Textual Data Science

Mathieu Roche

Textual Data Science

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**

Textual Data Science

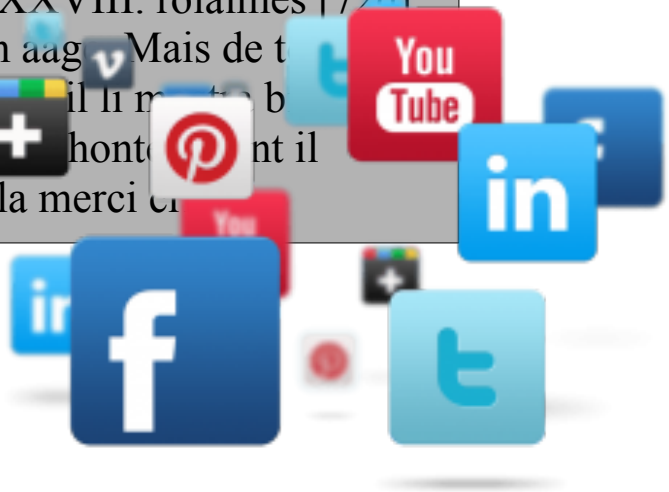
```
0/1 x B_1404 [WARNING]: "Asynchronous reset/set/load <%item> exists in module/unit"
0/1 x B_1405 [WARNING]: "<%value> asynchronous resets in this unit detected"
0/1 x B_1406 [WARNING]: "<%value> synchronous resets in this unit detected"
0/1 x B_1407 [ERROR]: "Do not use active high asynchronous reset/set/load"

//           Total Module Instance Coverage Summary

                                                    PERCENT
lines                                                  31.54
stateme                                              31.54

Policy:
     <vi                                          >]:<message>
     ---                                          ------------
     0/1                                          is not allowed to be used as
```

descovri son corage a Lancelot et dist
a guerre commença, baoit 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
grant honor sa hont nt il
roi Artu et il li ala merci c



Pluridiciplinary projects

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

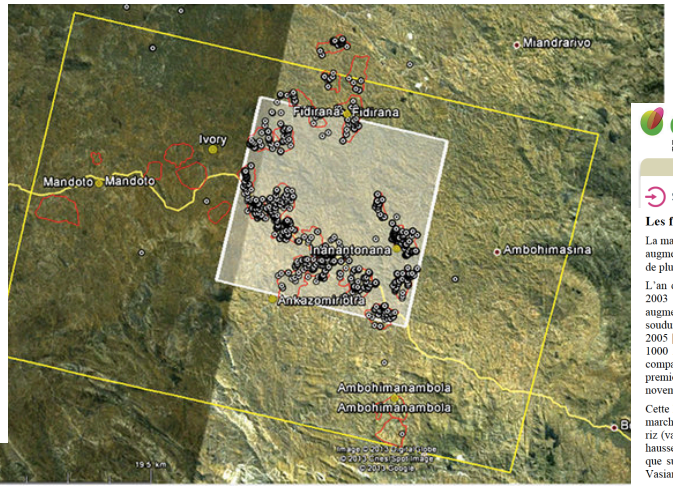
Vakinankaratra – L'agriculture de conservation lancée

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 Manitatra 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'Ankazomiriotra, d'Inanantonana de Vinany et de Fidirana, le projet Manitatra compte adhérer 1000 paysans, dont 200 femmes, sur la pratique de ce système de culture agro-écologique qui ne nécessite pas des 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'Ile. 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.

Angola Ny Avo



cirad LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT — Agritrop Archive ouverte des publications du Cirad

Recherche avancée | Auteurs CIRAD | Parcourir | 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.

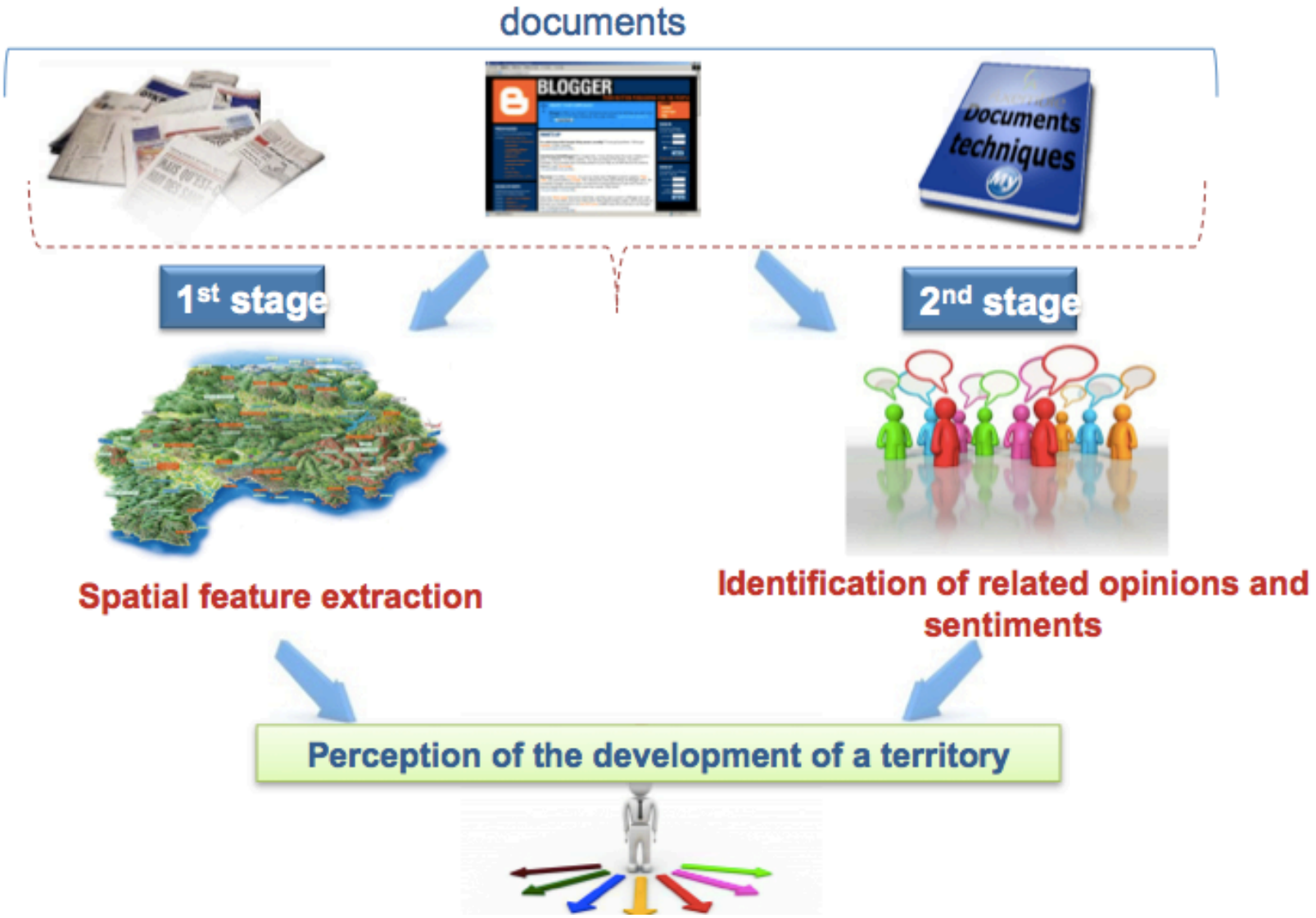
L'an dernier, les prix du riz ont normalement augmenté pendant la période de soudure, fin 2003 début 2004, mais ne sont pas redescendus en période de récolte, ils ont continué à augmenter à un rythme soutenu. La variation annuelle du prix du paddy entre récolte et soudure est habituellement de l'ordre de 50% au Lac Alaotra, elle a été de 150% en 2004-2005 [Minten et Ralison, 2005]. Le prix du riz national ou 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 Inanantonana (45 mn de piste en saison sèche), Vasiana (1h15mn), Mahasolo (2h30mn) ou Ambalanirana (4h).

Textual data and data bases – QuDoSSI *(since 2016)*

battre. Je vois un bon avenir.
**** *recit_2095 *sex_Masculin *age_17 *pays_Côte-d'Ivoire
Danané, ne se souvient pas (mai 2009), juillet 2009 C'est la situation économique de ma famille qui m'a encouragé à prendre la décision de partir. Ma famille n'était au courant de mon voyage. C'est même l'argent de mon père que j'ai volé pour entamer mon projet de voyage. C'est mon ami qui m'a parlé de cette route. Je ne peux pas vous parler de la route parce que si je le dis le chef va me punir. Le voyage s'est bien passé. Mon projet c'est d'aller jusqu'en Espagne. N.B. le mineur ne veut pas donner de détails parce qu'il a peur.
**** *recit_2097 *sex_Masculin *age_16 *pays_Côte-d'Ivoire
Lakota, ne se souvient pas, ne se souvient pas Je suis d'une famille à situation défavorable. Mon père veut qu'un enfant puisse aller en Europe et cela pouvait aider la famille. Je suis venu avec le fils d'un ami de mon père. Il a déjà fait la route. C'est lui qui me guidait. J'avais entendu parler de la route en causerie ave les amis. On est passé d'abord chez mon grand-père à Yamoussoukro. Lui aussi m'a donné de l'argent pour le voyage. C'est mon compagnon qui discutait le transport. A la demande de ma famille je lui ai donné tout mon argent. Il ne me faisait pas de compte des dépenses. C'est à cause de lui que je travaille parce qu'il a dépensé tout mon argent. Je n'ai plus d'argent pour continuer. Il est allé lui à Maghnia. Je suis resté seul ici. A Gao on a pris le pickup jusqu'en Algérie. Nous ne sommes pas passés par la frontière officielle. Je suis rentré en Algérie sans voir la police. Mon projet migratoire c'est de rentrer au Maroc. On m'a dit qu'au Maroc je peux rentrer en Espagne même si j'ai pas d'argent. Comme je suis jeune les gens sont beaucoup gentils avec moi. Tout le monde me donne à manger. Mes employeurs me demandent mon aide pour me prendre. Il

Senterritoire: Generic Process



Outline

Part 1 Data Science and Big Data

Part 2 Textual data and heterogeneity

Part 3 Applications in agriculture domain

Part 4 Conclusions and future work

Part 2

Textual data and heterogeneity

How to match documents?

- **Data and Issue**

- Hard Disc (157 188 files)

cirad irstea

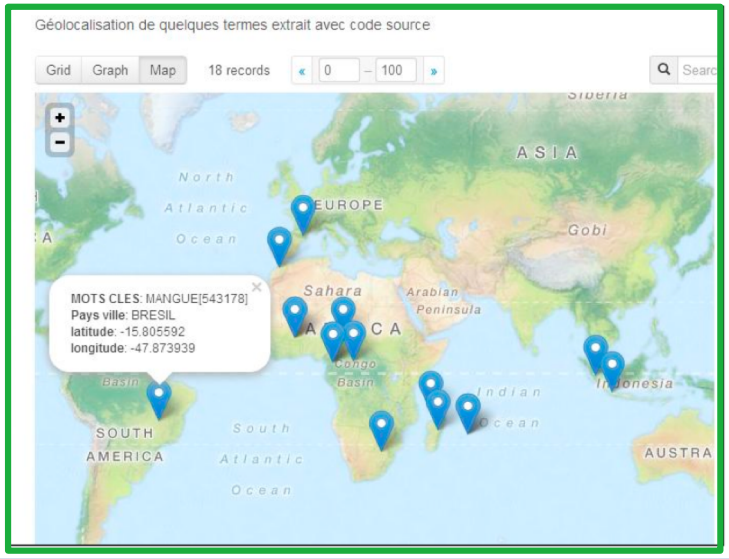
UNIVERSITÉ CONSTANTINE 2

How to match documents?

- **Method: Extraction of features** [Roche *et al.* CA'2015]

3 types of features:

- thematic
- spatial
- temporal



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 Manitatra 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 commue rurale d'Ankazomiriotra, d'Inanantonana, de Vinany et de Fidirana, le projet Manitatra compte adhérer 1000 paysans, dont 200 femmes, sur la pratique de ce système de culture agro-écologique qui ne nécessite pas des 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.

How to match documents?

- **(a) Extraction of features: thematic terms** [Lossio Ventura *et al.* ISWC'2014]

How to match documents?

- **(a) Extraction of features: spatial features (SF)**

Model

- **Global Model**: SF is composed of at least one Named Entity (NE) and one variable number of spatial indicators specifying its location. SF can then be identified in two ways:

- **Absolute spatial feature** (A_SF) one NE with a geo-localization, such as <(spatialIndicator)*, NE of Location> (ex: *the city of Montpellier*).

- **Relative spatial feature** (R_SF) one spatial with at least one SF (ex*: in the south of the city of Montpellier*).
  An R_SF is defined as <(spatial relation)$^{1..*}$, A_SF> or <(spatial relation) $^{1..*}$, R_SF>
  Five spatial relation types are considered: orientation, distance, adjacency, inclusion, and geometric which defines union or intersection linking two SFs.

How to match documents?

- **(a) Extraction of features: spatial features (SF)**

Methods [Kergosien *et al.*, IJGIS'2014]

- **Symbolic approach:** Using rules (*Text2Geo*) for extracting A_SF and R_SF

Basic patterns			Text2Geo paterns			
	A_SF	R_SF		R_SF	R_SF	OE
Precision	20%	**48%**	Precision	53%	84%	**92%**
Recall	**63%**	27%	Recall	**94%**	66%	35%
F-mesure	30%	**34%**	F-mesure	67%	**74%**	50%

- **Statistic approach:** Using context and IR methods for spatial feature disambiguation [Tahrat *et al.*, WIMS'2013]

How to match documents?

- **Disambiguation** between **location** and **organisation**

SVM		SF	OE
	SF	103	35
	OE	44	90
Accuracy	*70.96%*		

Naive Bayes		SF	OE
	SF	98	40
	OE	44	90
Accuracy	*69.12%*		

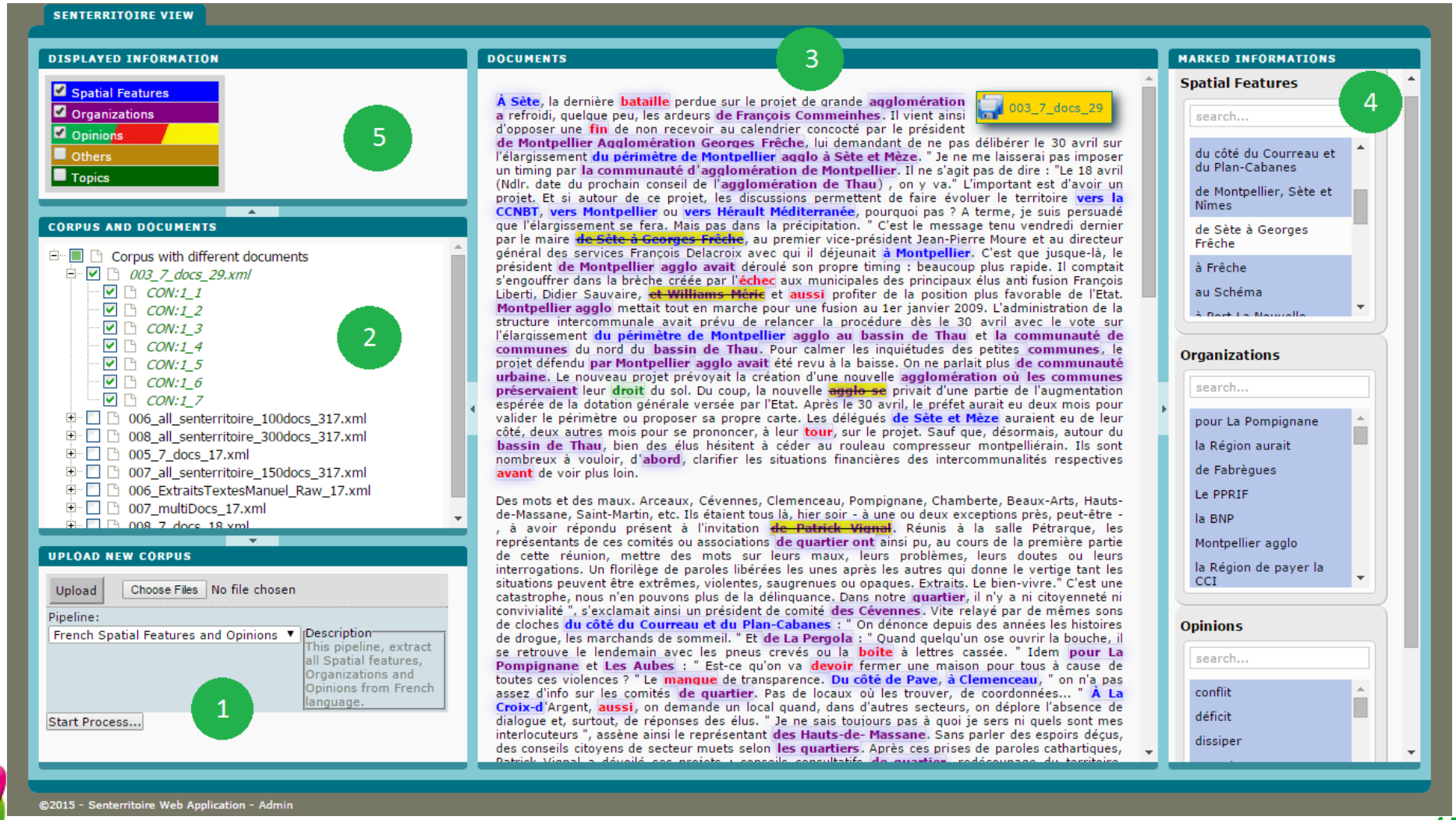
Features with ConceptOrg		SF	OE
	SF	108	30
	OE	47	87
Accuracy	*71.69%*		

Features with ConceptSpa		SF	OE
	SF	112	26
	OE	19	115
Accuracy	*83.45%*		

Both types of features		SF	OE
	SF	113	25
	OE	19	115
Accuracy	*83.82%*		

How to match documents?

- **(a) Extraction of features: spatial features (SF)**
  [Farvardin *et al.* Demo ISWC'2015]



How to match documents?

- **(c) Similarity**

$$Global\_Sim(vect1, vect2) = \alpha.cosT(vect1, vect2) + (1-\alpha).cosS(vect1, vect2)$$

with $\alpha \in [0,1]$

cosT: cosine based on **thematic features** (BioTex)
cosS: cosine based on **spatial features**

Perspective: adding temporal information

Part 3

Applications in agricultural domain

Animal disease surveillance

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

cirad INRA SCIENCE & IMPACT LABEX NUMEV

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.

OIE

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)
- **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

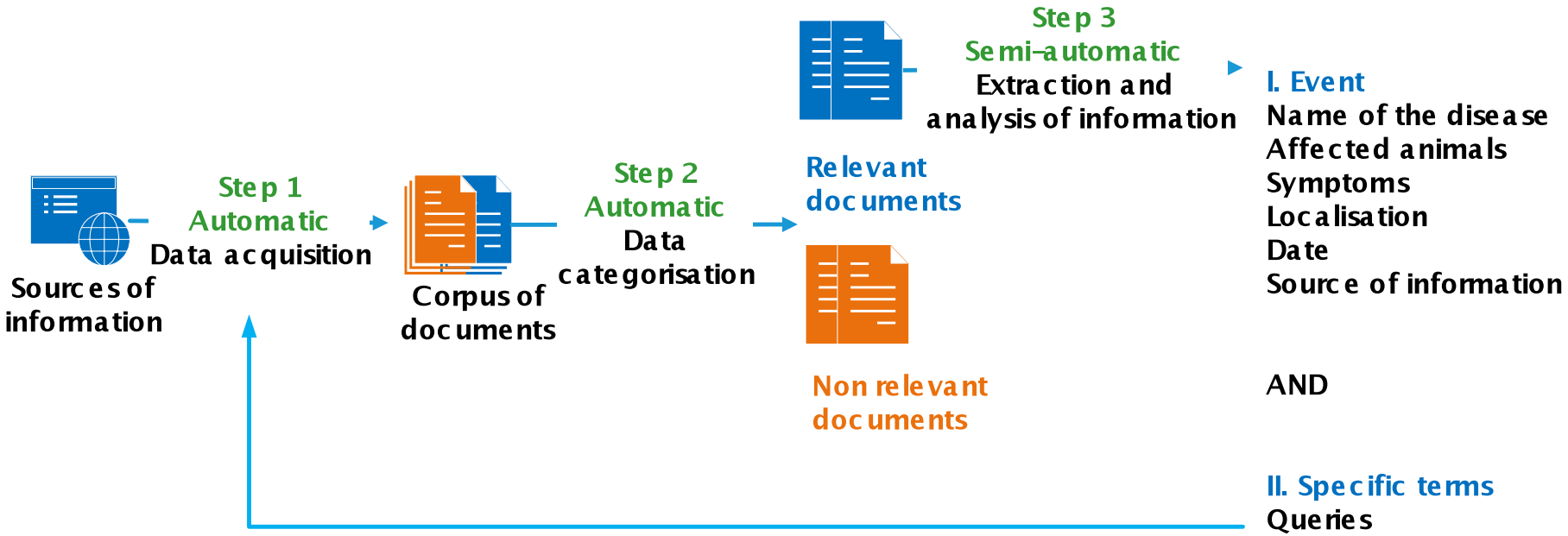
Polish local authorities said on Wednesday that preliminary tests have pointed to a case of African swine fever (ASF) among farm pigs in eastern Poland near the city of Bialystok.

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 at laying special mats, were being taken.

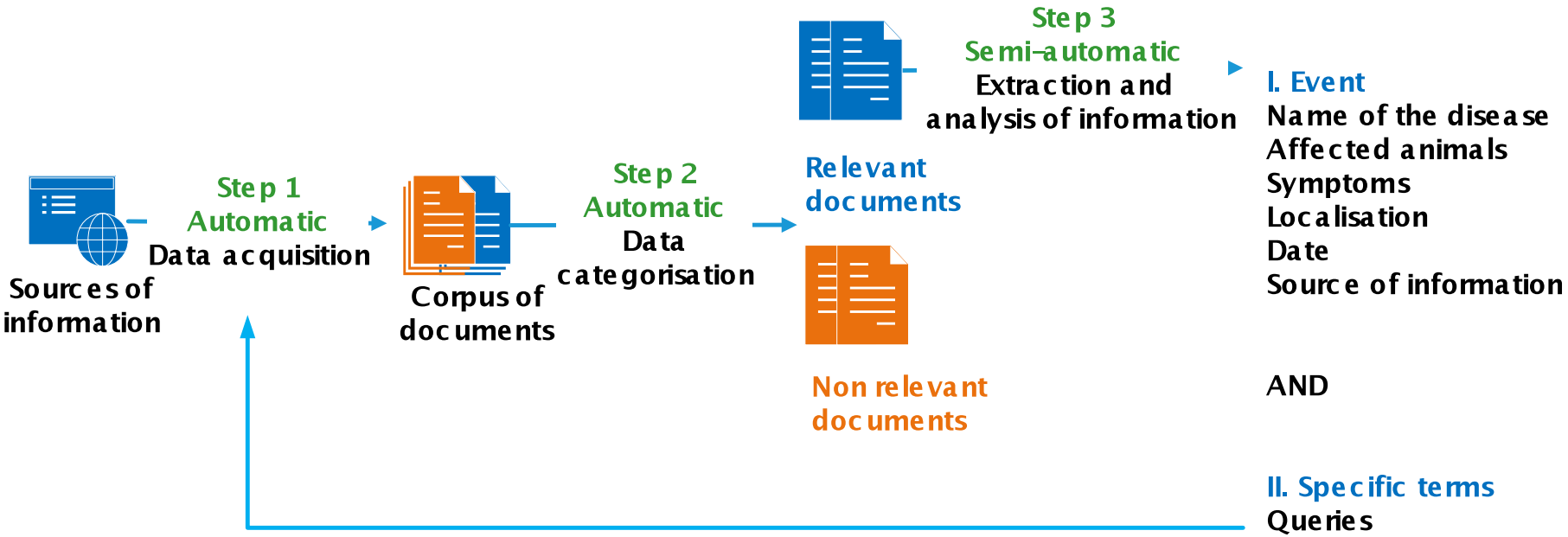
Poland's chief veterinary officer was unavailable for comment, while the county veterinary officer said a statement on the issue will be published later on Wednesday. (Reporting by Anna Wlodarczak-Semczuk; Writing by Marcin Goettig)

Methodology



Methodology - *step 1*

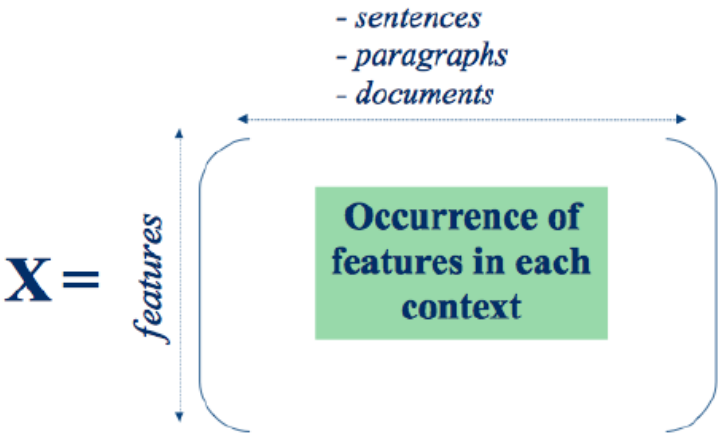
- **Step 1: Data acquisition**

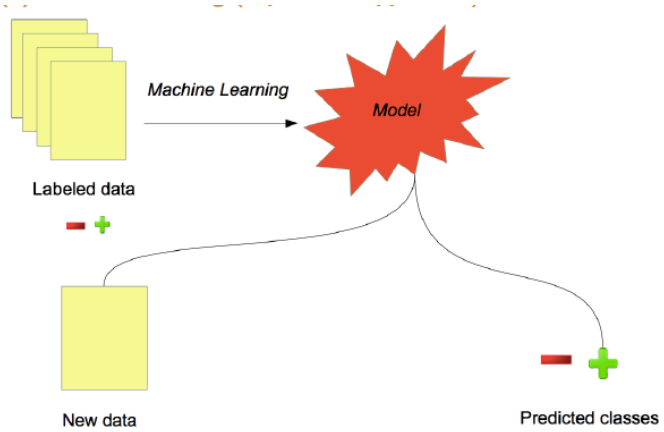


*https://news.google.com/news/feeds?pz=1&cf=all&ned=en&q=**Blue+tongue**&output=**rss***

Methodology - *step 2*

- **Step 2: Data classification**

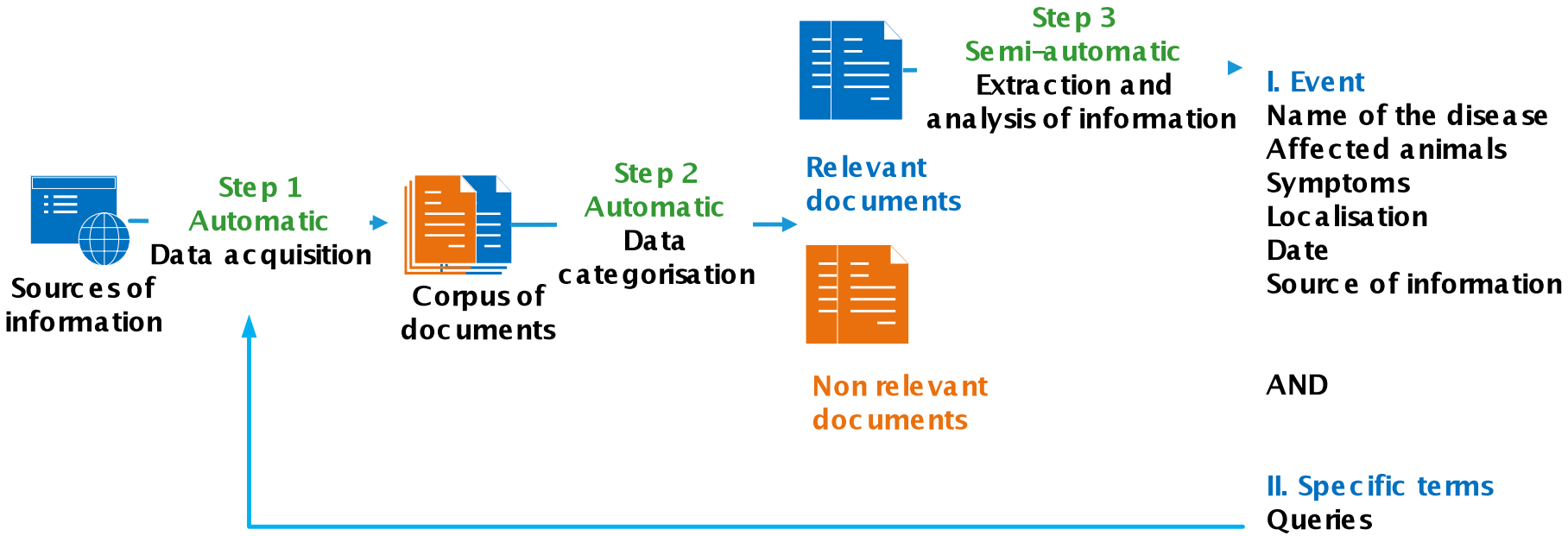




Classification algorithm		Naïve Bayes			Support Vector Machine		
Performance		Recall	Precision	F-score	Recall	Precision	F-score
Class	*disease*	0.724	0.766	0.744	0.657	0.68	0.669
	economy	0.478	0.530	0.503	0.489	0.726	0.584
	general	0.860	0.804	0.831	0.864	0.763	0.810
Weighted average		**0.750**	**0.745**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

Methodology - *step 3*

- **Step 3: Information extraction and management**



Methodology - *step 3*

- **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, HeidelTime, disease names, species names, etc.), and data mining techniques in order to learn extraction rules

Rules assiociated 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%**

Methodology - *step 3*

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

  - **First 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

Julien Rabatel,
LIRMM, Numev, France

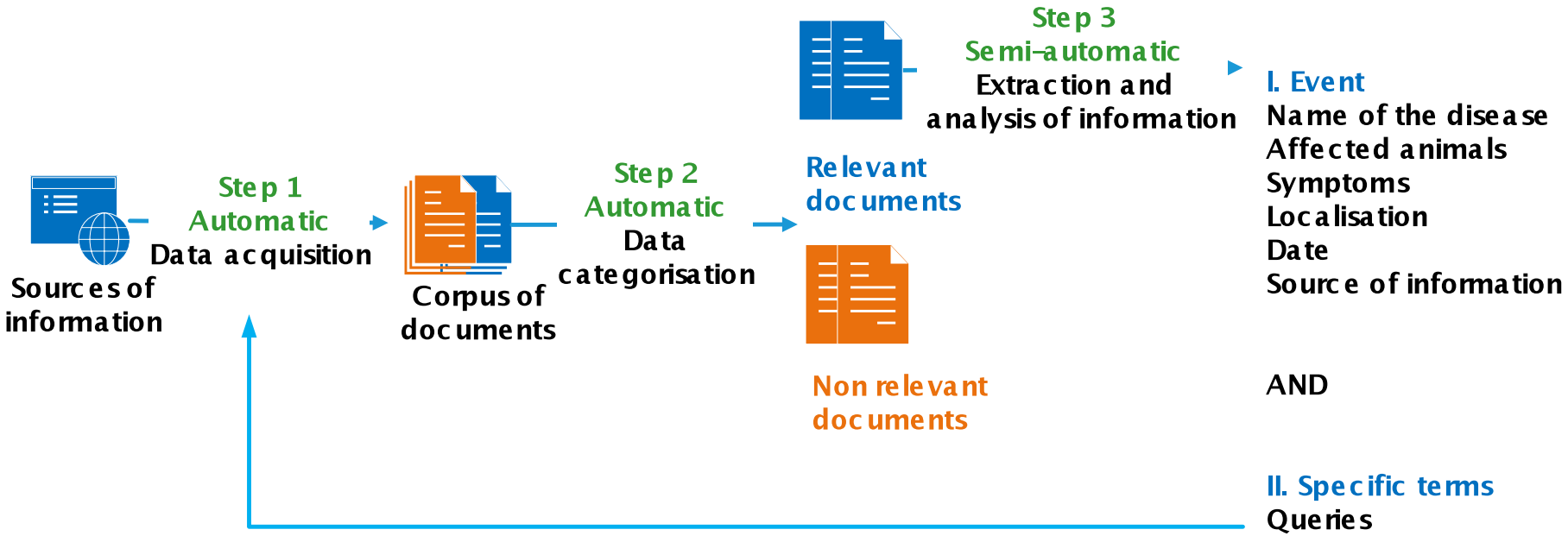
LABEX NUMEV

Methodology - *step 3*

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

Methodology - *step 3*

- **Step 3: Information management (II)**

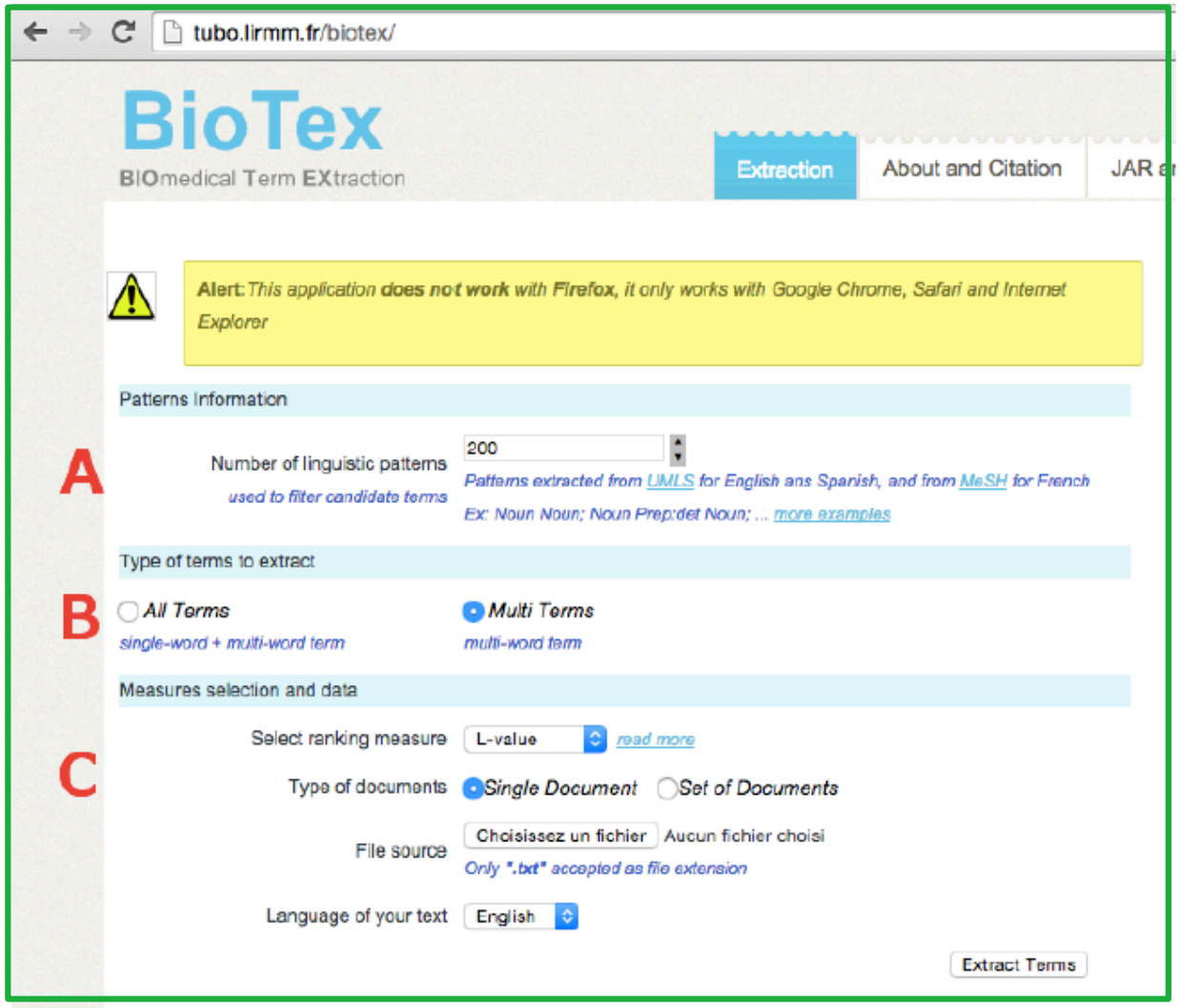


Methodology - *step 3*

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

SIFR project

LIRMM



Methodology - *step 3*

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

Statistics

- Frequency (TF) → **important** word

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Inverse Document Frequency (IDF) → **discriminant** word according the distribution in the corpus

$$IDF_i = log \frac{|D|}{|d_j : t_i \in d_j|}$$

- Global value:

$$TF\text{-}IDF_{i,j} = TF_{i,j} \times IDF_i$$

Methodology - *step 3*

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

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

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

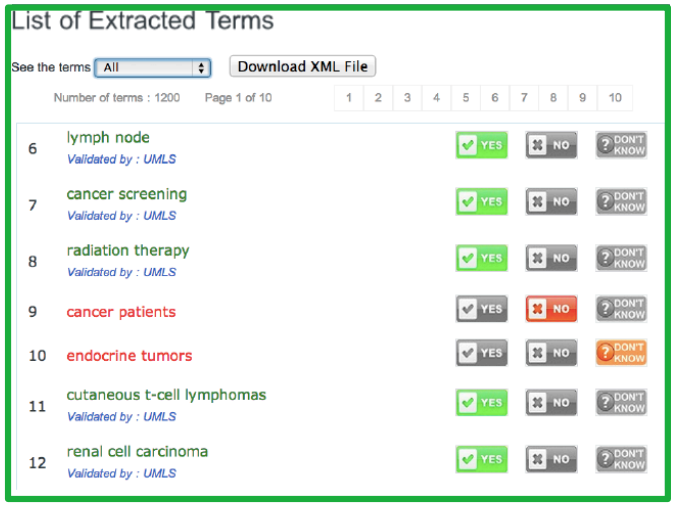
- A **new ranking function** to take into account the **heterogeneity** of the sources (S_i) [Arsevska *et al.* CEA'2016]:

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

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

Methodology - *step 3*

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



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*

Methodology - *step 3*

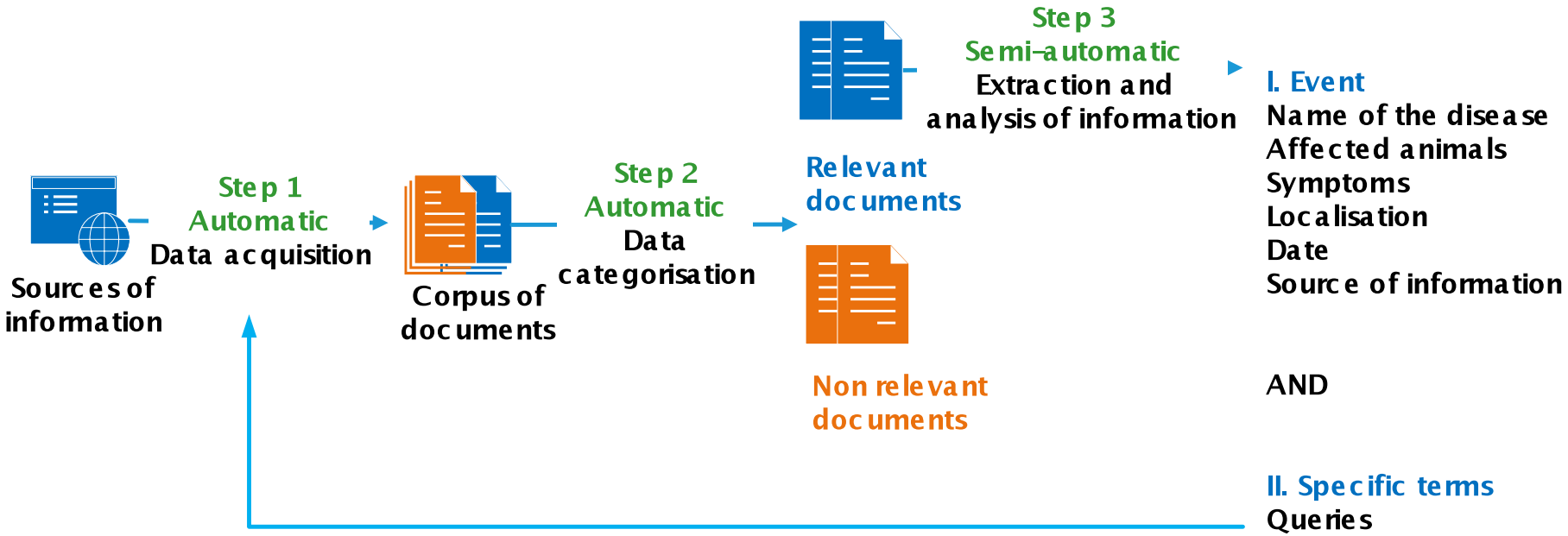
- **II. Querying the Web:** (d) ***Association of terms***

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

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

Rank	**Bluetongue** *hosts / clinical signs*	**Schmallenberg virus infection** *hosts/ clinical signs*
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

Methodology



Part 3

Applications in agricultural domain

Sentiment analysis

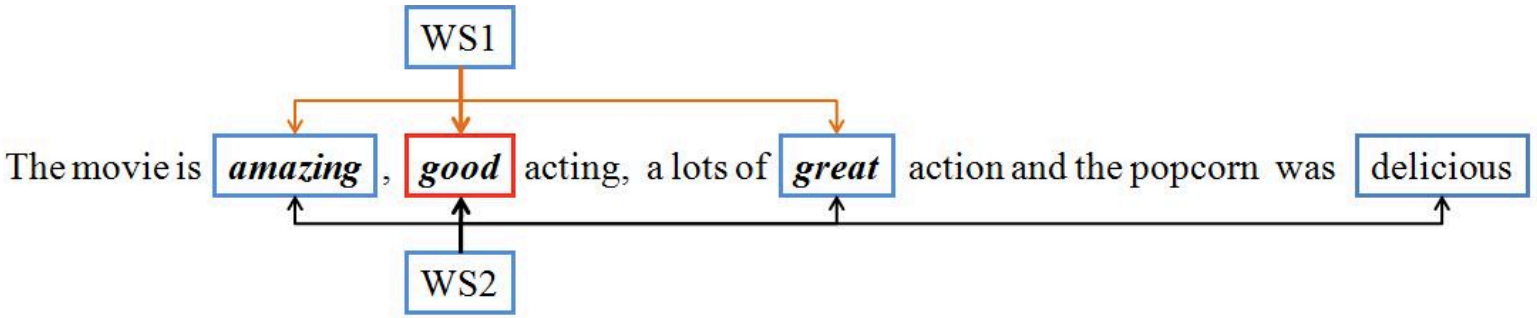
Sentiment analysis

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



Sentiment analysis

Step 3: Statistic selection and web mining

Statistic measures 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 Agustin University, Peru

Sentiment analysis

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.*, TALN'2016]

Part 4

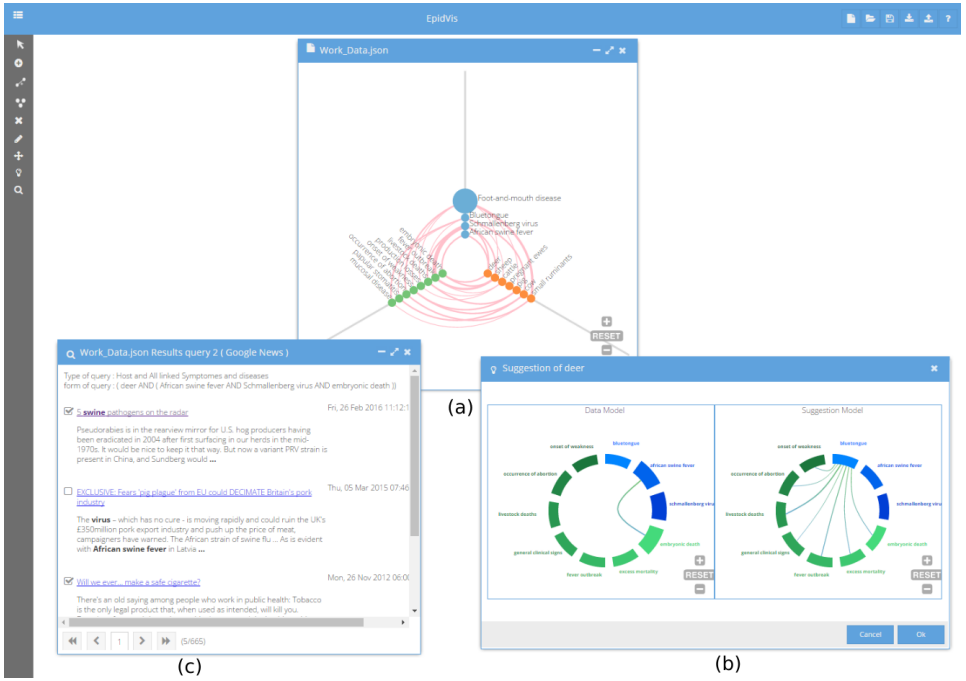
Conclusions and future work

Conclusions

New challenges of *Texual Data Science*:

- **Matching different types** of documents (image/text, video/text, and so forth)

- **Integration of visual analytics** skills [Fadloun, Inforsid'2016]



Conclusions

New challenges of *Texual Data Science*:

- Towards the valorisation of data: **Open Data**, **Data papers**

- From **Textual Data Science** to **Data Science** in pluridisciplinary context:

Unification of concepts and associated methods

For instance:
NLP: *n-grams, skip-grams, multi-word terms, syntactic relations, etc.*
Data mining: *association rules, sequential patterns, etc.*
Linguistics: *collocations, etc.*