Exploitation de la spatialité sur des données textuelles hétérogènes provenant de Madagascar

Journée de Variété des Données SHS
28 mai 2018

Jacques Fize, *Doctorant*, **TETIS**

Dirigé par Mathieu Roche et Maguelonne Teisseire

tetis
TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION INFORMATION SPATIALE

irstea

cirad
LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

UNIVERSITÉ MONTPELLIER

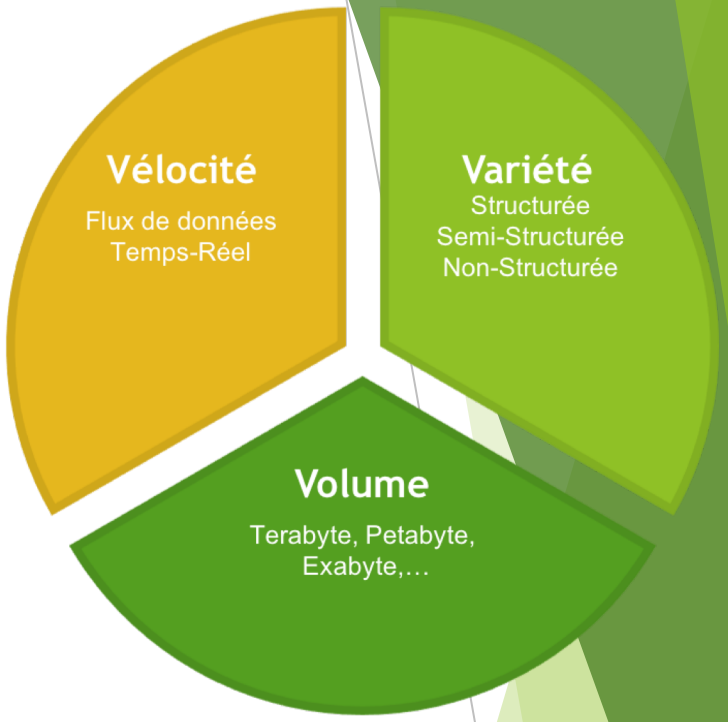
Sommaire

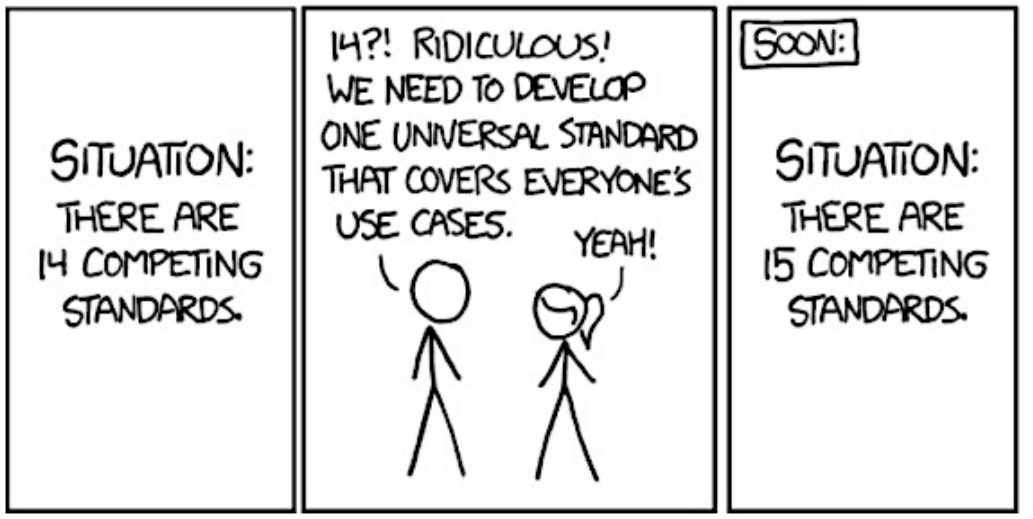
Introduction

L’ère du Big data

- Caractérisé par 3 dimensions
  - Volumétrie, Vélocité et Variété
  - … mais aussi : Véracité, Variabilité

- Augmentation constante de l’hétérogénéité de la **data**
  - **Modernisation** des différents secteurs
  - **Évolution** des formats en fonction des besoins
  - **Unification** des formats …





https://www.slideshare.net/InformaticaUCM/making-the-most-out-of-heterogeneous-chips-with-cpu-gpu-and-fpga

Exploiter l’hétérogénéité

- Augmenter le volume de données exploitables
  - Recherche d’Information
  - Extraction d’Information
- Différentes représentations d’une même information
  - Synthèse automatique
  - Découverte de connaissance

Hétérogénéité dans les textes

- L'hétérogénéité dans les données textuelles
  - Format
    - Texte brut, fichier de données (xml, sql,…), etc.
  - Langue
  - Vocabulaire, syntaxe
  - Structure

Différents formats, structures et vocabulaires

2016 PRESIDENTIAL ELECTION - EXIT POLLS VERSUS REPORTED VOTE COUNT

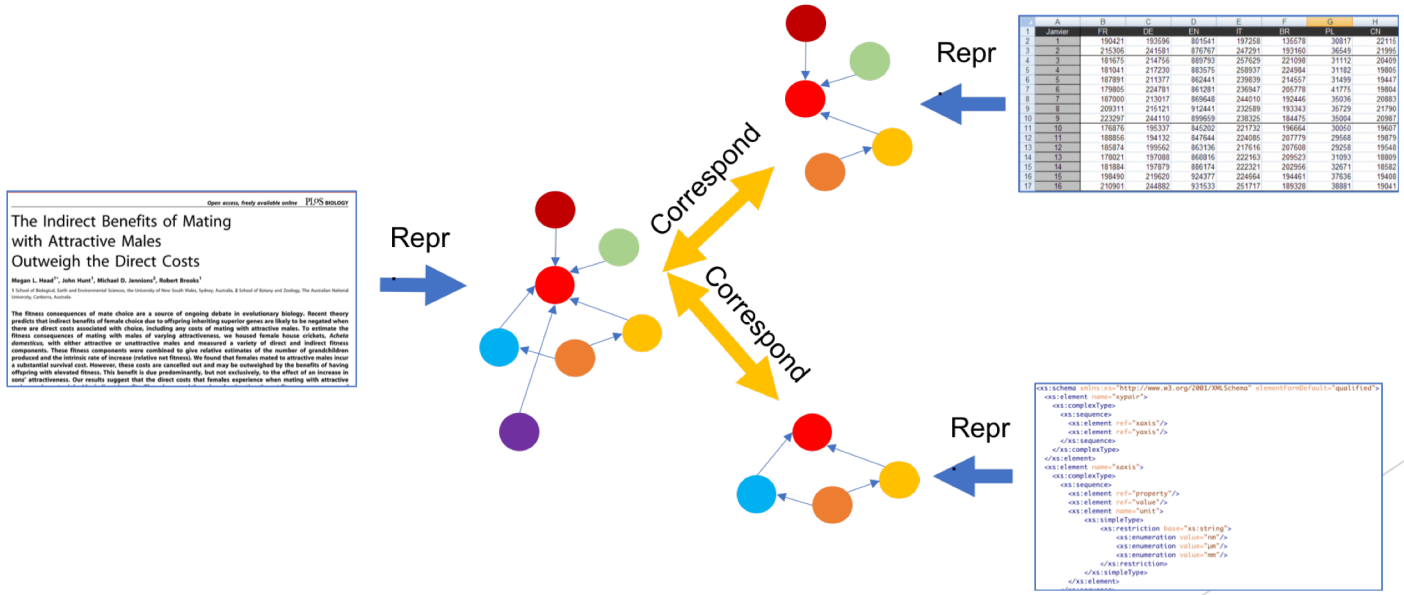
Presidential Election Clinton v. Trump	CNN PUBLISHED EXIT POLLS (EP) [1]				REPORTED VOTE COUNT (VC) [2]			EP / VC DISCREPANCIES	
	CLINTON EP	TRUMP EP	MARGIN TRUMP - CLINTON [3]	MOE [4] ON THE DIFFERENCE	CLINTON VC	TRUMP VC	MARGIN TRUMP - CLINTON [3]	MARGIN DISCREPANCY IN FAVOR OF TRUMP	DISCREPANCY GREATER THAN EP MOE
NEW JERSEY	59.8%	35.8%	-24.0%	5.8%	54.87%	41.94%	-12.9%	11.0%	5.2%
MISSOURI	42.8%	51.2%	8.4%	4.7%	38.01%	57.14%	19.1%	10.7%	6.1%
UTAH	32.4%	41.8%	9.4%	5.7%	27.81%	46.80%	19.0%	9.6%	3.9%
OHIO	47.0%	47.1%	0.2%	5.3%	43.51%	52.05%	8.5%	8.4%	3.1%
MAINE	51.2%	40.2%	-11.0%	5.0%	47.88%	45.13%	-2.8%	8.2%	3.2%
SOUTH CAROLINA	42.8%	50.3%	7.5%	4.9%	39.93%	55.63%	15.7%	8.2%	3.2%
NORTH CAROLINA	48.6%	46.5%	-2.0%	3.0%	46.70%	50.54%	3.8%	5.9%	2.8%
						51.78%	9.6%	5.7%	2.2%
						48.79%	1.1%	5.6%	1.8%
						47.35%	-0.2%	5.1%	0.5%
						57.17%	19.3%	4.9%	0.5%
						47.87%	0.9%	4.8%	1.4%
						51.33%	5.8%	4.4%	0.6%
						45.53%	-2.4%	3.5%	
						62.54%	29.8%	3.3%	
						44.96%	-4.8%	2.9%	
						44.80%	-2.1%	2.9%	
						49.06%	1.3%	2.6%	
						40.04%	-8.2%	1.8%	
						49.64%	4.3%	1.0%	
						40.91%	-11.0%	1.0%	
						33.25%	-28.2%	0.3%	
						47.60%	0.3%	0.2%	
						52.65%	9.3%	-0.2%	
						39.41%	-16.0%	-0.8%	
						45.37%	-1.5%	-1.6%	
						38.06%	-17.6%	-2.1%	
						37.44%	-21.4%	-5.4%	0.3%
						47.50%	-0.2%	3.0%	1.7%

re de Macedo Soares - www.tdmsresearch.com

```
<?xml version="1.0" encoding="utf-8" ?>
- <Dictionary xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  - <Words>
    - <Word>
        <Value>House</Value>
        <Definition>A place where people live which provides
          shelter</Definition>
      </Word>
    - <Word>
        <Value>Car</Value>
        <Definition>A device to transport people</Definition>
      </Word>
    - <Word>
        <Value>Wood</Value>
        <Definition>Part of a tree</Definition>
      </Word>
    - <Word>
        <Value>Serializer</Value>
        <Definition>An object for transforming another object to or from a
          linear sequence of bytes</Definition>
      </Word>
    </Words>
</Dictionary>
```

Mesurer la similarité entre documents hétérogènes

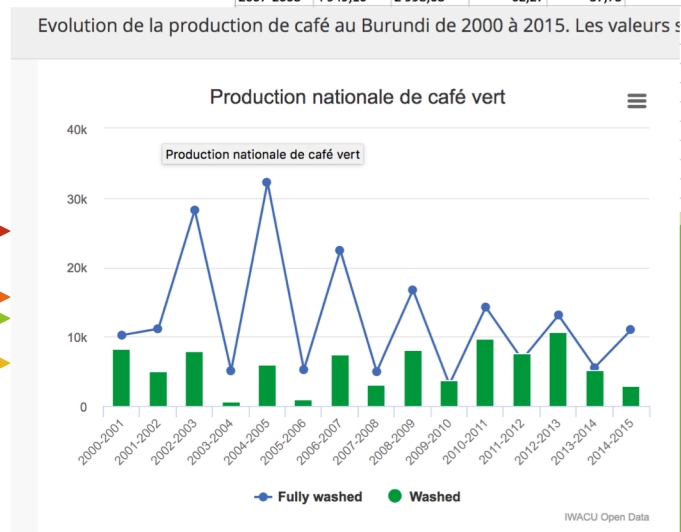
- Une grand nombre d'applications liées à la donnée hétérogène nécessitent de calculer la similarité entre les documents
- Pour mesurer la similarité entre ces documents, une représentation commune est nécessaire

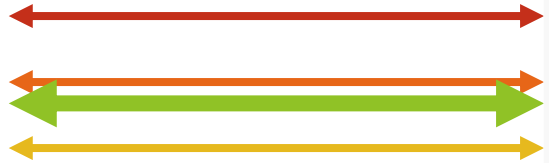


Question de recherche

- Quelles méthodes proposer pour **mettre en correspondance** des **données hétérogènes** et **textuelles** ? Approches exploitant différentes dimensions :
  - Thématique
  - Spatialité
  - Temporalité

A	B	C	D	E	F	G	H
Période	Production d	Production d	Production e	Production en %	par rapport au total (washed)		
2000-2001	10 212,48	8 286,06	55,21	44,79			
2001-2002	11 127,00	5 068,14	68,71	31,29			
2002-2003	28 235,75	7 990,08	77,94	22,06			
2003-2004	5 083,73	5 89,62	89,61	10,39			
2004-2005	32 246,04	5 956,92	84,41	15,59			
2005-2006	5 234,40	932,22	84,88	15,12			
2006-2007	22 433,40	7 512,90	74,91	25,09			
2007-2008	4 949,16	2 998,68	62,27	37,73			





REVIEW ARTICLE

Quality assessment of Arabica and Robusta green and roasted coffees – A review

Natalina Cavaco Bicho[1], Fernando Cebola Lidon[2*], José Cochicho Ramalho[1] and Antó nio Eduardo Leitã o[1]

[1]Grupo Interações Planta-Ambiente (PlantStress), Centro Ambiente Agricultura e Desenvolvimento (BioTrop), Instituto de Investigação Científica Tropical, I.P., Avenida da República, Quinta do Marquês, 2784-505 Oeiras, Portugal
[2]Departamento de Ciências da Terra, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus da Caparica, 2829-516 Caparica, Portugal

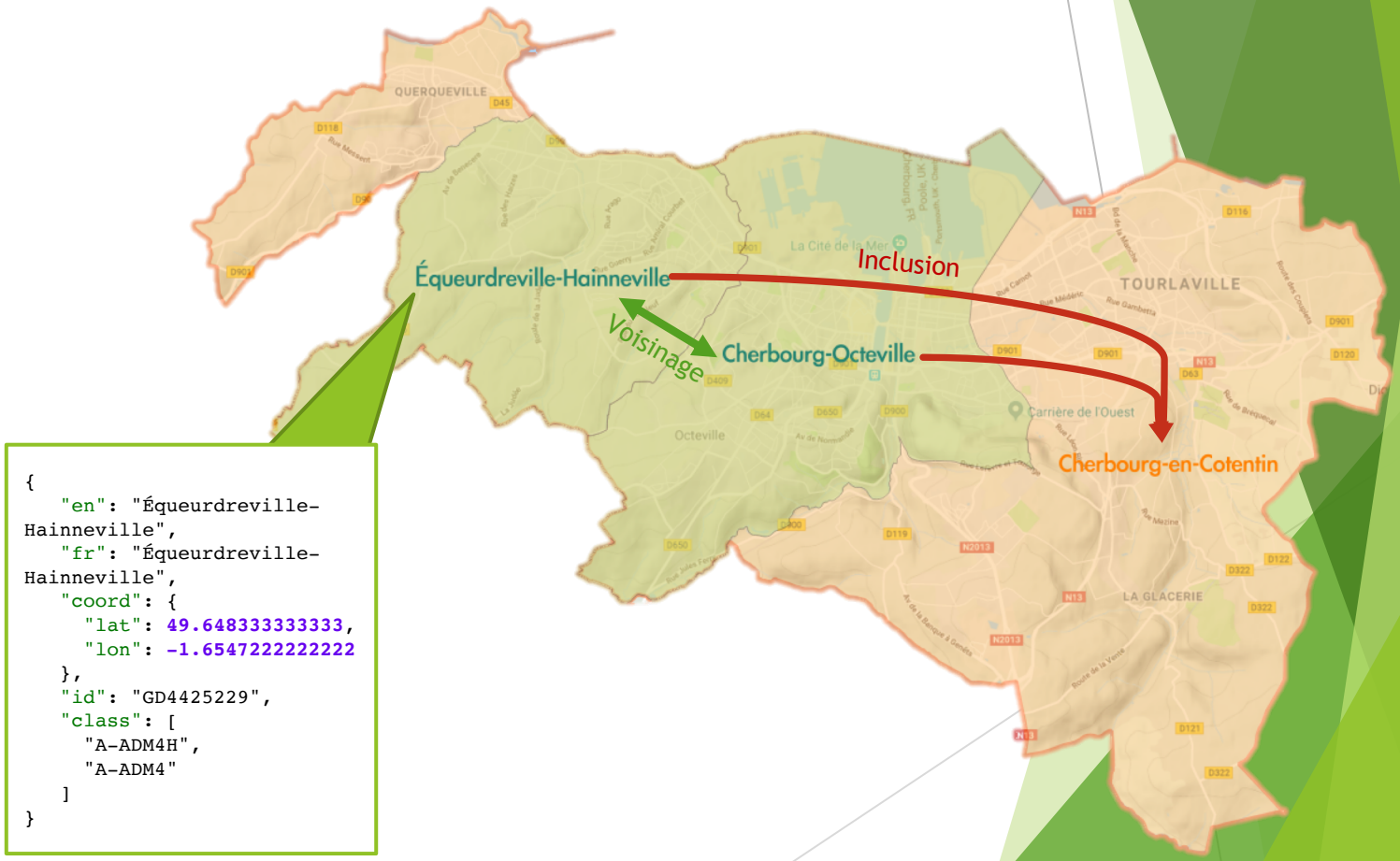
Abstract

This review is a synopsis on coffee quality assessment of green and roasted coffee beans of *Coffea arabica* and *Coffea canephora*. The particle size, medium sieve, most frequent sieve, share split, cumulative calibration, trade homogeneity, mass of 1000 beans, apparent density, strange bodies and defects, mass losses on drying, olfactory and visual parameters, chromatic parameters, soluble solids, pH and chemical characterization (chlorogenic acids, caffeine and trigonelline) is described and evaluated, considering the most important factor associated to the coffee trade, according to a technological perspective.

Spatialité dans les textes

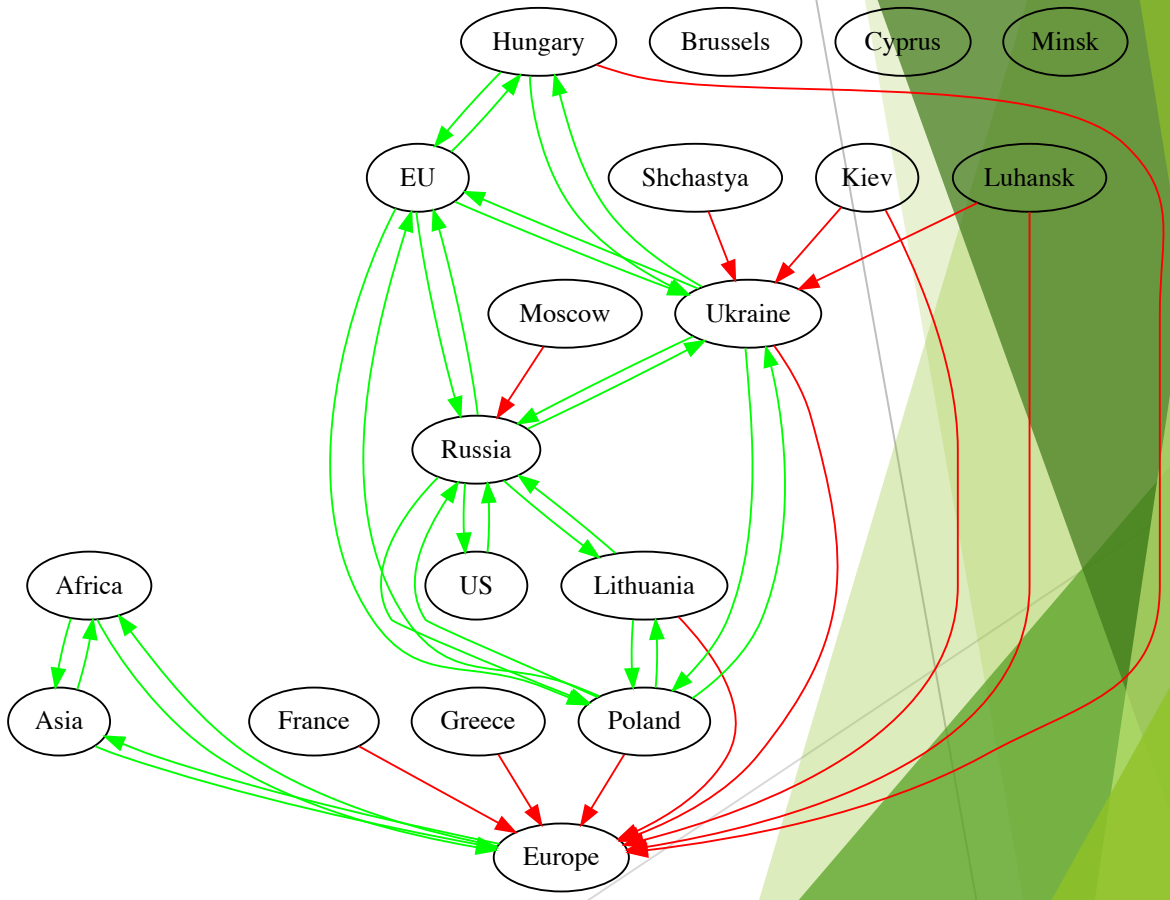
Quelle spatialité ?

- La spatialité dans les textes
  - Entités spatiales : entités localisées dans l'espace
  - Relations spatiales : inclusions, voisinages, …



STR ou Spatial Textual Representation

- Une structure de type **graphe** représentant **la configuration spatiale** d'un document
- **Entités spatiales** → *nommées* et *référencées* dans un jeu de données géographiques ou *gazetier* (Geonames, OSM, Dbpedia, …)
- **Relations spatiales** → inclusion et adjacence
- Perspective
  - Intégrer des relations thématiques entre les entités spatiales



Construction d'une STR

Document

Extraction de texte

Apache Tika

France, 2018. La ville de Paris reçoit le président de Madagascar. […] Le maire de Caen, originaire de la ville de Rots était invité.

Identification des toponymes

Stanford University NLP Natural Language Processing

France, 2018. La ville de Paris reçoit le président de Madagascar. […] Le maire de Caen, originaire de la ville de Rots était invité.

Résolution des toponymes

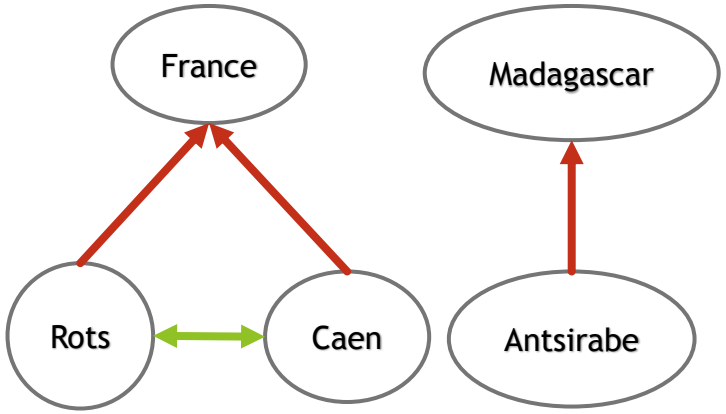
GeoNames

OpenStreetMap

WIKIDATA

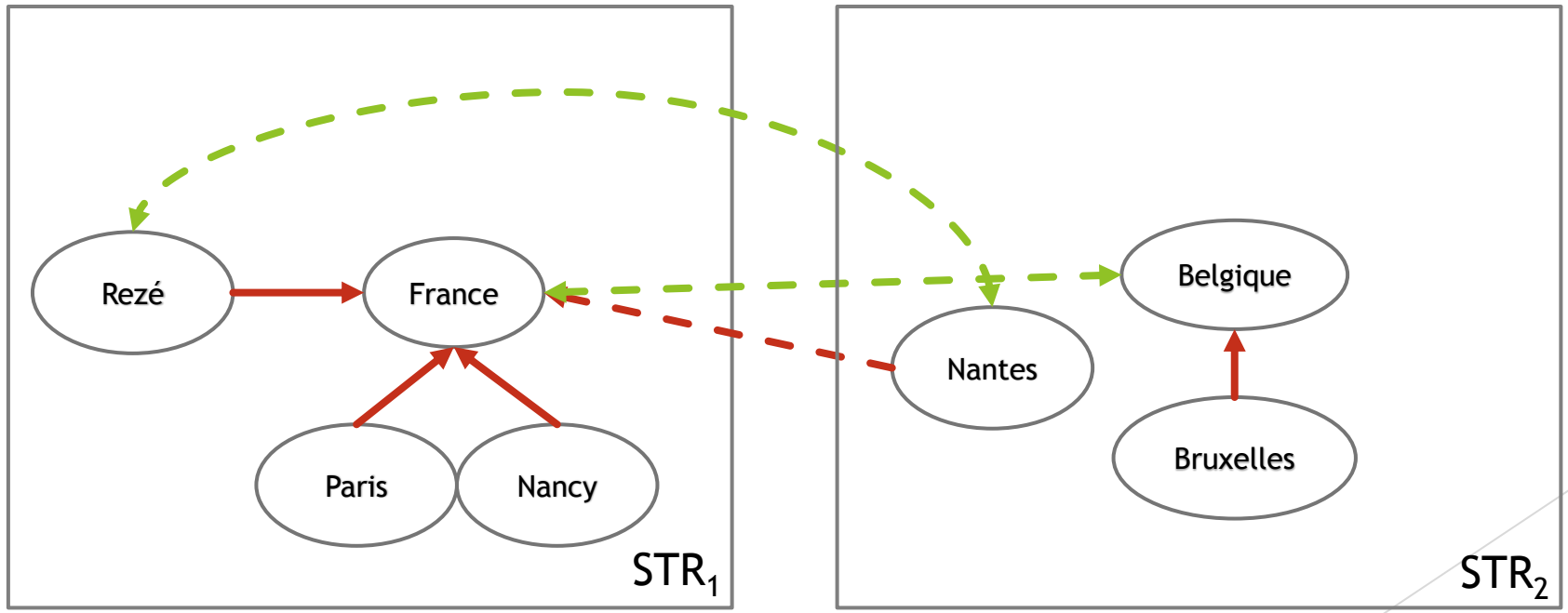
2988507 Paris
1062947 Madagascar
1069166 Antsirabe
…

Construction du graphe de la STR



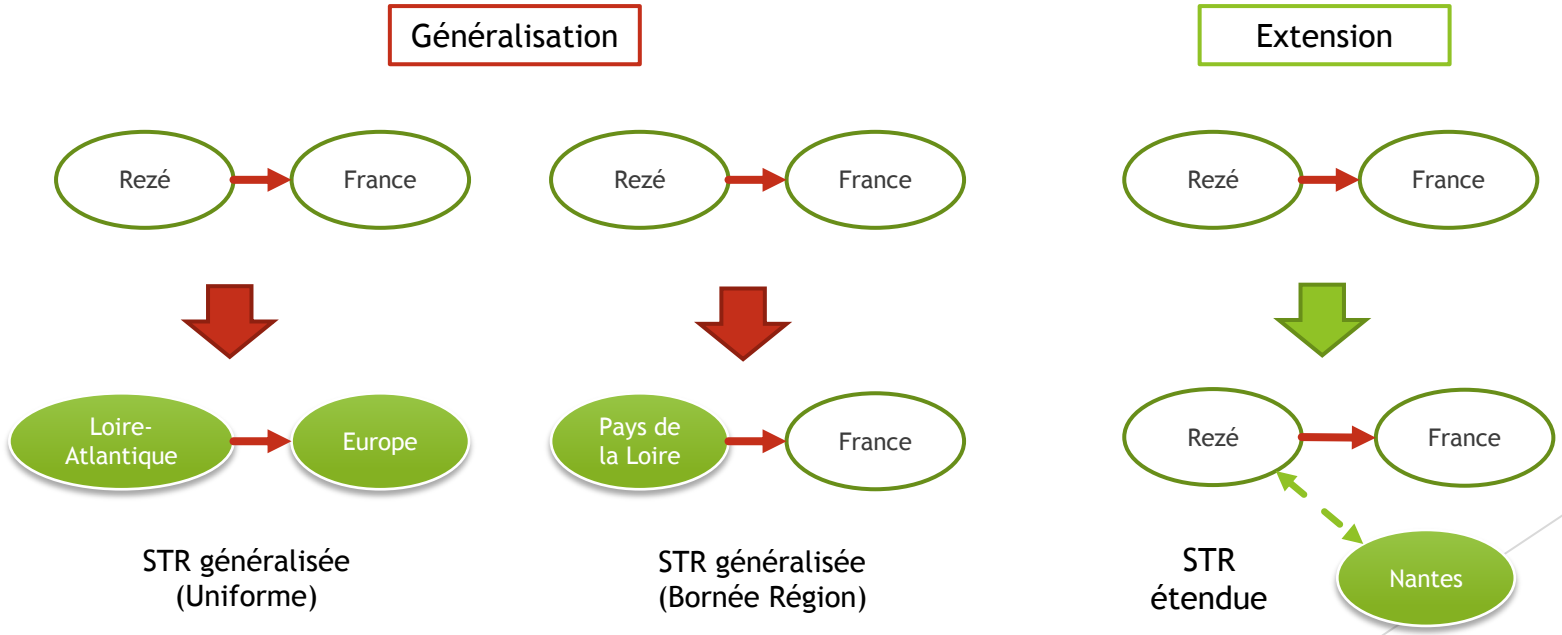
Une information complète ?

- L'information spatiale peut-être incomplète :
  - Référentiel incomplet
  - Élément omis par l'auteur du document (information implicite)
  - Élément manqué durant le processus de création



Différents types de STR

- **Objectif** : Augmenter la probabilité de correspondance entre deux STRs
- Deux transformations → Deux types de STR
  - Généralisation (uniforme et bornée) → STR Généralisée
  - Extension → STR Étendue

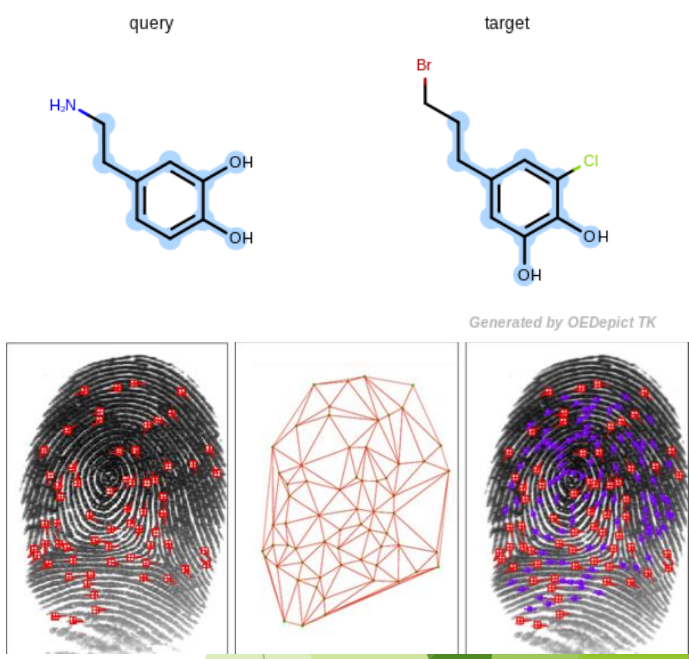


Comparaison spatiale

Comparaison de la spatialité (STR)

- Utilisation d'algorithmes de ***Graph Matching***
  - Mesure de similarité entre graphe
  - Différents algorithmes : MCS, VEO, Graph Edit Distance, Graph Kernels, ...
  - Implémentation de plusieurs algorithmes dans une librairie Python

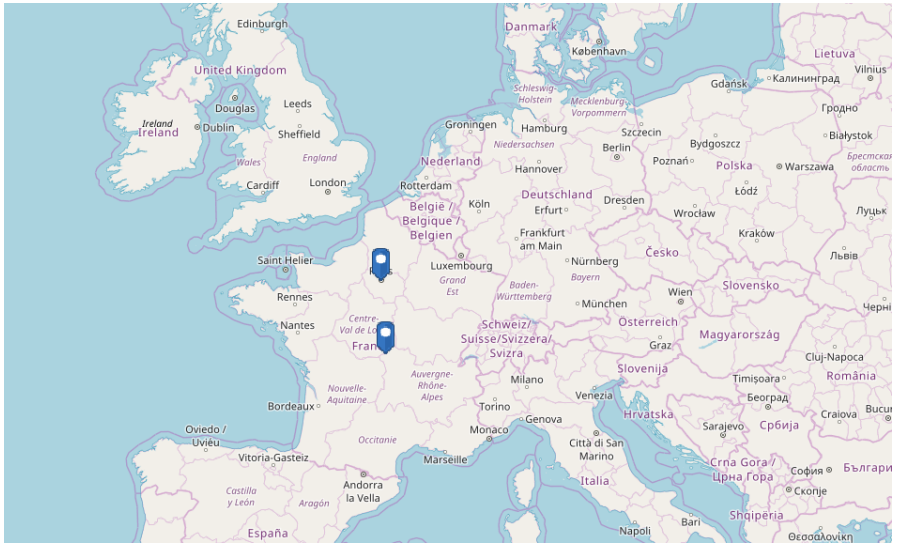
- Comment valider la correspondance entre deux configurations ?
  - Choix binaire → **Risque de biais**
  - Solution : **Utilisation d'un ensemble de critères**
    - Entités partagées ? Entités proches ?
    - Dispersion des entités similaires ?

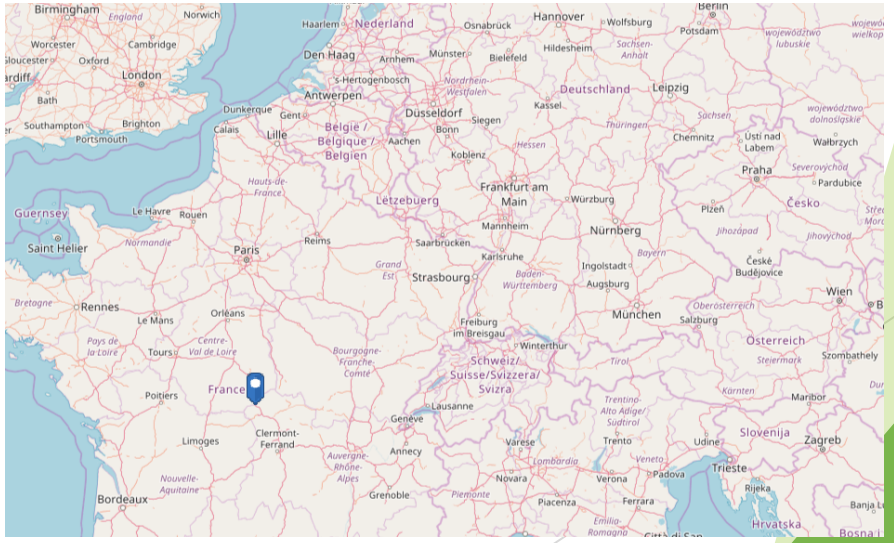


Entités Spatiales Communes (ESC) et Entités Spatiales Proches (ESP)

- Entités spatiales communes
  - Le critère est validé si les deux STR ont des entités spatiales communes.

- Entités spatiales proches
  - Le critère est validé si un couple d'entités spatiales, dans deux STR sont proches.
  - Proximité : relation spatiale ou faible distance géodésique





Emprise Spatiale Caractéristique (**ESSC**)

- Présence de groupes **significatifs d'entités spatiales proches.**
- Le critère est validé si : **Un** ou **plusieurs** de ces groupes appartiennent aux deux STRs.

L’Emprise Spatiale Stricte (ESS)

- Ce critère est validé si les deux STRs partagent une **répartition spatiale identique** de leurs entités spatiales.

Données hétérogènes

Corpus homogène – PADI-Web

- Corpus[1] construit et annoté durant la thèse de E. Arsevska
  - **Élaboration d'une méthode semi-automatique pour l'identification et le traitement des signaux d'émergence pour la veille internationale sur les maladies animales infectieuses**. Santé publique et épidémiologie. Université Paris-Saclay, 2017.
  - Pourquoi utiliser ces données ?
    - Beaucoup d'informations spatiales
    - Documents annotés

1. Rabatel Julien; Arsevska, Elena; de Goër de Hervé, Jocelyn; Falala, Sylvain; Lancelot, Renaud; Roche, Mathieu, 2017, "PADI-web corpus: news manually labeled", doi:10.18167/DVN1/KMTIFG, CIRAD Dataverse, V2

Corpus hétérogènes – AgroMada

- Ensemble de données produites par le CIRAD
  - Données fortement hétérogènes tant sur le :
    - Contenu (langue(s), structure du document, richesse, …)
    - Format (Voir Tableau)
  - Données massives : 13 742 documents (~13 Go)

format	Nombre de doc.
doc	6651
docx	838
html	3465
pdf	1931
ppt	228
pptx	40
sql	1
txt	157
xls	2544
xlsx	126
xml	29

Corpus hétérogènes –AgroMada

- Sélection de documents selon une thématique et un pays
  - **Agroécologie** et **Madagascar**
- Comment ?
  - Construction d'un dictionnaire de termes liés à l'agroécologie
    - Utilisation de sources INRA[1]
    - En français et en anglais
  - Construction d'un dictionnaire contenant les villes principales sur Madagascar
- Corpus résultant de la sélection
  - 5552 documents (anglais, français)

dicoAE
DICTIONNAIRE D'AGROÉCOLOGIE

[1] https://dicoagroecologie.fr/ et https://lovinra.inra.fr/2016/12/20/gestion-des-connaissances-en-agroecologie/

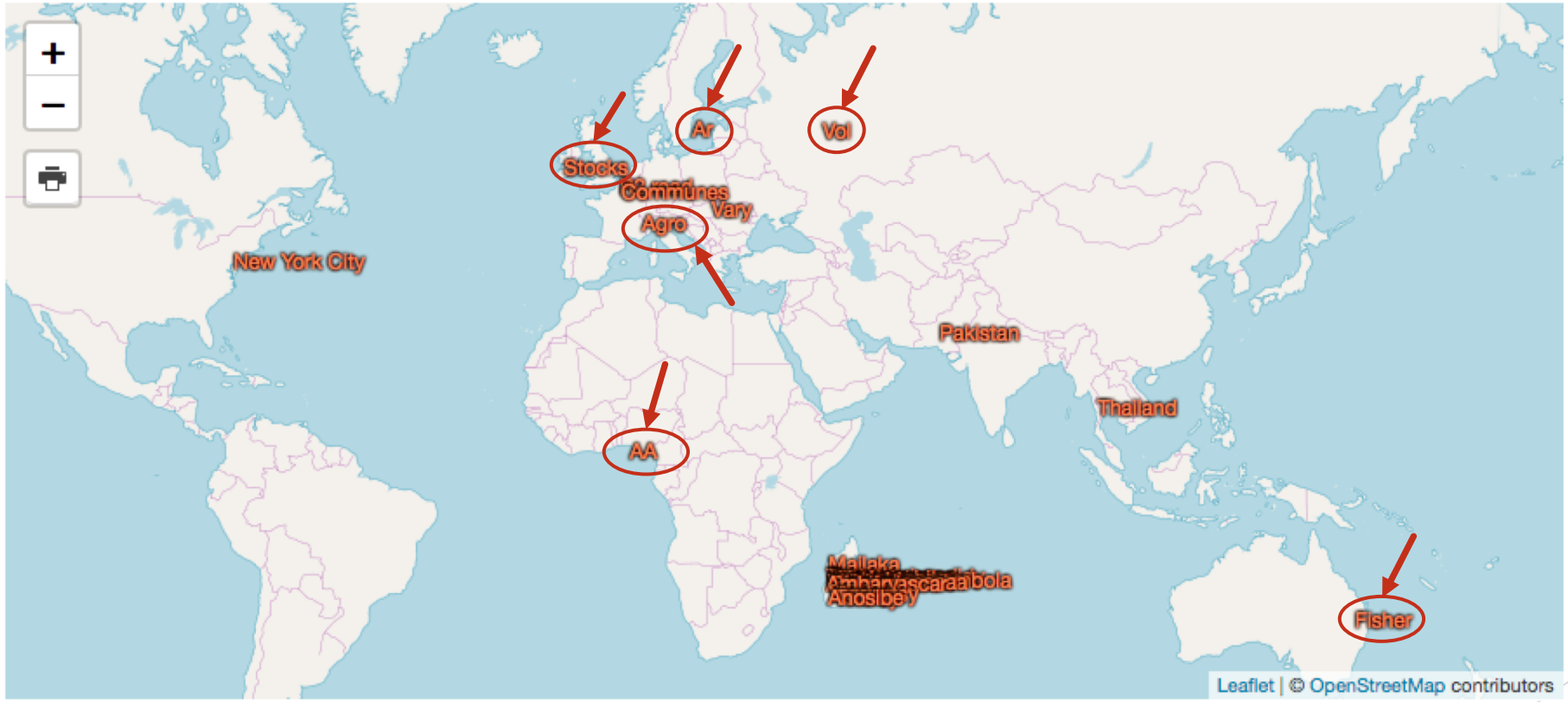
Conséquence de l'hétérogénéité dans le processus de création de la STR

- Outils de reconnaissance d'entités nommées (NER) : StanfordCoreNLP, NLTK, OpenNLP, Spacy …
  - Fonctionne sur des textes homogènes
  - … beaucoup moins sur des sources hétérogènes
- Cause
  - Méthodes basées sur la syntaxe
    - Utilisation de règles
    - Apprentissage → Modèles crées à partir de textes classiques
- Conséquences
  - Toponymes non-identifiés
  - … ou beaucoup de faux-positifs

Corpus Homogène (PadiWeb)			
NER	P	R	F1
StanfordNER	**0.59**	**0.77**	**0.67**
Polyglot	0.53	0.72	0.61
NLTK	0.42	0.66	0.52
Spacy	0.40	0.65	0.50

Corpus Hétérogène (Agromada)			
NER	P	R	F
StanfordNER	**0.31**	0.16	**0.22**
Polyglot	0.20	0.35	0.26
NLTK	0.13	0.15	0.14
Spacy	0.14	**0.84**	0.25

Exemple


Ar
Vol
Stocks
Communes
Vary
Agro
New York City
Pakistan
Thailand
AA
Fisher
Leaflet | © OpenStreetMap contributors

Expérimentations

Expérimentation

- Objectif : Définir **quelles mesures de similarité** et **types de STR** donnent les meilleures correspondances sur la dimension spatiale.
- Deux corpus utilisés
  - Homogène : brèves traitant d'épidémies animales variées
  - Hétérogène: documents produits à Madagascar (ppt, xls, doc, pdf) traitant de thématiques de recherche variées liées au projet BVLAC[1]
- Mesure de performance
  - **Précison@n:** pourcentage des STR **n**-plus similaires qui valide le critère (pour une STR)

1. Projet de Mise en Valeur et de Protection du Bassin Versant du Lac Alaotra)

PADI-WEB - Résultats

- Mesures respectant le mieux les critères, selon chaque type de STR (**Précision@3**)

Type	ESC	ESP	ESSC	ESS
Normal	BOC	MCS	MCS	BOWSE
Gen_country	VEO	VEO	VEO	VEO
Gen_region	VEO	MCS	MCS	MCS
Extension 1	MCS	MCS	MCS	VEO

- Scores associés aux meilleurs mesures indiquées précédemment (**Précision@3**)

Type	ESC	ESP	ESSC	ESS
Normal	0,872	0,772	0,424	0,376
Gen_country	0,756	**0,872**	**0,48**	0,368
Gen_region	**0,904**	0,796	0,464	**0,396**
Extension 1	0,896	0,796	0,456	0,384

- MCS, VEO, BOC donne les meilleurs correspondances entre STR
- Des mesures plus complexes donnent des résultats plus faibles
  - Nature « éparse » des graphes
- Les transformations de STR améliore de manière significative la qualité des correspondances

Données Madagascar - Résultats

- Mesures respectant le mieux les critères, selon chaque type de STR (**Précision@3**)

Type	ESC	ESP	ESSC	ESS
Normal	BOWSE/VEO	MCS	BOC	BOC
Gen_country	MCS	MCS	MCS	BOC
Gen_region	MCS/BOWSE/VEO	MCS	BOC	BOC
Extension 1	VEO	VEO	BOC	BOC

- Scores associés aux meilleurs mesures indiquées précédemment (**Précision@3**)

Type	ESC	ESP	ESSC	ESS
Normal	0,95	0,98	0,71	**0,35**
Gen_country	0,93	0,95	0,64	0,34
Gen_region	**0,95**	0,97	0,70	0,34
Extension 1	0,95	**0,99**	**0,73**	0,31

- MCS, VEO, BOC donne les meilleurs correspondances entre STR
- La transformation améliore la correspondance entre les STR
  - Moins significatif

Conclusion

- Proposition d'une représentation dédiée à la spatialité dans les textes
  - STR, une structure graphe composée d'entités spatiales liées par leur relation spatiale (adjacence, inclusion)
  - Utilisable sur des données textuelles hétérogènes
- Similarité spatiale entre documents
  - Des résultats encourageants !
- Perspectives
  - Ajout de la thématique dans la STR

Merci pour votre attention !