



Plateforme VESPA Mining pour accéder aux archives de l'épidémiosurveillance végétale.

Nicolas Turenne

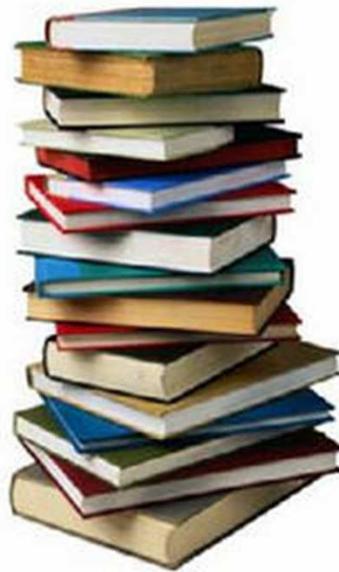
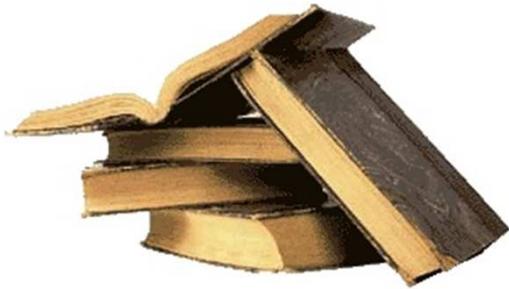
Université Paris-Est & INRA

UMR *LISIS*

(Joint Research Unit *Interdisciplinary Laboratory Sciences Innovations Societies*)

Wednesday, January 07th, 2015

Availability of Information



-500 BC 1500

1500 2000

2000 2015

+ 50 % of all information
informational deluge

Text and Corpus processing

a needle in a haystack

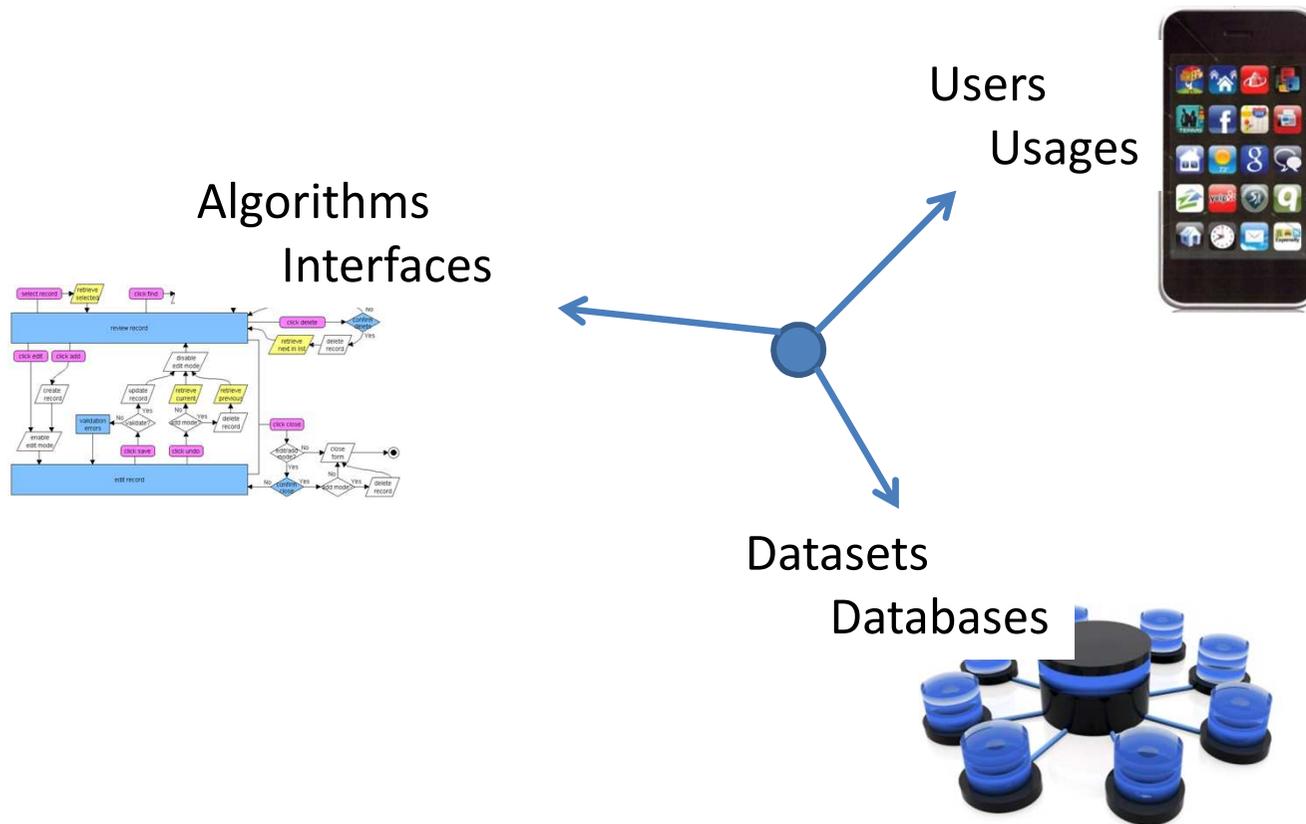
- Existing Research Communities since 1950
 - Natural Language Processing (Syntax Analysis, Logic)
 - Lexical Statistics
 - Information retrieval
- Different Tasks (basic usages)
 - Text Classification
 - Question/Answering
 - Automatic Translation
 - Ontology and Thesaurus Acquisition
 - Information Extraction
 - Open Survey Analytics
 - Opinion & Sentiment Analysis



Information Extraction

- IE = extracting information from text
- Extract entities
 - People, organizations, locations, times, dates, prices, ...
 - Or sometimes: genes, proteins, diseases, medicines, ...
- Extract the relations between entities
 - Located in, employed by, part of, married to, ...
- Extract scenario of relations
 - network reconstruction, figure out set of events

3 cornerstones



Information Extraction: relation extraction

Champagne-Ardenne

Bulletins Techniques des Stations d'Avertissements Agricoles n° 630 du 07 juillet 2004 - 2 pages

Relation Crop-Disease

Instance 1
wheat/ 'take-all-disease

Relation Crop-Pest

Instance
Beetroot/black aphid

Grandes Cultures

Blé

Stades : Grain pâteux mou à grain pâteux dur dans la majorité des parcelles de blé et d'orge de printemps.

Maladies

Avec le stress hydrique marqué des plantes dans certaines parcelles, les symptômes de piétin échaudage sont particulièrement visibles cette année (taches plus ou moins circulaires de pieds présentant des racines atrophiées et noires à la base, avec un manchon noir sur la tige).

• Sur les parcelles attaquées, le meilleur moyen de lutte est la rotation : éviter de remettre une céréale à l'automne (sauf l'avoine qui est résistante à ce champignon) et, avant de remettre une céréale, éviter les précédents favorisant l'expression de la maladie (maïs, ray-grass, luzerne ou soja). Éviter aussi les semis précoces et les fortes densités de semis qui favoriseront le développement du champignon. Certains traitements de semences (Jockey ou Latitude) ont une efficacité sur la maladie, mais ils ne peuvent enrayer les fortes infestations.

• Observer les parcelles : une application fongicide à base de tébuconazole ou metconazole est à prévoir contre la rouille lors du passage des premières pustules sur les étages intermédiaires.

Betterave

Stades : 90 à 100% de couverture du sol en terres colorées ; 70-100% en terres de craie.

Pucerons

Des colonies de pucerons noirs sont à présent visibles avec 88% à 100% des pieds sur la plupart des parcelles de notre réseau d'observation de la Marne et des Ardennes traitées depuis plus de 3 semaines (Isse-51, Cormicy-51, Bouchy-51, Bagneux-51, Barbey-08 ...). Des manchons couvrant la totalité des jeunes feuilles au cœur des betteraves sont parfois visibles sur plus de 10% des pieds !. Les parcelles gauchou ou imprimé sont également concernées.

Les interventions réalisées la semaine dernière avec du triazamate ont montré une très bonne efficacité, et permis le maintien des populations de micro-hyménoptères parasites des pucerons qui commencent à s'implanter dans les parcelles de betteraves. En revanche, les parcelles situées dans l'Aube ne semblent pas concernées par ces



Prochain bulletin prévu courant juillet, en fonction de l'actualité.



CEREALES
Piétin échaudage sur blé.

FEVEROLE
Rouille à surveiller

BETTERAVE
-Pucerons noirs
-Maladies encore

Information Extraction: NER our approach

- Dictionary domain-based
 - Crops, diseases, pests, auxiliaries, region, towns , chemicals

7 concepts

blé:**N**:blé:BLE:blés:Triticum:blé dur:blé tendre:

blé dur:**L**:BLE DUR:T. durum:Triticum durum:bles durs:blés durs:blé dur:

blé noir:**L**:BLE NOIR:f. esculentum:fagopyrum

esculentum:sarrasin:bles noirs:blés noirs:blé noir:sarrasins:

blé tendre:**L**:BLE TENDRE:T. aestivum:Triticum aestivum:blé froment:blés froments:ble froments:blé tendre:blés tendres:bles tendres:

wheat (species)
durum wheat (variety)

buckwheat (variety)

soft wheat (variety)

	entities						
	auxiliaries	crops	pests	diseases	chemicals	regions	towns
#entries	28	114	373	275	4968	26	33161
#leafs	28	103	334	241	4968	26	33161
#concepts	0	18	53	40	0	0	0
#lexems	107	727	2673	1846	4968	869	89603

Information Extraction: NER our approach

- Evaluation

37 annotated files (with entities and relations)

Annotation Cost: 1000 files ~5 months work by 1 person

CoNLL or BIO/BILOU format

Our format

Champagne-Ardenne
.
BSV
du
09/06/2011
--
semaine
23
A
RETENIR
CETTE
SEMAINE
.
TOURNESOL
:
Pucerons
:
fin
du
risque
.
Absence
de
maladies
.
MAÏS
:
Pyrale
:

REG

5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:r:\$:CHAMPAGNE-ARDENNE:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:\$:colza:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:b:\$:charançon de la tige du
colza:charançon de la tige du chou:Méligèthes:mouche du chou:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:d:\$:25.03.2010:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:v:\$:St
Dizier:Reims:Charleville:Langres:TROYES:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:n:\$:peu nuisible:à risque:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:s:\$:c2:c1:d1:f1:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:\$:colza:charançon de la tige du
colza:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:\$:colza:mouche du chou:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:\$:colza:Méligèthes:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:\$:colza:c2:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:\$:colza:d1:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:\$:colza:c1:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:n:\$:colza:charançon de la tige
du chou:peu nuisible:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:n:\$:colza:mouche du chou:à
risque:1

Information Extraction: NER our approach

- Evaluation

$$P = \frac{\#correct_answers}{\#produced_answers}$$

$$R = \frac{\#correct_answers}{\#possible_correct_answers}$$

$$F - score = \frac{(\beta^2 + 1)P.R.}{(\beta^2 R + P)}$$

SNER - Average						
Entity	P	R	F1	TP	FP	FN
BIO	92,66%	71,41%	80,52%	184,32	14,79	76,18
MAL	95,46%	77,38%	85,38%	98,46	4,68	30,39
PLA	93,99%	82,68%	87,94%	277,14	16,86	56,11
REG	93,20%	73,73%	81,92%	40,18	3,36	16,82
Totals	93,68%	76,85%	84,41%	600,11	39,68	179,50

X.ent						
Entity	P	R	F1	X_Y	X	Y
BIO	96,46%	95,52%	95,98%	299,00	313,00	310,00
MAL	96,97%	95,53%	96,24%	192,00	201,00	198,00
PLA	88,80%	98,67%	93,47%	222,00	225,00	250,00
REG	100%	100%	100%	37,00	37,00	37,00
Totals	94,33%	96,67%	95,48%	750,00	766,00	795,00

Information Extraction: NER our approach

- Rule-based

- other kinds of extraction, not specifically named entities

5 concepts

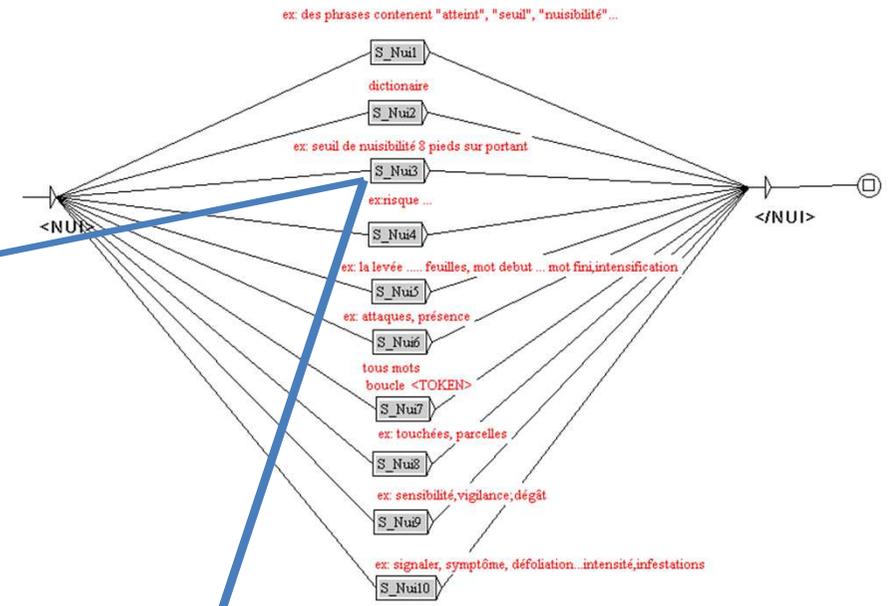
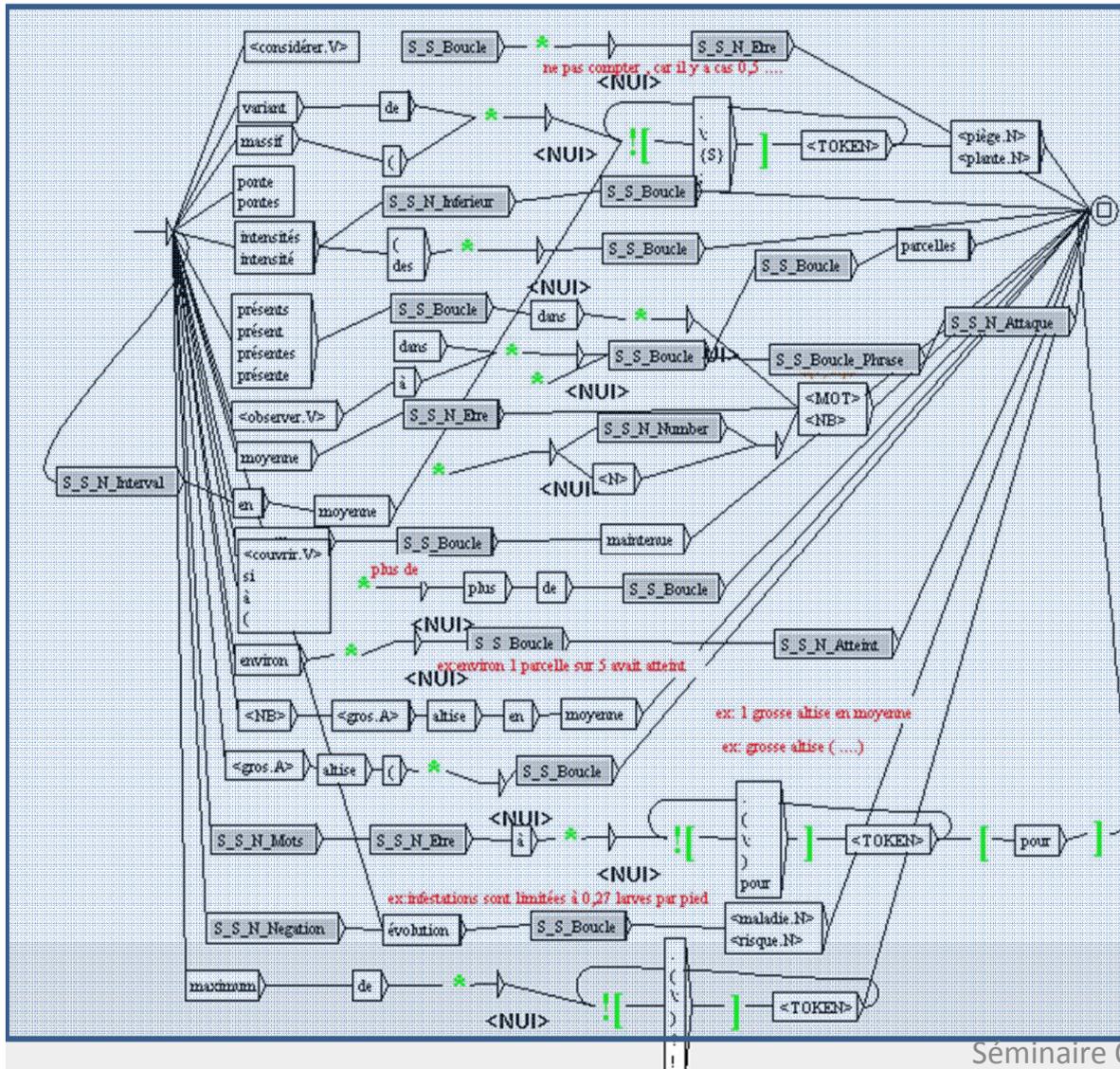
- Developmental stages, Risk assessment, Climat, Number of issue , Date
 - For towns , mixed approach rule-based and dictionary-based.

- No use of dictionaries but local grammar

- Handcrafted-rules with markers : such as « week » for *Number of issue*, of « xx {January | February...} xxxx » for a *date*
 - We use UNITEX tool to make rule . It is a Finite-State Automata Graphical Tool
 - Easy to make language models.

Information Extraction: NER our approach

- Unitex rules
 - Risk Assessment detections



infestations sont limitées à 0,27 larves par pied
 environ 1 parcelle sur 5 avait atteint
 1 grosse altise en moyenne

*infestations are limited to 0.27 larvae per foot
 about 1 parcel in 5 had reached
 1 large flea beetle in average*

Information Extraction: relation extraction

These are the main approaches for relation extraction.

Qualitative approaches:

- Exact analysis from lexical dictionary ('Exact Dictionary-Based Chunking')
- Hand-crafted pattern definition techniques

And optimisation approaches:

- Symbolic Learning Models ('inductive-logic programming')
- Statistical Learning Models ('Bayesian network analysis')
- Unsupervised Learning Models ('co-occurrence analysis')

Information Extraction: relation extraction

Our approach is a combination between :

- *Hand-crafted pattern definition techniques*

In the sense that we consider some **rules about the document design**.

&

- *Unsupervised Learning Models*

In the sense that we consider **detection of cooccurrence** without POS tagging attachment.

Information Extraction: relation extraction

Grandes Cultures

Champagne-Ardenne

Bulletins Techniques des Stations d'Alertes Agricoles n° 630 du 07 juillet 2004 - 2 pages

Rule 1:
Named Entity(crop) @ HEADER

Rule 2:
Named Entity(region) @ BEGIN

Rule 3:
Named Entity(issue) @ BEGIN

Rule 4:
Named Entity(date) @ BEGIN

Rule 5:
Reject Named Entity(*)
IN AVOID_PART

DRAF
Service Régional de la
Protection des Végétaux
Centre de Recherches
Agronomiques
Explanade Roux
Garny - BP 234
51096 REIMS Cedex
Tél : 03.26.77.36.40
Fax : 03.26.77.36.74
E-mail : srpv.draf@
champagne-ardenne@

Blé

Stress hydrique marqué des plantes dans la majorité des parcelles de blé et d'orge de printemps.

Maladies

Avec le stress hydrique marqué des plantes dans certaines parcelles, les symptômes de piétin échaudage sont particulièrement visibles cette année (taches plus ou moins circulaires de pieds présentant des racines atrophiées et noires à la base, avec un manchon noir sur la tige).

• Sur les parcelles attaquées, le meilleur moyen de lutte est la rotation : éviter de remettre une céréale à l'automne (sauf l'avoine qui est résistante à ce champignon) et, avant de remettre une céréale, éviter les précédents favorisant l'expression de la maladie (maïs, ray-grass, luzerne ou soja). Éviter aussi les semis précoces et les fortes densités de semis qui favoriseront le développement du champignon.

Certains traitements de semences (Jockey ou Latitude) ont une efficacité sur la maladie, mais ils ne peuvent enrayer les fortes infestations.

• Observer les parcelles : une application fongicide à base de tébuconazole ou metconazole est à prévoir contre la rouille lors du passage des premières pustules sur les étages intermédiaires.

Betterave

Certains traitements de semences (Jockey ou Latitude) ont une efficacité sur la maladie, mais ils ne peuvent enrayer les fortes infestations.

Pucerons

Des colonies de pucerons noirs sont très présentes sur la plupart des parcelles de notre réseau d'observation de la Marne et des Ardennes traitées depuis plus de 3 semaines (Basse-Saône, Comtois-Saône, Douce-Saône, Bagnac-Saône, etc.). Des manchons couvrant la totalité des jeunes feuilles au cœur des betteraves sont parfois visibles sur plus de 10% des pieds ! Les parcelles gauchon ou imprimé sont également concernées.

Les interventions réalisées la semaine dernière avec du triazamate ont montré une très bonne efficacité, et permis le maintien des populations de micro-hyménoptères parasites des pucerons qui commencent à s'implanter dans les parcelles de betteraves. En revanche, les parcelles situées dans l'axe ne semblent pas concernées par ces

Prochain bulletin prévu courant juillet en fonction de l'actualité.

- CEREALES
- Piétin échaudage sur blé.
- FEVEROLE
- Rouille à surveiller
- BETTERAVE
- Pucerons noirs
- Maladies encore

Information Extraction: relation extraction

Rule 5:
Reject Named Entity(*)
IN AVOID_PART

Gr

Fredec
Midi-Pyrénées

Service Régional de la
Protection des Végétaux
Cité Administrative
Bâtiment E
Bd Armand Duportal
31074 Toulouse Cedex
Téléphone : 05.61.30.62.72
Tél. abonnement :
05.61.38.62.78

Inspiré à la station
d'Essais Agricoles
de Midi-Pyrénées
Direction générale :
J.P. MOUREZIERES
Publication périodique
C.F.P.A. P n° 332AD
ISSN n° 07523052

Statut T.J.C. :
Avertissement Agricoles :
Couvert : 44 € - Fixe : 28 €
© Guide : 34 €

La colonisation des parcelles de colza est actuellement très faible et ne pourra intervenir de façon significative qu'avec le retour d'un temps doux et ensoleillé.

■ **Il est trop tôt pour intervenir.**
Attendre le retour de conditions douces et ensoleillées favorables à leur déplacement : 3 jours consécutifs avec des températures supérieures à 10°C.
Intervenir alors 8 à 12 jours après cette période de réchauffement, pour obtenir une efficacité optimale.

Céréales
Statut végétatifs
Plein à fin tallage pour la majorité des parcelles, épi 1 cm pour les plus précoces, début tallage pour les plus tardives.

Jaunisse nanisante
Depuis début janvier, le froid a limité la multiplication des pucerons en parcelle. Néanmoins, les notations de mi-février sur les isoriques confirment la pression pucerons ponctuellement importante sur les semis réalisés avant le 25 octobre coloni-

même dans les semis précoces.
Les premières simulations des modèles confirmeront la faible pression de maladies. Seuls quelques symptômes de septoriose sont observés sur les feuilles basses de semis précoces, avec une présence notable de symptômes d'*Ascochyta*.

Le développement du piétin verse a lui aussi été limité par les conditions climatiques. Toutefois, si le mois de mars est particulièrement pluvieux, les simulations du modèle piétin verse indiquent un risque potentiel non négligeable sur les semis précoces d'octobre levés avant le 10 novembre, dans les situations à risque (sol limoneux et retour fréquent du blé dans la rotation).
Les conditions climatiques de mars-avril seront déterminantes pour l'évolution des maladies sur céréales ; la situation actuelle globalement saine peut évoluer rapidement si la fin de l'hiver et le début du printemps sont pluvieux.

Colza
Charançon de la tige :
Ne pas intervenir trop tôt !!

Tournesol
Taupins :
Note commune CETIOM - SPV

Sur orges, seuls quelques symptômes limités d'helmintosporiose sont observés.
Nos bulletins feront régulièrement le point sur l'évolution sanitaire des cultures et vous indiqueront l'évolution des risques en fonction de votre situation.

Reasoning control against

wireworms

Instance: crop/wireworms is false

Raisonner la lutte contre les larves de taupins en tournesol
Note commune CETIOM - SPV : Février 2005

L'objectif de cette note commune CETIOM-SPV est de présenter le raisonnement de la lutte contre les larves de taupins sur tournesol, tel qu'il peut être proposé pour la campagne 2005, en l'état actuel de nos connaissances. De nouvelles données, recueillies dans le cadre d'un réseau de surveillance du risque taupins (dispositif présenté à la fin de cette note) mis en place dès cette année, sont susceptibles de faire évoluer ces conseils de lutte.

1/ Evaluation actuelle du risque taupin en tournesol

Dans le contexte actuel de production du tournesol, le risque d'attaque est globalement faible pour cette culture. Ce risque est limité pour deux raisons : le tournesol est peu attractif pour les larves de taupins et les situations favorables sont peu fréquentes dans les systèmes de culture actuels incluant le tournesol.

- **Une assez faible sensibilité du tournesol :**
Le tournesol fait partie des grandes cultures susceptibles de subir des attaques de larves de taupins. Cependant les références disponibles montrent, d'une part, que cette culture est faiblement attractive pour les larves et que, d'autre part, la période de sensibilité aux attaques est relativement brève (de la germination de la graine au stade cotylédons). La sensibilité globale du tournesol aux larves de taupins est donc assez faible, dans tous les cas bien inférieure au maïs.
- **Des systèmes de cultures incluant le tournesol peu favorables :**
Le tournesol est majoritairement cultivé dans des rotations à base de cultures annuelles. Dans ces rotations, le

Implementation: x.ent

cran.r-project.org/web/packages/x.ent/index.html

Google



x.ent: eXtraction of ENTity

x.ent is a system for extracting information (entities and relations between them) in text datasets. It also emphasizes results exploration with graphical displays. It is a rule-based system and works with hand-made dictionaries and local grammars defined by users. x.ent has been written in perl and use javascript to define user preferences through a browser. Local grammars are defined and compiled with the tool Unitex, developed by University Paris Est, and supporting multiple languages. See ?xconfig for an introduction.

Version: 1.0.6
Depends: R (\geq 3.0.0), [opencpu](#), [rJava](#)
Imports: [stringr](#), [xtable](#), [jsonlite](#), [venneuler](#), [ggplot2](#)
Published: 2014-11-03
Author: Nicolas Turenne, Tien T. Phan
Maintainer: Tien T. Phan <phantien84@gmail.com>
License: [GPL-3](#)
NeedsCompilation: no
SystemRequirements: Perl (\geq 5.0), Unitex (\geq 3.0 <http://www-igm.univ-mlv.fr/~unitex/>)
Materials: [README](#) [NEWS](#)
CRAN checks: [x.ent results](#)

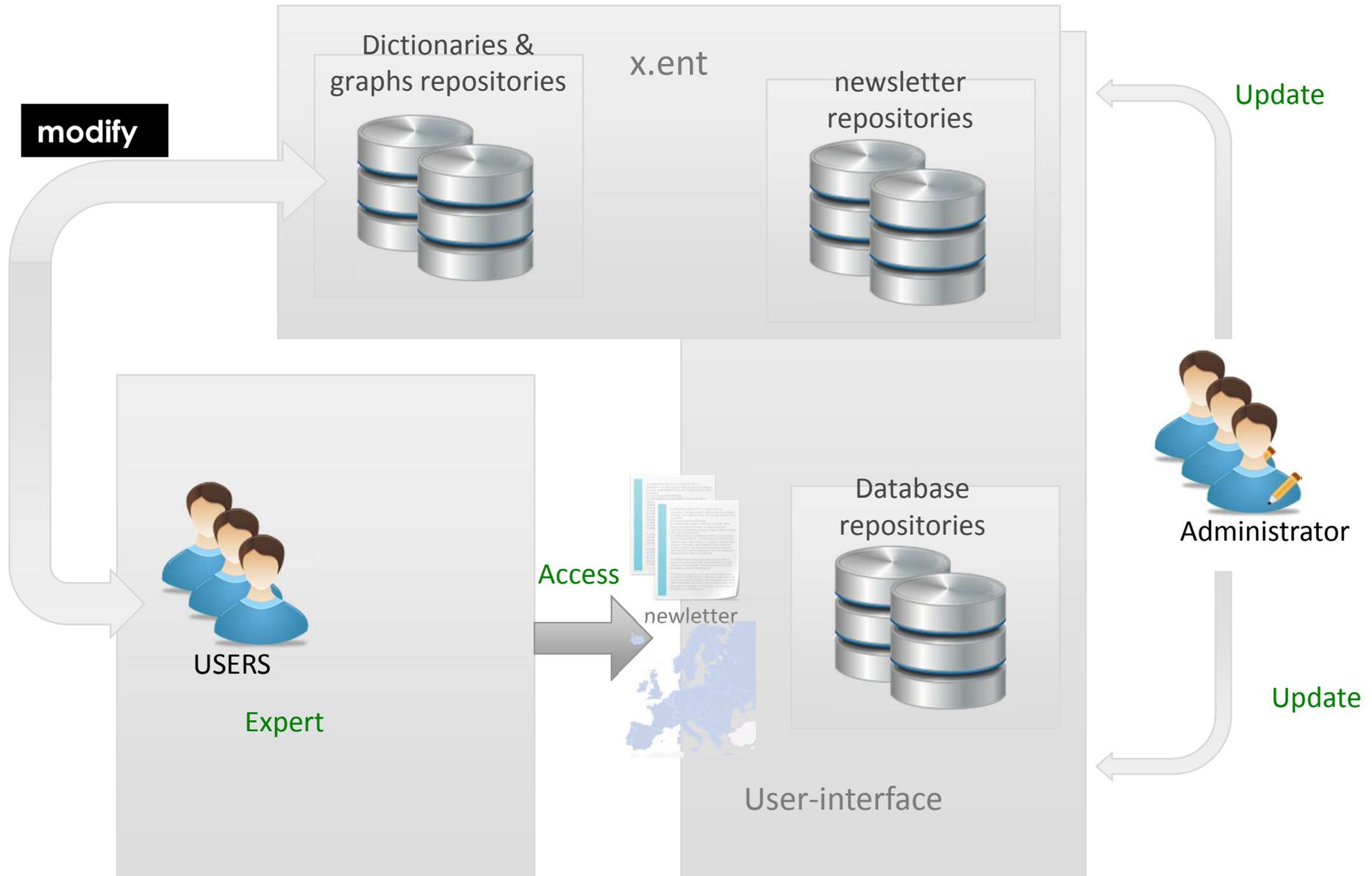
Downloads:

Reference manual: [x.ent.pdf](#)
Package source: [x.ent 1.0.6.tar.gz](#)
Windows binaries: r-devel: [x.ent 1.0.6.zip](#), r-release: [x.ent 1.0.6.zip](#), r-oldrel: [x.ent 1.0.6.zip](#)
OS X Snow Leopard binaries: r-release: [x.ent 1.0.6.tgz](#), r-oldrel: [x.ent 1.0.6.tgz](#)
OS X Mavericks binaries: r-release: [x.ent 1.0.6.tgz](#)

Relation Extraction: evaluation

	relations		
	crops-diseases	crops-pests	<i>total</i>
P	48.9	48.5	48.6
R	61.8	68.1	65.4
F-score	54.6	56.6	55.8

Vespa platform



VESPA platform

 Vespa Mining

Plante
X pomme de terre

Maladie

Ravageur

Date de début
02/11/1945

Date de fin
28/07/2011

Recherche Textuelle

→ LANCER LA RECHERCHE

Les Bulletins
Grandes Cultures

Années

Toutes

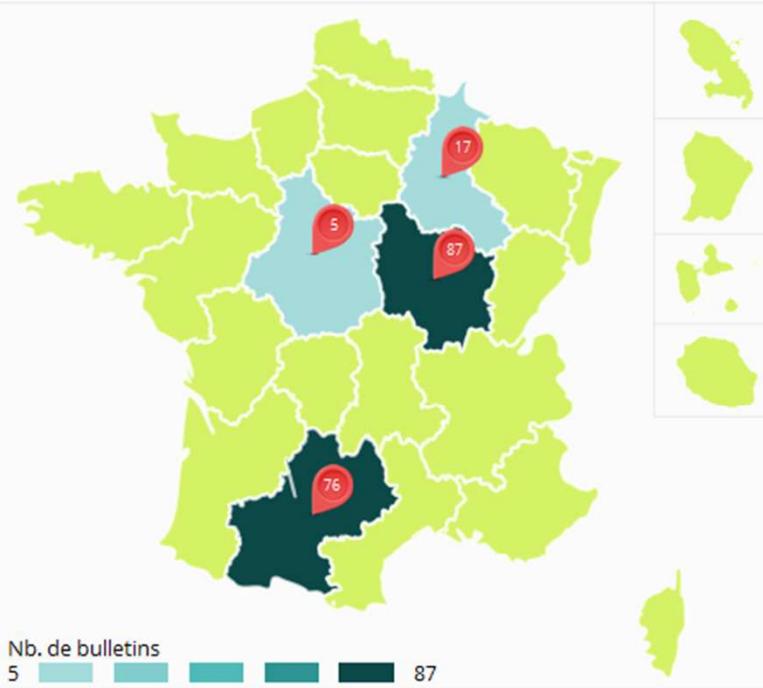
- 1956 (1)
- 1961 (1)
- 1963 (8)
- 1964 (4)
- 1965 (8)
- 1966 (9)
- 1967 (5)
- 1968 (10)
- 1969 (10)
- 1970 (7)
- 1971 (7)
- 1972 (7)
- 1973 (8)
- 1974 (6)
- 1975 (3)
- 1976 (2)
- 1977 (8)
- 1978 (12)

185/185

Trier : Nom Région Date

CHAMPAGNE-ARDEN...	09/08/2006
avisgc 697	
CENTRE	17/08/2005
gc centre 17.08.05 suite	
CHAMPAGNE-ARDEN...	10/08/2005
avisgc 666	
MIDI-PYRÉNÉES	28/02/2005
gc-05-05	
CENTRE	12/08/2004
gc centre 12.08.04	
CHAMPAGNE-ARDEN...	11/08/2004
avisgc 634	
CHAMPAGNE-ARDEN...	03/08/2004
avisgc 633	
CENTRE	29/07/2004
gc centre 29.07.04	
CENTRE	25/02/2004
avisgc 612	
BOURGOGNE	24/06/1998
aa_gc_bourgogne_franche_comt...	
BOURGOGNE	20/12/1995

Bulletins citant  pomme de terre du 02/11/1945 au 28/07/2011



VESPA platform

<http://213.229.108.100/Vespa/Site/Vespa.html>

DEMO



TEST 1

Crop: wheat

Disease: rust

Pest: ./.

On map : risk assessment
region burgundy

Culture: blé

Maladie: rouille

Ravageur: ./.

sur carte : nuisibilité
région bourgogne



TEST 2

Crop: rapeseed

Disease: ./.

Pest: cabbage maggot

On map: Date Sort
region centre

Culture : colza

Maladie: ./.

ravageur: mouche du chou

sur carte : tri par date
région centre



TEST 3

Crop: ./.

Disease: potato late blight

Pest: ./.

On map: region burgundy
crop potato

Culture : ./.

Maladie : mildiou

Ravageur: ./.

sur carte : région bourgogne
culture pomme de terre

Conclusion

- Concrete real-world issue about damage on crops
- Implementation of an original tool to extract relations (x.ent)
 - F-55%
 - crops/diseases & crops/pests
- Integration of the tool in a user-friendly platform with geolocalization and feedback to original documents

Perspectives

- X.ENT
 - Add a cooccurrence analysis for unformatted documents
 - Evaluation of relation with know set of relations (extern ontology)
 - Refine extraction of risk factor with a scale.
 - Add probabilistic approach for unknown relationship detection, with unknown named entities
- Vespa interface
 - Add a multiuser collaborative interface to modify ontology
 - Perhaps add other languages and documents (polish?)
 - Fusion with a meteorological ontology and database

Staff

Vespa platform

Nicolas Turenne

Research fellow



Tien Phan

Computer scientist



Alexandre Louchart

Computer scientist



Chloe Duloquin

designer



OCR dataset

Vincent Cellier

Research engineer



Mathieu Andro

PhD student & engineer

