INRAE

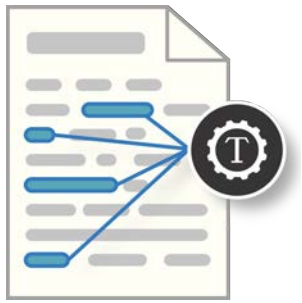
Extraction d'information à partir de textes

Robert Bossy & Claire Nédellec
MaIAGE, Jouy-en-Josas

ANF 2021, 17 novembre 2021

Des analyses numériques de documents complémentaires

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*

Extraction d’information et de connaissance - partage et réutilisation

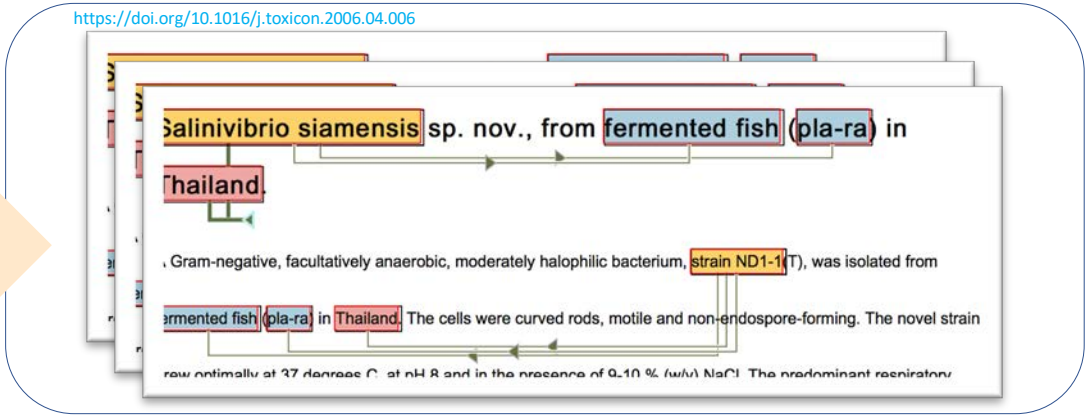
Notion de référentiel commun

Extraction automatique d'information : identifier, typer et relier les entités

Texte

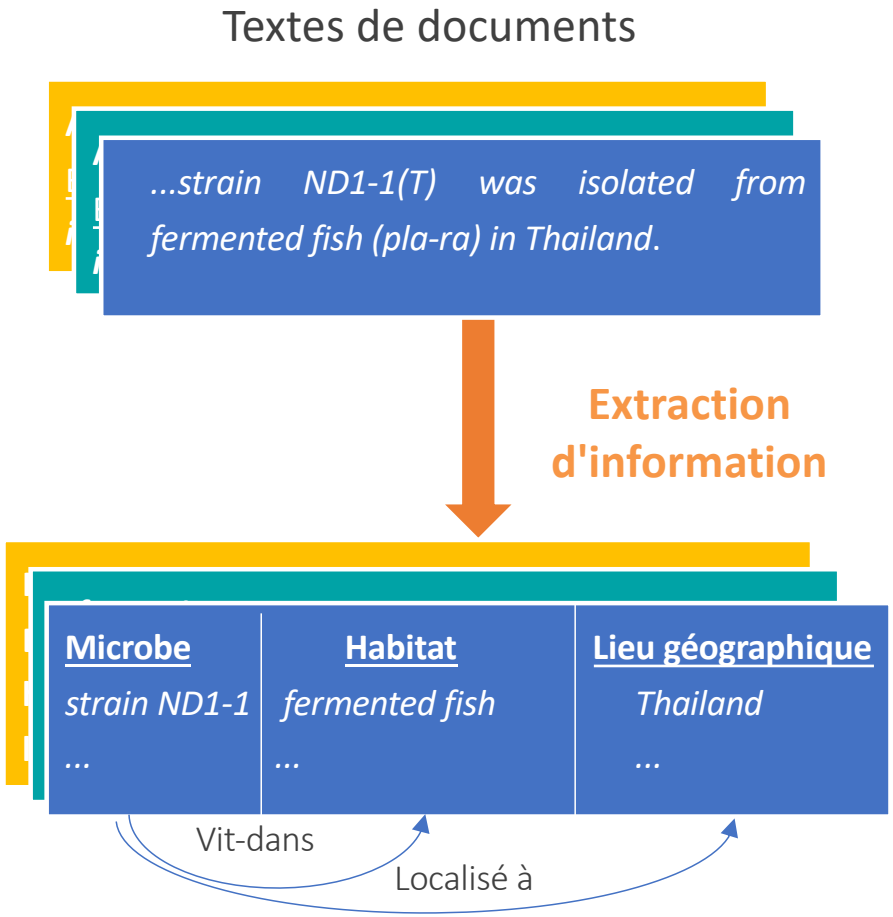
Reconnaissance des entités

Extractions des relations



Texte annoté par des entités et par des relations

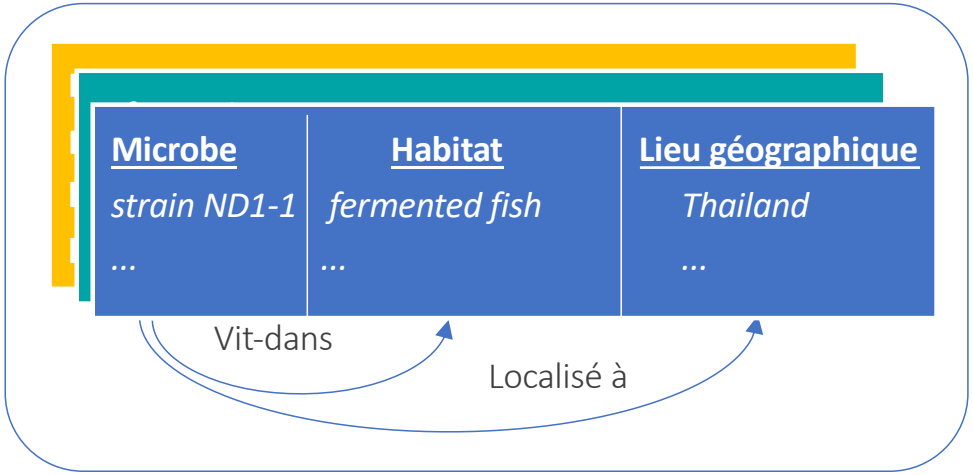
- Les entités de type microbe, habitat et lieu géographique
- Reliés par des relations orientées vit-dans, localisée-à



Extraction automatique d'information

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

Schéma d'annotation des informations



Les champs du formulaire correspondent au schéma des données

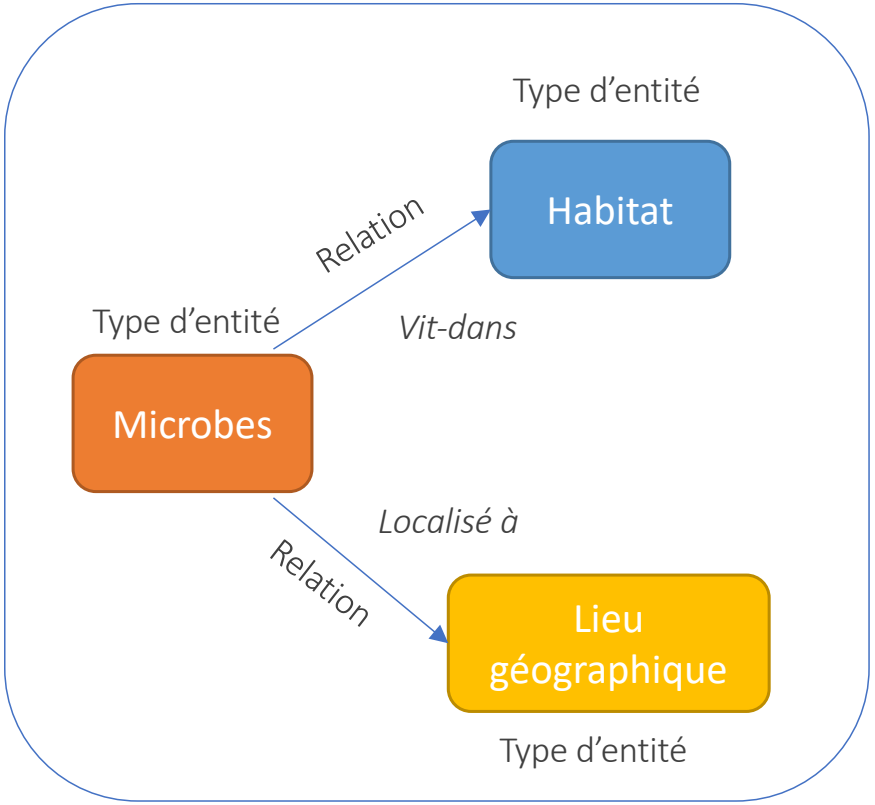
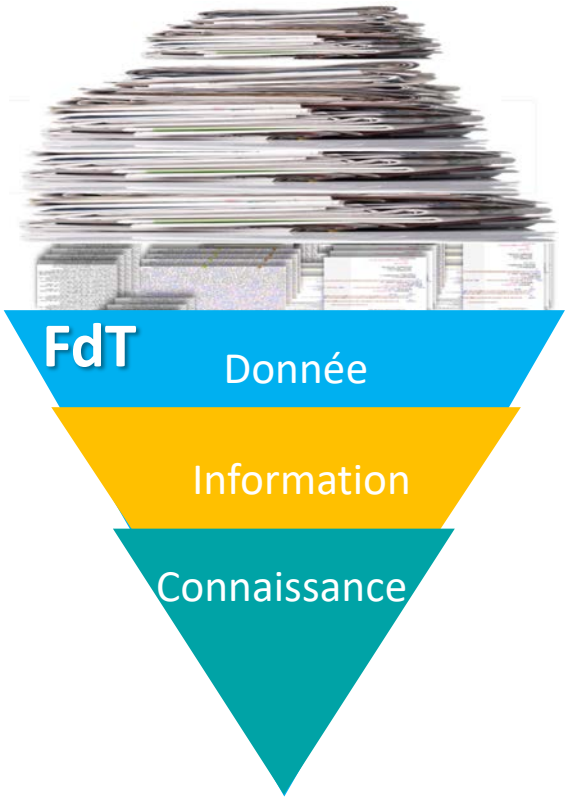


Schéma des données, entité - relation

- En général, des relations binaires entre deux entités
- Orientées
- Les arguments de la relation sont typés

Extraction automatique d'information et de connaissances
Partage et réutilisation



L'extraction d'information transforme

des données en information

Partager l'information nécessite

un référentiel de connaissance commun

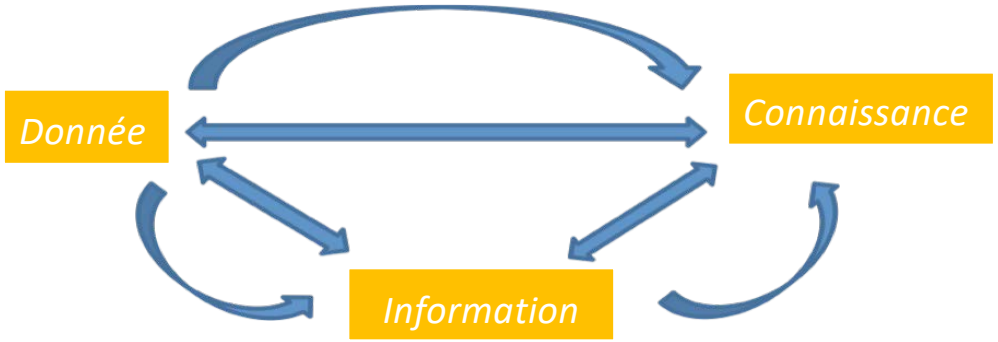
et de relier l'information aux connaissances connues

Données, informations, connaissances

Données. Eléments discrets, nombres, bases de données, texte

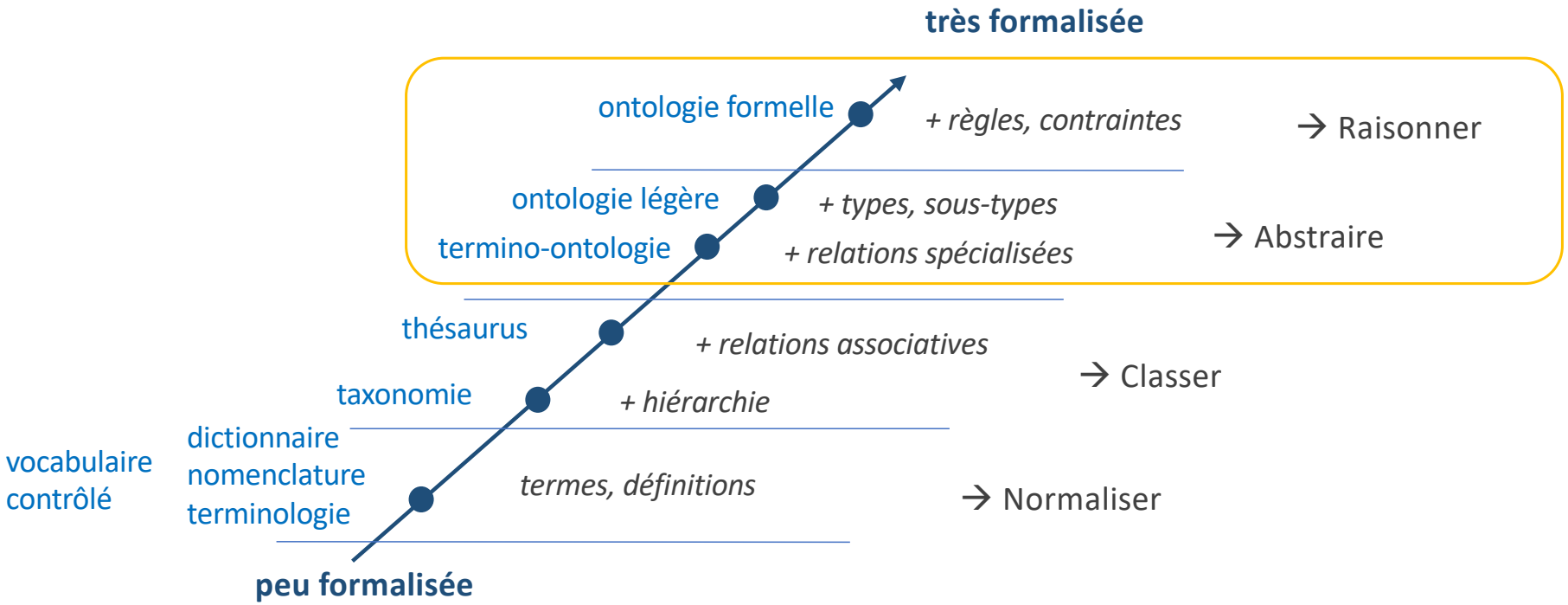
Informations. Données filtrées, contextualisées, interprétées et reliées localement à l'aide de la connaissance de façon à être réutilisables

Connaissances. Informations organisées, synthétisées et reliées à un modèle de connaissance : un système unifié de définitions formelles et de principes fondamentaux, pour la compréhension partagée et des traitements fiables (*sémantique formelle)*



Référentiels

D'après Sophie Aubin



Termino-ontologie

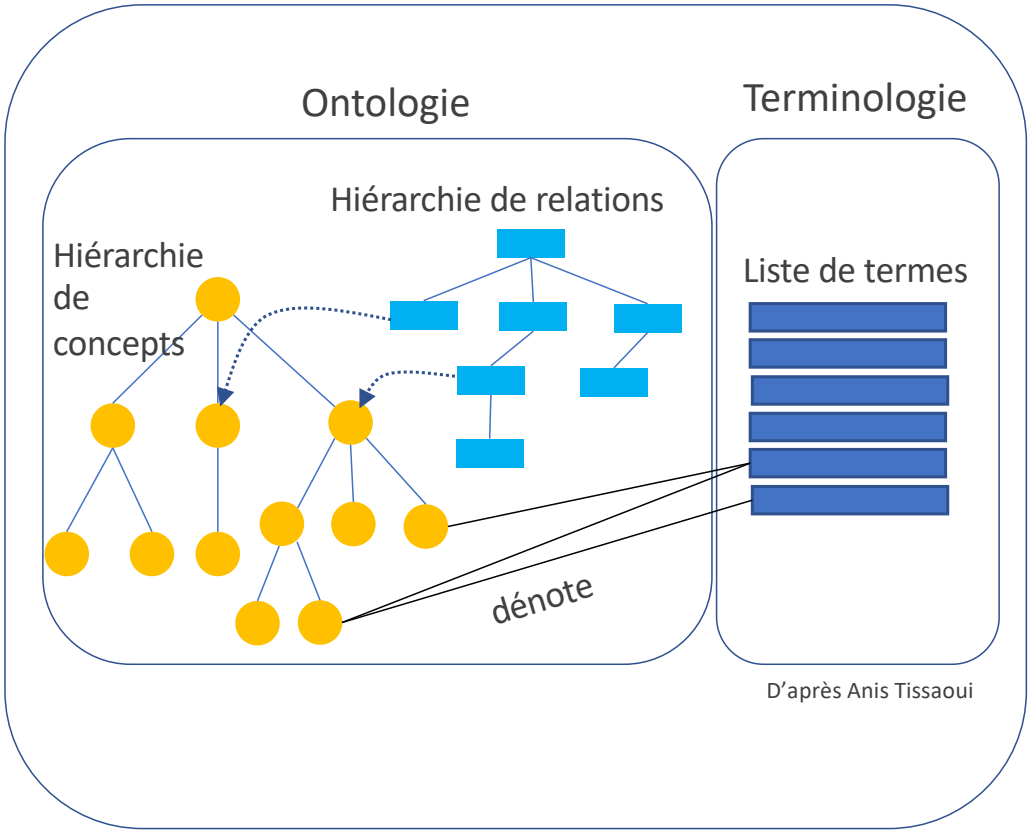
Chaque concept est

- Relié à ses **parents** (concepts plus généraux)
- Relié à ses **descendants**
  (concepts plus spécifiques)
- Relié à d'autres concepts par des **relations spécialisées**

Il est associé à

des **informations terminologiques**, label, synonyme

Éventuellement, les sens et usages du terme



D'après Anis Tissaoui

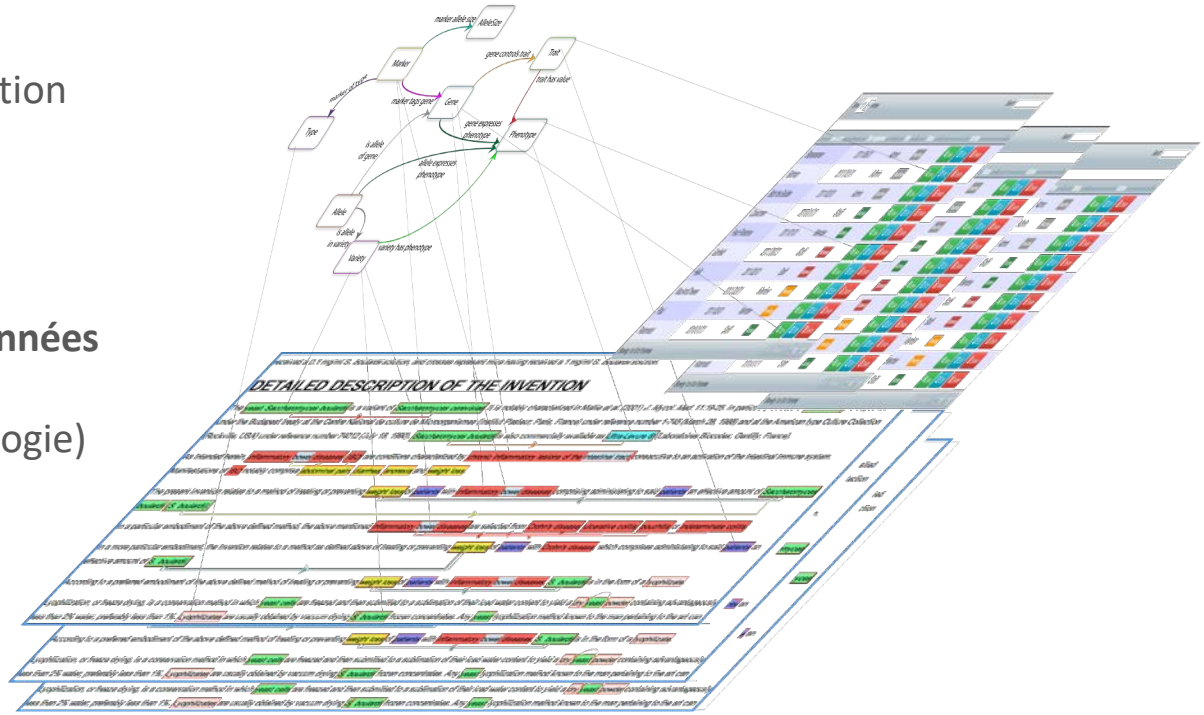
Extraction d'information *pour des bases de connaissance intégrées*

Objectif
Centraliser, structurer et standardiser l'information
Pour en faciliter l'accès et l'utilisation

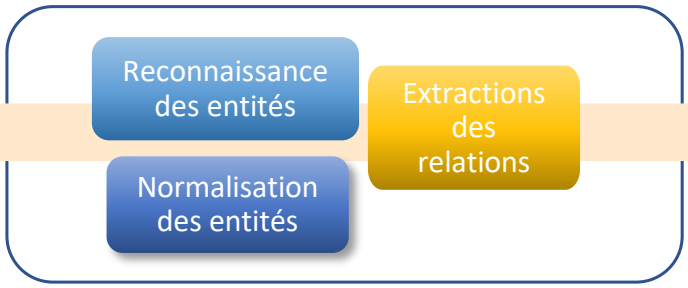
Données en **texte libre nombreuses**

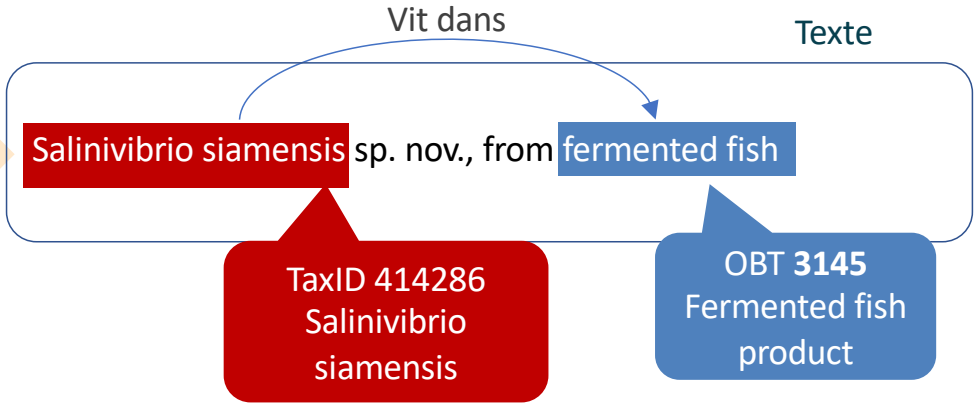
Les structurer pour les combiner à d'autres données
→ Détecter les informations
→ Les associer à une référence partagée (ontologie)
→ Les relier entre elles
→ Les mettre à disposition

Par des méthodes de *text mining*

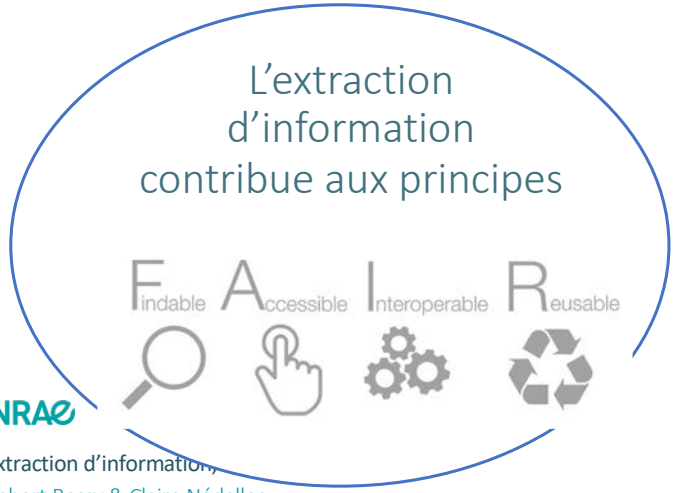


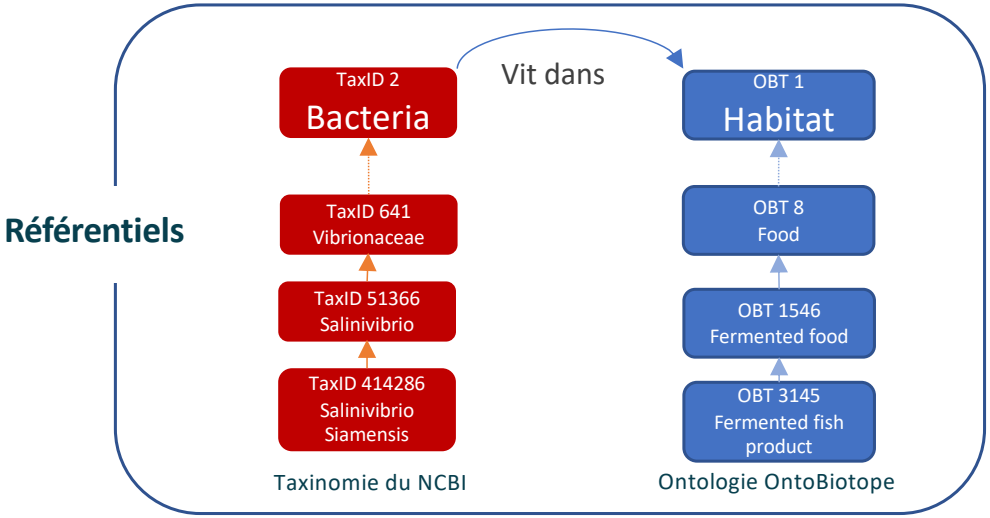
Formaliser l'association entre données du texte et référentiels





Extraction automatique d'information basée sur des ontologies





[Nédellec et al., EGC 2019]
AgroPortal ONTOBIOTOPE

Méta données des documents

Des entités emboitées, discontinues

PMID- 19329624
TI - Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand.
LID - 10.1099/ijs.0.001768-0 [doi]
AB - A Gram-negative, facultatively anaerobic, moderately halophilic bacterium, strain ND1-1(T), was isolated from fermented fish (pla-ra) in Thailand. The cells were curved rods, motile and non-endospore-forming. The novel strain grew optimally at 37 degrees C, at pH 8 and in the presence of 9-10 % (w/v) NaCl. [..]
Comparative 16S rRNA gene sequence analyses indicated that strain ND1-1(T) was closely related to Salinivibrio costicola, which comprises three subspecies, and Salinivibrio proteolyticus with gene sequence similarities of 98.3-98.6 %. [..]
On the basis of the physiological and biochemical characteristics and the molecular data presented, strain ND1-1(T) should be classified as a novel species of the genus Salinivibrio for which the name Salinivibrio siamensis sp. nov. is proposed. The type strain is ND1-1(T) (=JCM 14472(T)=PCU 301(T)=TISTR 1810(T)).

FAU - Chamroensaksri, Nitcha
AD - Department of Microbiology, Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok 10330, Thailand.
FAU - Tanasupawat, Somboon
FAU - Akaracharanya, Ancharida
FAU - Visessanguan, Wonnop
FAU - Kudo, Takuji
FAU - Itoh, Takashi
LA - eng
PT - Journal Article
TA - Int J Syst Evol Microbiol
JT - International journal of systematic and evolutionary microbiology
MH - Fishes/*microbiology
MH - *Food Microbiology
MH - Phylogeny
MH - RNA, Ribosomal, 16S/genetics
MH - Sequence Analysis, DNA
MH - Temperature
MH - Thailand
SO - Int J Syst Evol Microbiol. 2009 Apr;59(Pt 4):880-5. doi: 10.1099/ijs.0.001768-0.

Index MeSH thématique global au document

Métadonnées locales des informations extraites du document

Entités de type Habitat

Id	Forme	Position	Longueur
1	Fermented fish	73	12
2	pla-ra	87	6
...			

Entités de type Taxon

3	ND1-1

Entités de type Géo

4	Thailand

Relation

Type	Argument 1	Argument 2
Located at	3	4

Types et classes d'entités en extraction d'information

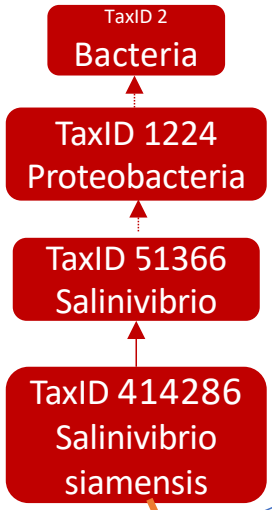
1. La reconnaissance d'entité nommée associe un type général aux entités reconnues
   Date, lieu, personne
   Microbe, habitat, maladie, symptôme, organe
   Entreprise, fonction

2. La normalisation associe aux entités un nom standard provenant d'un référentiel
   Microbe: nom scientifique
   Molécule : nom de la molécule dans le référentiel (ex CheBI)

3. L'annotation sémantique associe aux entités *une catégorie d'un référentiel, un concept précis de l'ontologie*
   Microbe: taxon dans la taxinomie (ex NCBI)
   Maladie : nom de la maladie dans la classification (ex CIM-10)

2. Normalisation et référentiel : associer un nom standard aux formes « figées »

Extrait de la taxinomie du NCBI



Extrait de Geonames

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

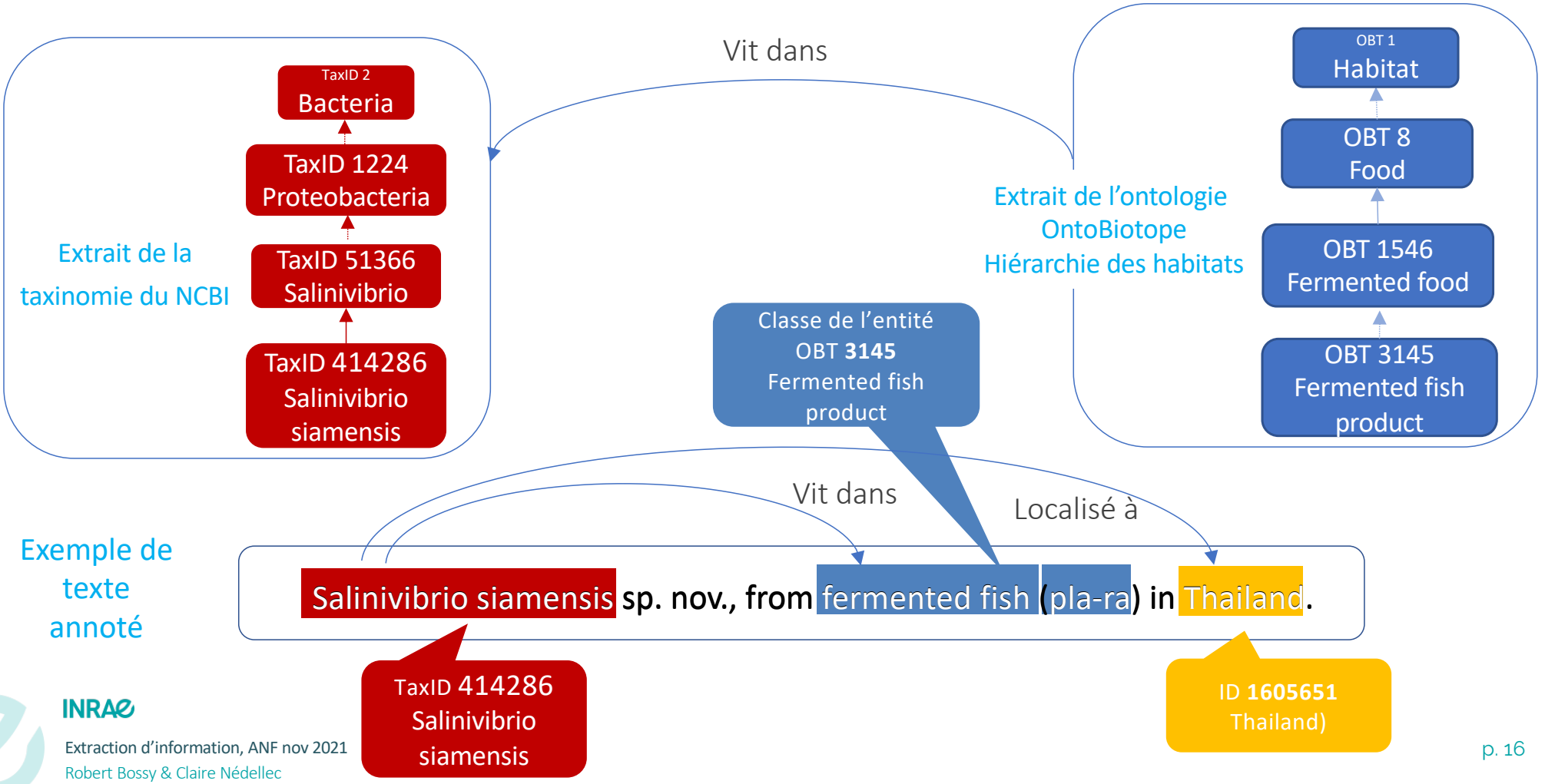
Exemple de texte annoté

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

TaxID 414286
Salinivibrio siamensis

ID **1605651**
Thailand)

L'annotation sémantique associe aux entités du texte *une catégorie précise et des relations du référentiel*



Différents référentiels pour différents types d'entités et de relations

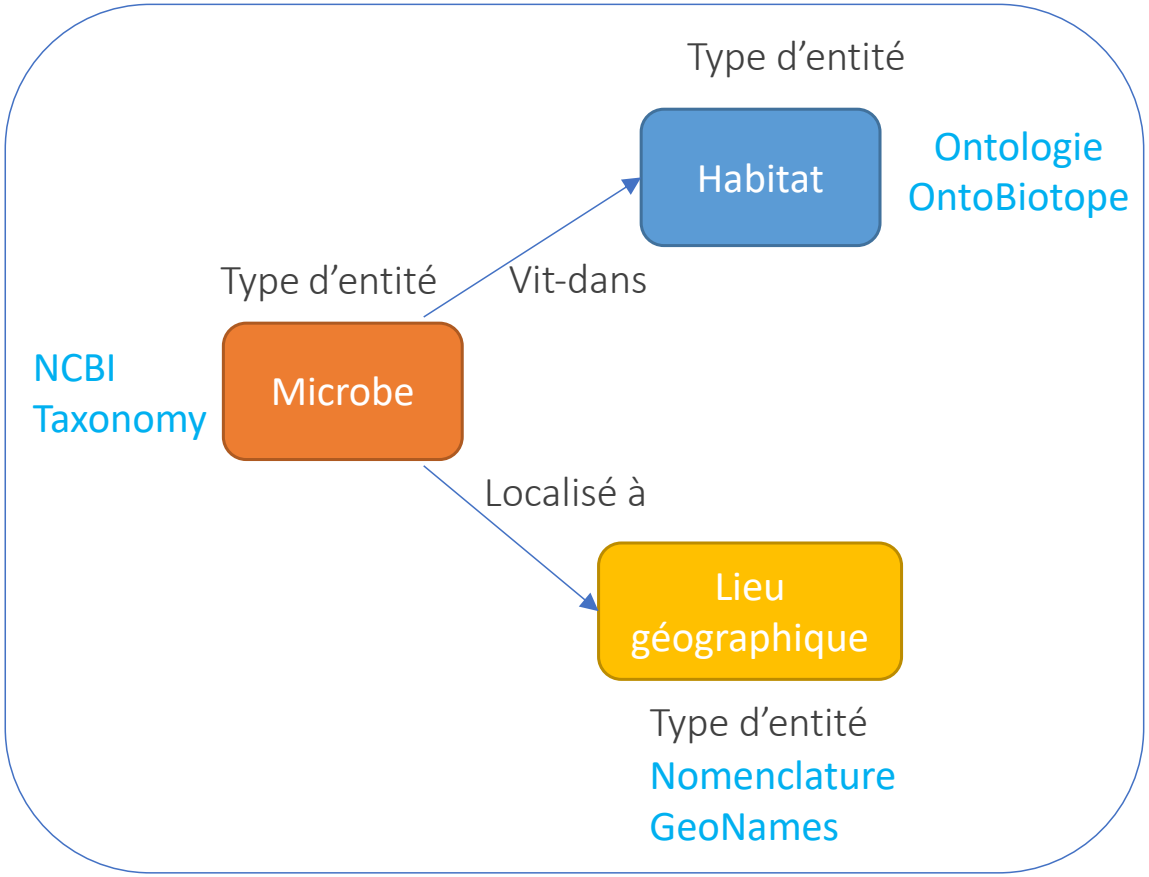
Le schéma d'annotation des données du texte peut faire référence à des référentiels **distincts** :

nomenclatures, dictionnaires, thesaurus, ontologies

Dans **l'exemple**

Une taxinomie : *NCBI Taxonomy*
Une nomenclature : *GeoNames*
Une ontologie : *OntoBiotope*
Des relations externes

Différents types de référentiels pour différents niveaux de standardisation et d'expressivité



Exemple de schéma d'annotation entité – relation et référentiels associés

Méta données des documents

Des entités emboitées, discontinues

PMID- 19329624
TI - Salinivibrio siamensis sp. nov., from fermented fish (pla-ra) in Thailand.
LID - 10.1099/ijs.0.001768-0 [doi]
AB - A Gram-negative, facultatively anaerobic, moderately halophilic bacterium, strain ND1-1(T), was isolated from fermented fish (pla-ra) in Thailand. The cells were curved rods, motile and non-endospore-forming. The novel strain grew optimally at 37 degrees C, at pH 8 and in the presence of 9-10 % (w/v) NaCl. [..]
Comparative 16S rRNA gene sequence analyses indicated that strain ND1-1(T) was closely related to Salinivibrio costicola, which comprises three subspecies, and Salinivibrio proteolyticus with gene sequence similarities of 98.3-98.6 %. [..]
On the basis of the physiological and biochemical characteristics and the molecular data presented, strain ND1-1(T) should be classified as a novel species of the genus Salinivibrio for which the name Salinivibrio siamensis sp. nov. is proposed. The type strain is ND1-1(T) (=JCM 14472(T)=PCU 301(T)=TISTR 1810(T)).

Index MeSH thématique global au document

FAU - Chamroensaksri, Nitcha
AD - Department of Microbiology, Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok 10330, Thailand.
FAU - Tanasupawat, Somboon
FAU - Akaracharanya, Ancharida
FAU - Visessanguan, Wonnop
FAU - Kudo, Takuji
FAU - Itoh, Takashi
LA - eng
PT - Journal Article
TA - Int J Syst Evol Microbiol
JT - International journal of systematic and evolutionary microbiology
MH - Fishes/*microbiology
MH - *Food Microbiology
MH - Phylogeny
MH - RNA, Ribosomal, 16S/genetics
MH - Sequence Analysis, DNA
MH - Temperature
MH - Thailand
SO - Int J Syst Evol Microbiol. 2009 Apr;59(Pt 4):880-5. doi: 10.1099/ijs.0.001768-0.

Métadonnées locales des informations extraites du document

Entités de type Habitat

Id	Classe	Forme
1	OBT:Fermented fish product	Fermented fish product
2	OBT:Fermented fish product	pla-ra
...		

Entités de type Taxon

3	NCBI tax :ND1-1	ND1-1

Entités de type Géo

4	Geoenames:Thailand	Thailand

Relation

Type	Argument 1	Argument 2
Located at	3	4

Que faire en recherche documentaire en fonction du référentiel

1. La **reconnaissance d'entité nommée** associe un type général aux entités reconnues

Retrouver tous les documents qui mentionnent un type d'entité

Produire la liste des entités dans un ensemble de documents

Ex. Trouver tous les documents qui mentionnent une maladie,

Ex. Identifier toutes les expressions du texte qui dénotent des maladies

2. **La normalisation** associe aux entités un nom standard provenant d'un référentiel

Retrouver précisément les données ou documents contenant une entité étant donné son nom,

Indépendamment des variations

Ex. Trouver tous les documents qui mentionnent la bactérie *Salinivibrio siamensis*

3. L'annotation sémantique associe aux entités *une catégorie précise de l'ontologie*

Retrouver des données ou documents contenant une entité, à différents niveaux de généralité

Sans connaître son nom

Ex. Trouver tous les documents qui mentionnent des types d'habitat, par exemple, des aliments

Aperçu des méthodes d’extraction d’information

Projection de lexiques

Rechercher dans le texte des fragments identiques à des entrées dans une liste préétablie de formes connues.

- Exploitation de ressources existantes : *gazetteers*, glossaires, annuaires, terminologies, ontologies, etc.
- Approche simple pour la reconnaissance d'entités, voire l'annotation sémantique si les étiquettes sont associées à un identifiant.

Limites

- Silence sur les entités hors-nomenclature.
- Peu robuste aux variations typographiques ou aux variations de langage.
- Entrées ambigües.

Capture des variations

Saturation automatique des lexiques pour les abréviations.

Escherichia coli	→	E. coli
Klebsiella pneumoniae	→	K. pneumoniae
Pseudomonas aeruginosa	→	P. aeruginosa
Streptococcus aureus	→	S. aureus
Listeria monocytogenes	→	L. monocytogenes

Recovery of Gram-Negative Bacteria from Aerobic Blood Culture Bottles Containing Antibiotic Binding Resins after Exposure to β-Lactam and Fluoroquinolone Concentrations.

1.4142135

Authors: *Iris H Chen David P Nicolau Joseph L Kuti*

2019 ***Journal of clinical microbiology***

Abstract Blood culture bottles containing antibiotic binding resins are routinely used to minimize artificial sterilization in the presence of antibiotics. However, the resin binding kinetics can differ between antibiotics and concentrations. This study assessed the impact of clinically meaningful peak, midpoint, and trough concentrations of meropenem, imipenem, cefepime, cefazolin, levofloxacin, and piperacillin-tazobactam on the recovery of Pseudomonas aeruginosa , Escherichia coli , and Klebsiella pneumoniae from resin-containing BacT/Alert FA Plus and Bactec Aerobic/F blood culture bottles. P. aeruginosa -inoculated bottles alarmed positive in 4/20 (20%), 16/20 (80%), and 20/20 (100%) of those with peak, midpoint, and trough concentrations of antipseudomonal agents, respectively ($P \leq 0.001$). E. coli was recovered from 8/24 (33%), 11/24 (46%), and 14/24 (58%) of bottles with peak, midpoint, and trough concentrations, respectively ($P = 0.221$). K. pneumoniae was recovered from 8/16 (50%) at all concentrations of the studied antibiotics ($P = 1.0$). BacT/Alert and Bactec bottles inoculated with antibiotics and P. aeruginosa had similar times to detection (TTD) ($P = 0.352$); however, antibiotic-containing BacT/Alert bottles had a shorter TTD compared with antibiotic-containing Bactec bottles for E. coli ($P = 0.026$) and K. pneumoniae ($P \leq 0.001$). Pathogen recovery in BacT/Alert FA Plus and Bactec Aerobic/F blood culture bottles containing antibiotic binding resins was greatly reduced in the presence of antibiotics, especially at higher concentrations. These data support the practice of drawing blood cultures immediately before an antibiotic dose to maximize the chances of pathogen recovery.

Lemmatisation pour les flexions.

plants	→	plant
plant	→	plant
Plant	→	plant
PLANT	→	plant

Comparison of microbial and transient expression (tobacco plants and plant-cell packs) for the production and purification of the anticancer mistletoe lectin viscumin.

1.4142135

Authors: *Benjamin B Gengenbach Linda L Keil Patrick Opdensteinen Catherine R Müschen Georg Melmer Hans Lentzen Jens Bührmann Johannes F Buyel*

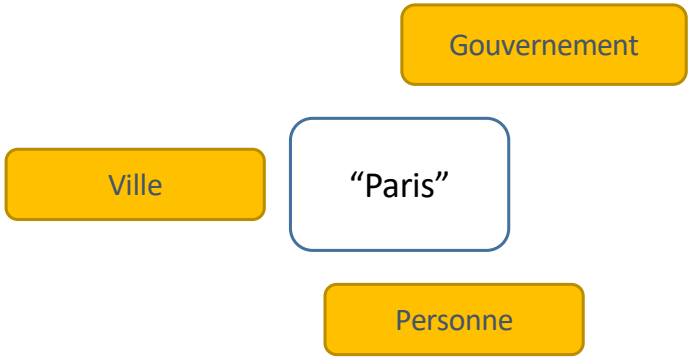
2019 ***Biotechnology and bioengineering***

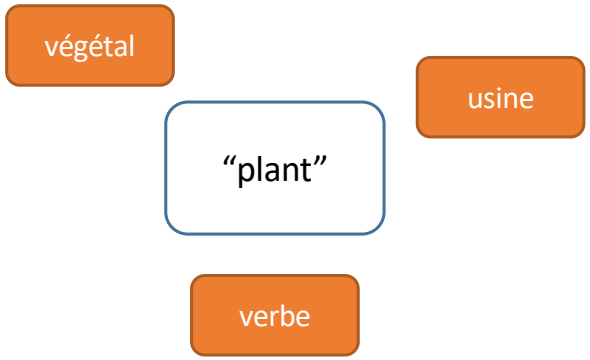
Abstract Cancer is the leading cause of death in industrialized countries. Cancer therapy often involves monoclonal antibodies or small-molecule drugs, but carbohydrate-binding lectins such as mistletoe (Viscum album) viscumin offer a potential alternative treatment strategy. Viscumin is toxic in mammalian cells, ruling them out as an efficient production system, and it forms inclusion bodies in Escherichia coli such that purification requires complex and lengthy refolding steps. We therefore investigated the transient expression of viscumin in intact Nicotiana benthamiana plants and Nicotiana tabacum Bright Yellow 2 plant-cell packs (PCPs), comparing a full-length viscumin gene construct to separate constructs for the A and B chains. As determined by capillary electrophoresis the maximum yield of purified heterodimeric viscumin in N. benthamiana was ~7 mg/kg fresh biomass with the full-length construct. The yield was about 50% higher in PCPs but reduced 10-fold when coexpressing A and B chains as individual polypeptides. Using a single-step lactosyl-Sepharose affinity resin, we purified viscumin to ~54%. The absence of refolding steps resulted in estimated cost savings of more than 80% when transient expression in tobacco was compared with E. coli. Furthermore, the plant-derived product was ~3-fold more toxic than the bacterially produced counterpart. We conclude that plants offer a suitable alternative for the production of complex biopharmaceutical proteins that are toxic to mammalian cells and that form inclusion bodies in bacteria.

Formes ambigües

Sources d'ambiguïté

- homonymie
- métonymie
- polysémie





Désambiguïsation par le contexte

Identifier des mots dans le contexte qui permettront de discriminer entre plusieurs sens.

Par exemple : *plant/factory* et *plant/organisme* ?

new strains can be rapidly implemented into existing infrastructures such as **bioethanol production**	**plants**	
Evaluation of methanogenic activity of **biogas**	**plant**	slurry
two reference points (indoor and outdoor) that are assumed not to be **contaminated** by the	**plant**	's activities
a phage can survive in a **cheese**	**plant**	for more than a year
The first large-scale **biogas**	**plant**	was put into operation

a problem that is present in the transformation of all	**plant**	**species**
The tendency of closely related	**plant**	**species** to share natural enemies
Rhizosphere is the complex place of numerous interactions between	**plant**	**roots**, microbes and soil fauna
an appropriate method of transgene introduction into a	**plant**	**cell**
common mechanisms for	**plant**	**cell** reprogramming during endosymbiosis

Exploitation du contexte par des règles d'extraction

Il existe des outils qui permettent d'exprimer des règles de désambiguïsation selon les mots du contexte.

("biogas" ou <nourriture>) "plant" → étiqueter "plant" comme une usine.

"plant" ("species" ou "cell") → étiqueter "plant" comme une plante.

Les mêmes types de règles peuvent servir aussi à extraire des relations.

<microorganism> "isolated from" <NOM>

<microorganism> "isolated from" <DET> <ADJ> <NOM>

... mycobacteria **isolated from** seals ...

... *Mycobacterium malmoense* **isolated from** soil ...

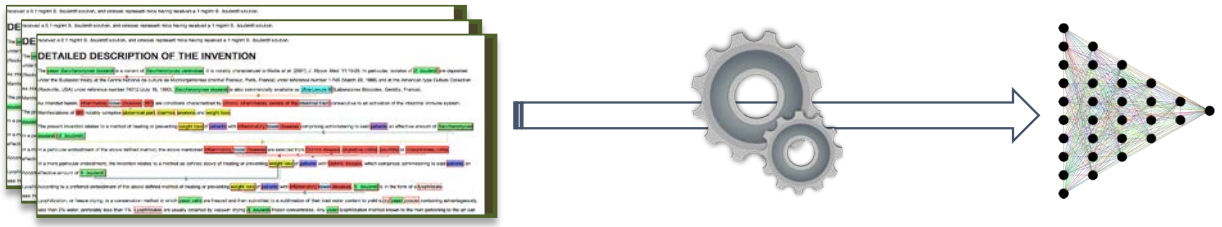
... *Nocardia africana* **isolated from** a feline mycetoma ...

Limites

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

Apprentissage supervisé

Les algorithmes d'apprentissage supervisé induisent des règles à partir d'exemples étiquetés.



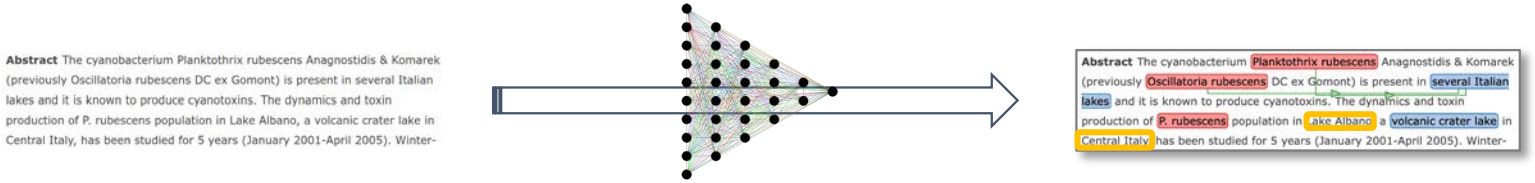
Corpus d'apprentissage annoté **Algorithme d'apprentissage** **Modèle**

1. Phase d'apprentissage

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

Apprentissage supervisé

Les algorithmes d'apprentissage supervisé induisent des règles à partir d'exemples étiquetés.



Corpus non-annoté | **Modèle** | **Corpus annoté automatiquement**

Phase de production

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

À savoir

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

Exercice d’annotation

Création d'un corpus d'apprentissage

"Gold Standard"

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

Valorisation

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

Création d'un corpus d'apprentissage : méthodologie

Intérêts de suivre une méthodologie d'annotation

- Garantir la conformité de l'annotation aux besoins.
- Contrôler l'homogénéité de l'annotation.
- Assurer la reproductibilité du "Gold Standard".

Éléments de

- Outils d'annotations spécialisés.
- Partage d'un document de spécifications, exemples et arbitrages : "**guide d'annotation**" ("annotation guidelines").
- **Double annotation** de chaque document, éventuellement en double-aveugle.
- **Accord Inter-Annotateur** : mesure de la variabilité de l'annotation.

Conclusion

Les différentes méthodes ont différentes qualités.

Projection de lexiques

Application de règles

Apprentissage supervisé

Conclusion

Les différentes méthodes ont différentes qualités.

Conclusion

Les différentes méthodes ont différents rôles.

Point de référence

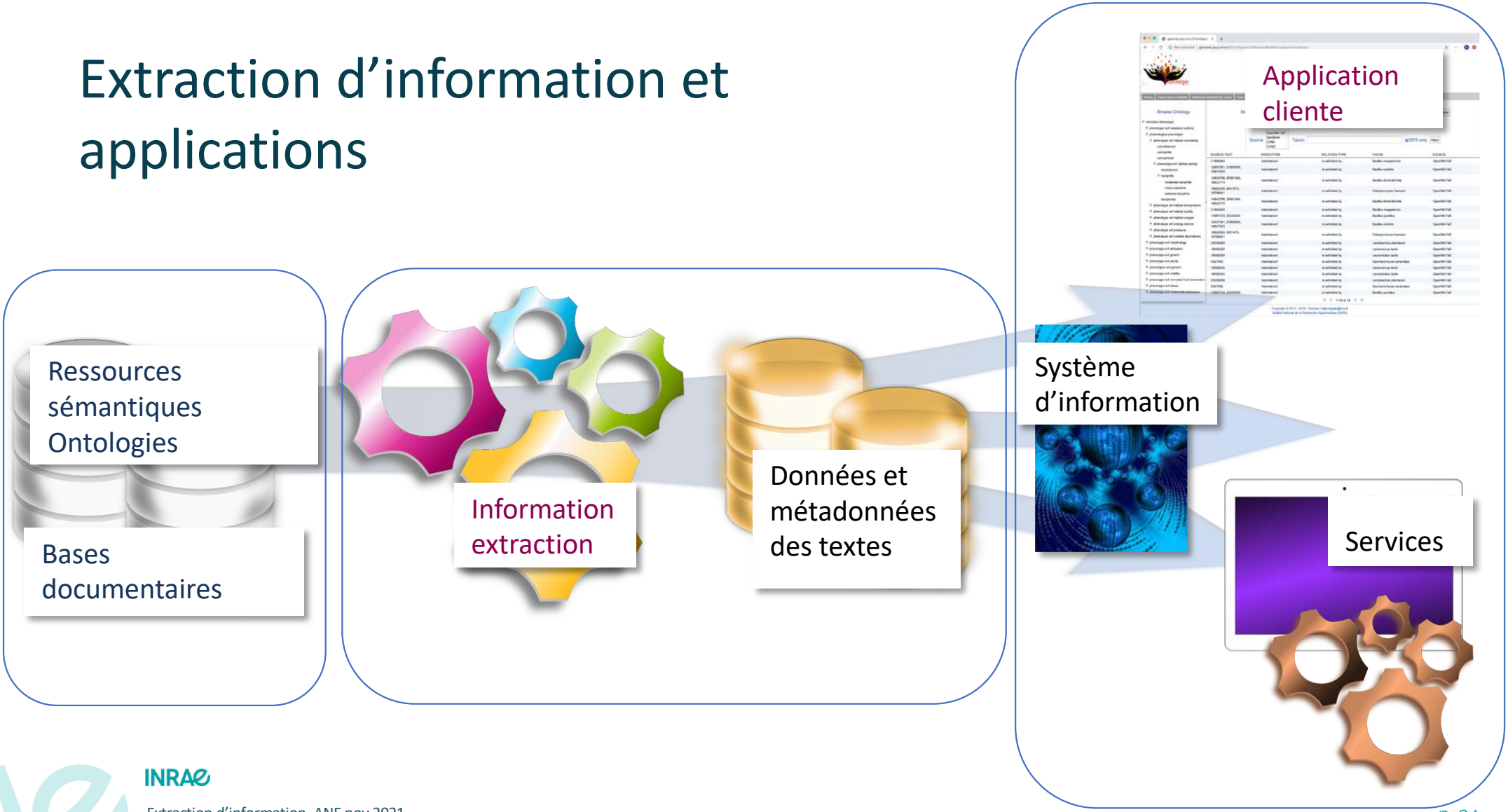
Correction d'erreurs
Production

Usage de plus en plus généralisé

Ressources souvent disponibles.

Modèles pré-entraînés quelquefois disponibles et applicables.

Extraction d'information et applications



Exemples d'applications

Recherche d'information

- Recherche sémantique de documents
- Intégration de données de sources diverses : bibliographie, données expérimentales, données de référence, etc.

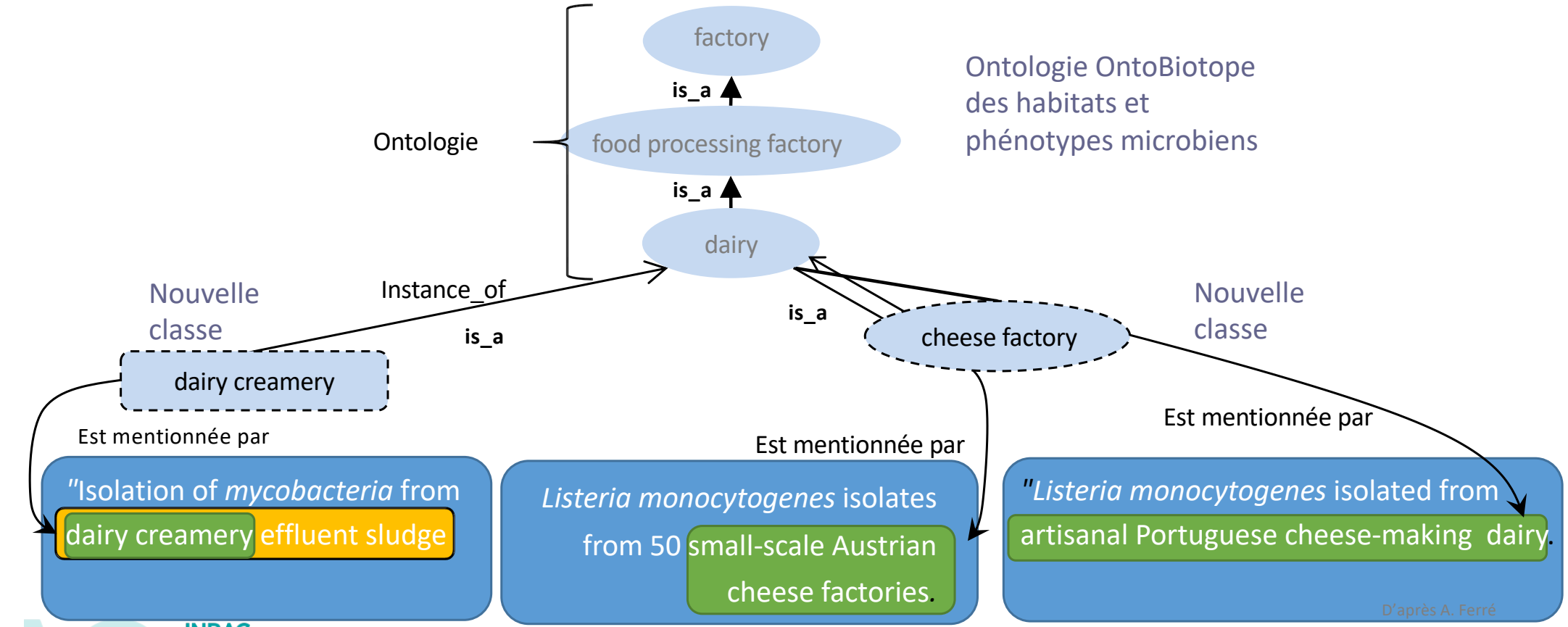
Aide à la décision

- Cartographie thématique (synthèse de corpus, tendances)
- Profilage d'experts (par ex. sélection de relecteurs)

Extension de référentiel

- Recherche de nouveaux synonymes ou concepts

Enrichir une ontologie par de nouvelles classes



Un moteur de recherche sémantique, AlvisIR

Exercice de recherche d’information

INRAE