

Mining Textual data for agriculture and environment

Mathieu Roche

Cirad – TETIS – Montpellier, France



Web : <http://www.textmining.biz>

Email : mathieu.roche@cirad.fr



Data Science

Volume

Velocity

Variety

3V of Big Data



**Variability, Véracity, Value,
Visualisation, Valorization**

→ **Pluridisciplinary** domain



Data Science

Volume

Velocity

Variety

3V of Big Data



**Variability, Véracity, Value,
Visualisation, Valorization**

→ **Pluridisciplinary** domain



Songes Project



Textual Data Science

```
0/1 x B_1404 [WARNING]: "Asynchronous reset/set/load <%item> exists in module/unit"
0/1 x B_1405 [WARNING]: "<%value> asynchronous resets in this unit detected"
0/1 x B_1406 [WARNING]: "<%value> synchronous resets in this unit detected"
0/1 x B_1407 [ERROR]: "Do not use active high a
```

```
// Total Module Instance Coverage Sum
```

```
lines
state
```

```
Policy:
<v>
---
0/1
```



Ma famille est très pauvre. On cultive l'igname toute l'année mais on ne gagne pas assez d'argent. Mon père dit que je suis courageux et que je peux réussir en Europe. C'est ainsi qu'après la récolte de l'igname il a réuni l'argent pour me donner pour voyager. C'est de mon père que j'ai eu l'information sur cette route. Je suis venu à Bamako chez des compatriotes. Ils étaient suffisamment informés sur la route. J'ai eu là tous les détails concernant sur la route. Le chef a contacté un passeur à Gao et j'y suis allé à son adresse. J'ai payé 45.000 FCFA et il m'a donné un passeport Malien qui m'a permis de passer la frontière Algérienne. Ceux qui n'avaient de passeport ont pas 500 CFA pour passer. Le voyage m'a épuisé parce que le véhicule 4x4 qui nous a transporté jusqu'en Algérie était très petit pour le monde qu'il contenait. Le voyage est la chose la plus difficile que j'ai connue dans ma vie. La route est mauvaise mais le conducteur roule à vive allure. Les gens crient de souffrance mais eux s'en foutent. J'ai passé dans les foyers à chaque étape. Ces séjours nous permettaient d'organiser les étapes suivantes. Les compagnons ne sont pas du tout solidaires. J'étais le seul Burkinabé et j'étais isolé. Mon projet c'est de pouvoir rentrer en Espagne. Je n'ai plus d'argent pour continuer donc je travaille pour en avoir. A chaque étape, je dois travailler pour avancer. Je ne dois pas décevoir mon père. La police ne m'a pas pris. Mais j'ai peur d'aller au refoulement. Je n'ai pas d'avenir en Afrique. En Europe on peut travailler pour se faire un avenir.

a guerre commença, baon il a tot le
e: et bien i parut, kar il fu a vint et cinc
puis conquist il .XXVIII. roialmes [72d]
ns fu la fin de son aag. Mais de t
st Lancelos arie il li m... b
grant honor sa... honte... nt il
e roi Artu et il li ala merci...



Pluridiciplinary projects (Mastodons)

Textual data and satellite images – ANIMITEX (2013-2014)

Vakinankaratra – L'agriculture de conservation lancée

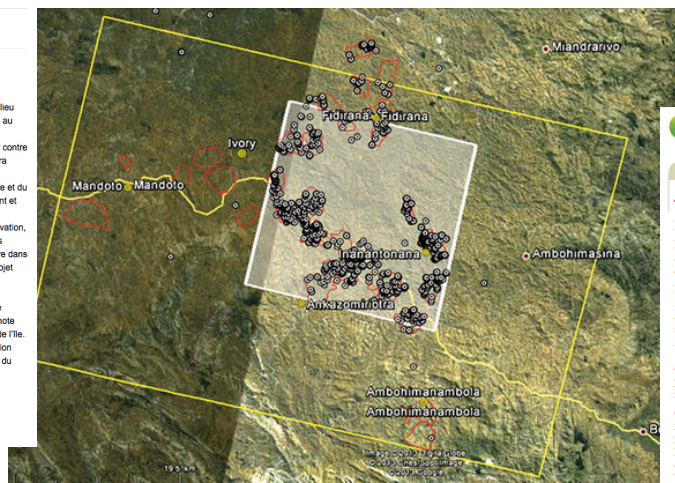
17.12.2014 | 7:18 | Non classé | 0



L'agro-écologie est une nécessité. Plus de 80% de la population malgache vit en milieu rural et opère en général dans l'agriculture. La croissance démographique associée au changement climatique provoque une forte destruction de l'environnement et une dégradation alarmante de la fertilité des sols. Afin d'y faire face et pour mieux lutter contre la malnutrition, le Groupement semis direct de Madagascar lance le projet Manitra dans quatre communes rurales du district de Betafo et de Mandoto, dans la région Vakinankaratra. Ce projet est réalisé en partenariat avec le ministère de l'Agriculture et du développement rural et sur financement de l'Association française du développement et du Comesa.

Le groupement qui focalise son activité sur l'agro-écologie et l'agriculture de conservation, sensibilise et incite les paysans des communes ciblées à pratiquer l'agriculture sous couverture végétale et la rotation culturale. Et afin d'assurer une sécurité alimentaire dans la commune rurale d'Ankazomiratra, d'Inanankaratra, de Vinany et de Fidirana, le projet Manitra compte adhérent 1000 paysans, dont 200 femmes, sur la pratique de ce système de culture agro-écologique qui ne nécessite pas de nombreux travaux et éreintant comme l'exige le labourage. « Il suffit que les paysans recouvrent le sol de végétaux et cultivent sans dépenser du temps et de l'argent pour l'achat d'outils », note Rakotondramanana, directeur exécutif du projet qui s'active aussi dans le Sud-Est de l'île. Des formations sur la régénération de la fertilité du sol et la lutte contre sa dégradation ainsi que l'introduction du système des légumineuses seront la priorité des activités du projet.

Angela Ny Avo






Archive ouverte des publications du Cirad

Recherche avancée Auteurs CIRAD Parcours Déposer

Se connecter Crise hier, opportunités aujourd'hui, défis pour demain : le cas de la filière riz à Madagascar : [Draft]

Les facteurs de la crise 2004-2005 sur le marché du riz

La manifestation la plus visible de la situation du marché du riz en 2004 et début 2005 est une augmentation sans précédent des prix de détail. Les causes de cette crise sont une conjonction de plusieurs facteurs, internes et externes : physiques, monétaires et politiques.

L'an dernier, les prix du riz ont normalement augmenté pendant la période de soudure, fin 2003 début 2004, mais ne sont pas descendus en période de récolte, ils ont continué à augmenter à un rythme soutenu. La variation annuelle du prix du paddy entre récolte et soudure est habituellement de l'ordre de 50% au Lac Alaotra, elle a été de 150% en 2004-2005 (Mintin et Raison, 2005). Le prix du riz national on importé dépassait historiquement 1000 Ar le kg entre septembre 2004 et février 2005 sur les marchés de la capitale. Si on compare l'évolution des prix du riz en 2001 et en 2004, on peut se rendre compte que les trois premiers mois de l'année il coûtait moins cher en 2004 qu'en 2001 et 2.5 fois plus cher en novembre.

Cette hausse des prix s'est généralisée dans tout le pays. Elle s'est répercutée dans l'espace : marchés urbains et ruraux, auprès de tous les agents de la filière et pour toutes les variétés de riz (vary gazy, makalioka, tsipala, riz pluvial...). A titre d'exemple, dans le Moyen-Ouest, la hausse des prix du riz a été aussi importante sur les marchés situés en bord de route nationale que sur les marchés plus enclavés comme Inanankaratra (45 km de piste en saison sèche), Vasiara (1h15mn), Mahasolo (2h30mn) ou Ambalamirana (4h).

Textual data and data bases – QDoSSI (since 2016)



Ma famille est très pauvre. On cultive l'igname toute l'année mais on ne gagne pas assez d'argent. Mon père dit que je suis courageux et que je peux réussir en Europe. C'est ainsi qu'après la récolte de l'igname il a réuni l'argent pour me donner pour voyager. C'est de mon père que j'ai eu l'information sur cette route. Je suis venu à Bamako chez des compatriotes. Ils étaient suffisamment informés sur la route. J'ai eu là tous les détails concernant sur la route. Le chef a contacté un passeur à Gao et j'y suis allé à son adresse. J'ai payé 45.000 FCFA et il m'a donné un passeport Malien qui m'a permis de passer la frontière Algérienne. Ceux qui n'avaient de passeport ont pas 500 CFA pour passer. Le voyage m'a épuisé parce que le véhicule 4x4 qui nous a transporté jusqu'en Algérie était très petit pour le monde qu'il contenait. Le voyage est la chose la plus difficile que j'ai connue dans ma vie. La route est mauvaise mais le conducteur roule à vive allure. Les gens crient de souffrance mais eux s'en foutent. J'ai passé dans les foyers à chaque étape. Ces séjours nous permettaient d'organiser les étapes suivantes. Les compagnons ne sont pas du tout solidaires. J'étais le seul Burkinabé et j'étais isolé. Mon projet c'est de pouvoir rentrer en Espagne. Je n'ai plus d'argent pour continuer donc je travaille pour en avoir. A chaque étape, je dois travailler pour avancer. Je ne dois pas décevoir mon père. La police ne m'a pas pris. Mais j'ai peur d'aller au reflux. Je n'ai pas d'avenir en Afrique. En Europe on peut travailler pour se faire un avenir.





Outline

- Part 1** Data Science and Big Data
- Part 2** Information extraction in documents
- Part 3** Sentiment Analysis
- Part 4** QDoSSI project
- Part 5** Future Work



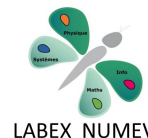


Part 2

Information extraction in documents

Animal disease surveillance

In collaboration with CMAEE lab
(Control of exotic and emerging animal diseases)





Why the need of Epidemic Intelligence?

More than **60% of the initial outbreak reports** come from unofficial informal and **heterogeneous sources**, including sources other than the electronic media, which **require verification** [Arsevska *et al.* ISVEE'2015]

INTERNATIONAL BUSINESS TIMES

MONDAY, JUNE 01, 2015 AS OF 2:24 PM CDT

Home Politics Economy Markets / Finance Companies Technology

TECHNOLOGY SCIENCE

Unknown Disease Kills Kazakhstan's Rare Saiga Antelopes, Scientists Baffled

By Kukil Bora [@KukilBora](#) on May 30 2015 7:21 AM EDT

News



African Swine Fever in Three Lithuanian Wild Boar

18 May 2015

LITHUANIA - Three wild boar found at two locations were confirmed with African swine fever last week.



Mysterious disease kills Nigerian patients within a day

The unknown disease has so far killed 17 people in a southeastern Nigerian town and officials have ruled out Ebola.

18 Apr 2015 21:32 GMT | [Health](#), [Nigeria](#), [Africa](#)



How to detect disease outbreak on the Web?

- **Four animal disease models:** African swine fever (ASF), Foot-and-mouth disease (FMD), Bluetongue (BTV), and Schmallenberg virus (SBV), Influenza
- **First studying model:** ASF



Industries | Wed Jul 23, 2014 9:54am EDT

Related: NON-CYCLICAL CONSUMER GOODS

Poland investigates suspected case of African swine fever in farm pigs

WARSAW, JULY 23

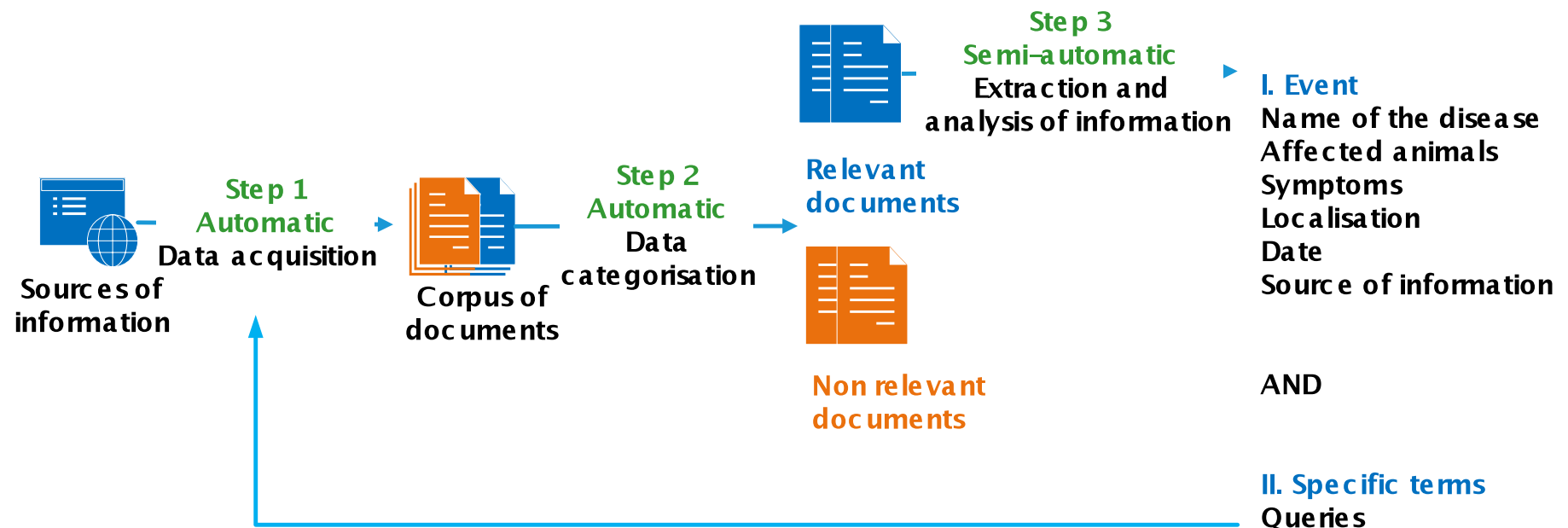
Polish local authorities said on Wednesday that preliminary tests have pointed to a case of African swine fever (ASF) among farm pigs in eastern Poland near the city of Bialystok.

The head of the Grodek county, Wieslaw Kulesza, told Reuters that preliminary results of tests showed that ASF was the cause of death of two-three farm pigs in the county.

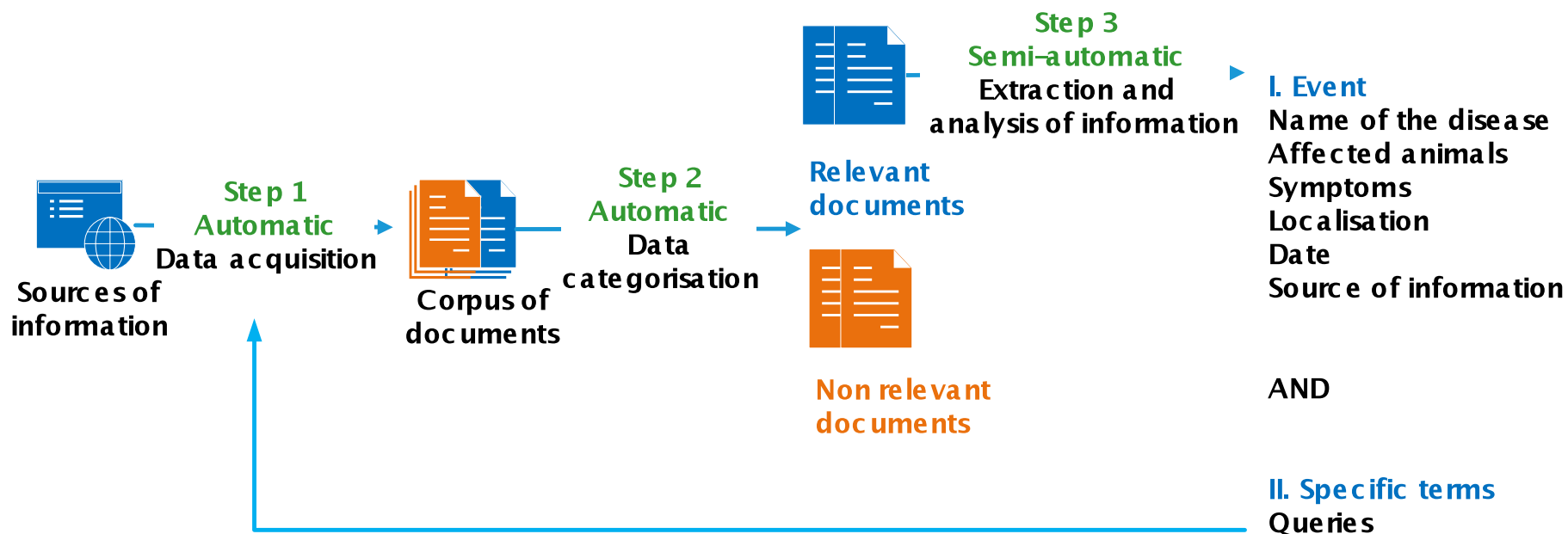
"We are marking the area," Kulesza said, adding that further steps, such as laying special mats, were being taken.

Poland's chief veterinary officer was unavailable for comment, while the county veterinary officer said a statement on the issue will be published later on Wednesday. (Reporting by Anna Wlodarczyk-Semczuk; Writing by Marcin Goettig)

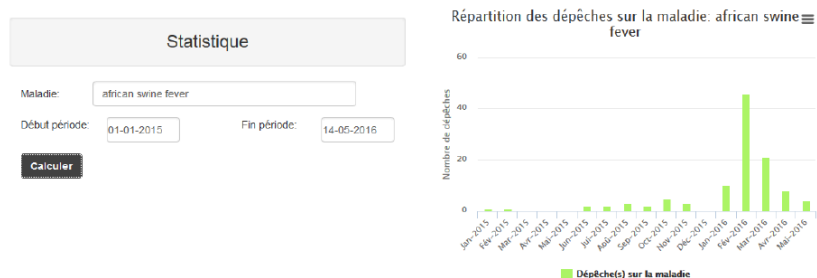




• Step 1: Data acquisition



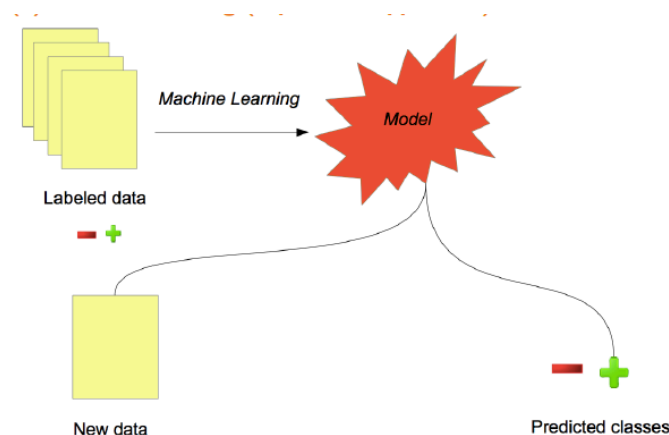
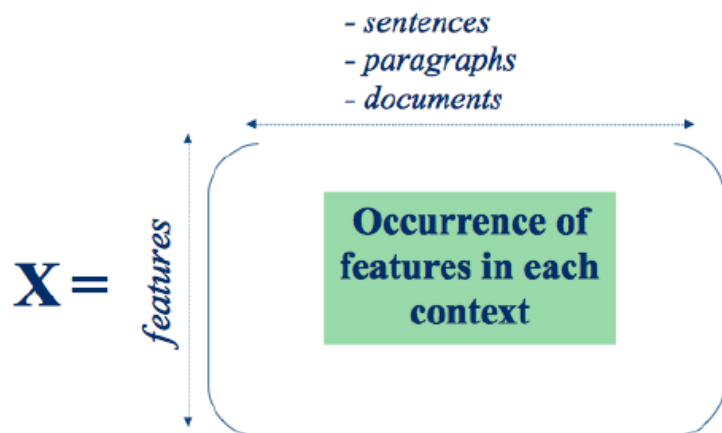
<https://news.google.com/news/feeds?pz=1&cf=all&ned=en&q=Blue+tongue&output=rss>



[Falala *et al.* IC-Demo'2016]



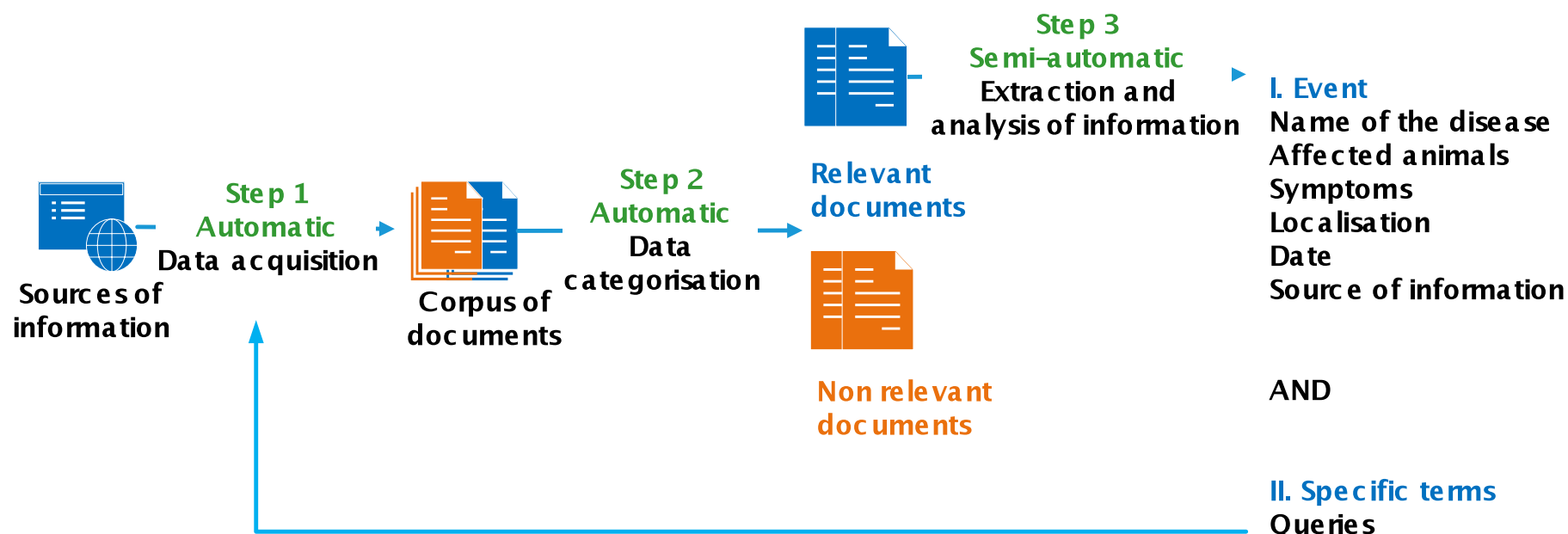
- Step 2: Data classification



Classification algorithm	Naïve Bayes			Support Vector Machine		
Performance	Recall	Precision	F-score	Recall	Precision	F-score
Class <i>disease</i>	0.724	0.766	0.744	0.657	0.68	0.669
<i>economy</i>	0.478	0.530	0.503	0.489	0.726	0.584
<i>general</i>	0.860	0.804	0.831	0.864	0.763	0.810
Weighted average	0.750	0.745	0.747	0.732	0.729	0.725



- Step 3: Information extraction and management



- **Step 3: Information extraction (I)**

Aim: Automatically detecting **key information** from **Web** news articles (country, species, diseases, number of cases, dates, ...)

“Since its initial appearance in **Poland** in **February 2014**, **72** cases of **African Swine Fever** have been detected in **wild boars** and there have been three outbreaks in **pigs**.” - <http://www.thenews.pl>

- Use **dictionaries** (Geonames, Heidelberg, disease names, species names, etc.), and data mining techniques in order to learn extraction rules

Rules associated with ‘case numbers’:

(number)(species_name,1-3) with frequency **26%** and confidence **83%**

(number)(species_name,1-2) with frequency **21%** and confidence **100%**



- **Step 3: Information extraction and disambiguation (I)**

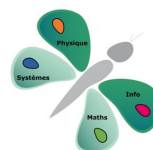
Spatial features taken into account in the learning approach:

- frequency of spatial entity (SE) candidate in the document related to candidates
- frequency of SE associated with the country of the candidate
- situation of the candidate in the documents
- number of candidates in the documents



- **Step 3: Information extraction (I)**
 - **Results for the rule-based approach on the annotated corpus**
 - **Classification based on SVM** (features are rules)
 - **3 classes:** correct, incorrect, partial
 - 10-fold cross validation

Type	Accuracy (%)
Locations	70.6
Dates	71.2
Diseases	93.6
Cases	78.1
Species	89.5



LABEX NUMEV

**Julien Rabatel,
LIRMM, Numev, France**



• Step 3: Information extraction (I)

Veille Sanitaire Internationale

Projet VSI | Accueil | Consultation ▾ | Paramétrage ▾

Annotation

Vous êtes connecté en tant que Elena ✕ Déconnexion

Jeux de données ▾

annotations

532 article(s) ▾ | Article précédent | Article suivant ▸

1. Rivers govt. eliminates chickens infected by flu 1 / 532 ✓ Tout a été annoté → Nouveau cas Bilan Non-pertinent

Date: 2015-01-20

The 🇳🇮 Rivers State Government said on 📅 Tuesday that it had killed # hundreds of 🐔 fowls infected by the 🦠 Avian Flu in a privately owned farm in 📍 Port 🇳🇮 Harcourt .
 The Commissioner for Agriculture , Emma 🇳🇮 Chinda , said that the farm had been quarantined and decontaminated .
 He also said no human infection had been recorded .
 " On 📅 January 🇳🇮 14 , we got a report from a farm that was worrisome . The report we got suggested that the farm may have been infected by the 🦠 highly pathogenic 🦠 avian influenza .
 According to the commissioner , samples of the flu were taken to the Veterinary Research Institute in Vom , 🇳🇮 Plateau State .
 " The result came out on 📅 January 🇳🇮 17 and it read positive of 🦠 highly pathogenic 🦠 avian influenza .
 " On the basis of that , we had to take necessary steps . Apart from quarantining the farm , we had to depopulate the 🐔 birds in the farm to stop further spread " .
 " Thereafter , we decontaminated the farm . We are containing the situation because officials of government and experts are on ground monitoring the situation " , he added .
 Mr. 🇳🇮 Chinda said there was no need to panic because government was well equipped to handle the situation .
 He said before the outbreak , they received information from the 🇳🇮 Federal Ministry of Agriculture on 🦠 avian influenza in 🇳🇮 Kano and a 🐔 bird market in 🇳🇮 Lagos .
 " We were very much on alert and when it happened here , we handled the situation " , he said .
 (NAN)

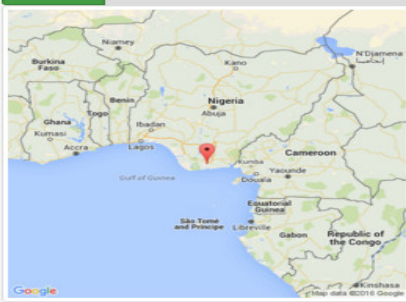
29 candidat(s) | Candidat précédent | Candidat suivant ▸

6 / 29

LOCATION

Port Harcourt

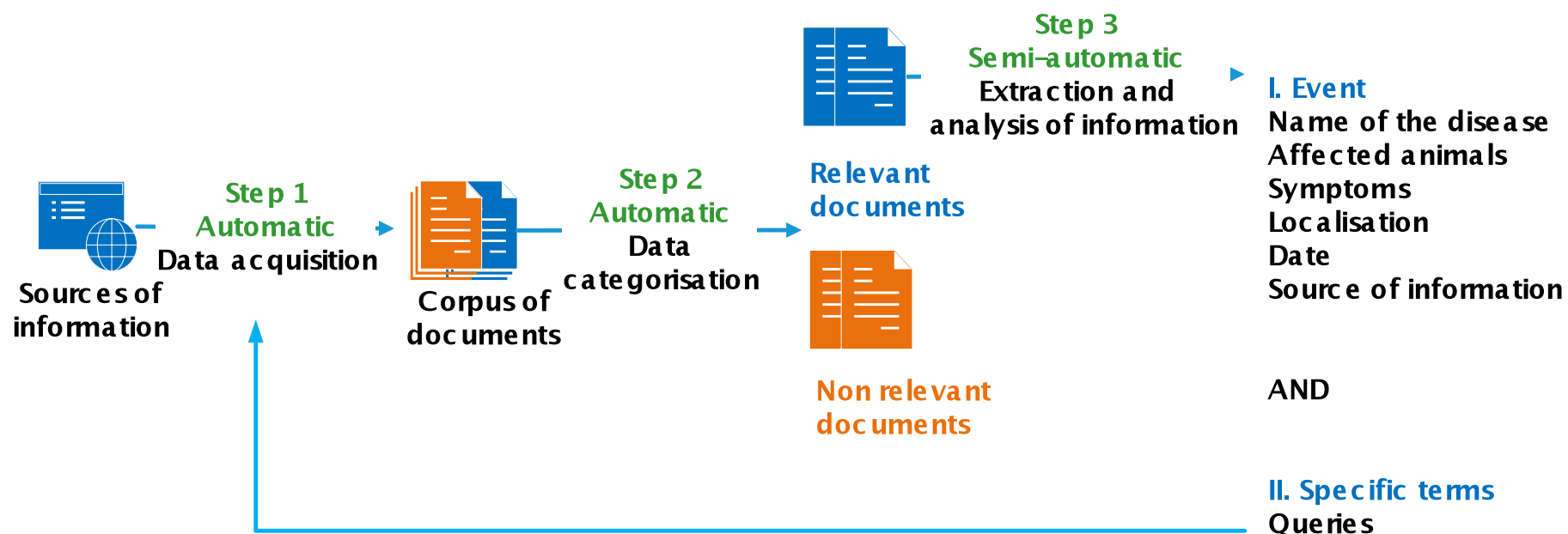
✓ Correct | 🇳🇮 Partiel | ✕ Incorrect | Non annoté



◀ Article précédent | Article suivant ▶

Max Devaud-Thomas Filloi Copyright 2015 All Right Reserved.

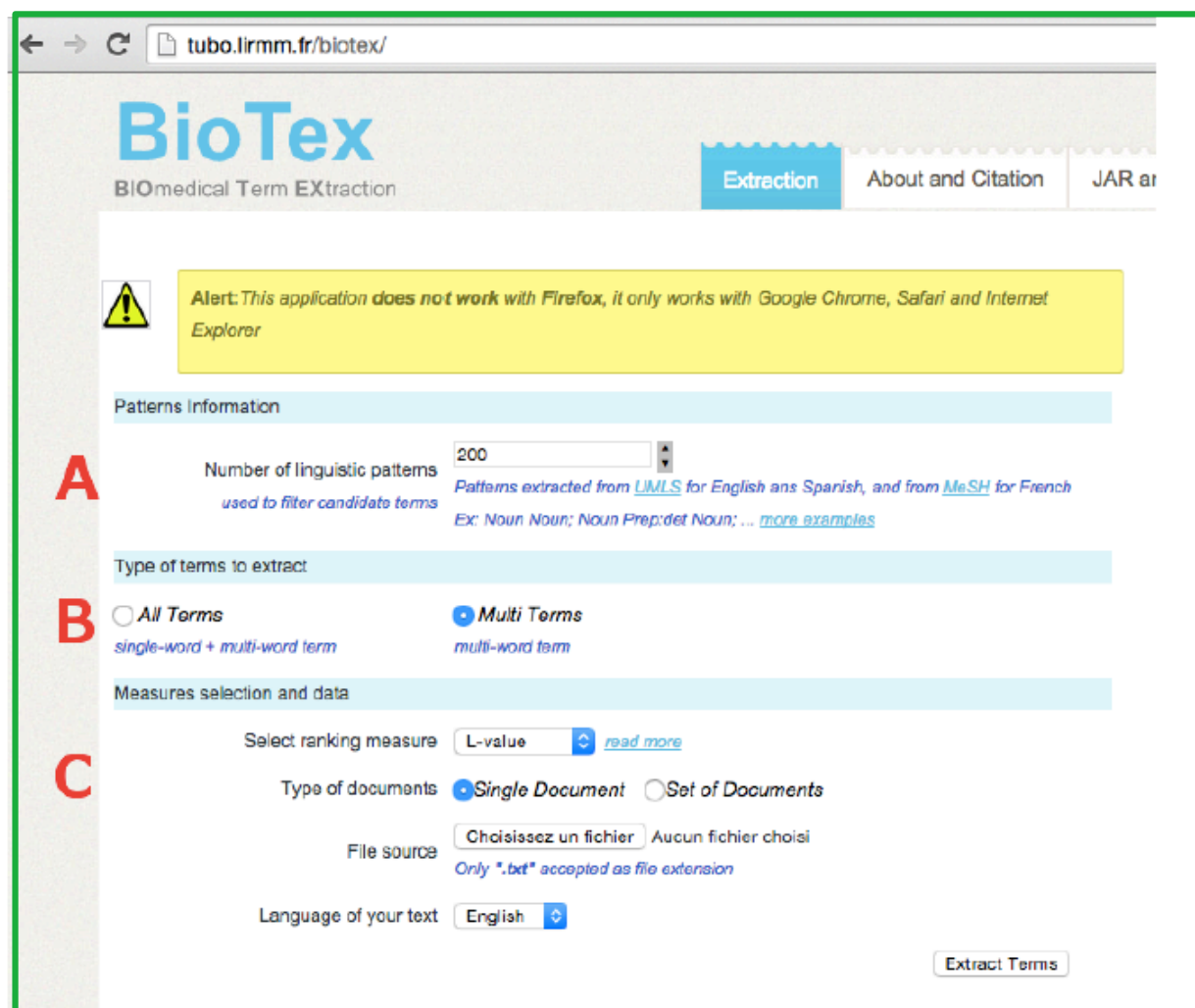
- Step 3: Information management (II)



- II. Querying the Web: (a) **Terminology extraction**

[Lossio Ventura *et al.* ISWC-Demo'2014]

SIFR
project

The screenshot shows the BioTex web application interface. The browser address bar displays 'tubo.lirmm.fr/biotex/'. The page has a header with the 'BioTex' logo and the subtitle 'BIOmedical Term EXtraction'. Navigation tabs include 'Extraction' (active), 'About and Citation', and 'JAR at'. A yellow alert box states: 'Alert: This application does not work with Firefox, it only works with Google Chrome, Safari and Internet Explorer'. The main form is divided into sections: 'Patterns information' with a dropdown set to '200' and a description of linguistic patterns; 'Type of terms to extract' with radio buttons for 'All Terms' and 'Multi Terms' (selected); and 'Measures selection and data' with a dropdown for 'L-value', radio buttons for 'Single Document' (selected) and 'Set of Documents', a file source selection, and a language dropdown set to 'English'. An 'Extract Terms' button is at the bottom right. Red letters A, B, and C are overlaid on the left side of the form sections.

- II. Querying the Web: (b) **Terminology ranking**

- BioTex Ranking [Lossio Ventura *et al.* IRJ'2016]:

$$LIDF-value(t) = P(t_{dom}) \times IDF(t) \times C-value(t)$$

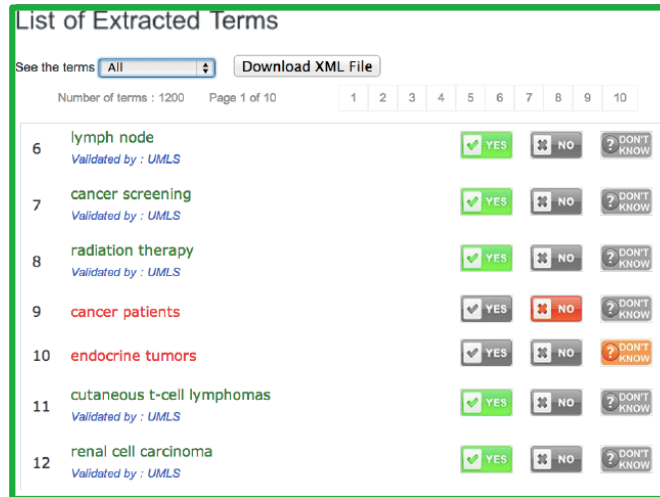
- A new ranking function to take into account the **heterogeneity** of the sources (S_i) [Arsevska *et al.* CEA'2016]:

$$w(t) = \sum \alpha_i \times \frac{1}{rank_{S_i}(t)}$$

$$\text{with } \alpha_i \in [0,1] \text{ and } \sum \alpha_i = 1$$



- II. Querying the Web: (c) *Terminology validation*



Term	Validated by	YES	NO	DON'T KNOW
6 lymph node	Validated by : UMLS	☒	☐	☐
7 cancer screening	Validated by : UMLS	☒	☐	☐
8 radiation therapy	Validated by : UMLS	☒	☐	☐
9 cancer patients		☐	☒	☐
10 endocrine tumors		☐	☐	☒
11 cutaneous t-cell lymphomas	Validated by : UMLS	☒	☐	☐
12 renal cell carcinoma	Validated by : UMLS	☒	☐	☐

Using of a **Delphi method** [Arsevska *et al.* LREC'2016]

Delphi method is to reach group consensus with experts (5 to 7 experts for each disease) when knowledge is not sufficient for a given scientific question



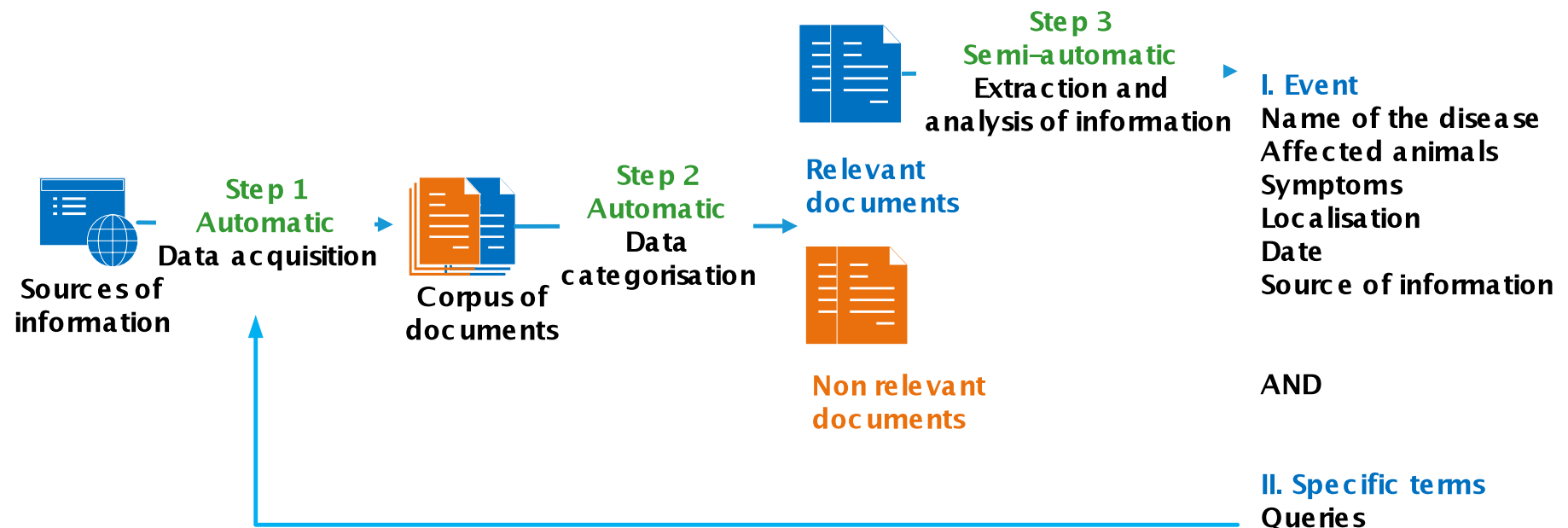
- II. Querying the Web: (d) **Association of terms**

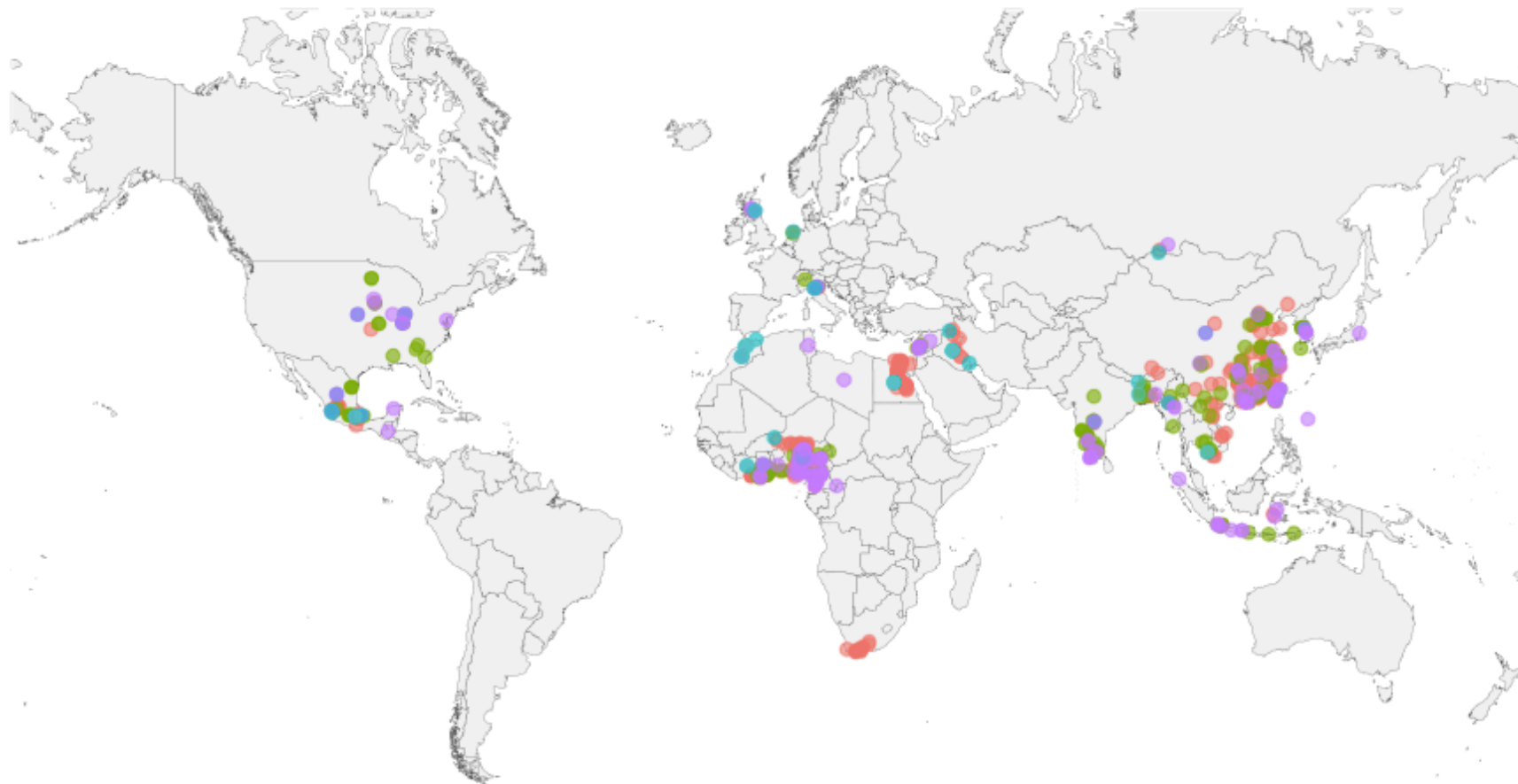
$$D_{web}^{AND} = \frac{2 \times \text{hit}(h \text{ AND } cs)}{\text{hit}(h) + \text{hit}(cs)}$$

[Roche and Prince Informatica'2010 ; Arsevska *et al.* IJAEIS'2016]

Rank	Bluetongue <i>hosts / clinical signs</i>	Schmallenberg virus infection <i>hosts/ clinical signs</i>
1	general clinical signs / pregnant ewes	stillborn bovine foetuses / camels
2	livestock deaths / sheep	stillborn bovine foetuses / bison
3	embryonic death / cow	aborted foetuses / sheep
4	general clinical signs / sheep	deformed offspring / sheep
5	livestock deaths / cow	stillborn bovine foetuses / deer
6	livestock deaths / deer	aborted foetuses / cattle
7	fever outbreak / sheep	deformed offspring / cattle
8	embryonic death / sheep	stillborn bovine foetuses / calves
9	fever outbreak / cow	deformed offspring / lambs
10	embryonic death / pregnant ewes	acute bronchopneumonia / bison







Source ● Empres-i ● VSI-web ● HealthMap|ProMED ● HealthMap|Autre



Scientific results in Epidemiology:

- **Spatial detection** larger than official surveillance (e.g. PPA detected in Ouganda in February and March 2016)
- Detection of **important information** related to diseases not given by official surveillance systems (e.g. preventive vaccinations)
- Detection of information about **other diseases** (e.g. SBV)

Scientific results in Computer Science:

- New **visualisation** systems
- New **ranking measures** to extract keywords from texts
- New **methods to extract information** in textual documents (i.e. new features based on sequential patterns)
- Implemented **system** in collaboration between LIRMM, TETIS, and CMAEE





Part 3

Sentiment Analysis





Methods in order to identify sentiments: *Towards a sentiment lexicon*

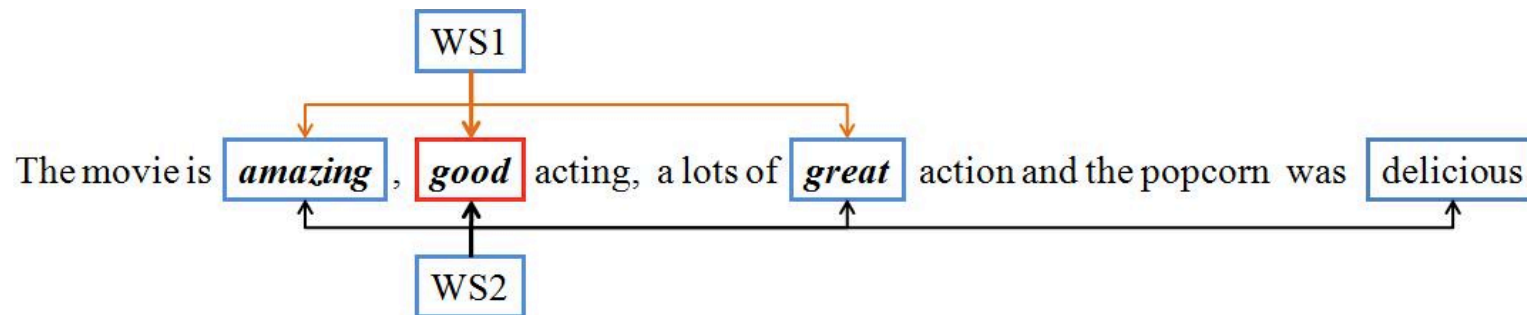
Step 1: choice of seeds related to opinions

$P = \{good; nice; excellent; positive; fortunate; correct; superior\}$

$N = \{bad; nasty; poor; negative; unfortunate; wrong; inferior\}$

Construction of 14 corpora related to a specific domain

Step 2: PoS, Association rules, choice of a window



Step 3: Statistic selection and web mining

Statistic measures (*Dice and MI*) that consist of measuring the association between **seed adjectives** and **candidate adjectives** association based on "hits" from the web (i.e. search engine) and contextual information

Examples of learnt adjectives:

great, hilarious, funny, happy, perfect, important, beautiful, amazing, complete, major, helpful

→ Agriculture domain:

{gmo; agricultural biotechnology; biotechnology for agriculture}



Examples of learnt adjectives: *green, healthy, enthusiastic, creative, etc.*



Laura Vanessa Cruz, San Agustín University, Peru

Work in progress: tweets and SMS (88milSMS corpus)



- Impact of **repetition of characters** for sentiment analysis: « *j'aaaadore IC'2016 !* » [Khiari *et al.*, JADT'2016]
- Identification of new **spatial entities** and **new spatial relations**: « *IC'2016 est organisé sur montpeul !* » [Zenasni *et al.*, MEDES'2016]





Part 4

QDoSSI

Data quality





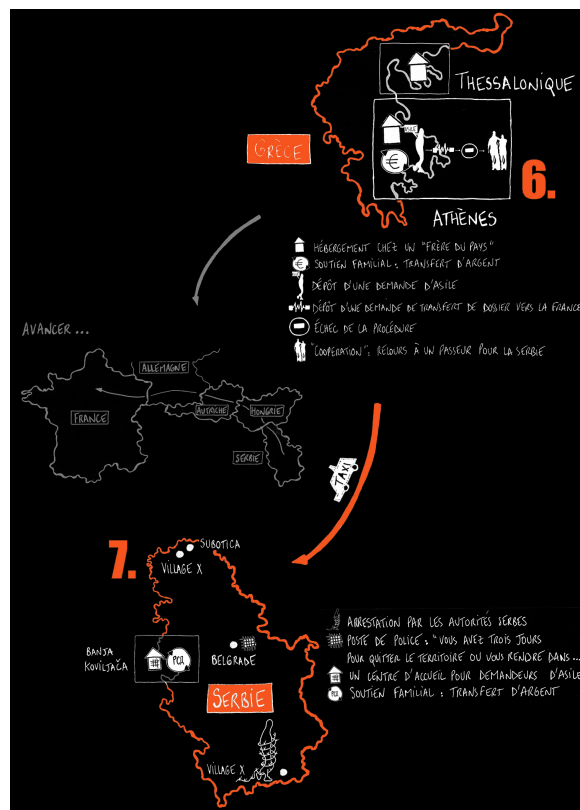
CNRS Mastodons 2016

Qualité des **D**onnées multi-**S**ources Un double défi pour les sciences **S**ociales et les sciences de l'**I**nformatique

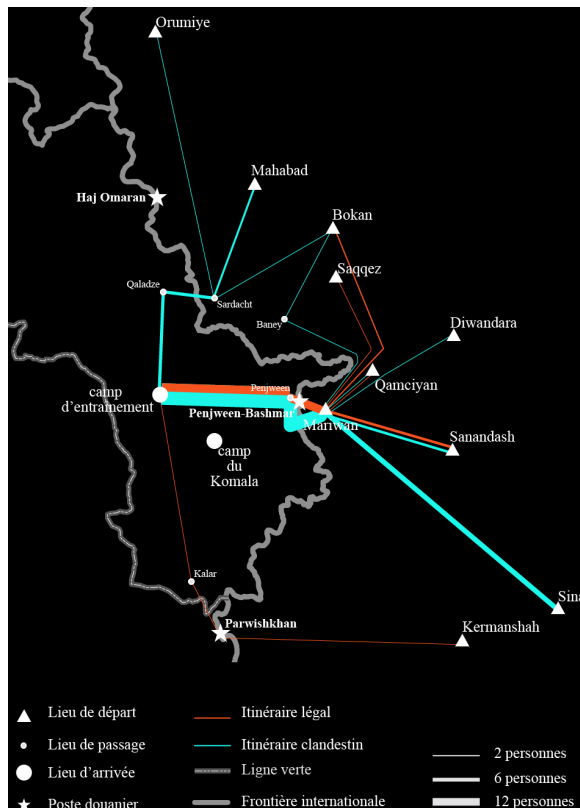


Different challenges

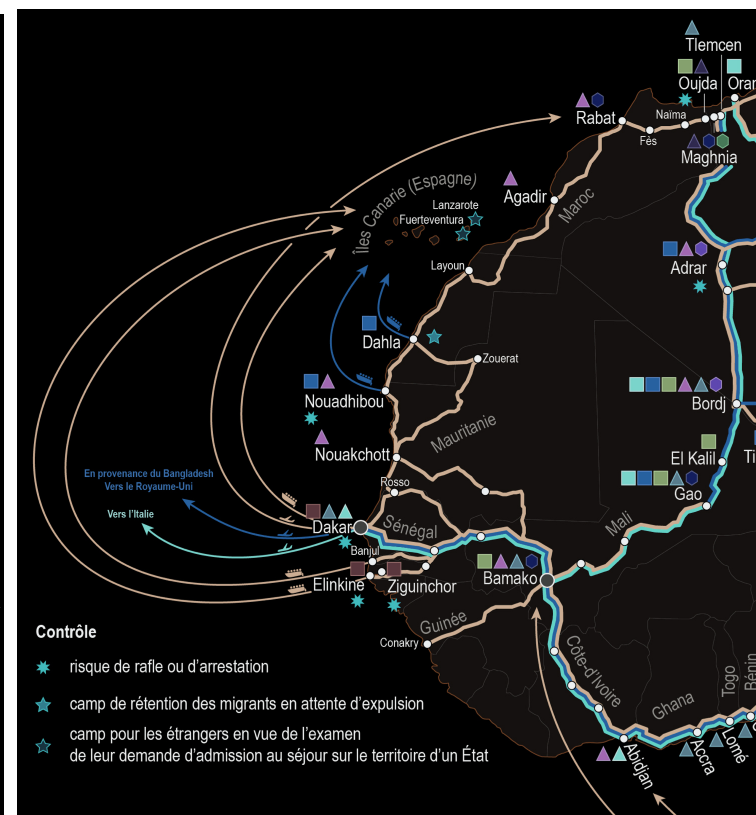




Balkan Corridor



Kurdistan



Trans-Saharan and sea lanes

Sources (de gauche à droite) : extraits de cartographies / Lucie Bacon, 2015, « Variations sur l'épisode migratoire de Kreuz. Cartographies mises en scène », *Mediapart* ; Cyril Roussel, 2013, « Les migrations des membres du Komala entre l'Iran et le Kurdistan d'Irak », *Atlas du Kurdistan d'Irak* ; Nelly Robin, 2013, « Mineurs en mobilité entre l'Afrique subsaharienne et l'UE », Programme Migrinter - Unesco



Text mining on interviews



Bobo-Dioulasso, juillet 2009, 06 septembre 2009. J'ai pris cette **route** parce que j'ai volé l'argent de mon **frère** pour fuir. C'est quand je suis arrivé à Koutiala au Mali on m'a parlé de cette **route**. C'est pour cette **raison** que j'ai tenté de venir ici. C'est des **amis** que j'ai **rencontrés** au Mali qui m'ont **informé** de cette **route**. Je suis passé par Koutiala, Bamako, Gao, Bordj, Riganne et Bordj où je suis resté. Je n'avais aucune **connaissance** sur les **routes** ni la **difficulté**. J'ai voyagé pour rentrer à Koutiala parce que j'ai **peur** que mon **grand-frère** me frappe. Je voulais me retourner à Bobo mais j'avais déjà dépensé beaucoup donc j'ai finalement opté pour le **voyage** sur l'Europe. Je suis resté avec une **famille** Malienne sous prétexte que je suis orphelin. Après que les **amis** m'ont parlé de la **route** j'ai pris un **minibus** pour aller à Bamako. Après, je suis parti à Gao. J'ai payé 75.000 FCFA (environ 110 Euros) pour passer en Algérie. J'ai emprunté le **pickup** avec une **dizaine** de **migrants**. On était très **serré** dans le **pickup**. J'étais fatigué et tout le monde criait et demandait au **chauffeur** d'arrêter le **pickup**. C'est un **très difficile voyage**. On a pris deux jours pour arriver à Bordj. Le **pickup** nous a laissés à environ deux kilomètres du poste frontière. On a marché pour aller au poste. Les **policiers** ont pris tous les **passesports** pour le contrôle. J'étais de ceux qui n'ont pas pu traverser la frontière **faute de passeport**. Alors le lendemain j'ai payé de l'argent à un **démarcheur** et j'ai finalement traversé la frontière. J'avais **besoin de passeport** pour continuer le parcours alors j'en ai payé un. J'ai pris le **transport commun** pour aller à Riganne. Je suis resté là-bas avec des Maliens. Je me suis enfin mis à travailler parce que je commençais déjà à en manquer. Un jour, des Maliens voulaient continuer le chemin et j'ai fait le chemin avec eux. On a emprunté un **autobus**. Comme on était au **nombre** de six **subsahariens** la **police** nous a contrôlé beaucoup de fois. Depuis que je suis arrivé ici je suis maintenant très **décidé** à continuer mon chemin jusqu'en Europe parce qu'il n'y a rien en Afrique. Je suis dans le **foyer** Malien (appartement de trois pièces). Je travaille dans la **construction** et je vis de cet **argent**. Les **gens** sont très sympas et j'ai pas de dérangements. Je n'ai pas grand **contact** avec les **algériens**. Je n'ai pas de **contact** avec mes **parents**. Ici la **police** rôde dans les environs mais ne font pas de **raffles**. C'est au Tchad (lieu où les **migrants** s'arrêtent pour chercher du **travail** que la **police** vient prendre les **gens**. Je ne sais vraiment pas ce qu'il faut penser de l'avenir.

Abidjan, janvier 2009, 19 juin 2009. J'ai **décidé** de partir à l'aventure parce que depuis la **guerre** en Côte-d'Ivoire la vie est devenue très chère et je suis d'une **famille** pauvre donc j'ai **décidé** de partir en Europe. Il y a un peu longtemps que j'ai appris que les **gens** passent par l'Algérie et le Maroc pour aller en Europe mais je ne me souviens pas au juste comment. Je suis venu avec un autre Ivoirien qui était plus âgé que moi. On a pris l'autobus jusqu'à Bouaké. De là on a pris un **minibus** pour venir à Ouangolo. Nous sommes **restés** à Ouangolo chez ses **parents** le temps que mon ami se prépare et s'entretient avec sa **famille**. Ensuite j'ai pris l'autobus avec lui pour continuer donc on était obligé de travailler à Bamako (vente d'eau glacée) avant de continuer à Gao. A Gao, on a aussi rencontré beaucoup d'autres **migrants** d'autres allaient vers la Lybie mais j'ai **décidé** avec mon ami d'aller au Maroc ensuite l'Espagne. Je n'étais pas bien **informé** sur les autres **routes**. L'avantage que j'ai eu sur cette **route** c'est que j'ai la **possibilité** de travailler pour continuer mon chemin. J'ai souffert durant mon parcours mais c'est l'axe Gao-Bordj qui m'a le plus fatigué. On était dans un **camion de marchandises**, **stockés** entre les **marchandises**. On était douze **personnes**. L'air manquait et la **chaleur** était **très vive**. Le **chauffeur** roulait vite et la **route** était mauvaise. J'ai regretté en ce moment de venir à l'aventure. Il y avait plus jeune que moi. C'est un **malien** qui avait 15 ans. Il a pleuré pendant des **heures de souffrance**. Je suis arrivé jusqu'à Maghnia, je crois que j'ai fait le plus difficile donc mon **projet** est de continuer jusqu'en Espagne. Je travaille dans un **jardin de fruits** avec deux **malien**s. C'est de ça que je vis. Je suis hébergé dans une **maisonnette** dans ce **jardin**. A part le **travail** je n'ai pas de **contact** avec les **algériens**. J'appelle très rarement mes **parents**. La **gendarmerie** nous **fatigue** beaucoup. Les **gendarmes** viennent rôder à tout moment et nous pourchassent. Lorsqu'ils nous attrapent on est refoulé à Tizawati (frontière Mali). Je n'ai pas encore été refoulé. Je doute de mon **avenir** parce que mon pays est très **mal**. L'Algérie où je suis n'est pas prête à m'accepter. L'Europe aussi où je veux partir ferme la **route**.

3 corpora:

- Balkan Corridor - Asylum seekers (in English and French)
- Sea lanes Atlantic - Economic migrants (in French)
- Trans-Saharan Africa - Isolated foreign minors (in French)



Methodology:

- Construction of multilingual corpora
- Automatic pre-processing of texts
- Adaptation of BioTex
- Manual analysis of extracted terms in order to identify (sub)concepts

Résultats: Multilingual thesaurus

- **3 levels** (e.g., Actors -> Authorities -> Police Authorities)
- **5 main concepts**: **Acteurs** (*famille, amis, autorités, organisation, acteurs clandestins, réseau de solidarité, population*), **Ressources** (*nourriture et hébergement, transport, communication, papiers, travail, argent, informations*), **Gestion des migrations, Motivations, Sentiments [+ Informations temporelles et spatiales]**
- More than **300 instances in English / 400 instances in French**: words (e.g, *road*), and multi-word terms (e.g., *airline tickets*).



Analysis of concepts *intra*-corpus and *inter*-corpus

In Croatia, they arrest you, and you came to Germany by **police**. You are fingerprinted, so if I go to France, I have to go back here. Me? In Switzerland. But I got fingerprinted. That was in the military **camp**. You have too fingerprinted. Germany need Syrian people... I don't know why. Like a **dream**... Now I am thinking that this is a **dream**. Because, suddenly I decided to come here. Two days after, we started the journey. Now, I think I'm like in the **dream**, waiting for someone who get me up. The only way for staying in Germany is to be a Christian. 27 people is not exactly the **family**. It's 3 **family** together. The **main family** is me, my **brother** and my **mother**. The others families, are **family** of my **mother** with **wife** and children, and there were also **family** of my **father**. Every **family** reached Germany. They all get **house**. So exactly we are 4 **family**. We decided to come here together. Me and my **brother**. I speak english and friends of my **brother**, actually, my **brother** has a **lot information** about the way. But he is a kid, he is just 18. But he has a lot of **information**. For example, in Athens he says "we have to buy simcard here, we have to buy something here... we have to change **document** here", he has a lot of **information** about the way. That was good, coming here. All the **problem** is to stay here. It's very hard. No, just me. In Iran, I lived in another city than my **family**. Here, I told you, to want to go Germany, you have to stay in Germany, and here, I want to leave alone, without my **family**. I like my **family**, but I don't want to stay with them. Every person, after 20 years like to leave alone. It's better. If you stay with your **family**, you can't go up. Like my **brother**. He stayed with my **mother**, so he doesn't have education, he doesn't speak english. So, in the **camp**, if I was alone, I could learn German quickly. But I am with my **family** and they are coming all the time, they want me to teach them. I have to **close** my

can... The rules [laughs]... it's very complicated... the rules say you have to stay in the **camp** for one week, waiting for medical **check**, you have to stay in XXX, because you are... for example, every city or every **state** get 1000 people, so those 100 people is for XXX. So you have to stay for medical **check**. You can't change your city. It's very dangerous, in the way, maybe **police** arrest you. In XXX, it was a **good camp**, but I told you, it was slow. You have to wait one week for the medical **check** and then you have to wait 3, 4 or 5 months before to get transferred. So, we decide to come here. Why? Because in this **camp** it's faster... No, with my **family**. No, no, just run. We just went. We get **train** and we came here. Yes, if **police** arrest you, it's **problem**. Actually, I saw **police** and, I go to the **police**... Because I thought, if **police** arrest us, it's finished. So I went to **police** and I ask the address of the **camp** and **police** forget to **check** my documents. After we came to Dresden, we went to the **police** station and I say "hey I'm here, I come from Afghanistan, I want to stay here", so **police** took us our fingerprints and after, we took **bus** to come here. So here, it's very fast. My **friend** had a **house** after one month. But we have to stay 5 months. Yes, fast. After I arrive here, you cannot go to another **camp**. You are not allowed but you can. For example, my **friend**, the first day he was



Extraction of spatial entities

Ma famille est très pauvre. On cultive l'igname toute l'année mais on ne gagne pas assez d'argent. Mon père dit que je suis courageux et que je peux réussir en **Europe**. C'est ainsi qu'après la récolte de l'igname il a réuni l'argent pour me donner pour voyager. C'est de mon père que j'ai eu l'information sur cette route. Je suis venu à **Bamako** chez des compatriotes. Ils étaient suffisamment informés sur la route. J'ai eu là tous les détails concernant sur la route. Le chef a contacté un passeur à **Gao** et j'y suis allé à son adresse. J'ai payé 45.000 FCFA et il m'a donné un passeport Malien qui m'a permis de passer la **frontière Algérienne**. Ceux qui n'avaient de passeport ont pas 500 CFA pour passer. Le voyage m'a épuisé parce que le véhicule 4x4 qui nous a transporté jusqu'en **Algérie** était très petit pour le monde qu'il contenait. Le voyage est la chose la plus difficile que j'ai connue dans ma vie. La route est mauvaise mais le conducteur roule à vive allure. Les gens crient de souffrance mais eux s'en foutent. J'ai passé dans les foyers à chaque étape. Ces séjours nous permettaient d'organiser les étapes suivantes. Les compagnons ne sont pas du tout solidaires. J'étais le seul Burkinabé et j'étais isolé. Mon projet c'est de pouvoir rentrer en **Espagne**. Je n'ai plus d'argent pour continuer donc je travaille pour en avoir. A chaque étape, je dois travailler pour avancer. Je ne dois pas décevoir mon père. La police ne m'a pas pris. Mais j'ai peur d'aller au refoulement. Je n'ai pas d'avenir en **Afrique**. En **Europe** on peut travailler pour se faire un avenir.



- **Analysis of concepts** (data mining + visualisation)



- **Vocabulary of the spatiality:** *forêt proche, ville là-bas, lieu de départ, big forest*
- **Sentiment analysis based on emotions:** *total disaster, horrible trip, lovely connection, good camp, nice room, stress, panic*
- **Construction of itineraries.**
Main issues: agregation of itinearies, integration of semantic knowledge
- **And... other data, other issues**



Part 5

Conclusions and future work



- **Matching different types** of documents (image/text, video/text, and so forth)
- **Integration of visual analytics** skills [Fadloun, Inforsid'2016]



- Data and Issue



- Hard Disc (157 188 files)



The screenshot shows the Agritrop website interface. At the top, there's a search bar and navigation links like 'Rechercher', 'Aide', and 'Liens utiles'. Below the header, there's a search bar with the text 'Mots (titre, résumé, mot-clé) contient "Alaotra"'. The results show a list of documents with titles, authors, and links. The first document is 'Short- to mid- term impact of conservation agriculture on yield variability of upland rice: Evidence from farmer's fields in Madagascar' by Bruelle Guillaume, Naudin Krishna, Scopel Eric, Domas Raphaël, Rabearisoa R. Lilla, Tittoneil Pablo. The second document is 'Trade-offs around the use of biomass for livestock feed and soil cover in dairy farms in the Alaotra lake region of Madagascar' by Naudin Krishna, Bruelle Guillaume, Salgado Paulo, Penot Eric, Scopel Eric, Lubbers M., De Ridder Nico, Giller Ken E.. The third document is 'Le technicien propose, le paysan dispose. Le cas de l'adoption des systèmes de culture sous couverture végétale au lac Alaotra, Madagascar' by Penot Eric, Domas Raphaël, Fabre Joana, Poletti Sarra, Mac Dowall Colombar, Dugué Patrick, Le Gal Pierre-Yves. The fourth document is 'Évaluer la durabilité de systèmes de culture en agriculture de conservation à Madagascar (région du lac Alaotra) avec MA SC-Mada' by Sester Mathilde, Craheix Damien, Daudin Gabriel, Sirdey Ninon, Scopel Eric, Angevin Frédérique. The website also has a sidebar with 'Se connecter' and 'Affiner la recherche' options.

