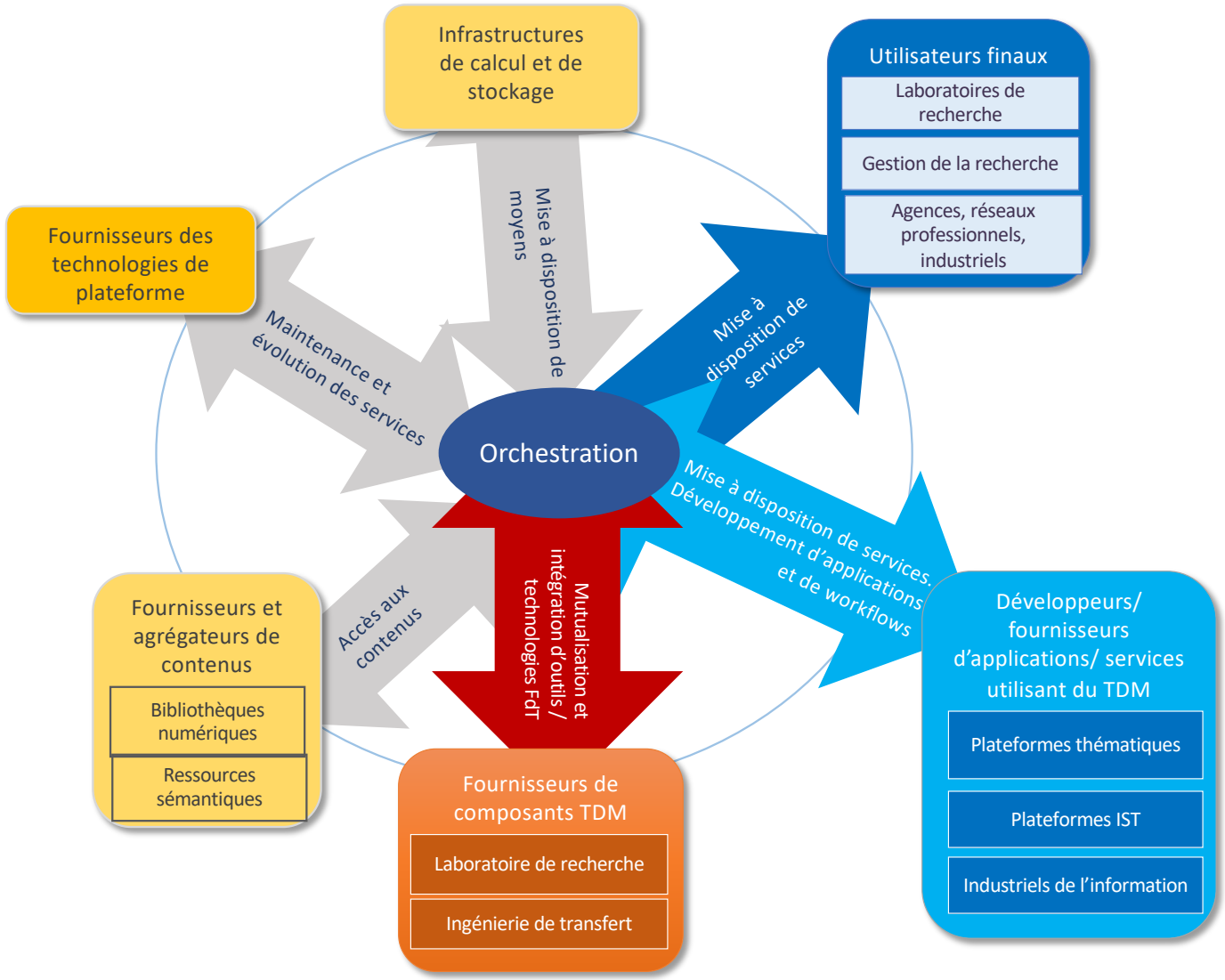
- Au travers de son réseau de compétences,
  elle assure le **lien entre la plateforme technique et ses utilisateurs**
- Elle facilite son **appropriation** et s'assure qu'elle **réponde aux besoins** des différentes communautés : utilisateurs finaux et intermédiaires, fournisseurs de composants et de contenu.
- Elle contribue à développer **une communauté de pratique**
  dans le domaine de l'utilisation de la fouille de texte.
- Elle promeut l'harmonisation des standards, représentations et pratiques, dans les **principes FAIR** auprès des différentes communautés concernées
  (données, publications scientifiques, ontologies, outils logiciels).

Enjeux des services innovants en Traitement Automatique de la Langue

- **Transfert** des produits de la recherche : composants de TAL et contenus (publications, ressources sémantiques).
- **Développement d'e-infrastructures** et de services coordonnés (fouille de texte, bibliothèques, ressources sémantiques, services métiers).
- Rendre les développeurs d'application **autonomes** dans l'exploitation d'outils de TAL
- Nouveaux **métiers**, nouvelles **compétences** : développeurs informatiques spécialisés, concepteurs d'applications, IST, ingénieurs de la connaissance, ... enjeu majeur.




Acteurs et organisation



Stratégie partagée : rationaliser – mutualiser les données

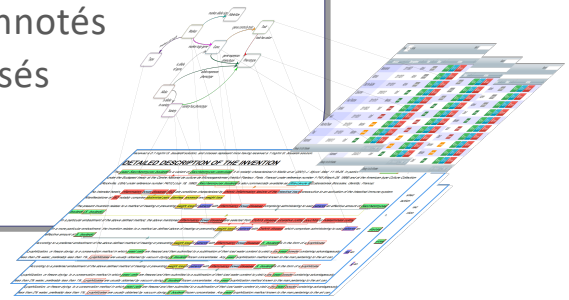
S'inscrire dans une stratégie de développement d'infrastructures et de service mutualisés et coordonnés

- **Rationaliser et mutualiser** l'accès aux **sources documentaires et sémantiques**.
- Réduire le coût d'ingénierie de la conception de corpus
- Rendre l'accès ouvert, transparent, adapté aux besoins et fiable (qualité, sécurité juridique, caractérisation)
- Intégrer les données, résultats de TAL avec les autres données

Comment

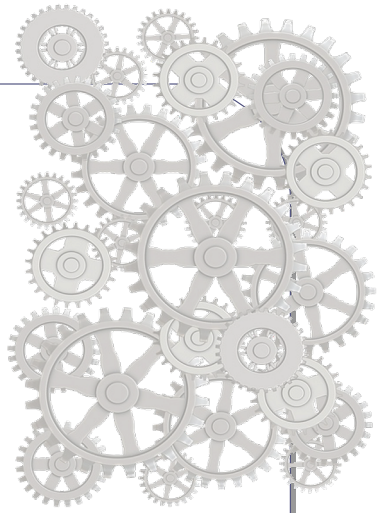
- Changer les **pratiques de publication**
- Rendre les publications accessibles et réutilisables dans des **formats standards**
- **Agréger**, centraliser, partager et réutiliser des **corpus** bruts, prétraités, annotés
- Développer et partager des **ressources et modèles sémantiques** spécialisés
- **Combiner et réconcilier** les données par des ontologies de référence

…



Stratégie partagée : rationaliser – mutualiser les traitements

S'inscrire dans une stratégie développement d'infrastructures et de services mutualisés et coordonnés, de données et de traitements



- **Réutiliser, combiner** les outils de TAL pour les adapter aux besoins
  (aujourd'hui, des milliers d'outils et des plateformes techniques).
- Créer des environnement favorables à **l'expérimentation et la reproduction**

Comment

- **Plateformes** de traitement et service
- **Interconnecter** durablement les sources de données, les traitements et les services
- **Calcul** haut débit
- **Interopérabilité** des composants de TDM
- Fournir aux utilisateurs sur leurs postes de travail des **modes d'interaction adaptés** à leurs profils
- Pour plus de finesse et d'**adaptation au besoin**, mieux intégrer l'apprentissage automatique et les ressources sémantiques. Gérer le compromis généricité-adaptation / coût-valeur ajoutée.

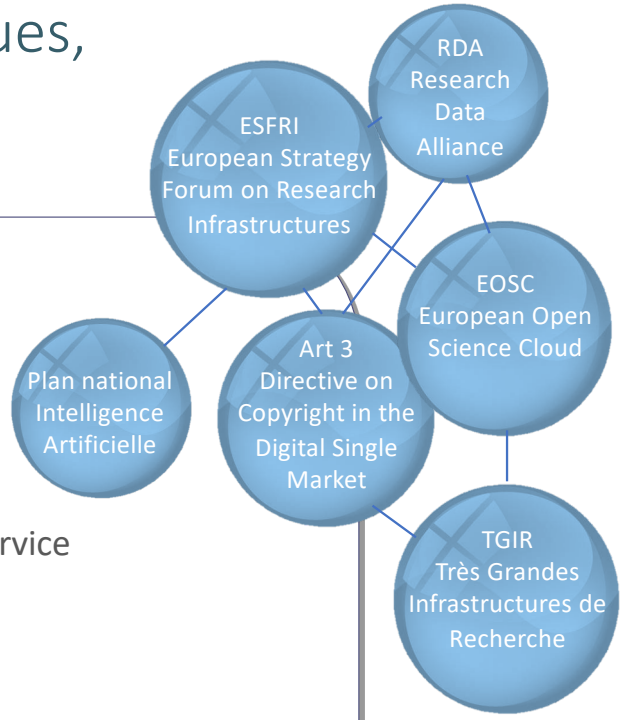
Concevoir une stratégie académique, par les académiques, pour les académiques ?

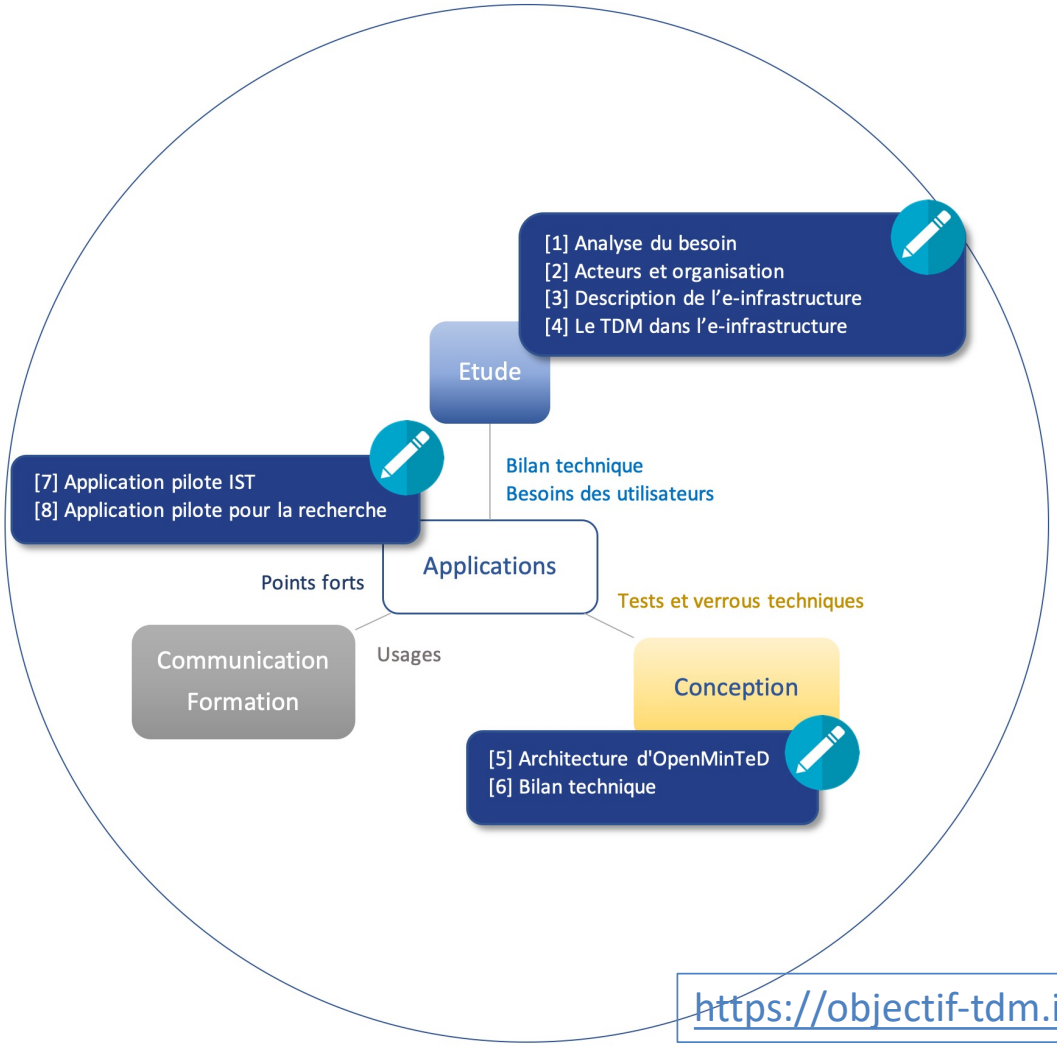
- Mutualisation et partage des résultats de la recherche en TAL pour la recherche académique : une **problématique collective**

- Des solutions de services de TAL académiques pour **des besoins non rentables** pour les entreprises du secteur : dans les domaines de production de la connaissance, non marchands
  - Coût de l'ingénierie documentaire, expertise du domaine, maintenance du service
  - Relevant des missions des organismes de recherche

Et parallèlement sont un levier d'innovation
Ainsi, l'accès aux publications
Le transfert des prototypes vers des produits

Solutions industrielles et académiques : complémentaires et supplémentaires.
Modèles mixtes à inventer
Services des PME innovantes exploitant des ressources libres





Rapports des projets OpenMinTeD, Future TDM et VisaTM

Mutualisation des services sur des données ouvertes
Pour des utilisateurs non spécialistes

https://project.futuretdm.eu/publications/

http://openminted.eu/deliverables/

https://objectif-tdm.inist.fr/2019/11/18/rapports-publics-du-projet-visa-tm/