Data mining applied to environmental and agriculture domains

Mathieu Roche

Cirad – TETIS and LIRMM – Montpellier, France

Web: http://www.textmining.biz
Email: mathieu.roche@cirad.fr

Outline

Part 1 Data Science and Big Data

Part 2 Heterogeneity and textual data

Part 3 Applications in agriculture domain

Part 4 Application in environmental domain

Part 5 Conclusions and future work

Part 1

Data Science and Big Data

Big Data

Volume

Velocity

Variety

3V of Big Data

Variability, Véracity, Value,

Visualisation, Valorization

Part 1

Part 2

Heterogeneity and textual data

Introduction

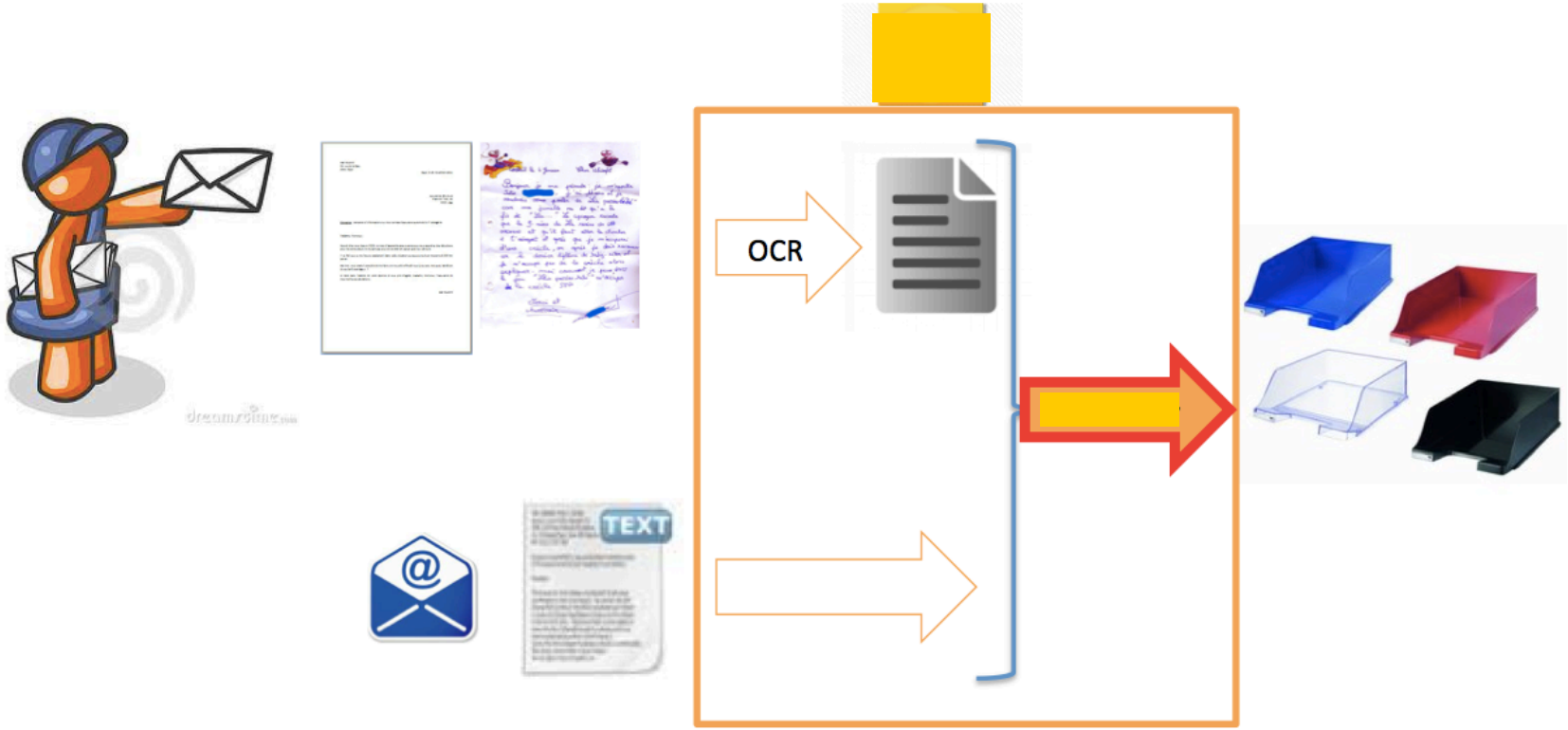
```
0/1 x B_1404 [WARNING]: "Asynchronous reset/set/load <%item> exists in module/unit"
0/1 x B_1405 [WARNING]: "<%value> asynchronous resets in this unit detected"
0/1 x B_1406 [WARNING]: "<%value> synchronous resets in this unit detected"
0/1 x B_1407 [ERROR]: "Do not use active high asynchronous reset/set/load"

//           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 aage. Mais
st Lancelos ari et il li m
grant ho a gra te, q
roi Artu et il li ala i crier...

Heterogeneity of content of textual data

Information Retrieval and heterogeneity

[Bouillot *et al.* ISMIS'2014]



Heterogeneity of content of textual data

Information Extraction and heterogeneity

[Saneifar *et al.* JUCS'2015 ; Saneifar *et al.* CI'2015]

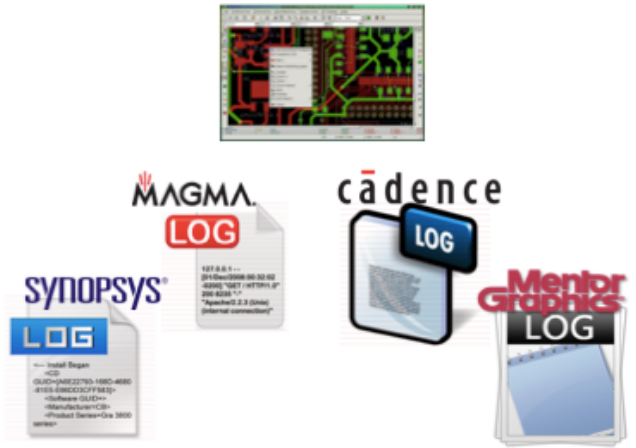
```
0/1 x B_1404 [WARNING]: "Asynchronous reset/set/load <%item> exists in module/unit"
0/1 x B_1405 [WARNING]: "<%value> asynchronous resets in this unit detected"
0/1 x B_1406 [WARNING]: "<%value> synchronous resets in this unit detected"
0/1 x B_1407 [ERROR]: "Do not use active high asynchronous reset/set/load"

//            Total Module Instance Coverage Summary

                    TOTAL         COVERED          PERCENT
lines                 501           158                31.54
statements            501           158                31.54

Policy: DESIGN  Ruleset: RESETS
    <violated>/<checked> x <label> [<severity>]:<message>
    ---------- ------------------- ---------------------
    0/1 x NTL_RST04 [ERROR]: "A reset signal is not allowed to be used as
```

- **Log A**

TOTAL	COVERED	PERCENT
Lines	10	11	12
statements	20	21	22

- **Log B**

  EC : 2.1%



Heterogeneity of content of textual data

Sentiment analysis and heterogeneity

[Roche and Poncelet RNTI'2009]

Example:

88,000 French text messages collected [Panckhurst *et al.* Epistémé'2013]: http://88milsms.huma-num.fr

- *SimBig is an **attractive** conference! (article or report)*
- *SimBig is an **aaaaattractive** conference! (SMS or tweet)*

Accuracy based on 5 classes and **machine learning** (SVM):
- 30% with repetition of characters
- 46% without repetition of characters

=> We need specific process

Heterogeneity of content of textual data

Other tasks that address **heterogeneity issue** in **SMS** and **tweets**

- **Anonymisation**
  [Accorsi *et al.* LI'2012]

- **Lexical creation in SMS**: *numb3rs, mc2, 106ounette, 3615ma-vie, Ar5ggggggh, a2min lami* [Lopez *et al.* Tranel'2015]

- **Lexical creation in tweets**: new and complex built hashtags [Tisserant *et al.* KDIR'2014]

Heterogeneity of content of textual data

Challenge: How to deal with the **heterogeneity** of the content of SMS/tweets?

Examples (http://88milsms.huma-num.fr):

- Coucou! Jai laissé Le médecin malgré lui ds le salon, tu pourras le prendre à Laborde? **Gracias**

- **gracias** meuf ;)

- Avec <PRE_5> hahahaha ! **Muy caliente** les revisions !

- **Es muy triste**

- T'as une secrétaire toi ?! Huuum, c'est **muy muy caliente** ca !

- Jai dit la meme chose ptdr!! ^^ **no se que hacer en mi vida** Moi jai mis sa ^^

Heterogeneity of types of documents

Textual data and satellite images

[Roche *et al.* SI'2014]

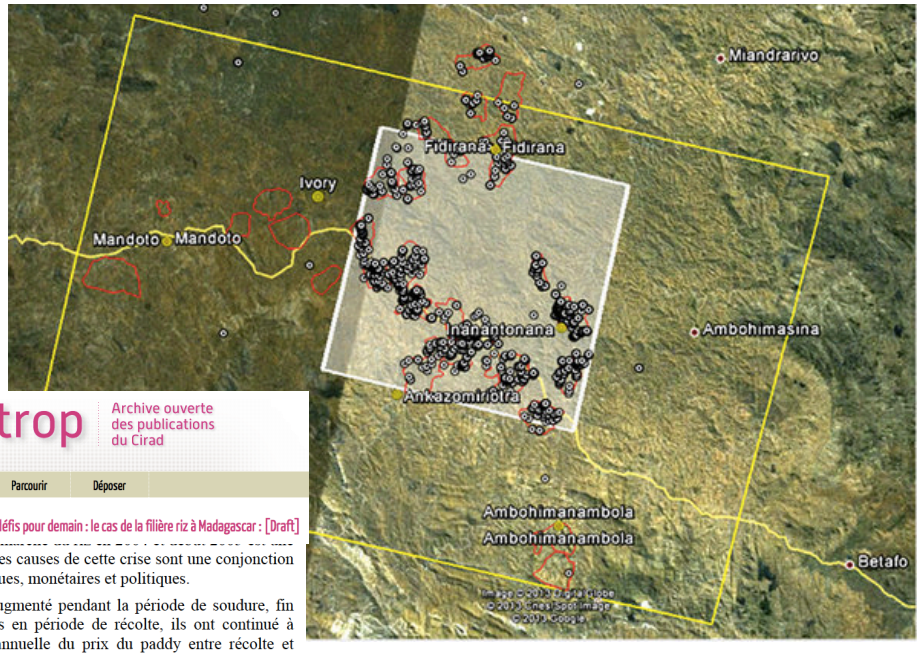
Vakinankaratra – L'agriculture de conservation lancée

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 da 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 », no Rakotondramanana, directeur exécutif du projet qui s'active aussi dans le Sud-Est de Des formations sur la régénération de la fertilité du sol et la lutte contre sa dégradatio ainsi que l'introduction du système des légumineuses seront la priorité des activités d 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]

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).

How to match documents?

- **Data and Issue**

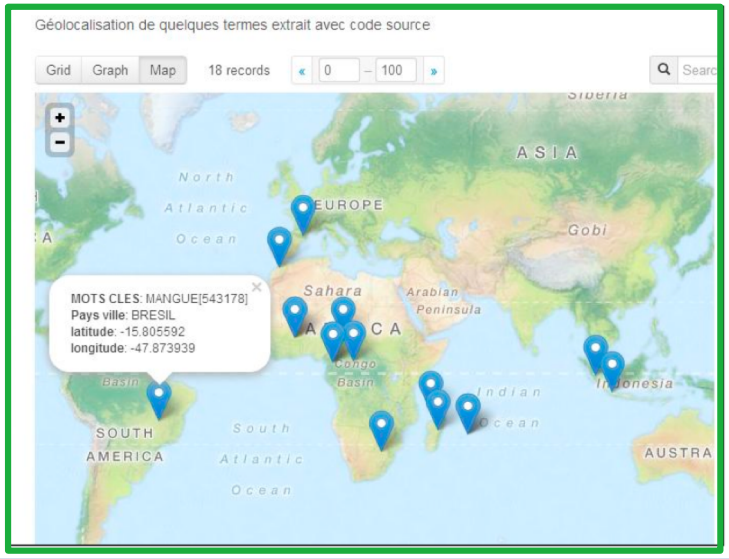
- Hard Disc (157 188 files)

How to match documents?

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

3 types of features:

- thematic features
- spatial entities
- temporal entities



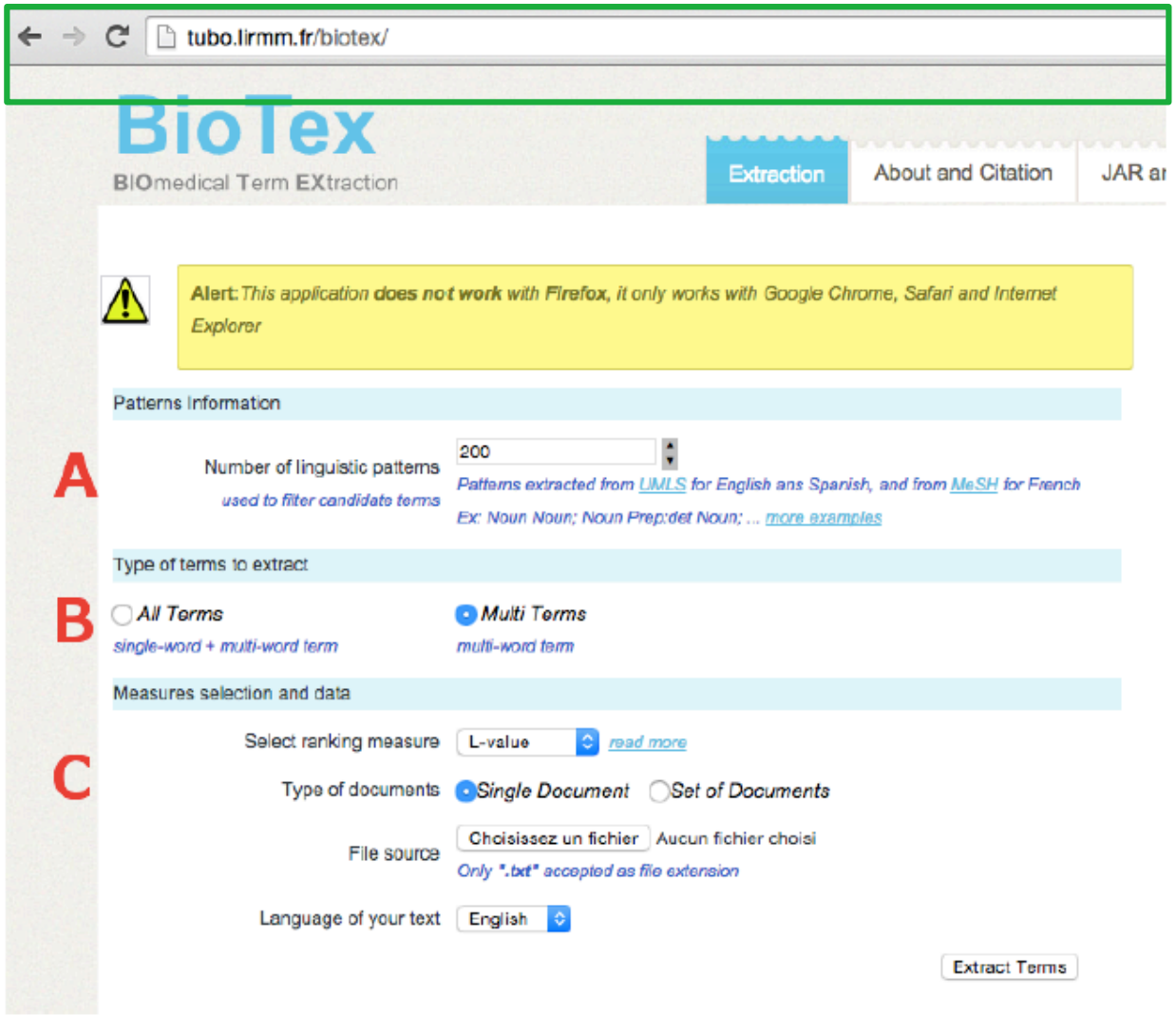
17.12.2014 | 7:18 | Non classé | 0

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 Lima*).

- **Relative spatial feature** (R_SF) one spatial with at least one SF (ex*: in the south of the city of Lima*).
  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)**

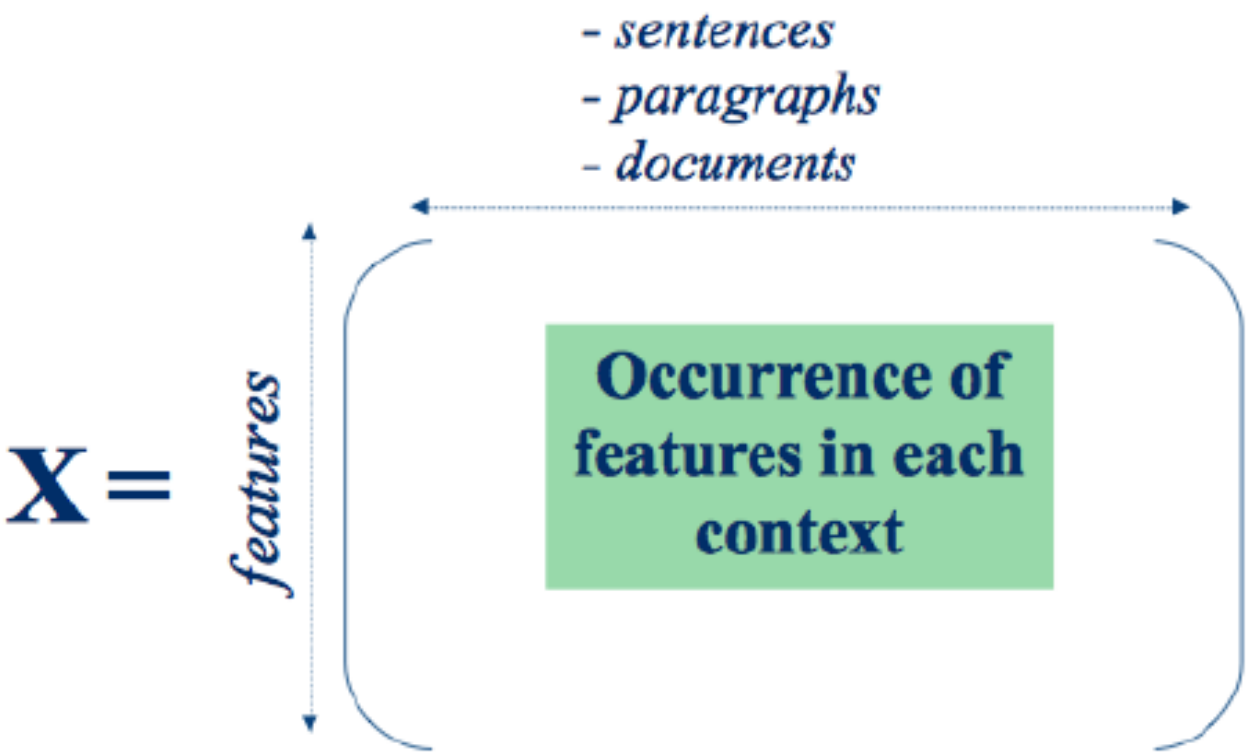
Methods [Kergosien *et al.*, IJGIS'2014]

- **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 features disambiguation

- **(b) Representation of documents**



How to match documents?

- **(c) Similarity**

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

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

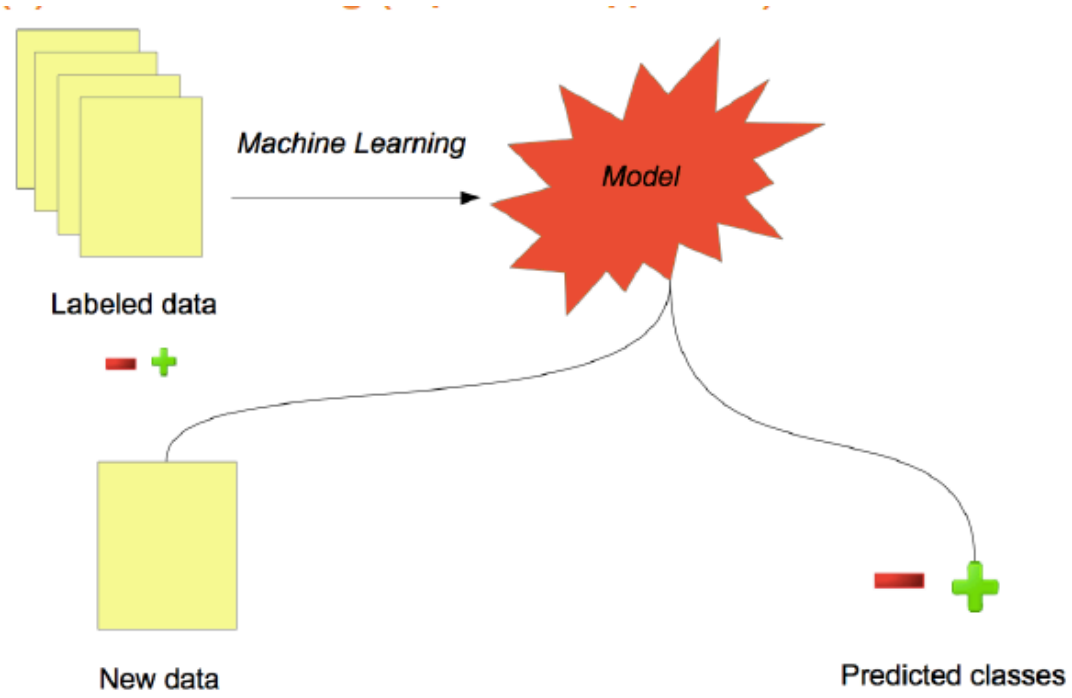
cosT: cosine based on **thematic features** (BioTex)
cosS: cosine based on **spatial features**

Perspective: adding temporal information

How to match documents?

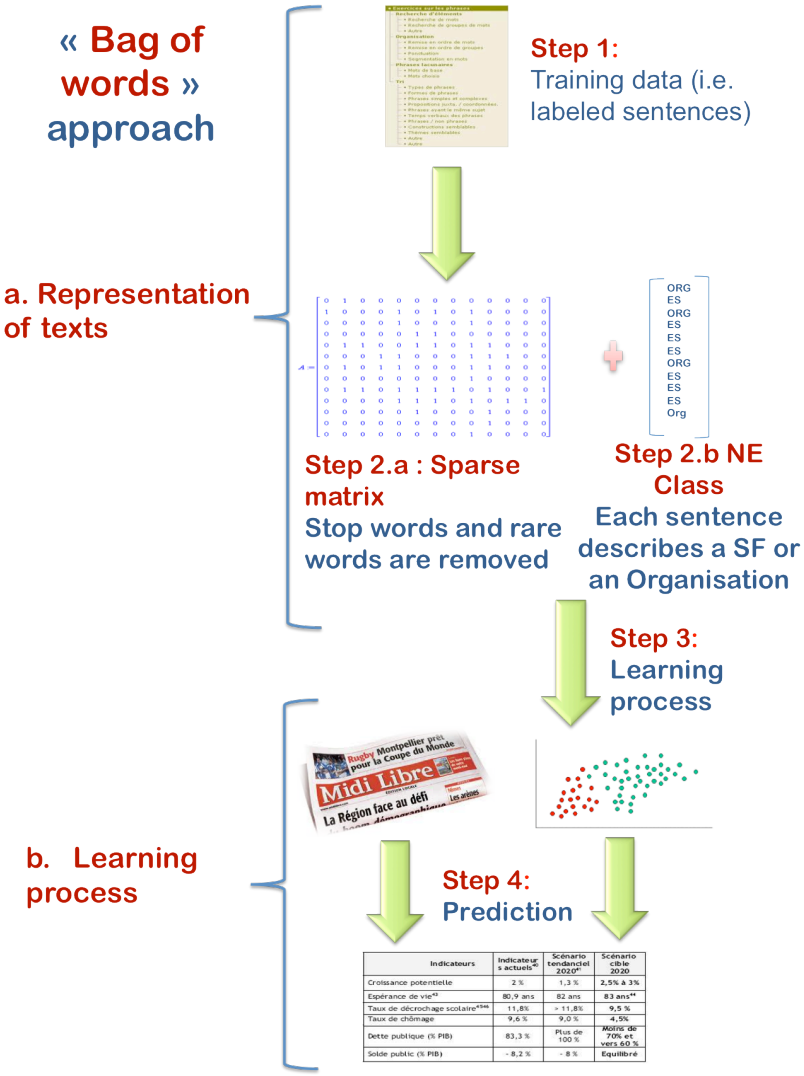
- **Extension:** How to analyse document with more precision?

Example: Disambiguation between **location** and **organisation** [Tahrat *et al.* WIMS'2013]



How to match documents?

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



How to match documents?

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

SVM			Naive Bayes		
	SF	OE		SF	OE
SF	103	35	SF	98	40
OE	44	90	OE	44	90
Accuracy	*70.96%*		*Accuracy*	*69.12%*	

Features with ConceptOrg			Features with ConceptSpa			Both types of features		
	SF	OE		SF	OE		SF	OE
SF	108	30	SF	112	26	SF	113	25
OE	47	87	OE	19	115	OE	19	115
Accuracy	*71.69%*		*Accuracy*	*83.45%*		*Accuracy*	*83.82%*	

Part 3

Applications in agricultural domain

Animal disease surveillance

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

Why the need of Epidemic Intelligence? (1/3)

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

Why the need of Epidemic Intelligence? (2/3)

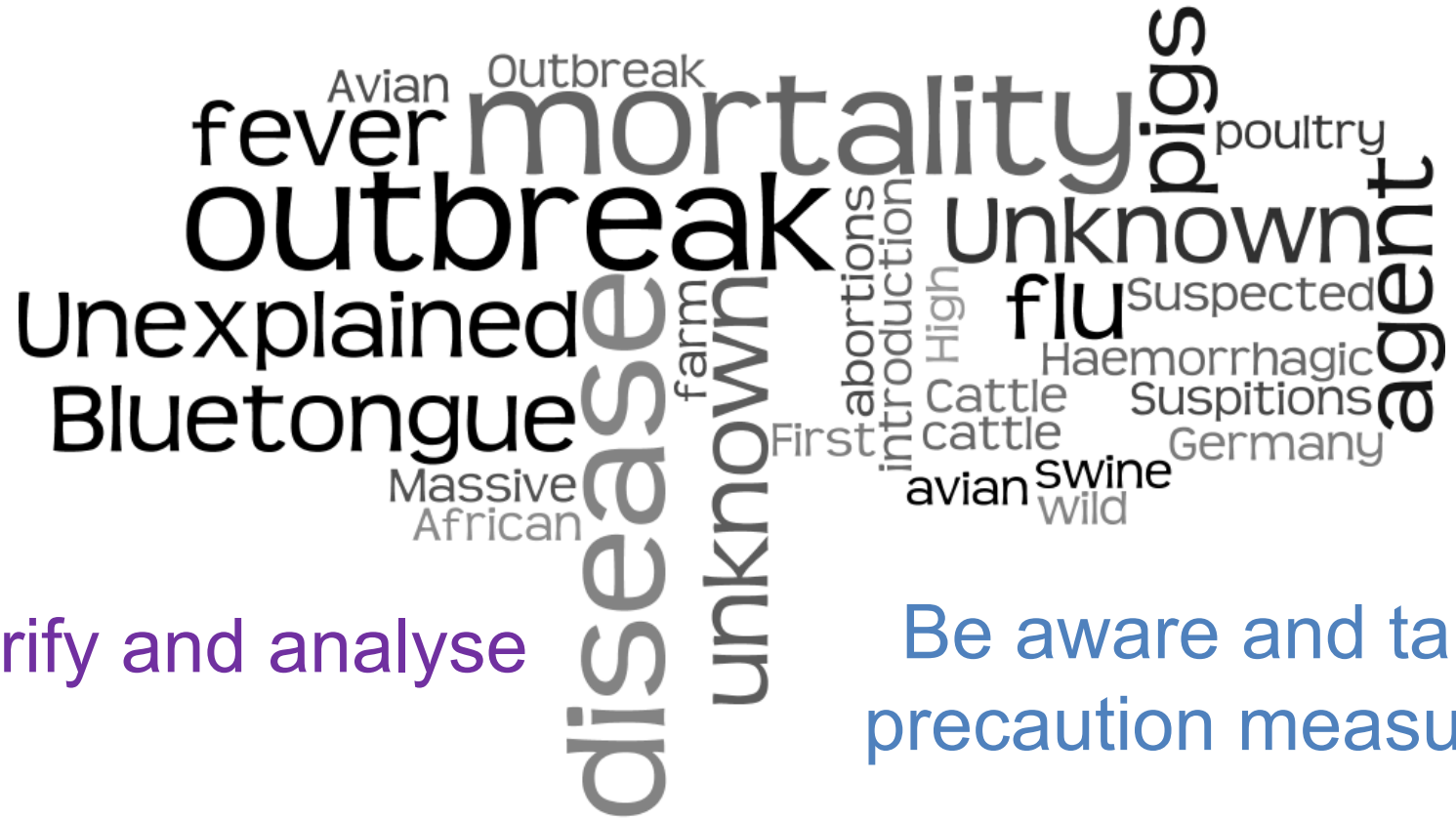
Experts on the field have a **crucial role** in verification and exchange of sanitary **multilingual** information

Timeliness in identification, verification and exchange of information is essential for taking **preventive measures** against introduction of animal health risks

Primary focus on **exotic infectious diseases** – big impact on animal health if introduced

Why the need of Epidemic Intelligence? (3/3)

Identify signals of new and exotic animal diseases



Verify and analyse

Be aware and take precaution measures

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 model to study**: 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

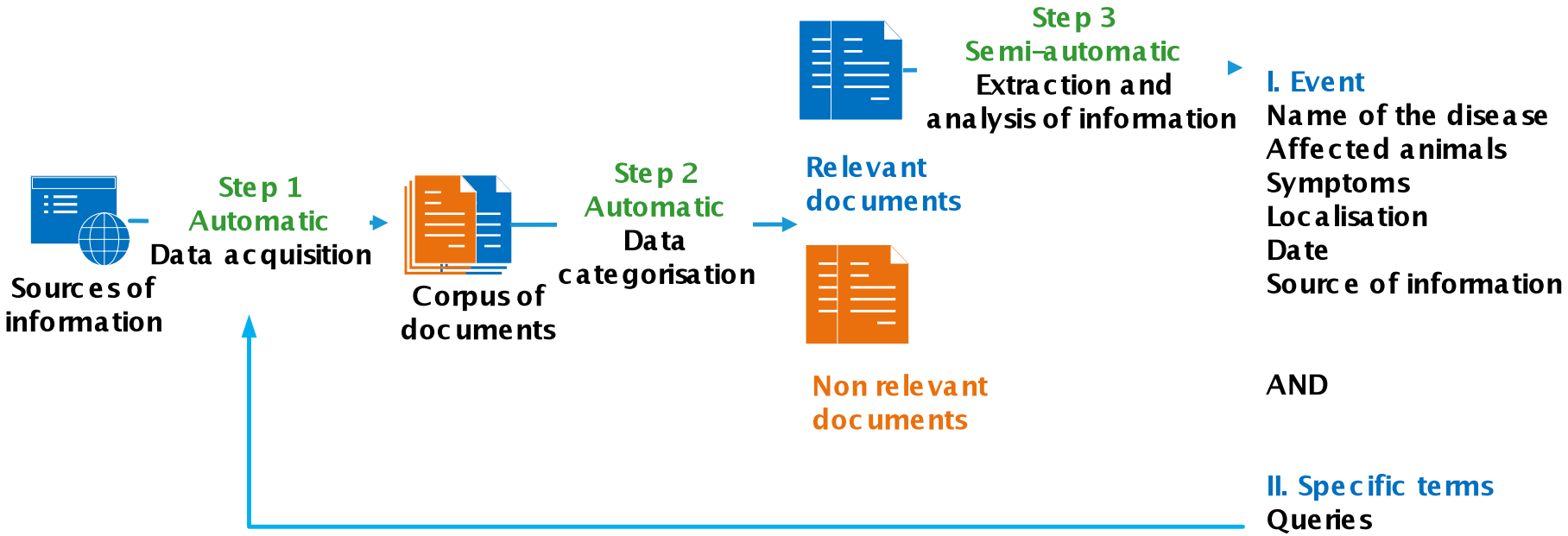
Polish local authorities said on Wednesday that preliminary tests have pointed to a case of African swine fever (ASF) among farm pigs in eastern Poland near the city of Bialystok.

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.

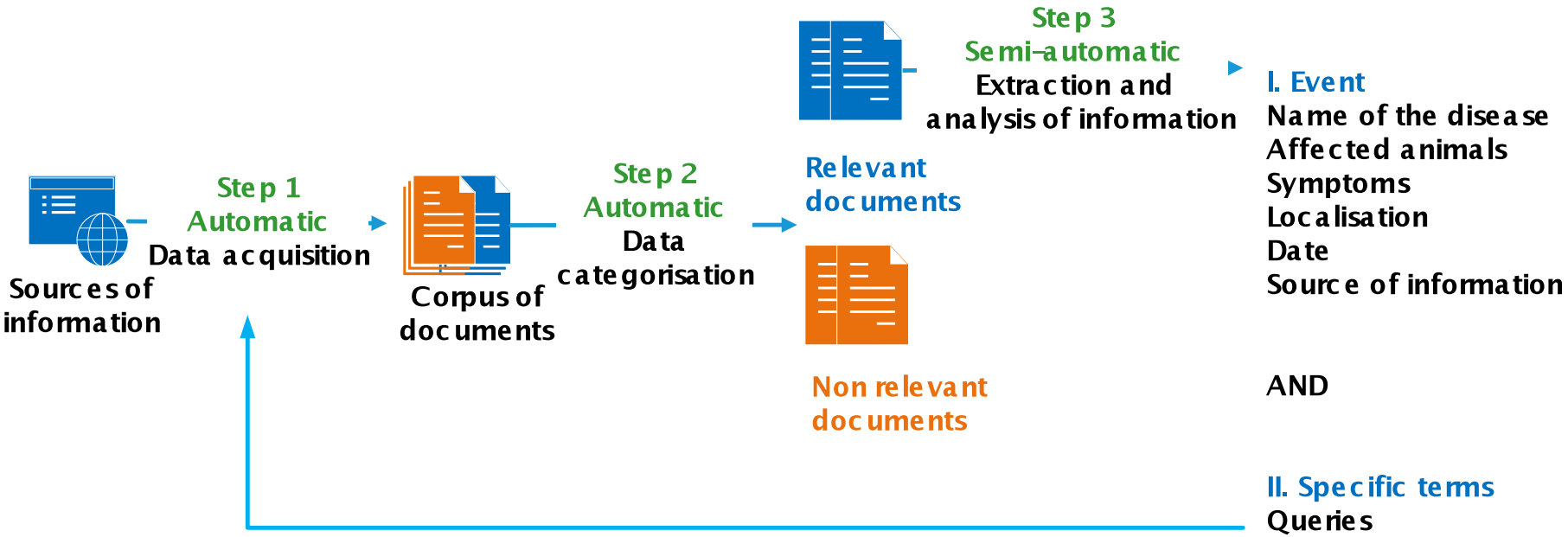
Poland's chief veterinary officer was unavailable for comment, while the county veterinary officer said a statement on the issue will be published later on Wednesday. (Reporting by Anna Wlodarczak-Semczuk; Writing by Marcin Goettig)

Methodology



Methodology - *step 1*

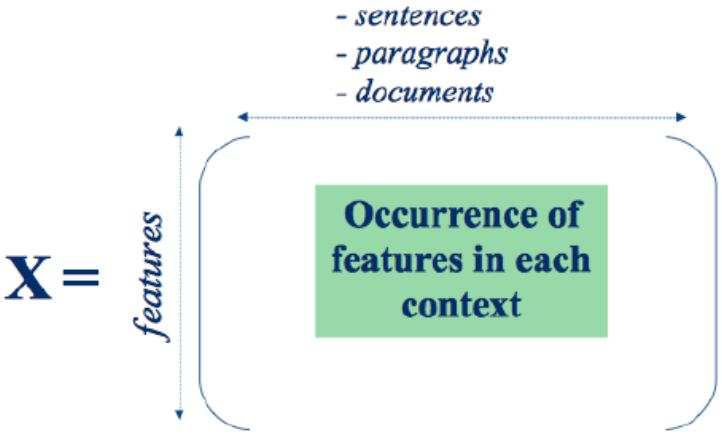
- **Step 1: Data acquisition**

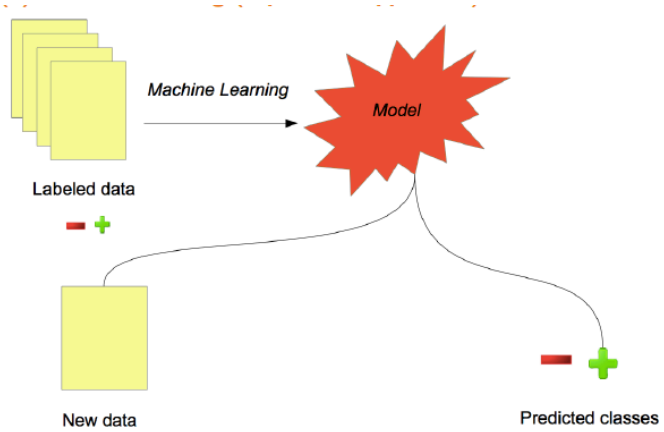


*https://news.google.com/news/feeds?pz=1&cf=all&ned=en&q=**Blue+tongue**&output=**rss***

Methodology - *step 2*

- **Step 2: Data classification**

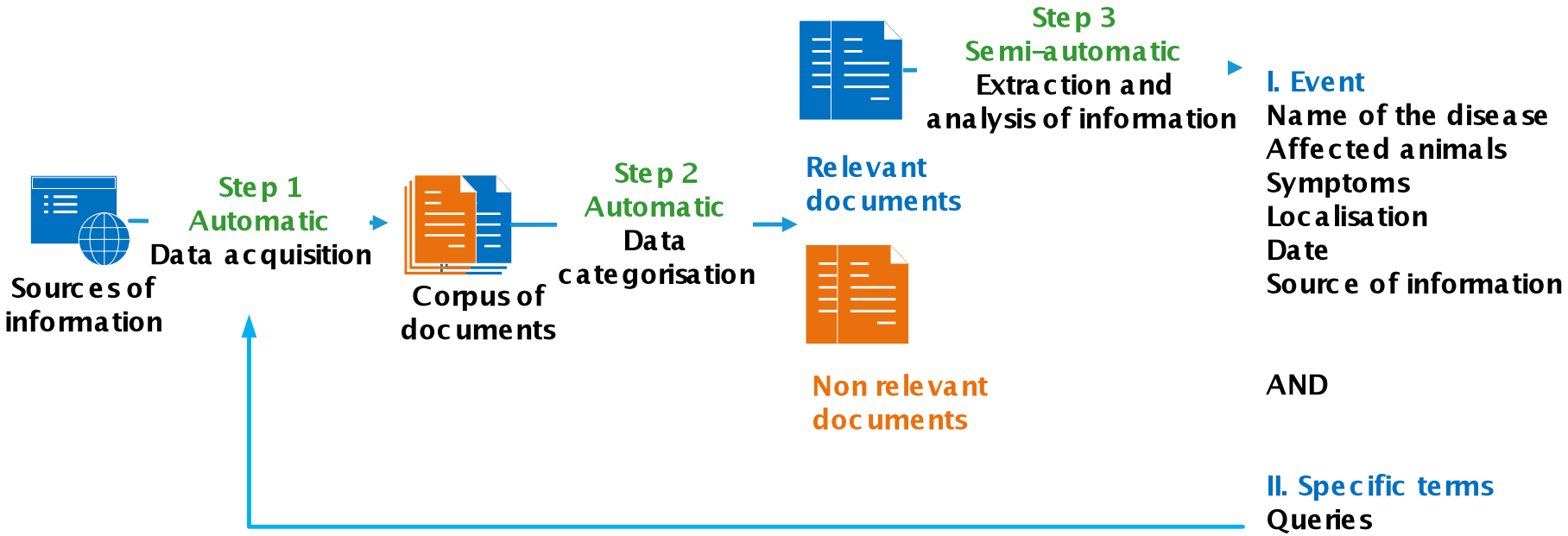




Classification algorithm		Naïve Bayes			Support Vector Machine		
Performance		Recall	Precision	F-score	Recall	Precision	F-score
Class	*disease*	0.724	0.766	0.744	0.657	0.68	0.669
	economy	0.478	0.530	0.503	0.489	0.726	0.584
	general	0.860	0.804	0.831	0.864	0.763	0.810
Weighted average		**0.750**	**0.745**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

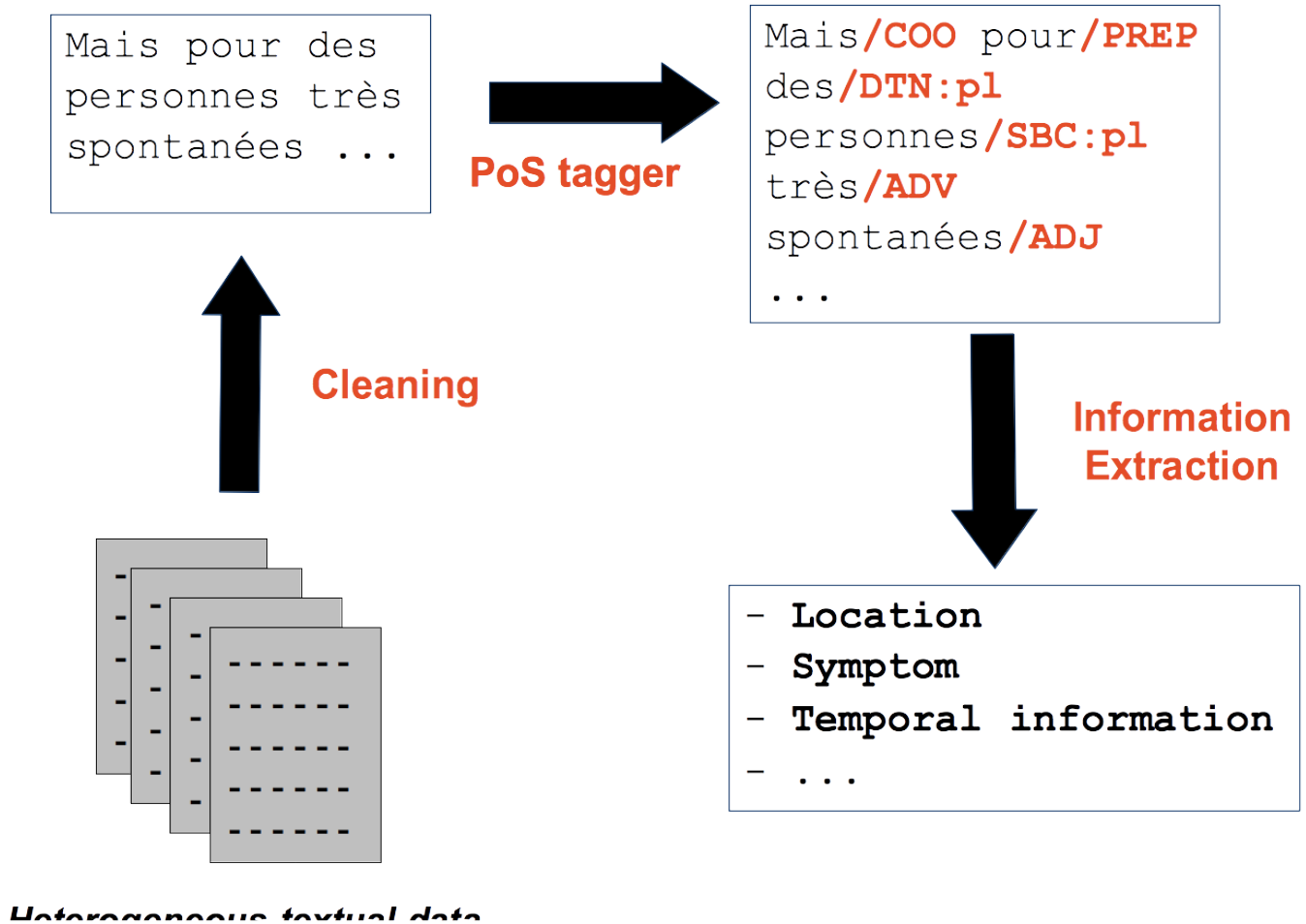
Methodology - *step 3*

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



Methodology - *step 3*

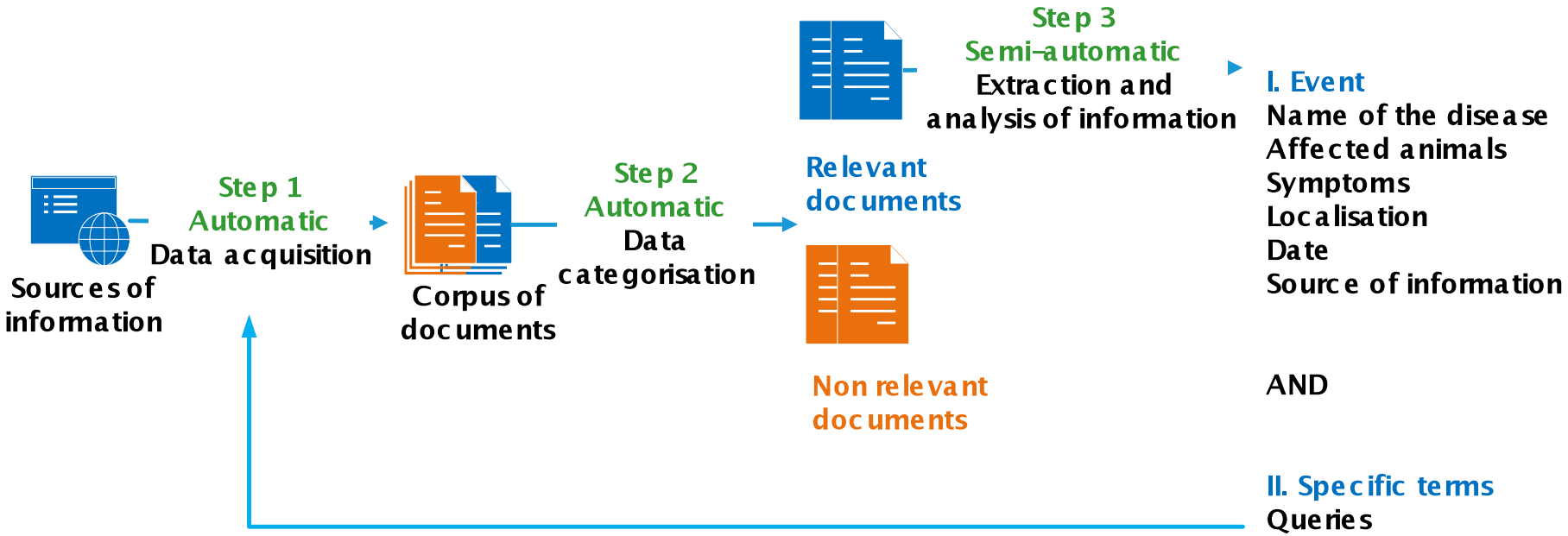
- **Step 3: Information extraction (I)**



Heterogeneous textual data

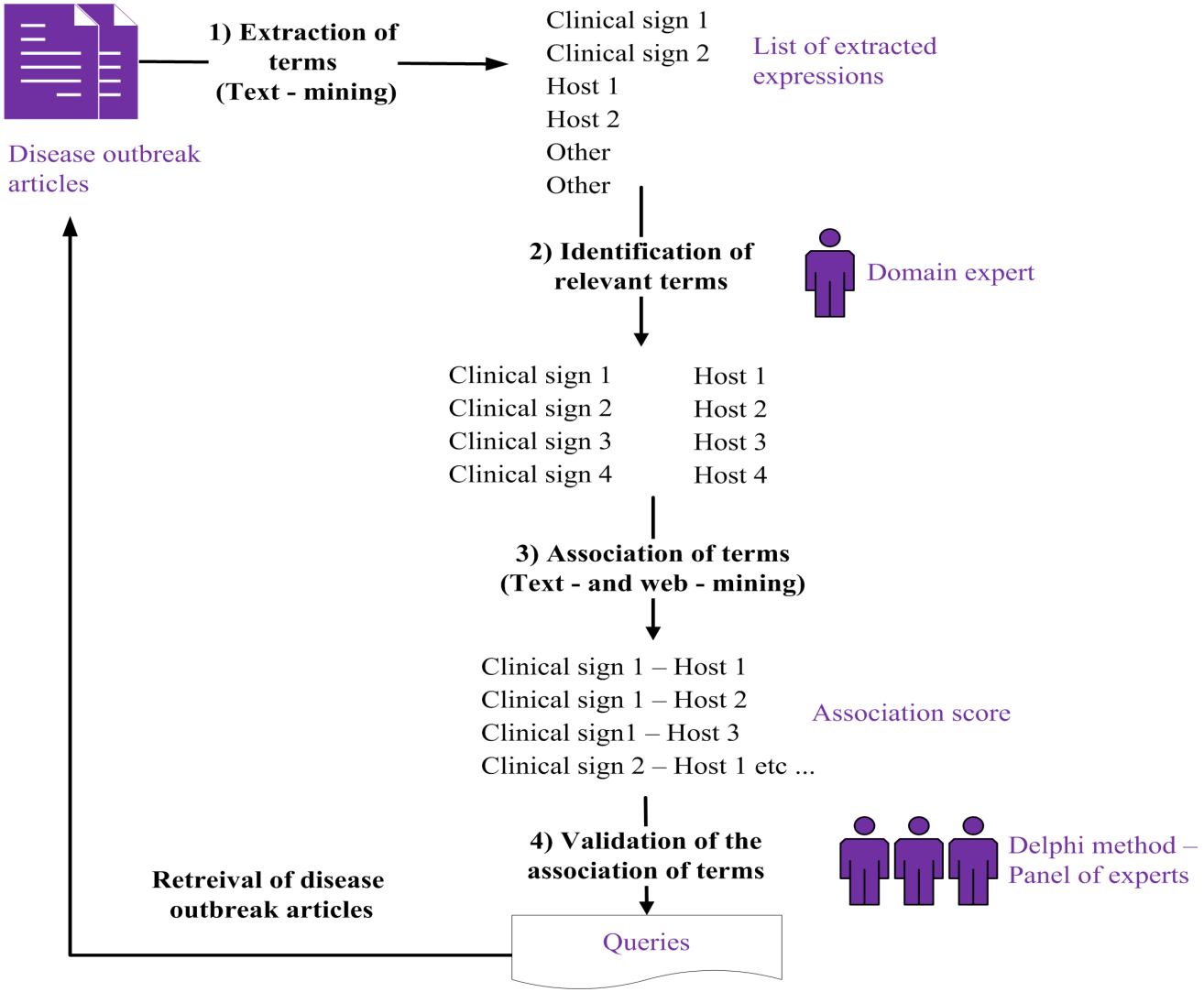
Methodology - *step 3*

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



Methodology - *step 3*

- **II. Querying the Web**



Methodology - *step 3*

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

Methodology - *step 3*

- **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*

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

- BioTex Ranking [Lossio Ventura *et al.* IRJ'2015]:

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

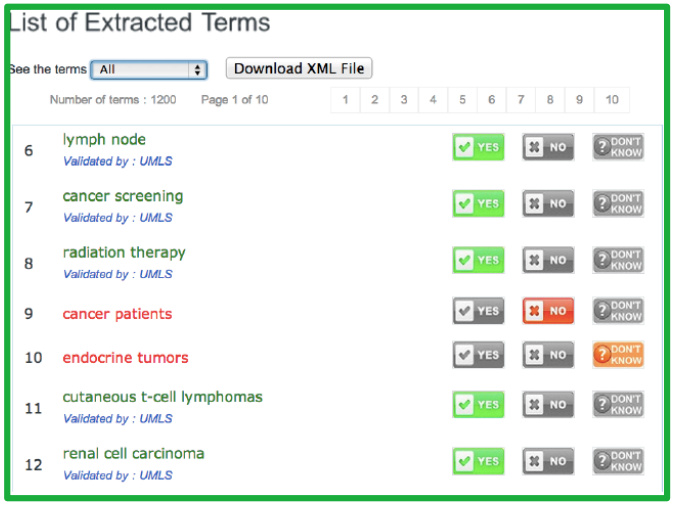
- A **new ranking function** to take into account the **heterogeneity** of the sources (S_i) [Arsevska *et al.* MSR'2014]:

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

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

Methodology - *step 3*

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



How to evaluate ranking functions?

Using of a **Delphi method**.

*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*

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

List of **extracted terms** identified to characterize Bluetongue virus (BTV) emergence.

Clinical signs	**Term**
General	**livestock deaths**, **general clinical signs**, **onset of weakness**, excess mortality, **fever outbreak**
Reproductive	**embryonic death**, **reproductive disorders**, **occurrence of abortion**
Hosts	**Term**
	red deer, adult **sheep**, **cattle** herds, **roe deer**, cattle population, **newborn calves**, new born dairy calves, dairy calves, dairy ewes, **pregnant ewes**, cattle and **goats**, small ruminants

In bold are the terms proposed to experts for evaluation

Methodology - *step 3*

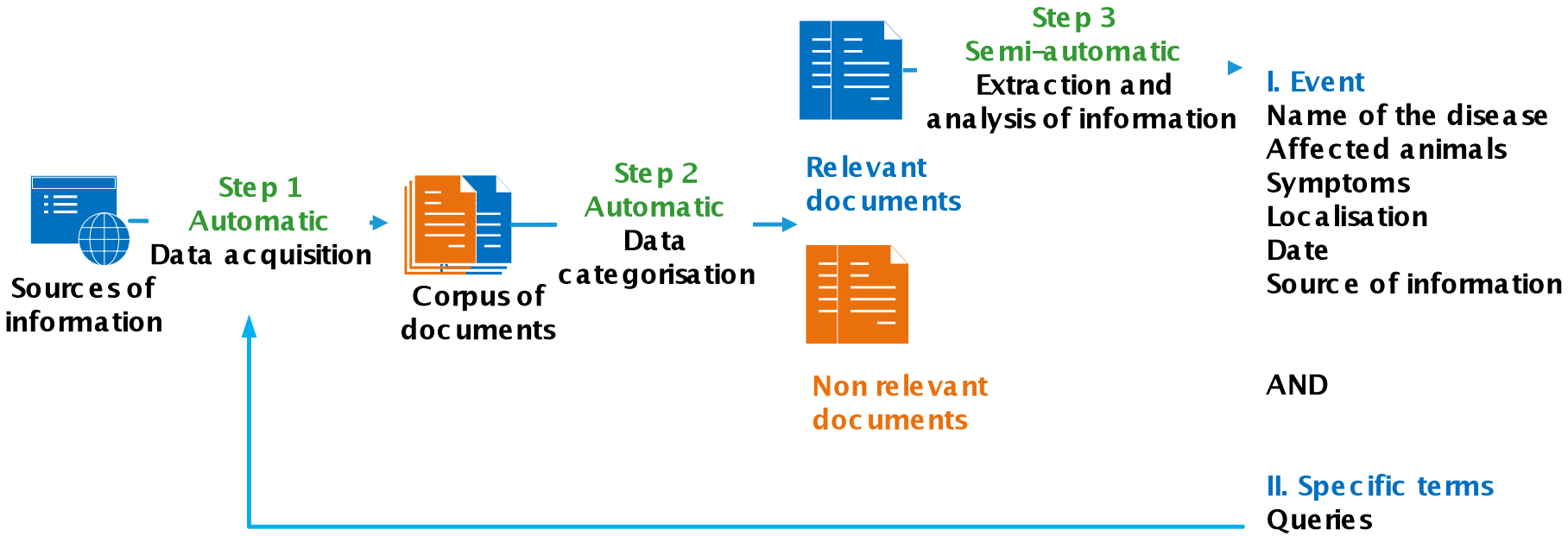
- **Querying the Web:** (d) ***Association of terms***

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

[Roche and Prince Informatica'2010]

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

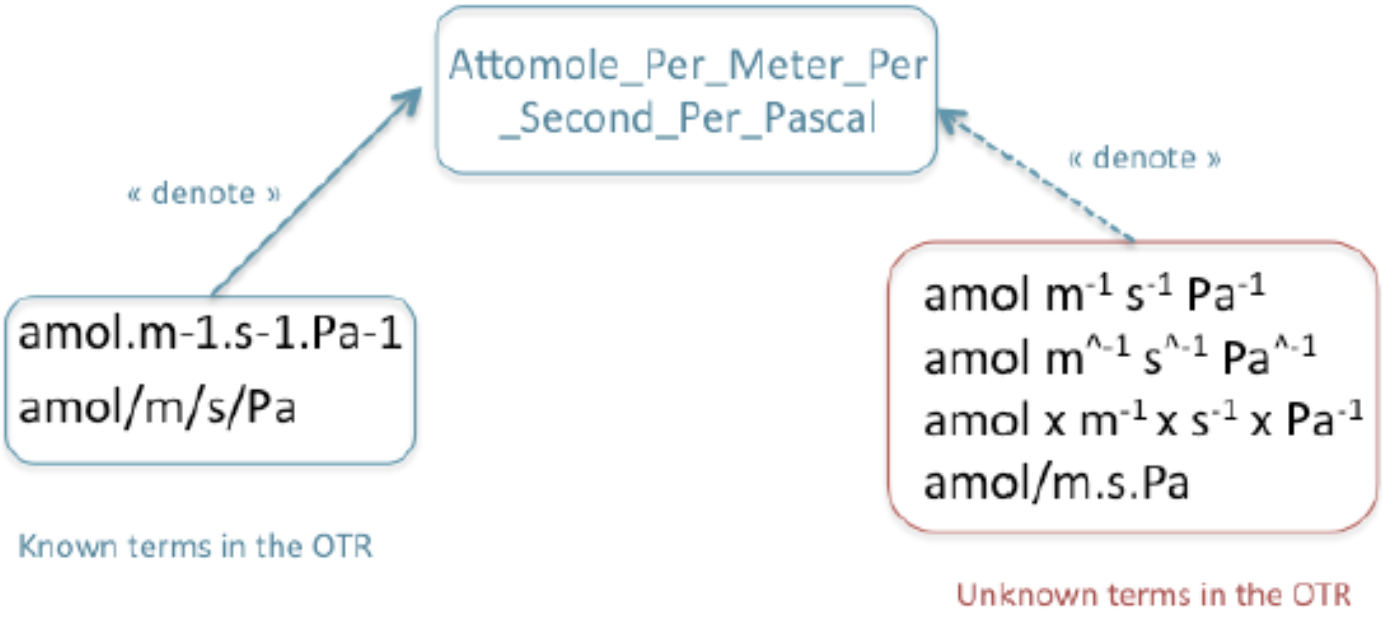
Applications in agriculture domain

Information Extraction from experimental data

Food science domain

Aim: Knowledge management in food science domain

Challenging issue: Unit recognition and extraction
[Berrahou *et al.* KDIR'2013]



Food science domain

Method:

- **Locating unit** (*machine learning*)

- **Extracting unit** (*lexical similarity*)

  Measures based on
  - **Levenshtein** distance
  - **n-grams** of characters

Food science domain

Method:

- **Extracting unit** (*lexical similarity*)

$$SM_{DL}(u1,u2) = max[0; \frac{\min(|u1|,|u2|) - DL(u1,u2)}{min(|u1|,|u2|)}]$$

$$\in [0;1]$$

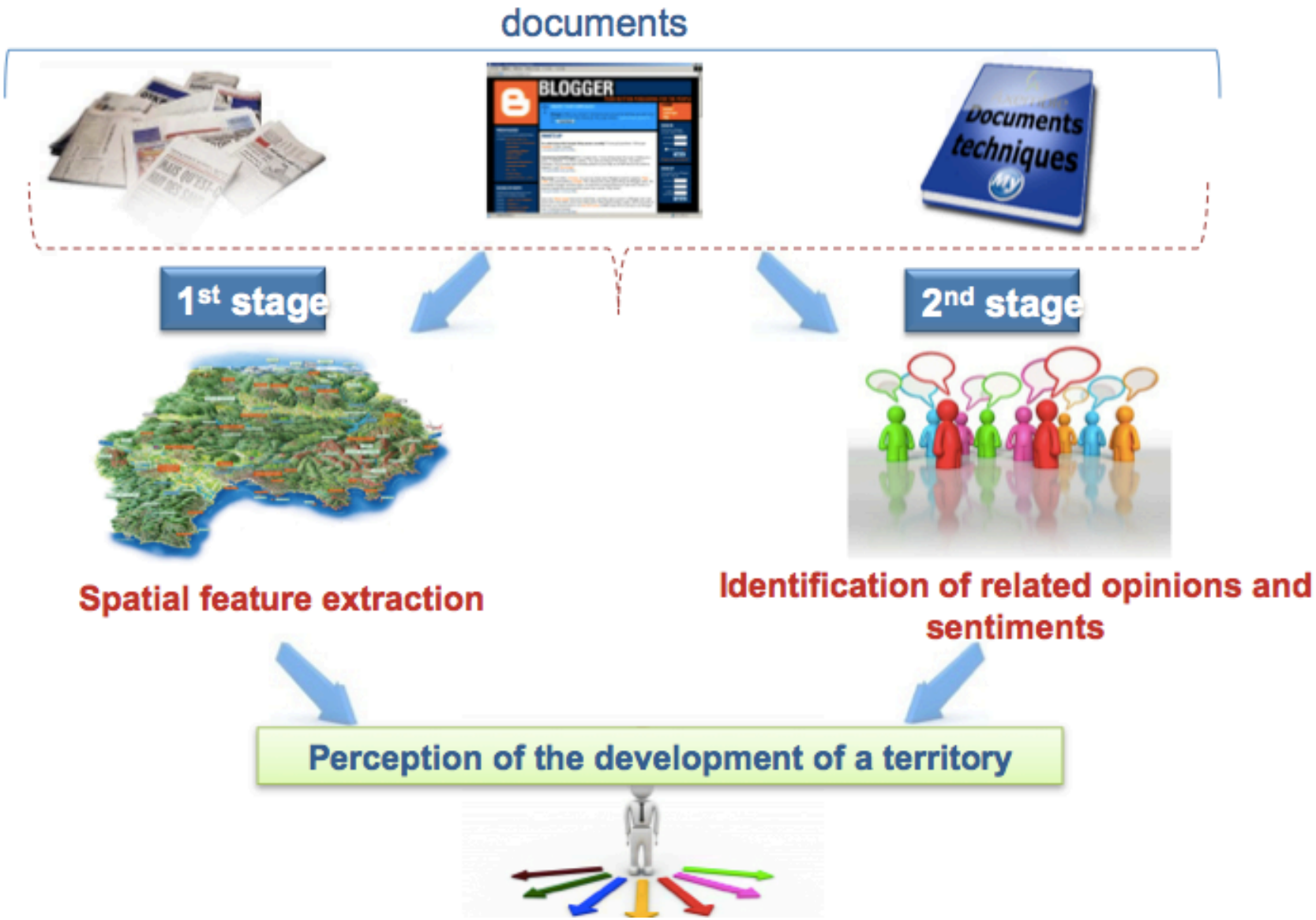
Variant term	Reference	SMDc	SMDb
10e10 (cm3.m-1.sec-1.Pa-1)	10e10.cm3.m-1.sec-1.Pa-1	0.87	1
10e-14(cm3/m.s.Pa)	10e-14.cm3/(m.s.Pa)	0.89	1
10e-16cm3.cm/cm.cm2.s.Pa	(10e-16cm3.cm)/(cm2.s.Pa)	0.76	0.8
10e18 (mol.m/Pa.sec.m2)	10e18.mol.m/(Pa.sec.m2)	0.87	1
amol.m-1.s-1.Pa-1	amol.s-1.m-1.Pa-1	0.88	0.75
amol/m.s.Pa	amol/(m.s.Pa)	0.84	1
amol/m.sec.Pa	amol/(m.s.Pa)	0.69	0.75
cm3.um/m2.d.kPa	cm3.μm/(m2.d.kPa)	0.77	0.8

Part 4

Applications in environmental domain

Opinion mining and land-use planning

Senterritoire



Part 5

Conclusions and future work

Conclusion

New challenges of *Big Data*:

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

- **Multilingual textual analysis** with precision

- **Integration of visual analytics** skills