



Méthodes et Outils pour l'Open Data
Le texte, une source de connaissance

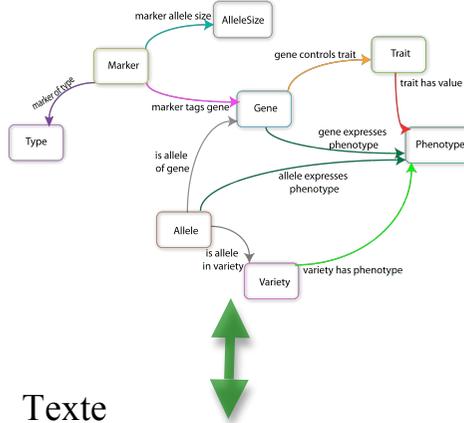


Montpellier – 18 décembre 2014

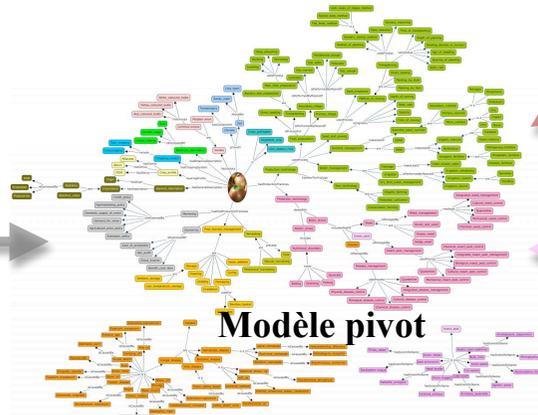
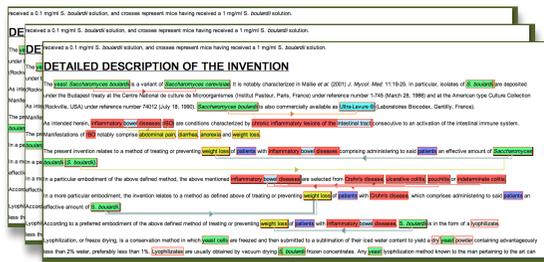
Modélisation de connaissances à partir de sources hétérogènes

Définir les rôles des modèles, leur représentation et leur articulation
 Caractériser les sources de connaissance (quoi, avec quelle qualité)
 Identifier les redondances et les complémentarités

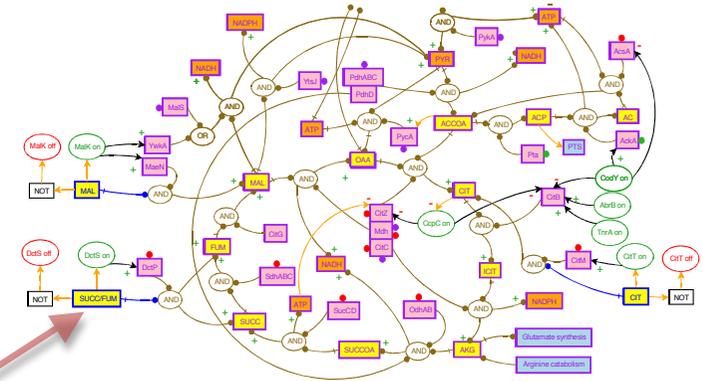
Modèle de connaissances du texte



Texte



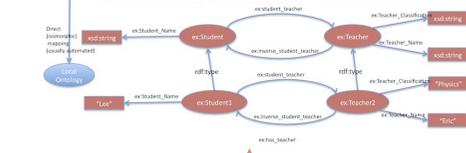
Modèle dynamique



Modèle de données

Option 1: Direct Mapping (no domain ontology involved)

StudentID	Name	StudentID	TeacherID	TeacherID	Name	Classification
1	Lee	1	2	1	Michael	Math
2	Juan	2	1	2	Eric	Physics



Données

Country	Unique Audience (000)	Time per Person (hh:mm:ss)
United States	142,052	6:09:13
Japan	46,556	2:50:21
Brazil	31,345	4:33:10
United Kingdom	29,129	6:07:54
Germany	28,057	4:11:45
France	26,786	4:04:39
Spain	19,456	5:30:55
Italy	18,256	6:00:07
Australia	9,895	6:52:28
Switzerland	2,451	3:54:34

Source: The Nielsen Company

Des modèles de connaissance associés aux données

Identification, Normalisation, indexation

Archiver, retrouver des données

Compréhension locale

Interpréter : replacer la donnée dans son contexte de production, rechercher des données similaires

Compréhension globale

Raisonner

Déduire de nouvelles connaissances

Identifier des contradictions

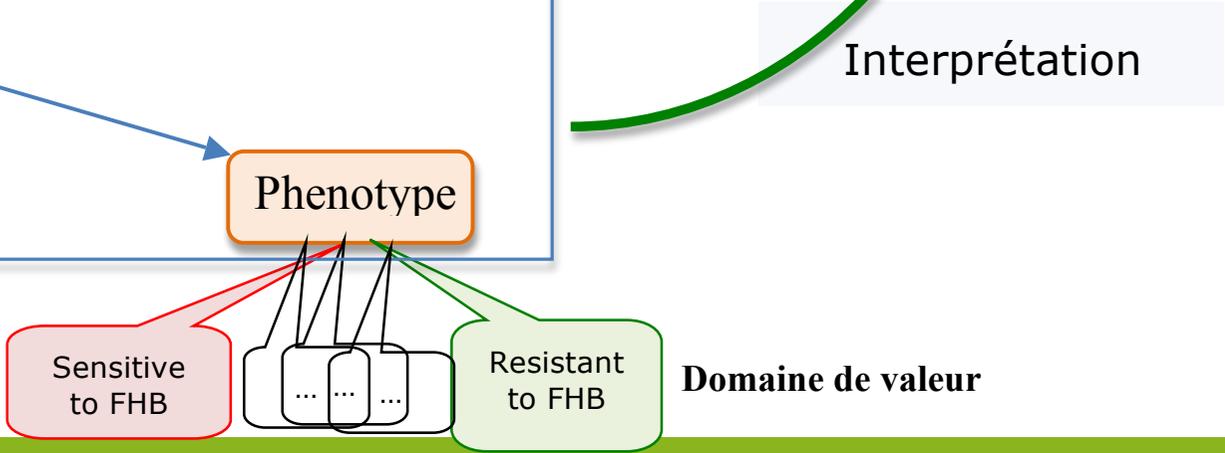
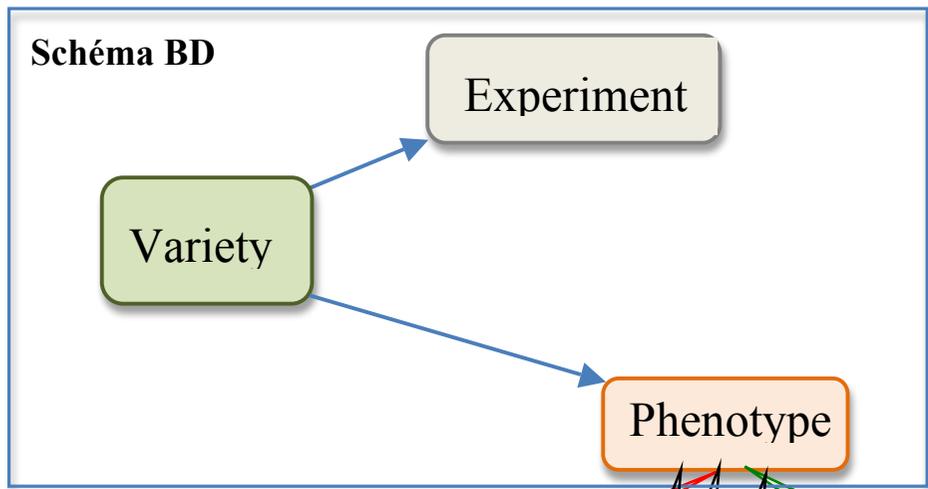
Prédire, simuler.

Produire des hypothèses

Grâce à diverses sortes de métadatas et de références

Archiver, retrouver des données

VARIETY	EXPERIMENT	PHENOTYPE
Apache	104100	Resistance to <i>Fusarium spp.</i>
Expert	104100	Sensitive to <i>Fusarium oxysporum</i>
Concerto	104100	High resistance to <i>Fusarium graminearum</i>
Chinese Spring	104100	FHB resistant



Retrouver des données similaires

Tous les *vibrio* isolés dans des « fermes aquacoles » ?

Requête

Documents réponses

Microorganisms

facet value	freq.	doc.
Vibrio vulnificus	46	18
Bacteria	29	15
Vibrio	14	11
Vibrio anguillarum	17	9
Vibrio alginolyticus	18	7
Vibrio harveyi	7	5
Vibrio splendidus	4	3
Vibrio parahaemoly	8	3
Vibrionaceae	2	2
Vibrio scophthalmi	9	2

Habitats

facet value	freq.	doc.

1
Occurrence of *Vibrio vulnificus* in mussel farms from the Varano lagoon environment.
 2010 *Letters in applied microbiology*
Abstract Monitoring the occurrence of the human pathogen *Vibrio vulnificus* in a mussel farm located in the lagoon of Varano (Italy).

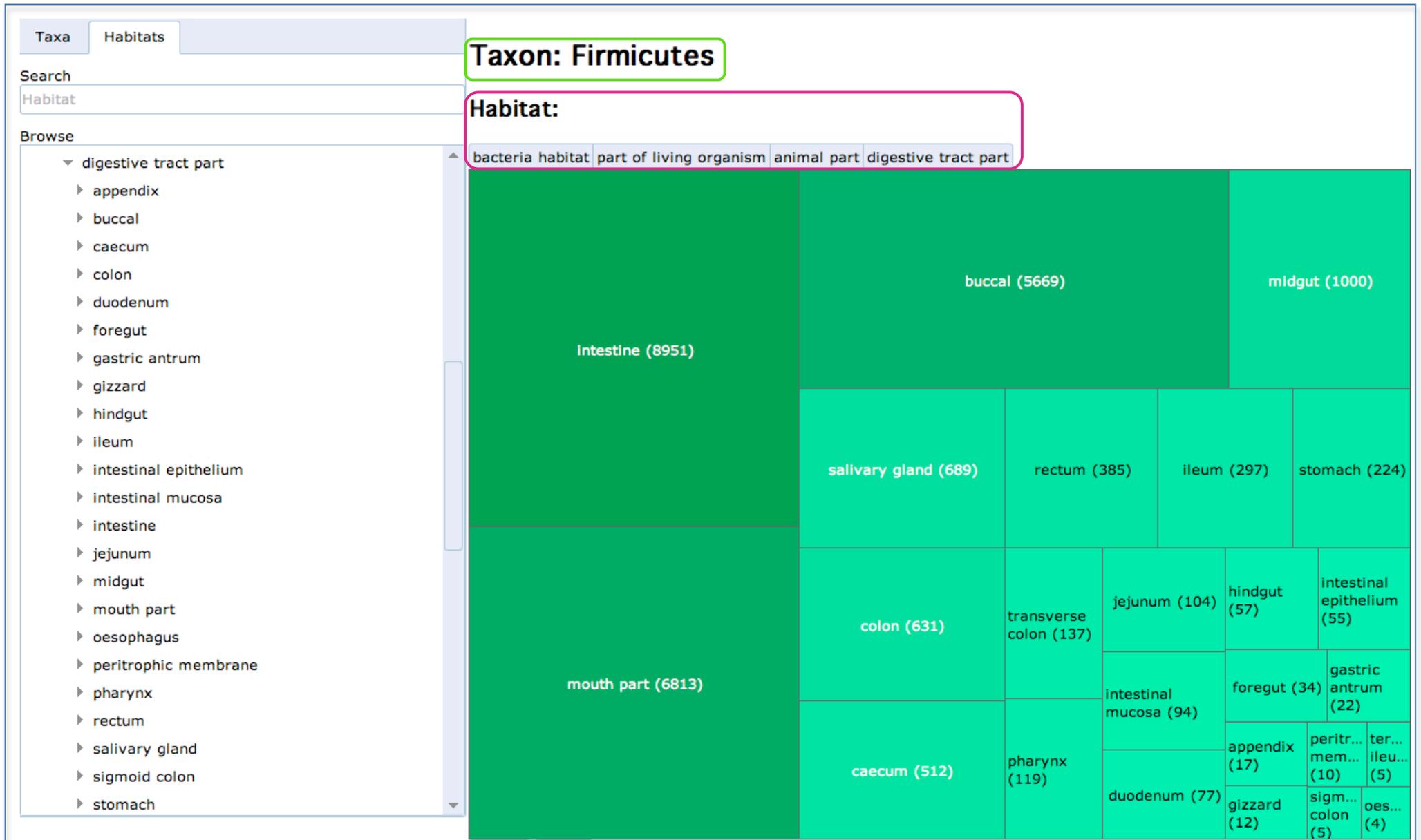
2
Indole-positive *Vibrio vulnificus* isolated from disease outbreaks on a Danish eel farm.
 1999 *Diseases of aquatic organisms*
Abstract *Vibrio vulnificus* was isolated in 1996 from 2 disease outbreaks on a Danish eel farm which used brackish water. A characteristic clinical sign was extensive deep muscle necrosis in

Vibrio (microorganism) (4196)
 ▶ Synonyms (15)
 ▶ Sub-concepts (5)

loc (Relation) (29)

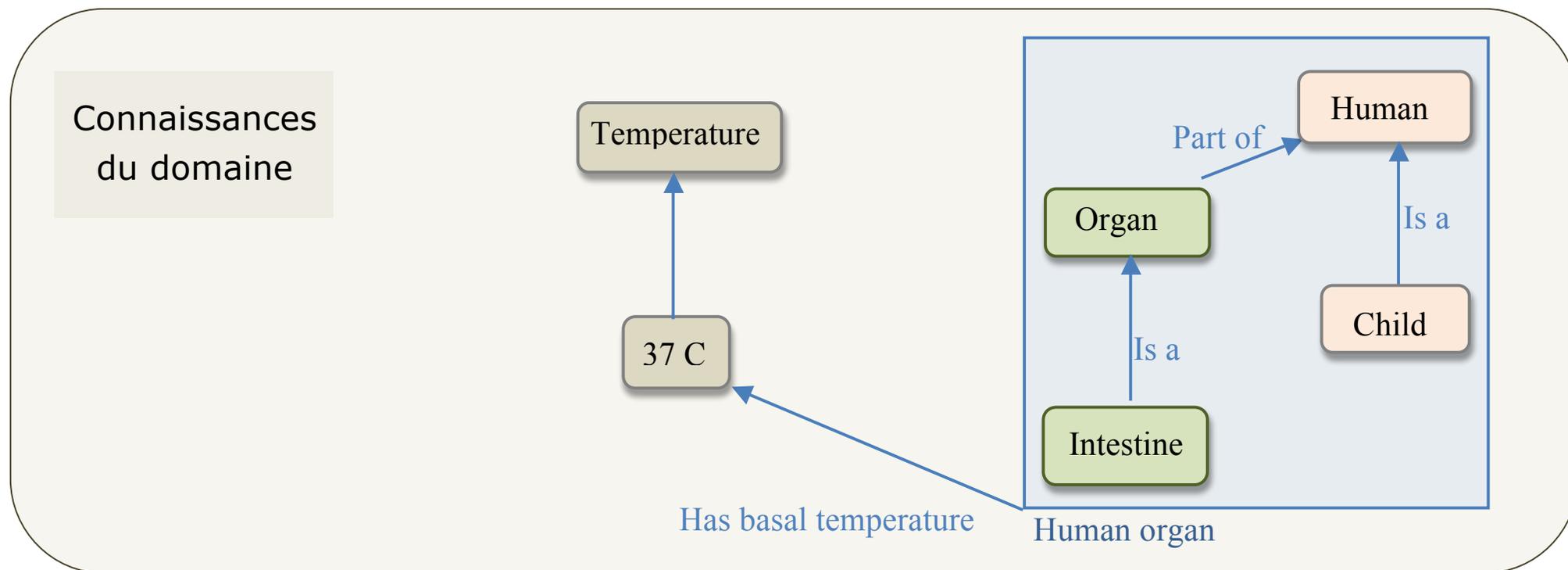
mariculture farm (habitat) (157)
 ▶ Synonyms (1)
 ▶ Sub-concepts (2)
 fish farm
 mussel farm

Modèle de connaissance hiérarchique

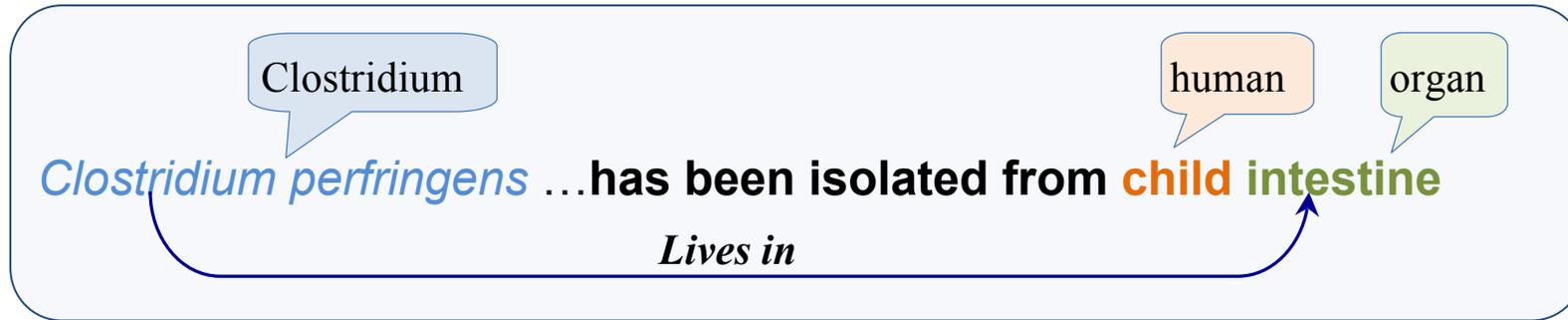


Raisonner

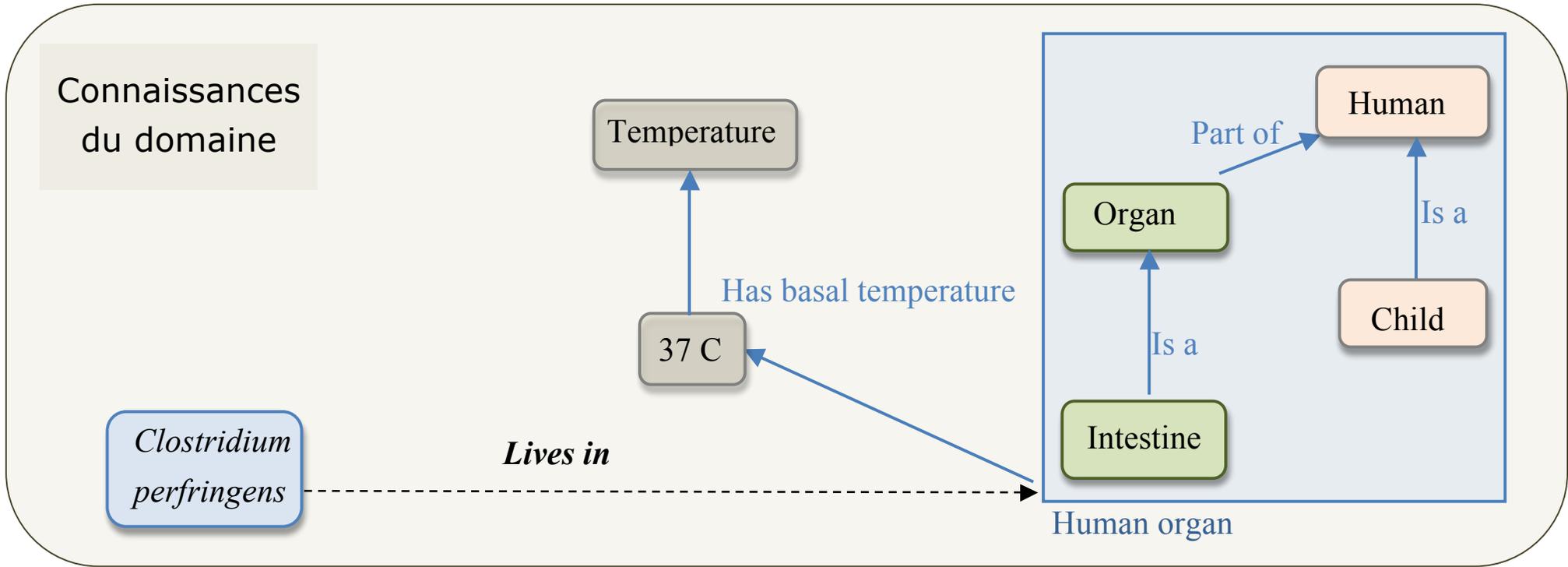
Domaine : biotopes bactériens



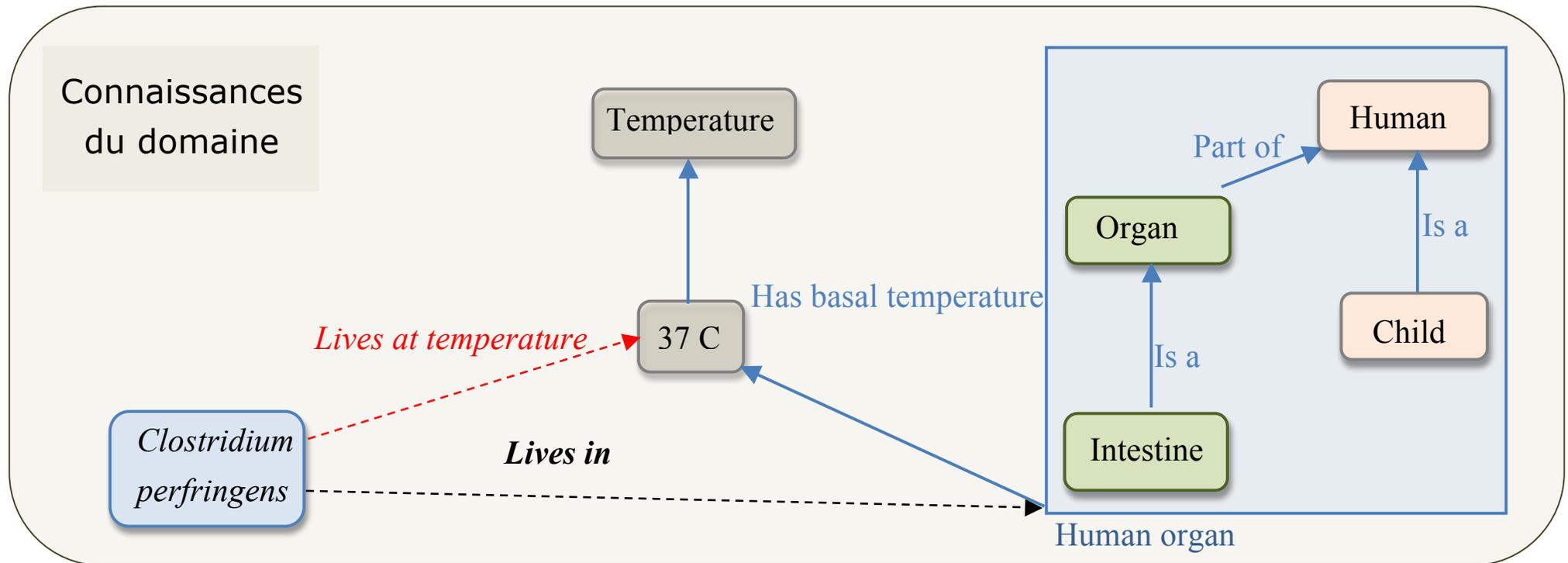
Raisonner, réviser



Information textuelle



Raisonner, déduire



Le texte, une source de données et de connaissance pour l'Open Data

Genre d'intérêt

Documents de domaines scientifiques et techniques

Dont l'objet est la communication de nouvelles connaissances

Contenus

A. Des données nouvelles

B. Des informations nouvelles

C. Des connaissances fondamentale du domaine, paradigmatique, partagées et connues

Plusieurs types de contenu reliés les uns aux autres

Leur rôle dépend de la tâche

Des données

Ontologie ATOL

- milk trait
 - ■ milk composition trait
 - ← ■ milk carbohydrate trait
 - ← ■ milk cell count
 - ← ■ milk dry matter concentration
 - ← ■ milk energy concentration
 - ← ■ milk fat trait
 - ← ■ milk hormone concentration trait
 - ← ■ milk mineral trait
 - ← ■ milk nitrogen trait
 - ← ■ milk organic acid trait
 - ← ■ milk acetic acid concentration
 - ← ■ milk citric acid concentration
 - ← ■ milk formic acid concentration
 - ← ■ milk hippuric acid concentration
 - ← ■ milk lactic acid concentration
 - ← ■ milk neuraminic acid concentration
 - ← ■ milk orotic acid concentration
 - ■ milk organoleptic traits
 - ■ milk structure trait
 - ■ milk technological trait
 - ■ milk yield

milk citric acid concentration

Uni				
Jap				
Bra				
Uni	Japi	United States	142,052	
Gei	Braz	Japan	46,558	2:50:10
Fra	Unit	Brazil	31,345	4:33:10
Spa	Gerl	United Kingdom	29,129	6:07:54
Ita	Frar	Germany	28,057	4:11:45
Aus	Spai	France	26,786	4:04:39
Swi	Italy	Spain	19,456	5:30:55
Aus	Ausl	Italy	18,256	6:00:07
Swit	Austra	Australia	9,895	6:52:28
Sour	Swit	Switzerland	2,451	3:34:34

Source: The Nielsen Company

Base de données

Animal (2009), 3:5, pp 710–717 © The Animal Consortium
doi:10.1017/S1751731109004042

Paper extract

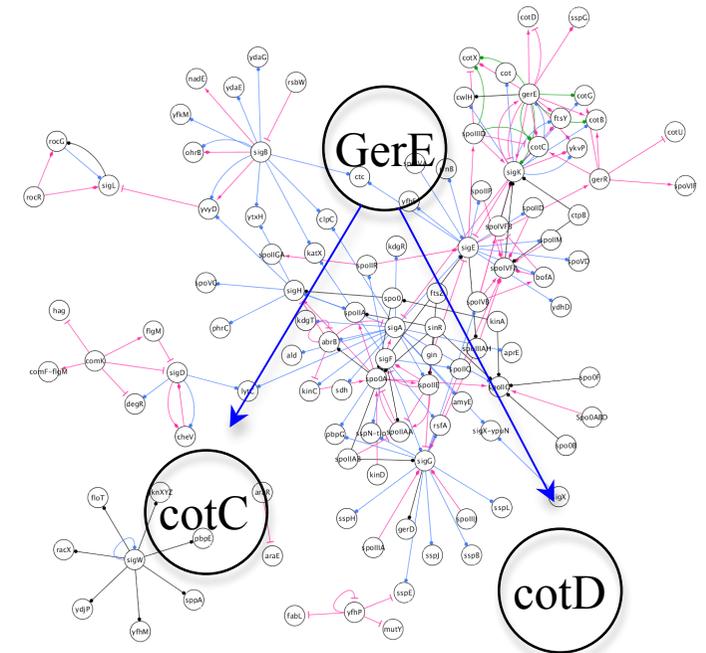
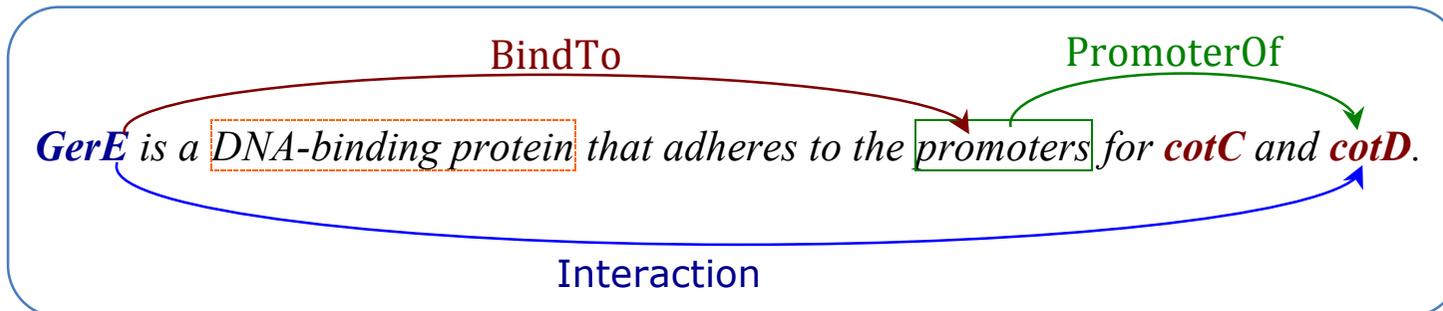
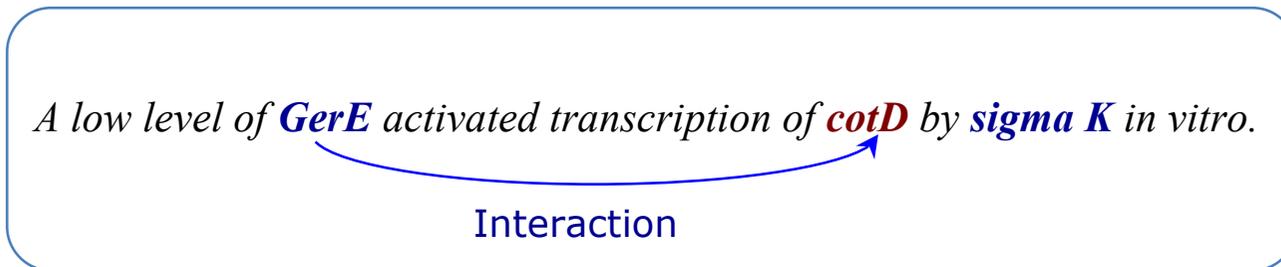
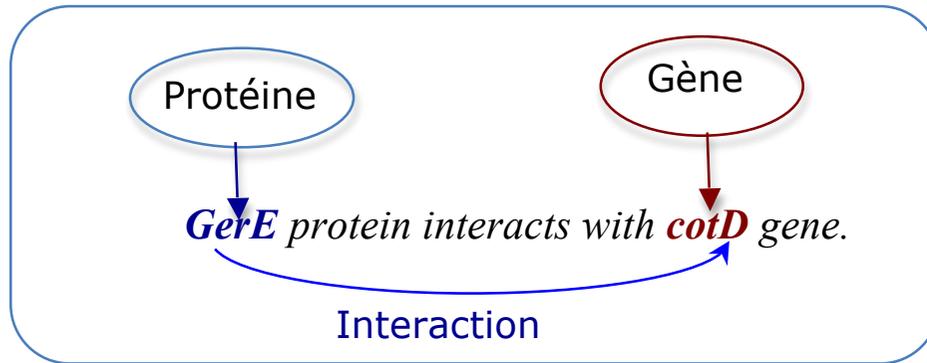
Table 1 Composition and yield of milk (LS means) of cows in a composite milk trial, of hours since last milking and the fix effects group and lactation stage, according to

Parameter	Group 1 [‡]
Number of cows	49
Milk yield/milking (kg)	11.16 ^a ± 0.37
Fat (%)	4.77 ^a ± 0.13
Total protein (%)	3.55 ^a ± 0.04
Casein (%)	2.61 ^a ± 0.03
Whey protein (%)	0.94 ^a ± 0.02
Casein number	0.74 ^a ± 0.00
Lactose (%)	4.55 ^A ± 0.03
Citric acid (%)	0.16 ^a ± 0.00
Log SCC (cells/ml)	4.50(32 [†]) ^A ± 0.05

SCC = somatic cell count.
[†] × 1000, displayed as antilogarithmic values.
[‡] The groups were: 1 – cow composite SCC < 100 000 cells/ml; 2 – cow composite SCC > 100 000 cells/ml.
[§] The lactation stage was described as discrete with the intervals: A = lactation week 1–4; B = lactation week 5–8; C = lactation week 9–12.
^{a–c} Significant differences ($P < 0.05$).
^{A–C} Significant differences ($P < 0.01$).

Des informations

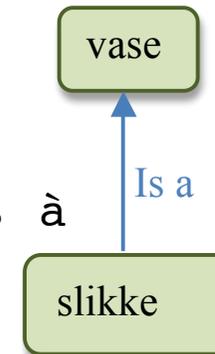
Régulations génétiques (source : PubMed)



Des connaissances, concepts et relations paradigmatique

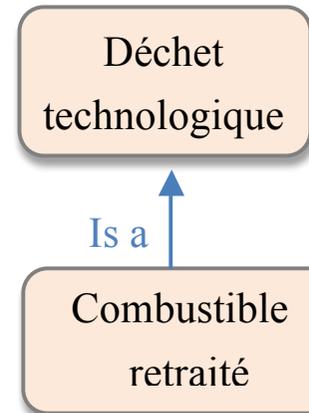
Énoncé définitoire

La vase peu colonisée, recouverte plusieurs heures à chaque marée, se nomme une slikke.



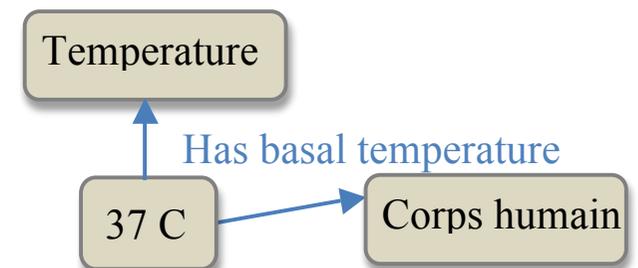
Coordination et marqueur de spécification

... des combustibles retraités et autres déchets technologiques ...



Rôles sémantiques

La température basale usuelle du corps humain est de 37,0 °C.





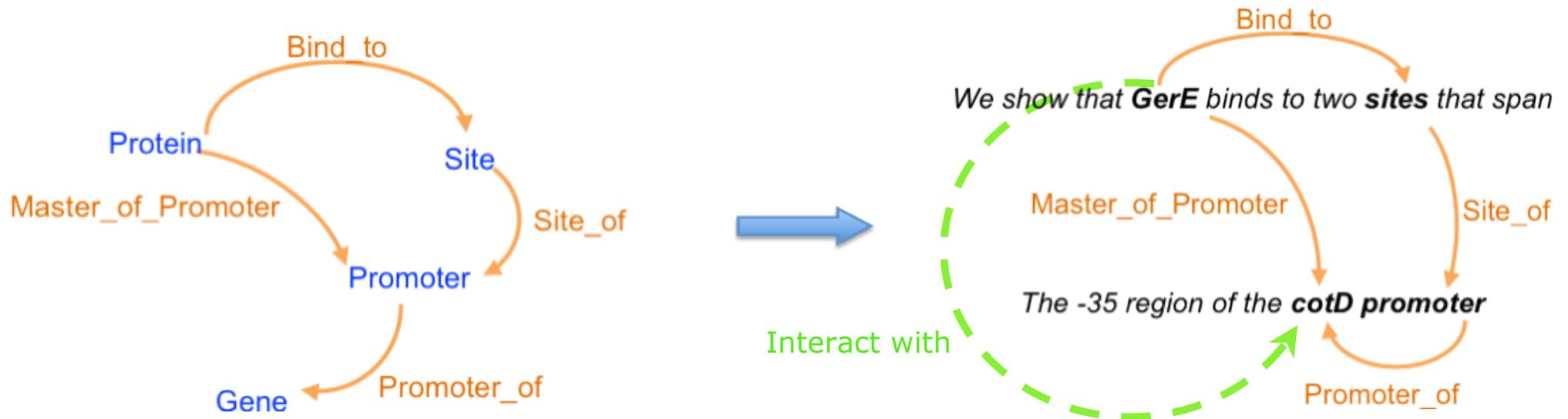
Extraction d'information guidée par la tâche

Extraction d'information

Une même approche pour extraire et formaliser à tous les niveaux de contenu

Guidée par la tâche

Qu'est-ce qui doit être extrait et comment cela s'intègre-t-il dans le modèle existant





Extraction d'information, des objectifs divers

A différents niveaux de connaissance

Des connaissances du texte pour l'acquisition d'ontologie et de modèles de connaissance

Des informations pour la *population* d'ontologie

Des données pour des bases de données

Pour différents services

Moteur de recherche sémantique

Question-réponse

Cartographie thématique

...



Méthodes d'extraction d'information

De façon transversale, des méthodes pour extraire des

Entités

Nommées ou non

Named entity recognition (NER)

Relations

Relations n-aires entre des entités

Relation Extraction (RE)

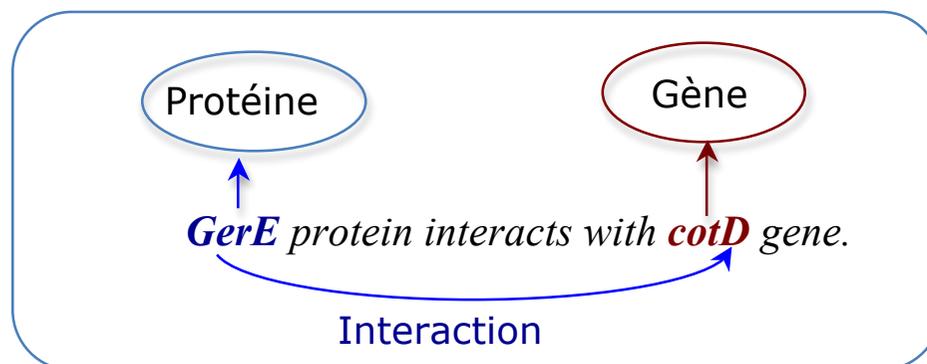
Evénements

Relations entre relations et entités



Méthodes d'extraction d'information

Cas simple



Protein : Un nom de 4 lettres, commençant par une majuscule et suivi du mot *protein*, est un nom de protéine

Gene : Un nom de 3 lettres, finissant par une majuscule et suivi du mot *gene*, est un nom de gène

Interact(Protein, Gene) : Si un nom de protéine est sujet du verbe *interact* et qu'un nom de gène est l'objet du verbe, alors la protéine interagit avec le gène

Règles
d'extraction
basées sur
des indices
linguistiques

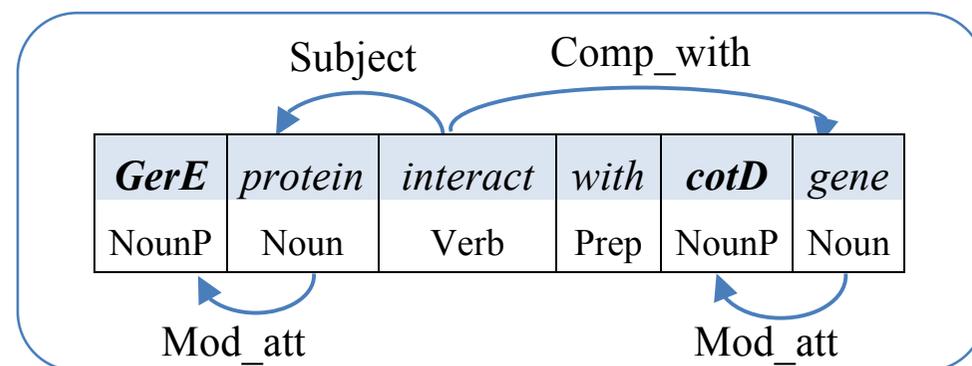
Traitement automatique de la langue naturelle

Analyse profonde pour prendre en compte

- les variations,
- le contexte pour désambigüiser,
- les relations entre les éléments du texte

- **Segmentation** en phrases, en mots, en caractères
- **Lemmatisation** : les mots sous forme canonique
- **Etiquetage** morpho-syntaxique : noms, verbes
- Analyse des **dépendances syntaxiques** :
sujet de, objet de, adjectif de

...



Aujourd'hui, de nombreux *pipelines* : UIMA, GATE, Alvis, NLTK, CoreNLP

Et **outils** : TreeTagger, Genia tagger, ..., Stanford Parser, McClosky, CCG, ... Banner, Cocoa



Apprentissage automatique

Pour généraliser l'extraction quand il n'existe pas de solution "simple".

Apprentissage supervisé à partir d'exemples de la connaissance à extraire (CRF, SWM, ME)

Apprentissage non supervisé regroupe les éléments du texte par similarité (LDA, LSA, ...)

Des boîtes à outil riches et complexes

Weka, MLT, ... qu'il faut toujours adapter

La représentation, question centrale

Quelle connaissance apprendre à prédire

A partir de quelles informations du texte et du domaine

Dans quelle représentation

Apprendre à prédire

Beaucoup d'expertise

Tous les cas sont différents

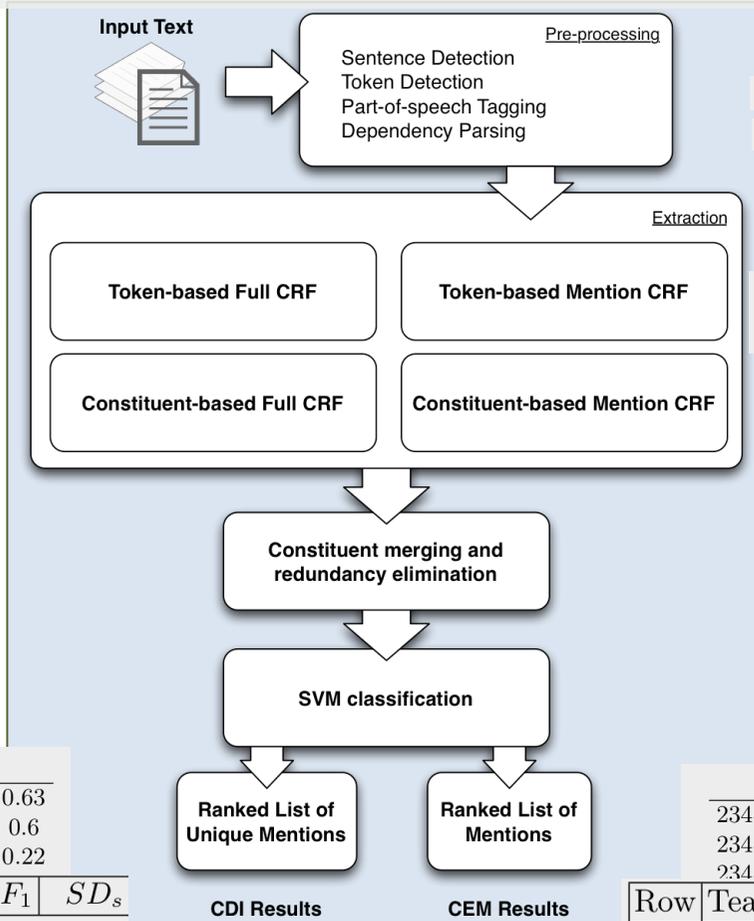
Démarche empirique

Exemple d'extraction d'entités nommées

BioCreative Challenge, ChemdNER 2013 : médicaments et composés chimiques

- Abréviation
- Famille
- Formule
- Identifiant
- Multiple
- Systématique
- Trivial
- Pas de classe

The lystabactins are composed of serine (Ser), asparagine (Asn),



bactins are composed of serine (Ser), asparagine ...
MILY 0 0 0 TRIVIAL 0 FORMULA 0 0 TRIVIAL ...

The lystabactins serine Ser asparagine Asn
FAMILY TRIVIAL FORMULA TRIVIAL FORMULA

“DT The NNS lystabactins VBP are VBN
composed IN of NN serine -LRB- (NNP Ser -RRB-

The lystabactins are composed of serine (Ser)
0 M 0 0 0 M 0 M 0

The lystabactins serine Ser asparagine Asn two for
M M M M M

CDI results

23444833	serine	1	0.63
23444833	lystabactins	2	0.6
23444833	nonproteinogenic	3	0.22

Row	Team	P	R	F ₁	SD _s
A	231	87.02%	89.41%	88.20%	0.30%

CEM results

23444833	A:318:324	1	0.5
23444833	A:289:301	2	0.5
23444833	A:448:464	3	0.5

Row	Team	P	R	F ₁	SD _s
A	173	89.09%	85.75%	87.39%	0.37%

Exemple de catégorisation riche des entités

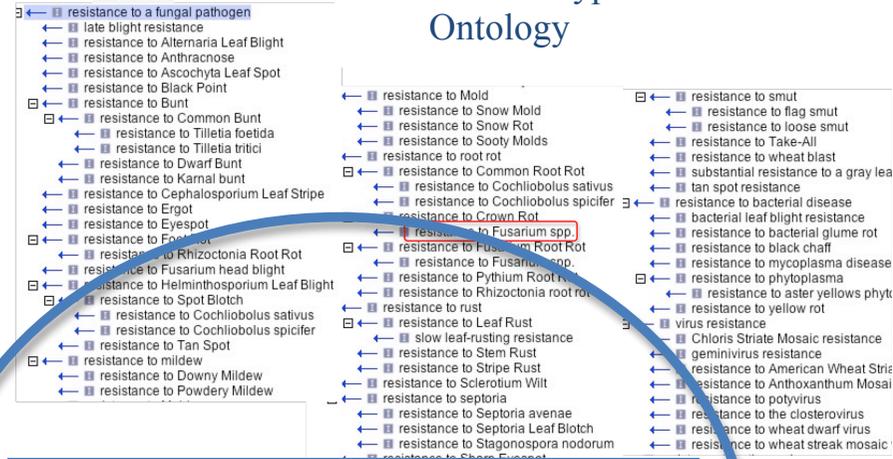
Entités désignées par des termes

Méthodes MetaMap, TaxoMap, ToMap

Participant	Precision	Recall	F1
Irisa-TeXMex	48	72	57
Boun	59	60	59
LIPN	61	61	61
LIMSI	62	35	44
BioYaTeA+ToMap	83	64	72

BB sub-task 1

WheatPhenotype Ontology



Normalisation terminologique

Cultivar	Gene	Phenotype	Pathogen
Shangai 3/CatBird	2DL-gwm265	Type I resistance to Fusarium head blight	Resistant to FHB Fusarium spp.
Allezy	227	resistant to leaf rust	Resistance to leaf rust Puccinia recondita
Yitpi	Tsn1	ToxA-sensitive	Toxin sensitive NA
Jagger	Yr17	high level of stripe rust resistance	Resistant to yellow rust Puccinia striiformis
Hongyanglazi	PmHYLZ	resistant to Bgt isolate E09	Resistant to Take-All Gaeumannomyces graminis f. sp. tritici
Solitär	Stb6	resistant to STB in the field	Resistant to STB Septoria tritici

BD phénotype du blé tendre (FSOV)

Catégorisation riche par analyse terminologique

1. Extraction terminologique (BioYateA, Syntex, ...)

2. Un terme du texte désigne un phénotype, si sa tête syntaxique (mot principal) est égale à la tête d'un terme de l'ontologie *WheatPhenotypes*

[[*Type I resistance*] to [*Fusarium head blight*]] → resistance to Fusarium head blight
= resistance to FHB

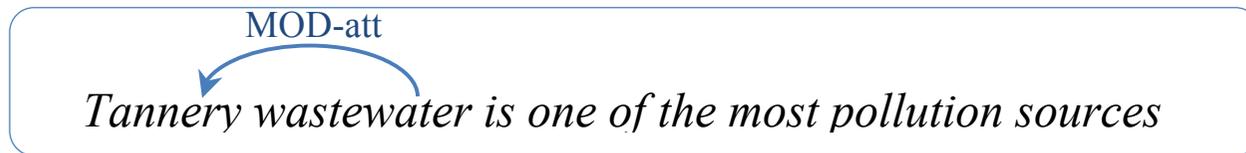
3. Un terme du texte désigne un phénotype si

- Sa tête appartient à la liste des têtes ambiguës, définie préalablement (*number, level, plant, size, time, index*)
- Et que la tête de son sous-terme est une tête d'un terme de l'ontologie *WheatPhenotypes*

[[*high level*] of [*stripe rust resistance*]] → stripe rust resistant
= resistant to yellow rust

Exemple, extraction de synonymes par sémantique distributionnelle

Former des classes sémantiques par *clustering* à partir de cooccurrences syntaxiques



	waste	wastewater	sludge	effluent	site	estate
pesticide plant			X			
refinery			X			
olive-mill industry			X			
soda ash industry			X	X		
silk industry		X				
antibiotic production		X				
sugarcane industry		X				
oxytetracycline production		X				
pulpmill		X		X		
tannery	X	X		X		
slaughter house	X	X	X			
distillery	X	X	X	X		
industrial	X	X	X	X	X	X
oilfield			X			
oil			X			
municipal sewage			X			
fish pond			X			
animal	X					
swine	X					

Exemple : extraction de relations par apprentissage

Complexité de l'extraction de relation, fonction de

- La variabilité des expressions
- De la localisation de l'information (inter-phrases)
- Du nombre d'entités candidates

Whereas *F. columnare* and *F. psychrophilum* are sometimes capable of causing *branchial* lesions in several species of *fish*, a third species, *F. branchiophilum*, is mainly responsible for *gill* disease [...], a pathology which primarily affects *salmonides*.

Posé comme un problème de **classification supervisée**

(*F. columnare*, branchial) (+)

(*F. psychrophilum*, branchial) (+)

(*F. columnare*, fish) (+)

...

(*F. columnare*, gill) (-)

(*F. branchiophilum*, fish) (-)

(*F. columnare*, salmonides) (-)

...

Team	Precision	Recall	F-measure
TEES-2.1	82	28	42
IRISA-TextMex	36	46	36
Boun	38	21	27
LIMSI	19	4	6
<i>SPGAK-adapt</i>	<i>51.5</i>	<i>70.0</i>	<i>59.3</i>

Méthode de classification supervisée pour l'extraction de relations

Classification

Support Vector Machine (SVM)

Représentation des exemples

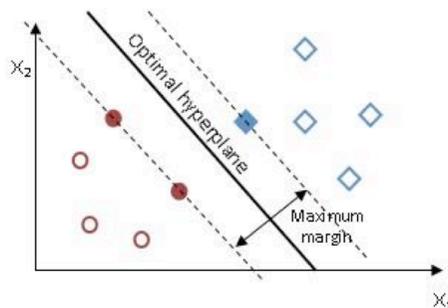
Chemin syntaxique entre les candidats

Exemple d'apprentissage

Annotés manuellement (Brat, AlvisAE)

Similarité entre exemples

Optimisation globale de l'alignement



The type strain of *B. garinii* was isolated from *Ixodes ricinus* in France.

prep pobj subj prep pobj
B. garinii ← of ← strain ← isolated → from → *Ixodes ricinus*

C. coli is usually isolated from *pigs*, environmental surface water.

<i>B. garinii</i>	prep	of	pobj	strain	subj	isolated	prep	from	pobj	<i>Ixodes ricinus</i>
<i>C. coli</i>					subj	isolated	prep	from	pobj	<i>pigs</i>
0	GP	GP	GP	GP	1	1	1	1	1	0

* GP=gap penalty



Construire une application

Importance de la conception

Conception du *pipeline* : pour chaque application, une analyse et une conception différente

Choix des données (corpus, ressources existante), paramètres

Compétences en IC, TAL, ML, SI

Pour un problème "classique" où le pipeline est connu

Choix des outils, paramétrage, intégration

Part d'imprévisibilité du texte

Pour un problème nouveau

Importance de l'analyse des besoins en fonction de la faisabilité, maquettes, itérations

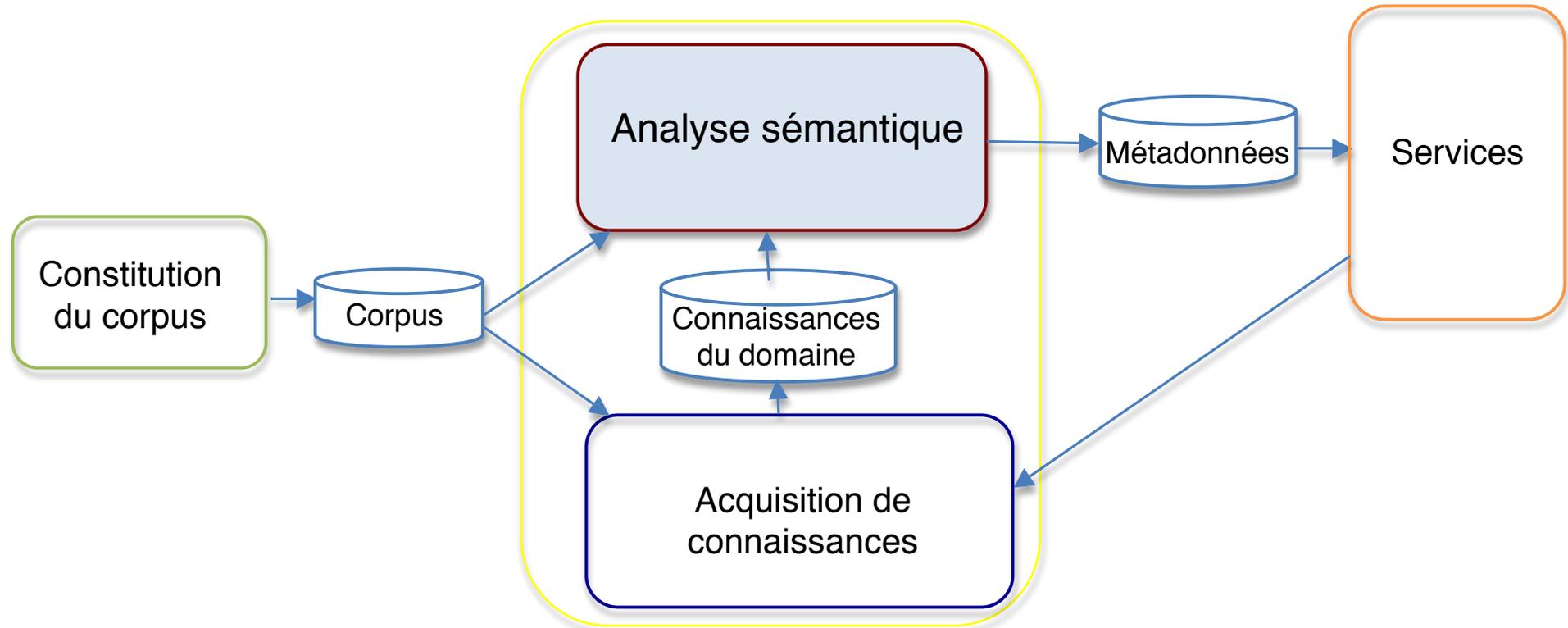
Ingénierie des connaissances

Contrôle, validation interprétation, intégration des résultats

Besoin d'interfaces homme-machine adaptées

Editeur d'annotation, éditeur de terminologie, éditeur d'ontologie, cartographie

Architecture d'analyse sémantique de texte



Term grid (1) – free search – Formation ATOL

Semantic Classes of term withdrawal response

Filter

ignore case

Form: include inferred terms: include dismissed terms:

Lemma: include unparsed phrases:

Syntactic category: Word count: >= <=

Head: Nb occurrences: >= <=

Expansion: show only class members: show only class representative:

Prevalidation: Justification: all users:

producer: OBO_1 FastR_2 FastR_1 YaTeA_1 Validation: D D? V? V all users:

Semantic Classes Tree Window

Formation ATOL withdrawal

- psychoneuroendocrinological state trait
 - behavior trait
 - biological rhythm trait
 - cognitive functions trait
 - emotional functions trait**
 - metabolism trait
 - pain responses trait
 - antalgic posture
 - emotional functions trait**
 - nociception
 - withdrawal response**

OccurrenceInContext Window – Candidate 8880651

Filename	#	Context
/bibdev/corp...	1	05) , while it tended to be correlated with the withdrawal response when approached from the front (P 0 .
/bibdev/corp...	1	However , the withdrawal response of the sow when in the farrowing crate was observed by another stockperson .
		This withdrawal response was strongly correlated with the other behavioural responses such as nervousness of sow in the crate before and around farrowing

43 rows

Surface form	Nb occ...	Nb doc.	Head	△ Expansion	Nb words	Syntactic categ...	ClaireNedellec
'withdrawal crate'	1	1	crate'	'withdrawal	2	JJ NN	
withdrawal movements	1	1	movement	withdrawal	2	NN NNS	
withdrawal period	2	2	period	withdrawal	2	NN NN	
withdrawal reaction	4	3	reaction	withdrawal	2	NN NN	
withdrawal response	5	4	response	withdrawal	2	NN NN	V
withdrawal test	4	1	test	withdrawal	2	NN NN	

Interface TyDI



Conclusion

Open Data

Diversité des sources, besoin critique de référentiel : modèle de connaissance

Normalisation, mais aussi raisonnement

Besoin de modèles riches dans des langages de représentation expressifs

Extraction d'information

Méthodes pour construire automatiquement des référentiels spécialisés *et* produire des données

Associe automatiquement le texte au référentiel

Aligne les labels des modèles

Perspectives

Représentations plus riches pour l'extraction d'information

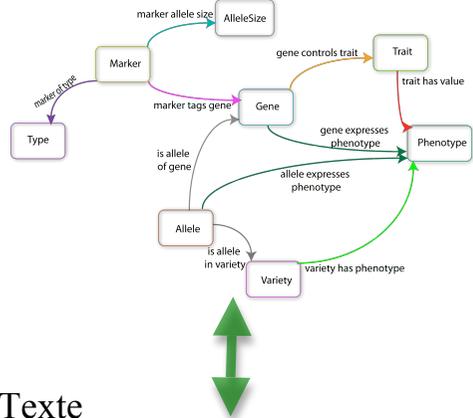
- Relations n-aires
- Appariement m-n texte / concept
- ...

Meilleure intégration des démarches d'EI et de modélisation des connaissances, raisonnement

Modélisation de connaissances à partir de sources hétérogènes

Définir les rôles des modèles, leur représentation et leur articulation
 Caractériser les sources de connaissance (quoi, avec quelle qualité)
 Identifier les redondances et les complémentarités

Modèle de connaissances du texte



Texte

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a method of treating or preventing **infectious diseases** or **infectious conditions** comprising administering to said **subject** an effective amount of **antibodies**.

In a particular embodiment of the above defined method, the above mentioned **antibodies** are selected from **antibodies** which comprise administering to said **subject** on **antibodies**.

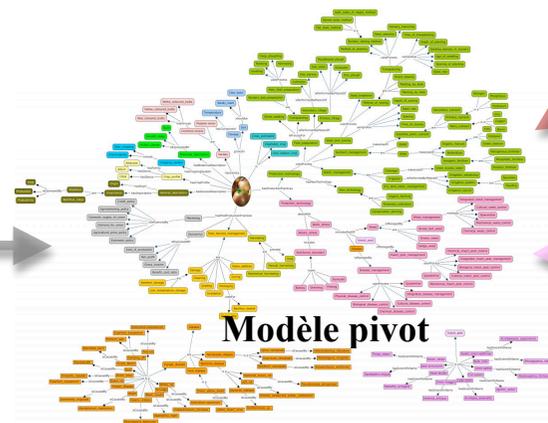
In a more particular embodiment, the invention relates to a method as defined above of treating or preventing **infectious diseases** or **infectious conditions** which comprise administering to said **subject** on **antibodies**.

According to a preferred embodiment of the above defined method of treating or preventing **infectious diseases** or **infectious conditions**, in the form of **antibodies**.

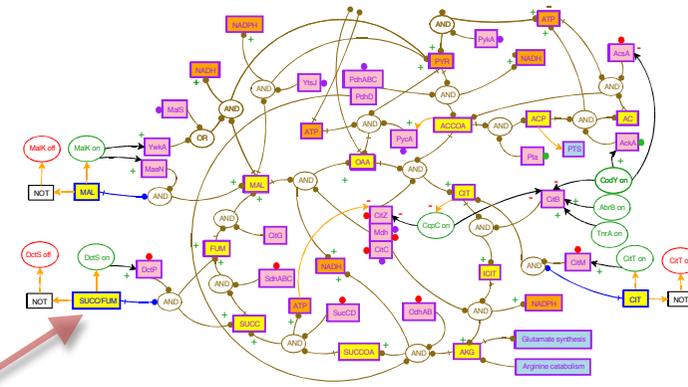
According to a further embodiment, in a composition method in which **antibodies** are present and their administration to a subject of the present invention is in the form of **antibodies** containing an adjuvant.

According to a further embodiment, in a composition method in which **antibodies** are present and their administration to a subject of the present invention is in the form of **antibodies** containing an adjuvant.

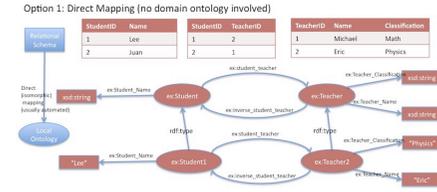
According to a further embodiment, in a composition method in which **antibodies** are present and their administration to a subject of the present invention is in the form of **antibodies** containing an adjuvant.



Modèle dynamique



Modèle de données



Données

Country	Unique Audience (000)	Time per Person (hh:mm:ss)
United States	142,052	6:08:13
Japan	46,558	2:50:21
Brazil	31,345	4:33:10
United Kingdom	29,129	6:07:54
Germany	28,057	4:11:45
France	26,786	4:04:39
Spain	19,456	5:30:55
Italy	18,256	6:00:07
Australia	9,895	6:52:28
Switzerland	2,451	3:54:34

Source: The Nielsen Company



Exemple d'éditeur d'annotation sémantique, AlvisAE

BTID-10086

Water

Bordetella petrii DSM 12804

Description

Bordetella petrii strain DSM12804. Bordetella petrii strain DSM12804 was initially isolated from river sediment

Unlike other members of the genus, this organism is not known to be associated with humans or other warm-blooded animals. Bordetella petrii also differs from other Bordetella species in that it is a facultative anaerobe. This strain is the type strain for the species and will be used for comparative genomics with other Bordetella species.

Annotations Text selection

Id	Annotation Set	Ki	Type	Details	V
4d450...ec	claire's annotation		Host	warm-blooded animals	
88cbf...a1	claire's annotation		Host	humans	
4c991...4E	claire's annotation		Bacteria	Bordetella petrii strain DSM12804	
162b9...01	claire's annotation		Bacteria	Bordetella petrii strain DSM12804	
5b053...5a	claire's annotation		Water	river sediment	
1021d...eE	claire's annotation		localization	bacteria (Bacteria Bordetella petrii strain DSM12804) + foundIn (Water river sediment)	

Identification des termes du texte par BioYaTeA

Clostridium acetobutylicum

Ecology

While the type strain of *C. acetobutylicum* was isolated from soil, *C. acetobutylicum* is ubiquitous. It has been found in lake sediment, well water, and clam gut. In addition, it has been recorded in a number of different feces specimens, including human, bovine, and canine feces. A search of the literature reveals that pathogenic or symbiotic relationships are not documented. [MicrobeWiki]

- Groupes nominaux ou adjectivaux (en jaune) entre les frontières prédéfinies (en rouge)
- Sous-termes extraits récursivement en fonction de leurs occurrences dans les textes
- Filtrés automatiquement sur des critères linguistiques

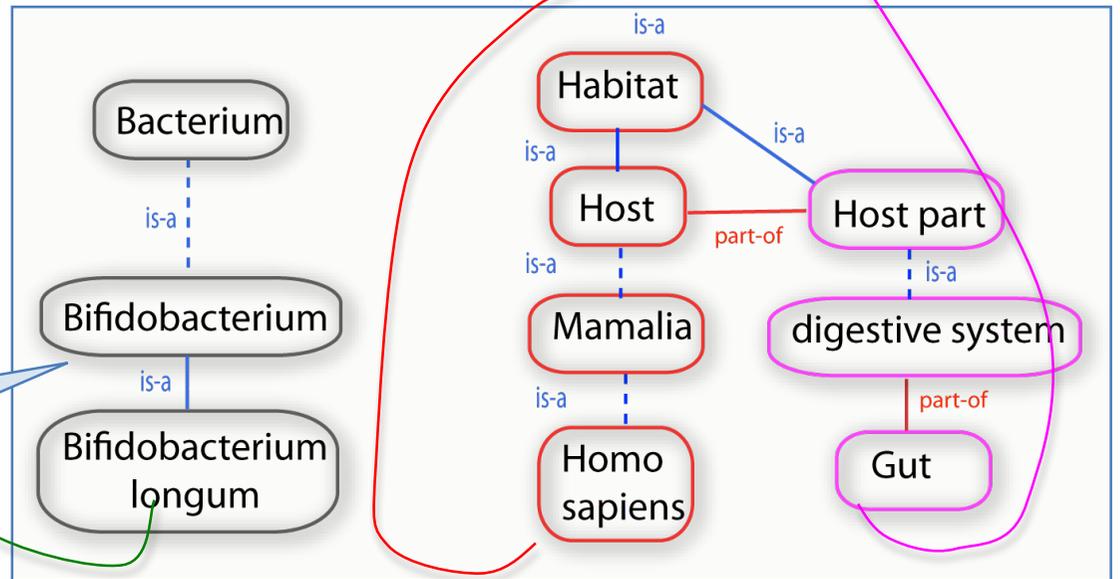
Exemple : in a *number of [different [feces [specimens]]]*, including
⇒ *feces specimen, feces, specimen*

Annotation sémantique de base de données

Acc. num	Length	Organism	Taxid	Pubmed	Isolation source
GQ380695	1398	Bifidobacterium longum	216816	19701668	Human intestinal microflora

Entrée
GenBank

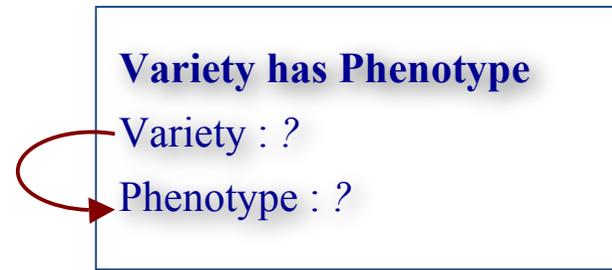
Ontologie
des habitats
microbiens



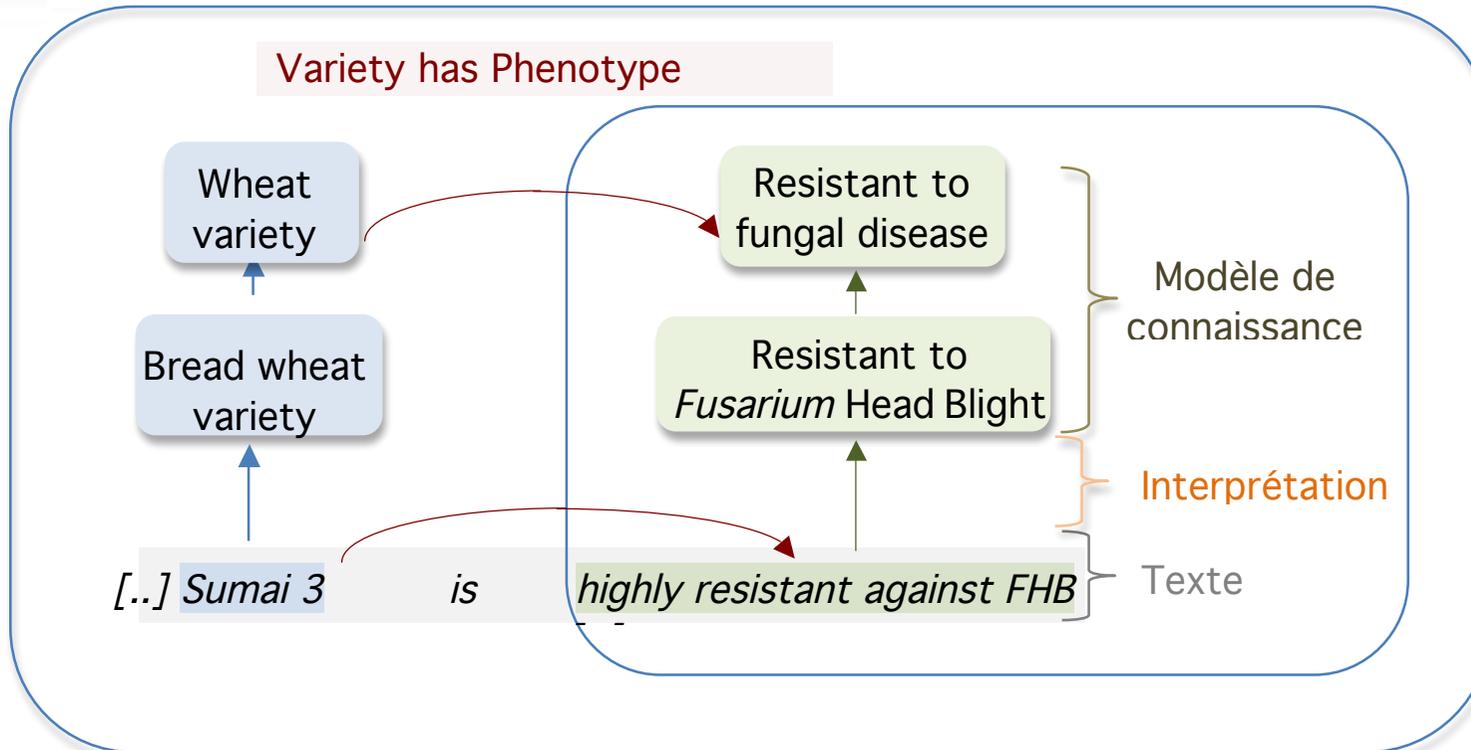


Reconnaissance des événements dans le texte

Événements prédéfinis composés d'entités reliées par des relations orientés et typés



1. Identification et normalisation des entités du texte par des concepts
2. Analyse des dépendances syntaxiques entre les entités et extraction des relations



Variety has Phenotype
Variety : Sumai 3
Phenotype : highly resistant against FHB