Text Mining et annotation sémantique pour l'Information Scientifique

Agnès E. Ricroch, Sophie Aubin, Claudine Mader, Audrey Boisron

▶ **To cite this version:**

Agnès E. Ricroch, Sophie Aubin, Claudine Mader, Audrey Boisron. Text Mining et annotation sémantique pour l'Information Scientifique. Présentation des Résultats Obtenus avec l'Outil Luxid sur le corpus Pest and Crops, Apr 2013, Paris, France. 47 p. hal-02810218

HAL Id: hal-02810218
https://hal.inrae.fr/hal-02810218v1

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Text Mining et annotation sémantique pour l'Information Scientifique

Exemple : Pests & Crops, la place de la transgenèse

18 / Avril / 2013

SOMMAIRE

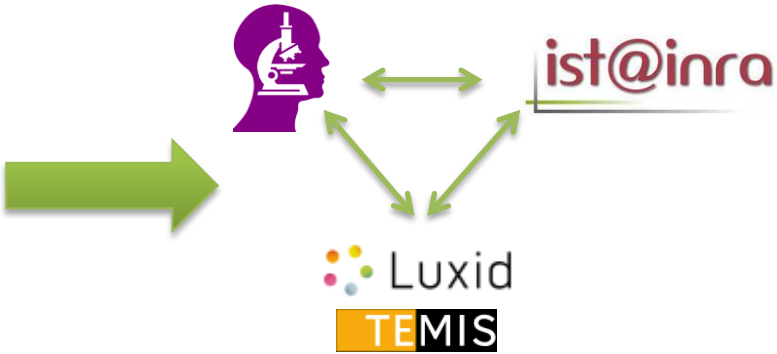
Constat

+

+ des outils inadaptés

Nouveau besoin: produire des analyses intelligentes

1. Sémantiser les ressources
2. Formaliser les connaissances scientifiques
3. Capitaliser les connaissances scientifiques
4. Avoir des outils d'exploitation et de restitution adaptés



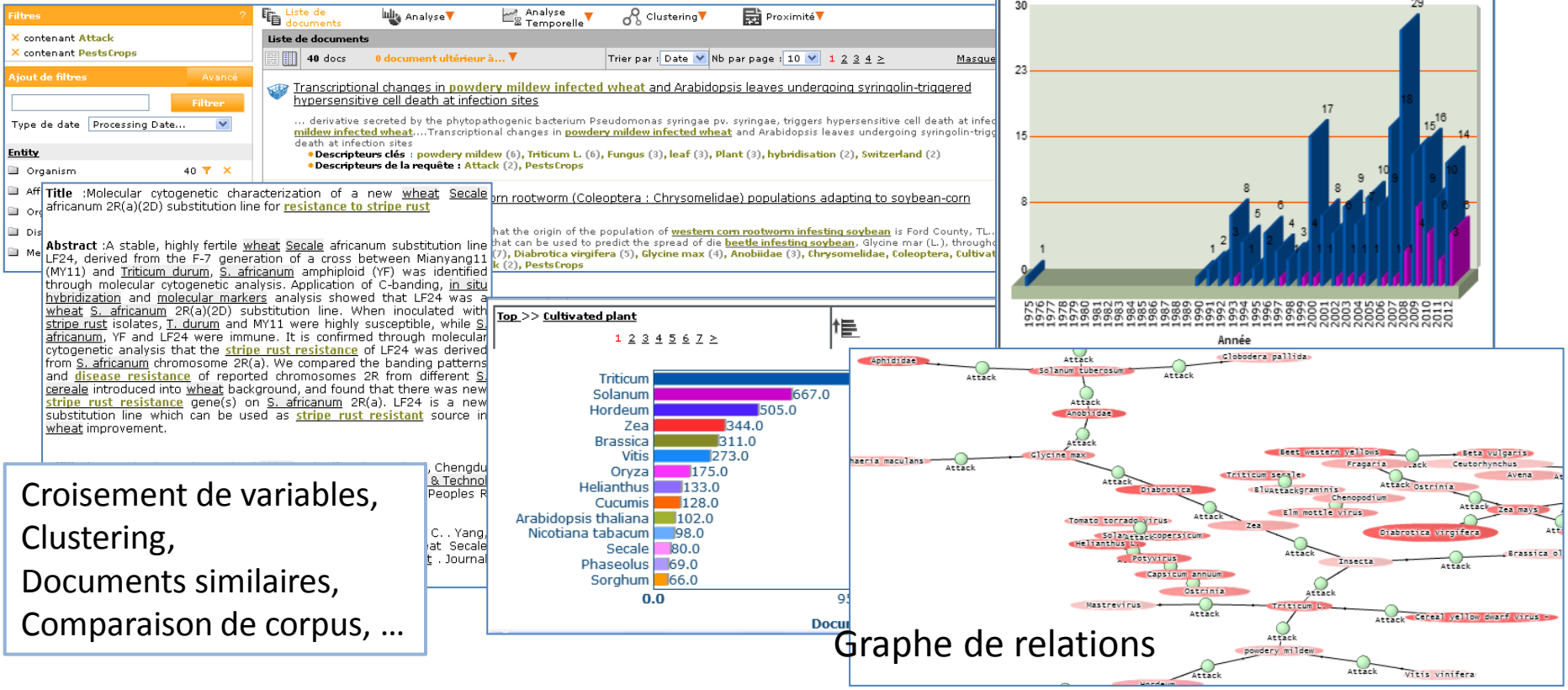
_01

Outils et ressources

Luxid: technologies et connaissances pour l'analyse de contenu

Outils et ressources

Luxid: une interface pour la découverte et l'analyse de contenu

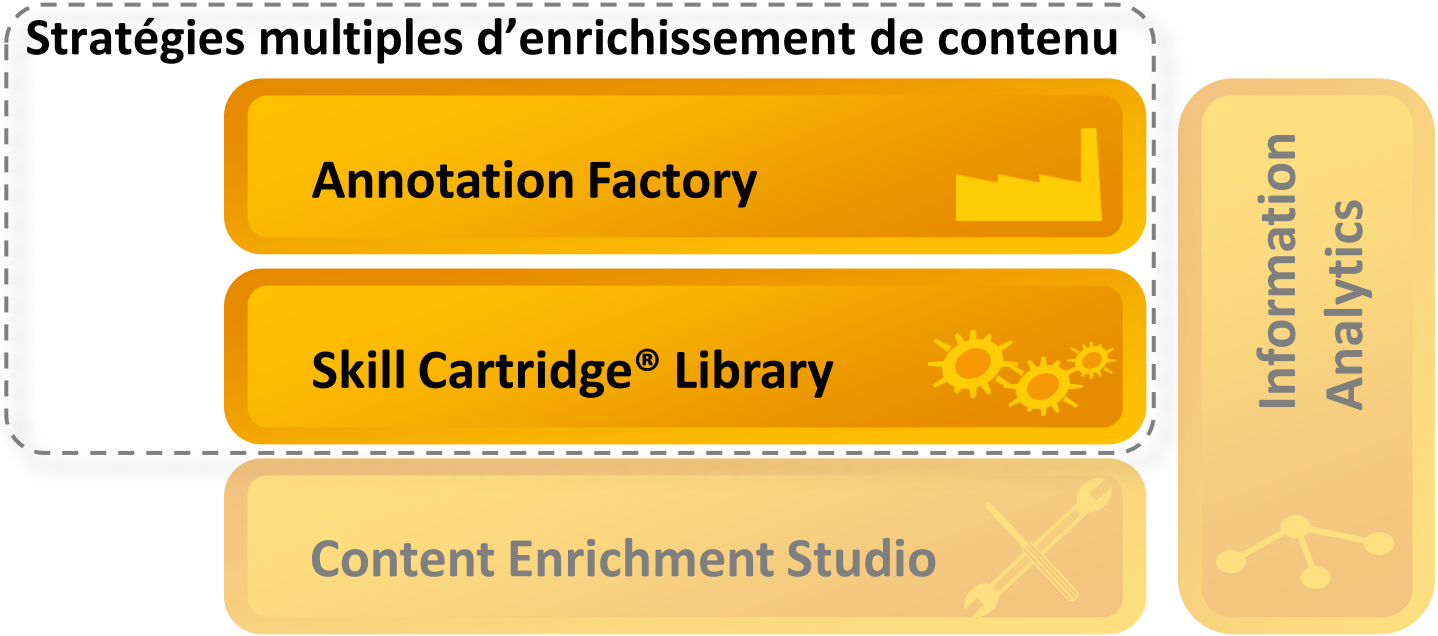


Croisement de variables,
Clustering,
Documents similaires,
Comparaison de corpus, …

Contenu emprunté à TEMIS

Outils et ressources

Luxid: un environnement intégré



Contenu emprunté à TEMIS

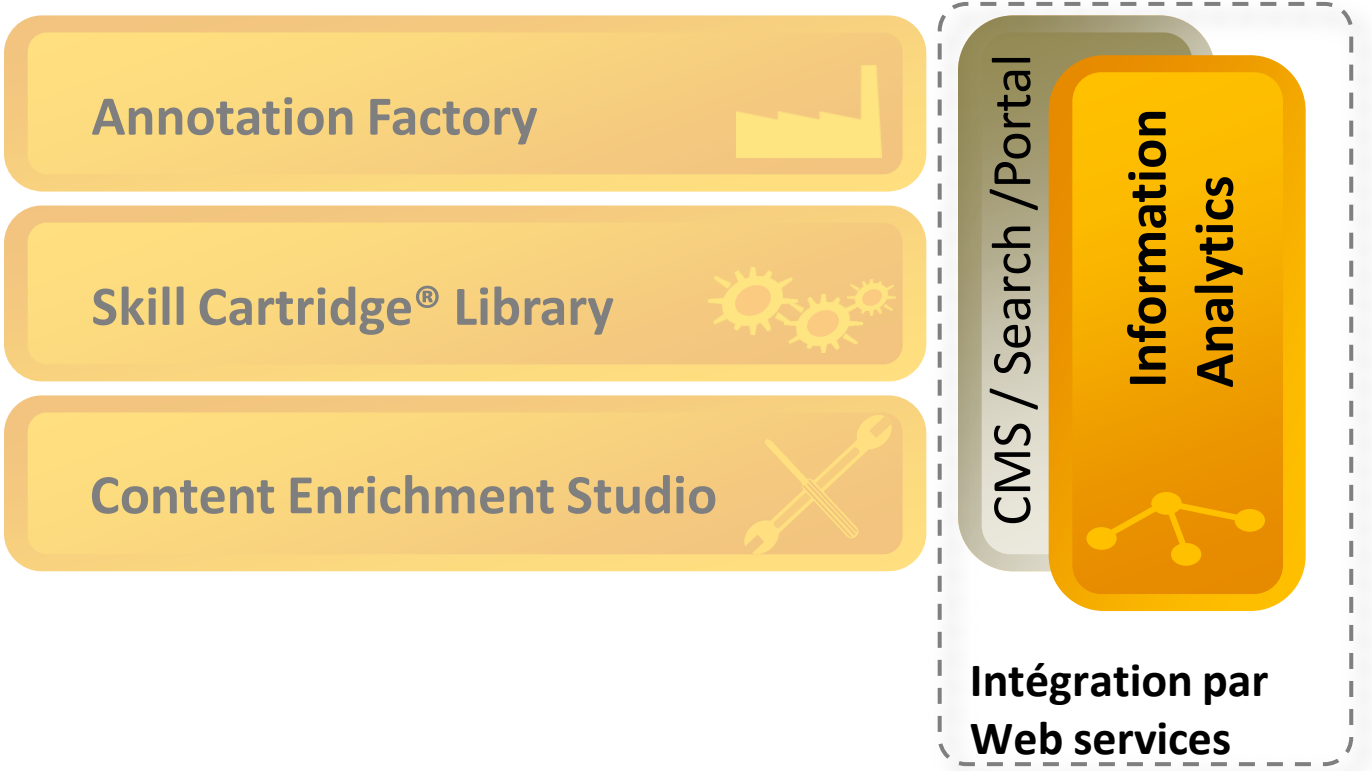
Outils et ressources

Luxid: un environnement intégré

Contenu emprunté à TEMIS

Outils et ressources

Luxid: un environnement intégré



Outils et ressources

Cartouches de connaissances

transversal

thématique

TEMIS

- Relevant Term Finder (RTF)
- Text Mining 360° (TM360)
- Sentiment Analysis
- Competitive Intelligence (CI)
- Biological Entity Relationships (BER)
- Medical Entity Relationships (MER)
- Chemical Entity Recognition (CER)

adaptation

INRA SCIENCE & IMPACT ist@inra

- AgroActors

création

- AgriOrganisms
- AgroStudies

transversal

thématique

Outils et ressources

Cartouches de connaissances

transversal

thématique

TEMIS

- Relevant Term Finder (RTF)
- Text Mining 360° (TM360)
- Sentiment Analysis
- Competitive Intelligence (CI)
- Biological Entity Relationships (BER)
- Medical Entity Relationships (MER)
- Chemical Entity Recognition (CER)

Luxid COMMUNITY

INRA SCIENCE & IMPACT ist@inra

- AgroActors
- AgriOrganisms
- AgroStudies

transversal

thématique

_02

Démonstration

Pests & Crops

Projet « Pests & Crops »

Un projet pilote sur 44 couples plante-maladie/ravageur en Europe

Des ressources d'annotation développées avec la DV-IST au cours de l'année 2012-13

Une interface sous Luxid pour faire des analyses

Corpus « Pests&Crops » = **42 349** publications

période sélectionnée 2007-2012
21 008 publications

Question initiale : Quelle est la place de la transgénèse pour répondre à ces défis agricoles ?

Les niveaux d'analyse

MACRO

1. Les 44 problématiques identifiées sont-elles traitées par la recherche académique ?
2. Quelles techniques en *plant breeding* sont-elles utilisées ? Et pour quels problèmes ?
   - Transgénèse
   - 8 Nouvelles Techniques en Plant Breeding (NTPB)

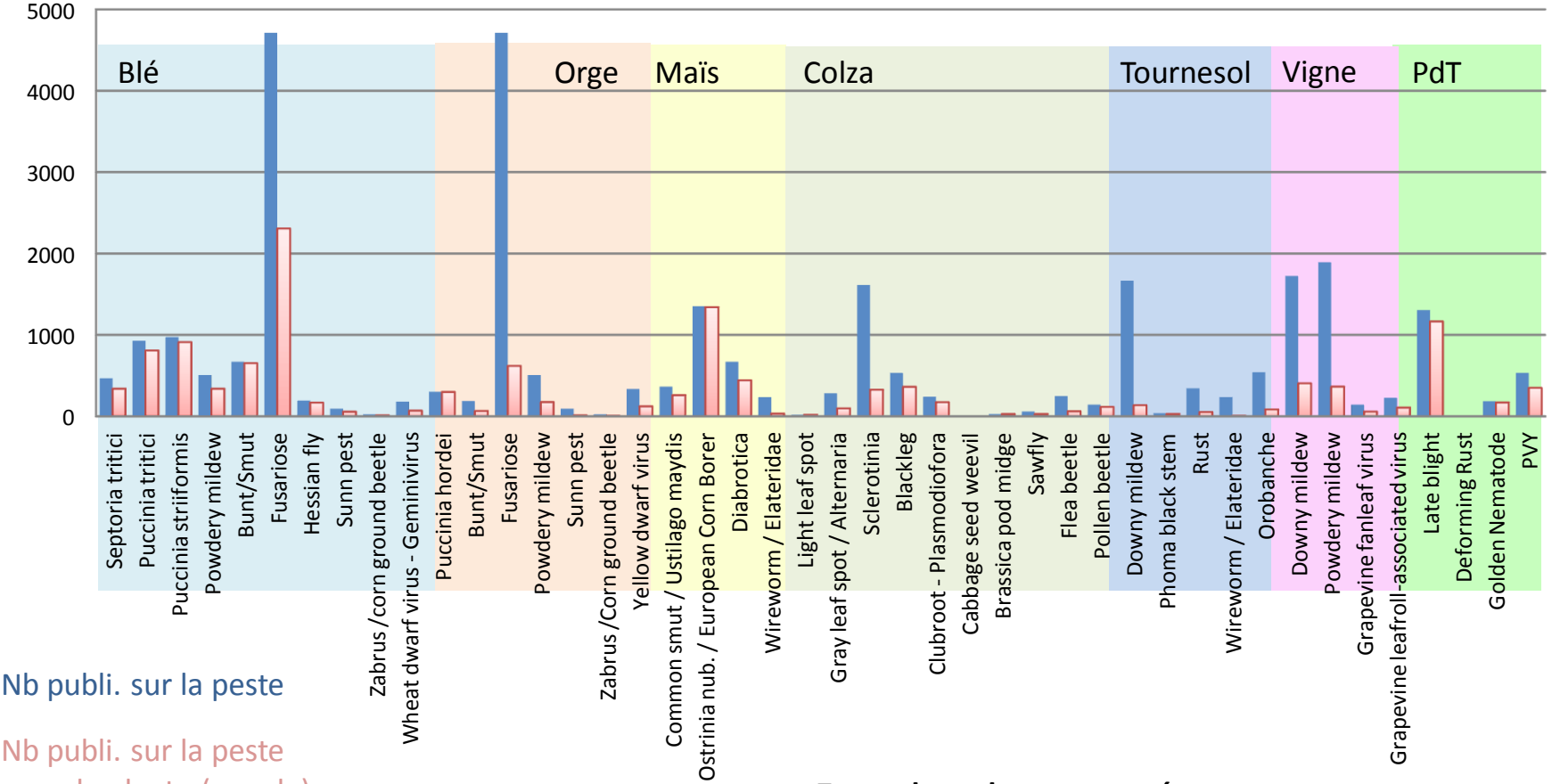
     Cisgenèse, Intragenèse, Oligonucleotide Directed Mutagenesis (ODM), Reverse breeding, Agroinfiltration, Zinc Finger Nuclease (ZFN), RNA-dependent DNA methylation (RdDM), Grafting on GM rootstock
   - RNAi
   - MAS (sélection assistée par marqueurs)
3. Quels sont les pays concernés par ces recherches ?
4. Quelles conclusions précises en tirer sur les 44 problématiques, une par une ?

MICRO

Exploration des résultats

Sur trois questions

QUESTION 1 Les 44 problématiques identifiées sont-elles traitées par la recherche académique ?



Nb publi. sur la peste

Nb publi. sur la peste avec la plante (couple)

7 grandes cultures européennes

Tableur Excel

Valeur ajoutée de Luxid

- Traitement d'un corpus de grande taille
- Aide lexicale pour l'identification des couples
- Outil **'Guessed species'** permettant d'identifier des coquilles et des espèces voisines.
- Exploration du corpus « non sélectionné » (avec la peste mais pas la plante),
- Possibilité d'analyse par plante pour observer si la peste est étudiée sur des problématiques transposables (d'où l'intérêt d'un corpus global sur la maladie)

Développements souhaités

- Mise en relation systématique du nom latin avec le nom vernaculaire anglais
- Optimisation 'Guessed species'

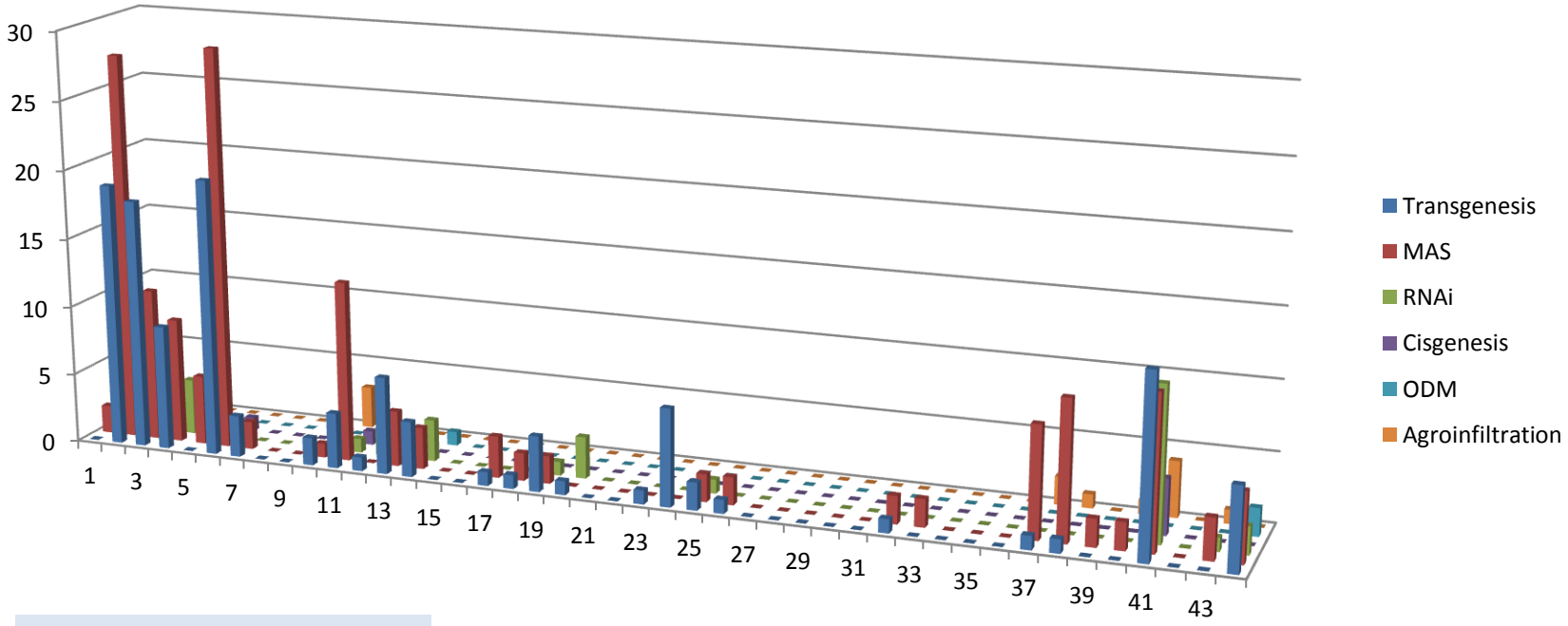
Vigilance

- Certains résultats proposés par Luxid ont du être vérifiés **manuellement** : l'observation experte connaissant bien le sujet conduit à demander des améliorations.

QUESTION 2

Quelles techniques sont-elles mobilisées ?
Quelle est la place de la transgénèse ?

Glissement progressif de notre recherche : Transgénèse + NTPB + RNAi + MAS



Fait manuellement avec EXCEL

Etiquette	Pests & Diseases	Crop	Nb de pub concernées par la plante	Transgenesis	MAS	RNAi	Cisgenesis	ODM	Agroinfiltration
1	Septoria triti	Wheat	338	0	2	0	1	1	0
2	**puccinia** triti	Wheat	808	19	28	0	0	0	0
3	**puccinia** strii	Wheat	911	18	11	2	0	0	0
4	powedery m	Wheat	337	9	9	4	0	0	0
5	bunt/smut	Wheat	653	0	5	0	0	0	0
6	**fusariose**	Wheat	2307	20	29	0	1	0	0
7	hessian fly	Wheat	168	3	2	0	0	0	0
8	sunn pest	Wheat	56	0	0	0	0	0	0
9	Zabrus /corn	Wheat	12	0	0	0	0	0	0
10	Wheat dwarl	Wheat	68	2	1	0	0	0	3
11	puccinia hord	Barley	298	4	13	1	1	0	0
12	bunt/smut	Barley	64	1	0	2	0	0	0
13	fusariose	Barley	618	7	4	0	1	0	0
14	powdery mil	Barley	174	4	3	3	0	1	0
15	sunn pest	Barley	12	0	0	0	0	0	0
16	Zabrus /corn	Barley	5	0	0	0	0	0	0
17	Yellow dwarl	Barley	121	1	3	0	0	0	0
18	Common sm	Maize	259	1	2	0	0	0	0
19	Ostrinia nubi	Maize	457	4	2	1	0	0	0
20	Diabrotica	Maize	442	1	0	3	0	0	0
21	Elateridae /	Maize	32	0	0	0	0	0	0
22	Light leaf spo	OSR	17	0	0	0	0	0	0
23	Gray leaf spo	OSR	96	1	0	0	0	0	0
24	Sclerotinia	OSR	326	7	0	0	0	0	0
25	Blackleg - ma	OSR	361	2	2	1	0	0	0
26	Clubroot - pl	OSR	173	1	2	0	0	0	0
27	cabbage see	OSR	57	0	0	0	0	0	0
28	brassica pod	OSR	26	0	0	0	0	0	0
29	sawfly	OSR	26	0	0	0	0	0	0
30	flea beetle	OSR	61	0	0	0	0	0	0
31	pollen beetl	OSR	115	0	0	0	0	0	0
32	Downy milde	Sunflower	136	1	2	0	0	0	0
33	Phoma	Sunflower	27	0	2	0	0	0	0
34	Rust	Sunflower	52	0	0	0	0	0	0
35	Wireworm /	Sunflower	6	0	0	0	0	0	0
36	Orobanche	Sunflower	83	0	0	0	0	0	0
37	Downy milde	Grapevine	405	1	8	0	0	0	2
38	Powdery mil	Grapevine	364	1	10	0	0	0	1
39	Grapevine fa	Grapevine	57	0	2	0	0	0	0
40	Grapevine le	Grapevine	106	0	2	0	0	0	1
41	Late blight	Potato	1165	13	11	11	4	0	4
42	**Deforming R**	Potato	0	0	0	0	0	0	0
43	Golden Nem	Potato	169	0	3	1	0	0	1
44	PVY	Potato	349	6	5	2	0	2	0

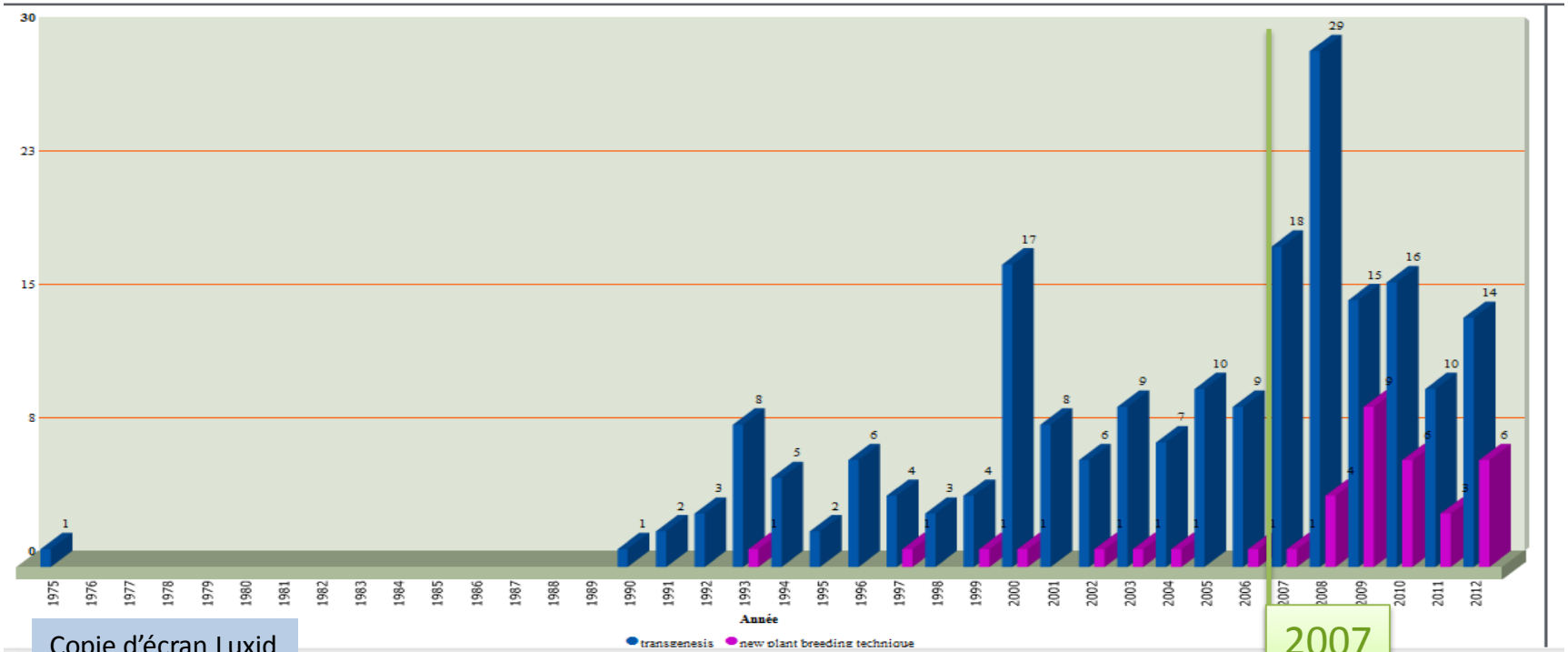
Valeur ajoutée de Luxid

- Garder l'historique dans les tableaux de bord.
- Visualisation graphique à l'intérieur d'un corpus.
  Mais pour représenter visuellement les résultats sur les 44 couples, il a fallu transférer les résultats dans un tableur Excel.
- Sélection rapide de sous-corpus pour une méta-analyse.
- Aide à la lecture rapide de l'article
  (surlignage des mots-clés)

Améliorations souhaitées

- Meilleure présentation des résultats
- Conservation des critères de recherche
- Ergonomie de la notice

Evolution temporelle des publications sur la Transgénèse et NTPB

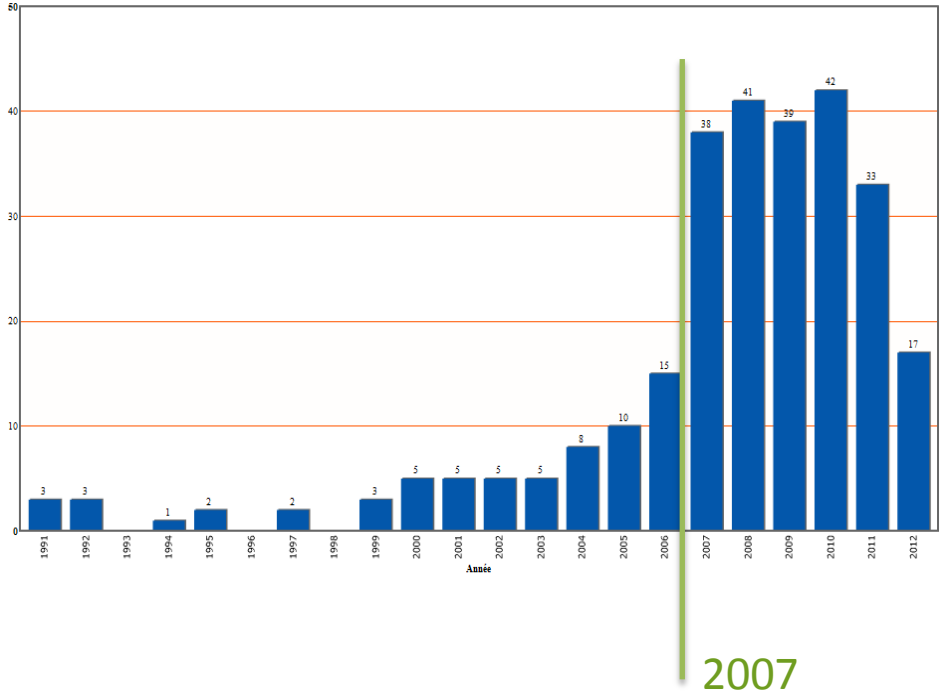


Copie d'écran Luxid

Evolution temporelle de la MAS et des RNAi

MAS

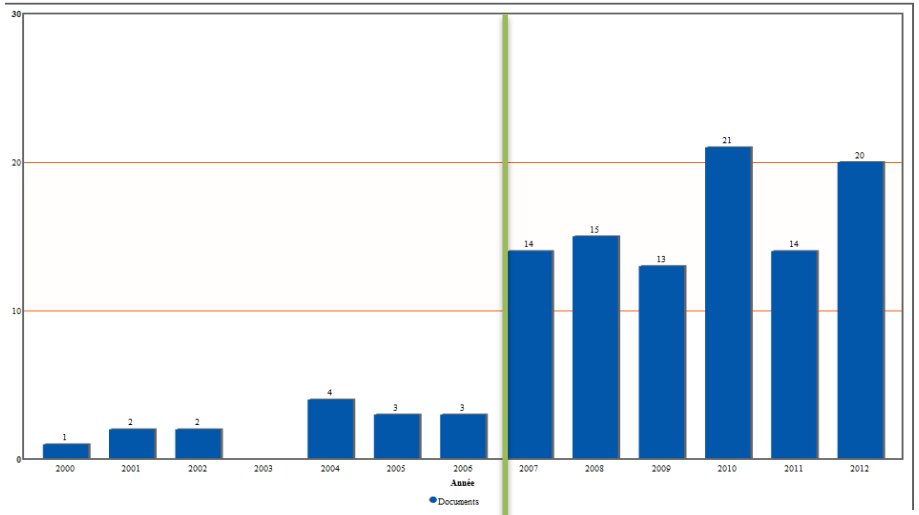
1991 à 2012



2007

RNAi

2000 à 2012

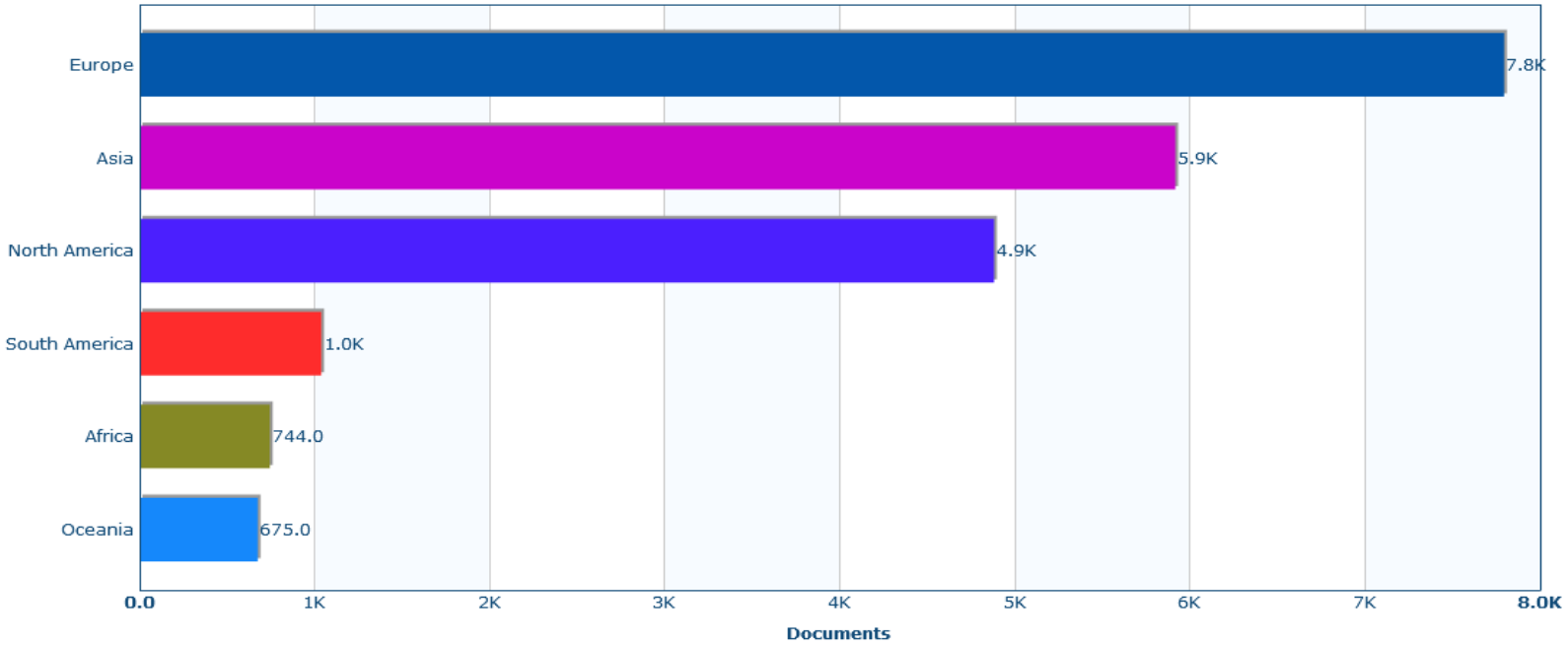


2007

QUESTION 3

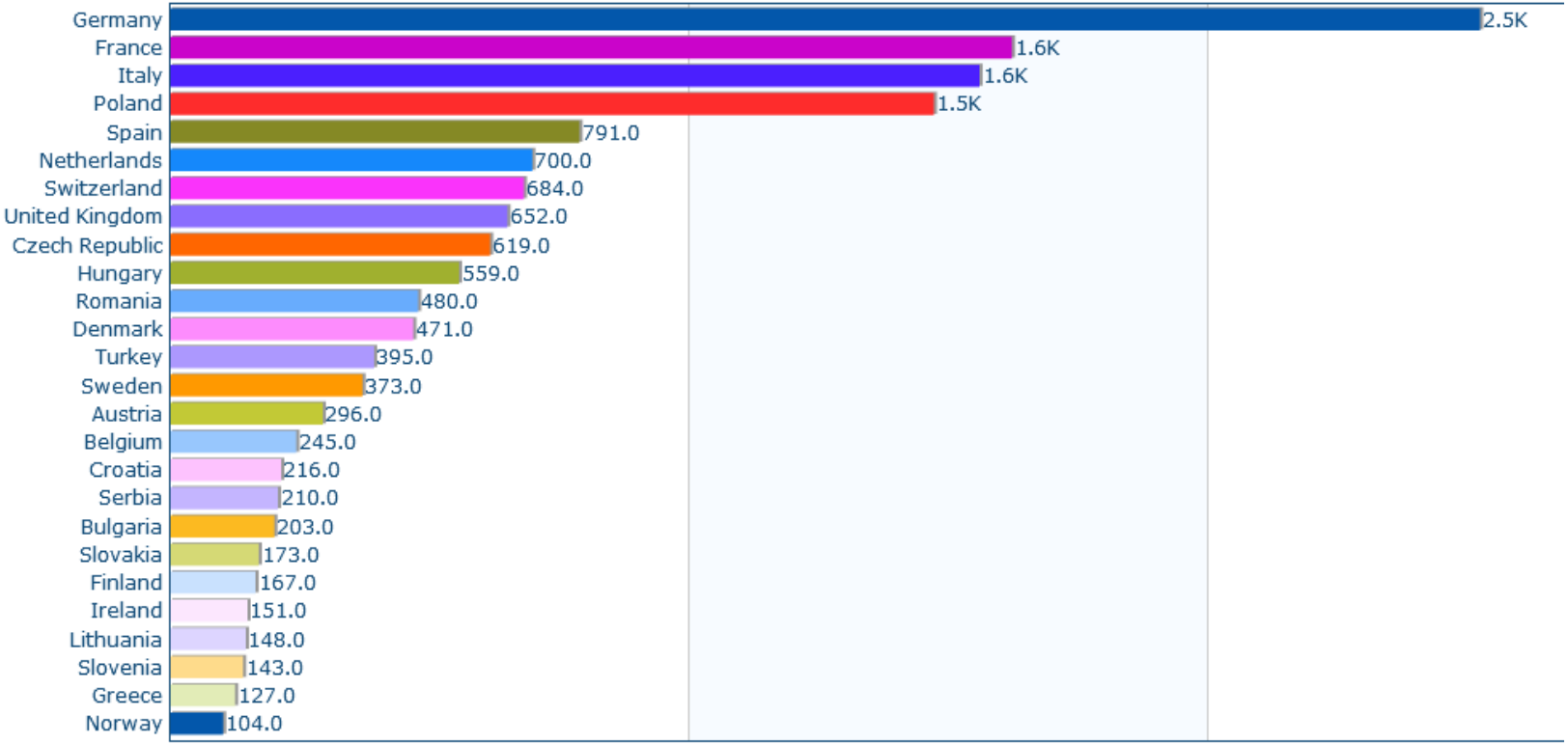
Quels sont les pays concernés par ces recherches ?

publiant sur les 44 couples étudiés

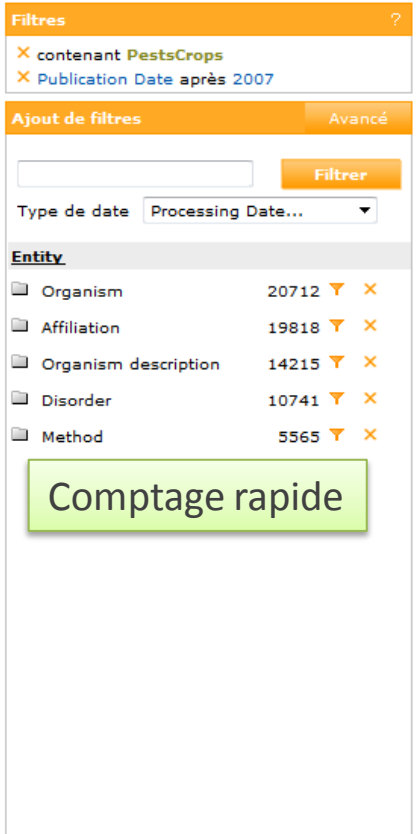


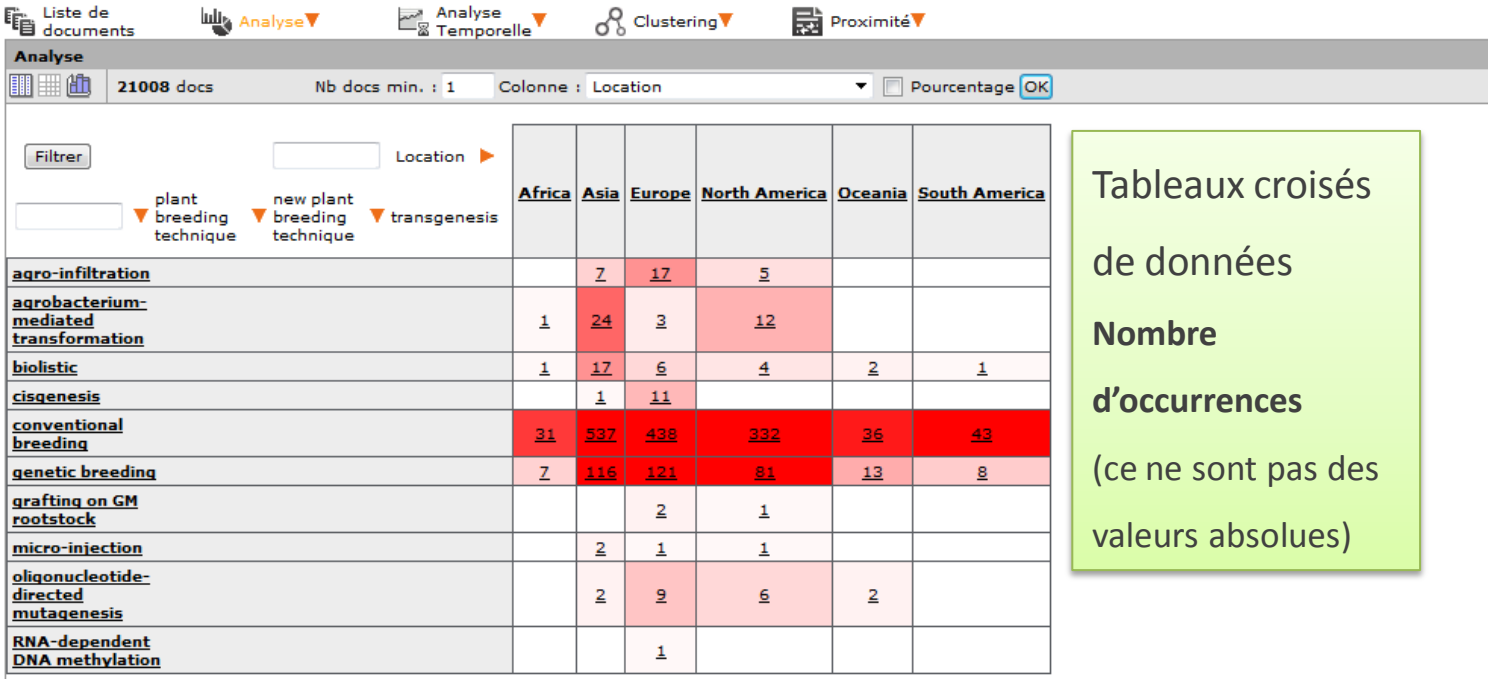
Pays d’origine des auteurs européens

publiant sur les 44 couples étudiés - Nombre d’occurrences



Visualisation des résultats



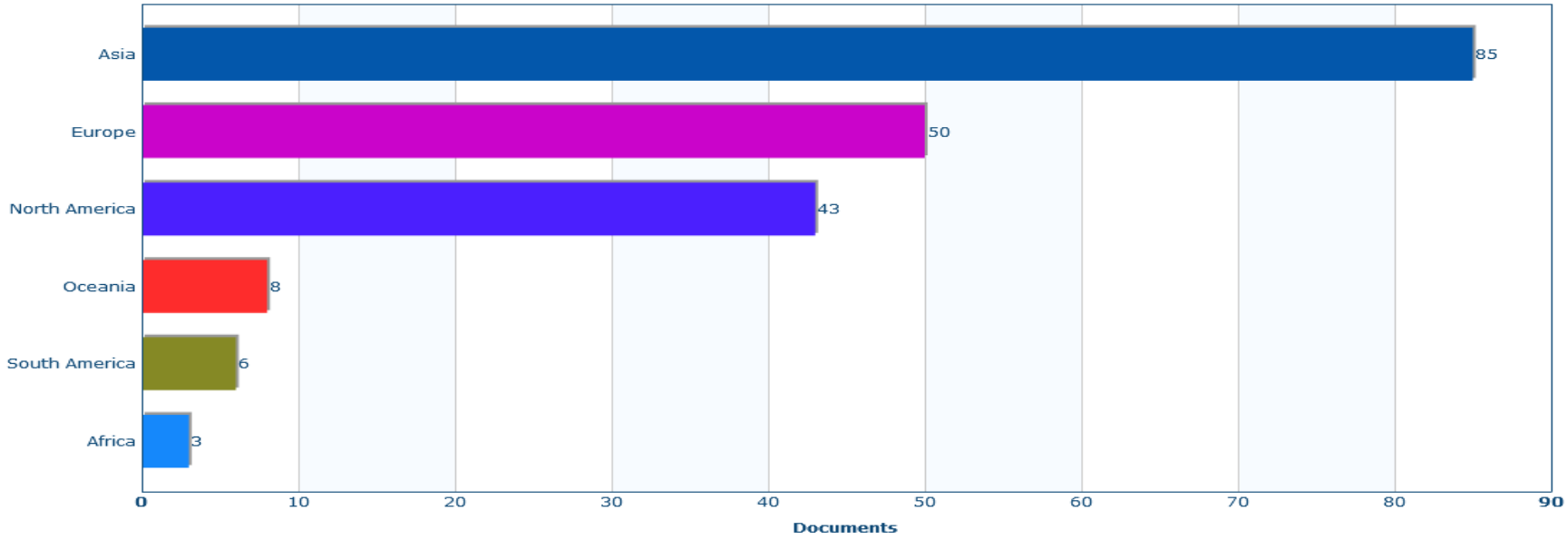


	Africa	Asia	Europe	North America	Oceania	South America
agro-infiltration		7	17	5		
agrobacterium-mediated transformation	1	24	3	12		
biolistic	1	17	6	4	2	1
cisgenesis		1	11			
conventional breeding	31	537	438	332	36	43
genetic breeding	7	116	121	81	13	8
grafting on GM rootstock			2	1		
micro-injection		2	1	1		
oligonucleotide-directed mutagenesis		2	9	6	2	
RNA-dependent DNA methylation			1			

Tableaux croisés de données **Nombre d'occurrences** (ce ne sont pas des valeurs absolues)

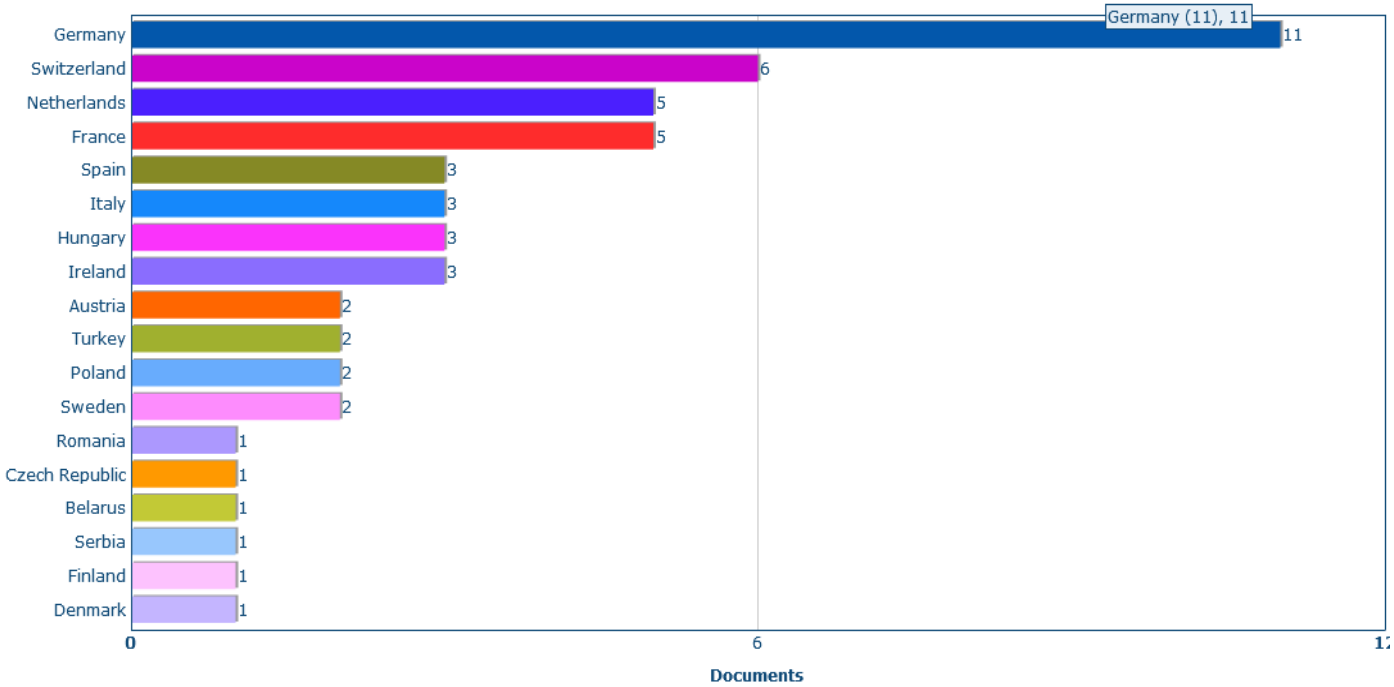
Continents d'origine des auteurs traitant de Transgénèse

Nombre d'occurrences



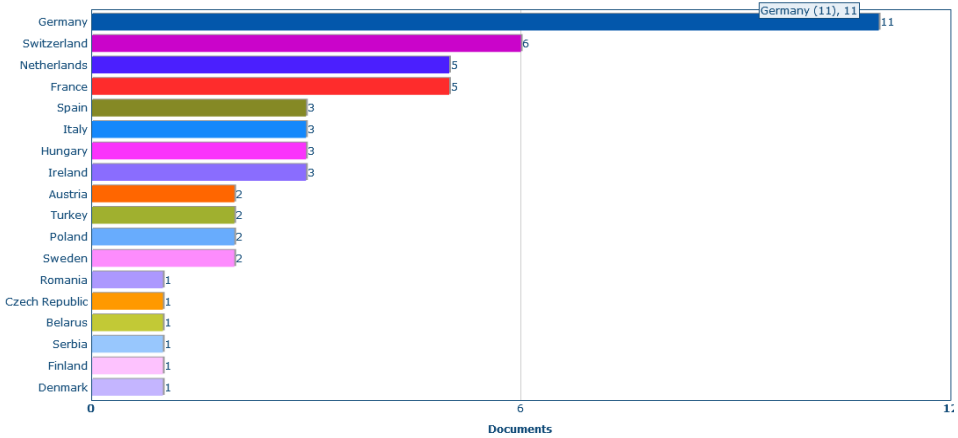
Prédominance de l'Asie

La Transgénèse en Europe



Corrections demandées sur l'entité OGM : ex. « gene transfer »
Nouvelle annotation pour inclure les corrections
(1 seule annotation a été faite en décembre 2012)

Intérêt de coupler la base GM et la base Pests&Crops



Aucune publication du Royaume-Uni ? expert

Quid du John Innes Center (JIC) qui a un programme de recherche ?

WoS – 13 publications retrouvées du JIC sur les PGM

Pests&Crops : 87 articles du JIC mais aucun sur la transgénèse
Recherche rapide des articles par auteurs issus du JIC dans toute la base

BergéRicrochGM : publications retrouvées sur les OGM et la transgénèse sans traiter des pestes ciblées

Valeur ajoutée de Luxid

- Outil d'analyse de résultats de veille
- Progression intéressante de nos hypothèses au fur et à mesure (aller plus loin dans les observations grâce au gain de temps)
- Identification de documents d'intérêts,

  (y compris ceux à signal faible)

Vigilance

- Exigence de fiabilité
- Interprétations des résultats et validation par un utilisateur expert
- Résultats obtenus en phase de test
- Améliorations à apporter
- Mais il faut des scientifiques pour effectuer une veille et faire évoluer l'outil

Observations par plantes

pour les 44 couples étudiés

Blé & Orge

- Problématiques **très étudiées par la transgénèse** ;
- Peu par les NTPB : les champignons (cisgénèse et ODM) ou les virus (agroinfiltration)

Maïs

- Beaucoup de PGM déjà développées contre les insectes ravageurs des cultures (CRY) ;
- Pourtant **peu de publications avec la transgénèse** : probables recherches par le secteur privé.

Colza & Tournesol

- **Peu d'études** (ni transgénèse, ni NTPB)
- Des insectes gérés par pesticides : attention aux résistances**.**

Vigne

- Problématique majeure des virus ; **perte du leadership de la France**
- Changement de méthode (GM Rootstock→Agroinfiltration)
- avec positionnement des USA et la Chine

Pomme de Terre

- Recherche **NTPB importante** pour le *Late Blight* (mildiou) avec cisgenèse et agroinfiltration
- Pays : USA ; Pays-Bas ; Chine

Résultats sur les **Virus**

Name	CROP	Nb pub (2007-12)	Nb pub with crop studied	Methods				
				Transgenesis	New Techniques of Plant Breeding			
					Cisgenesis	ODM	Agroinfiltration	Grafting on GM rootstock
Wheat dwarf virus	Wheat	180	68	3	0	0	**3**	0
Yellow dwarf virus	Barley	334	121	1	0	0	**1**	0
Grapevine fanleaf virus	Grapevine	143	57	0	0	0	0	**3**
Grapevine leafroll-associated virus	Grapevine	228	106	0	0	0	**1**	**5**
PVY	Potato	533	349	6	0	**2**	0	0
TOTAL		1418	701	10	0	2	5	8

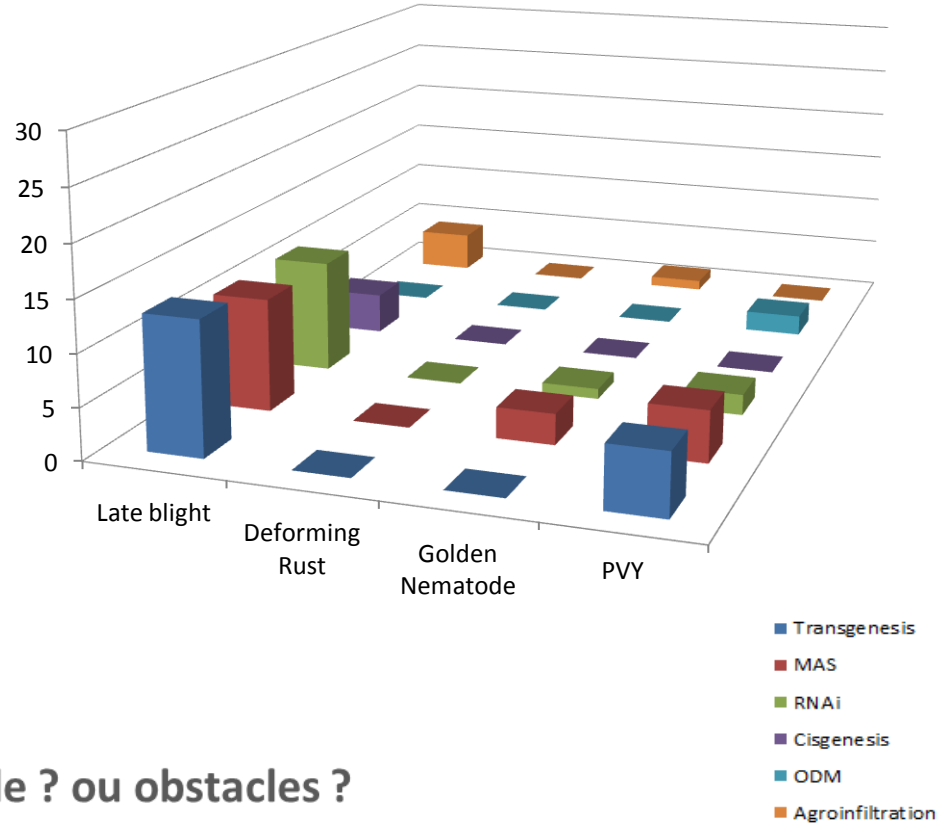
avant 2007
en France

La transgénèse donne peu de résultats contre les virus.

Techniques utilisées sur la **pomme de terre**

Beaucoup de techniques différentes sur *Late blight*

Aucune publication sur le *Deforming Rust*



Transposition des techniques possible ? ou obstacles ?

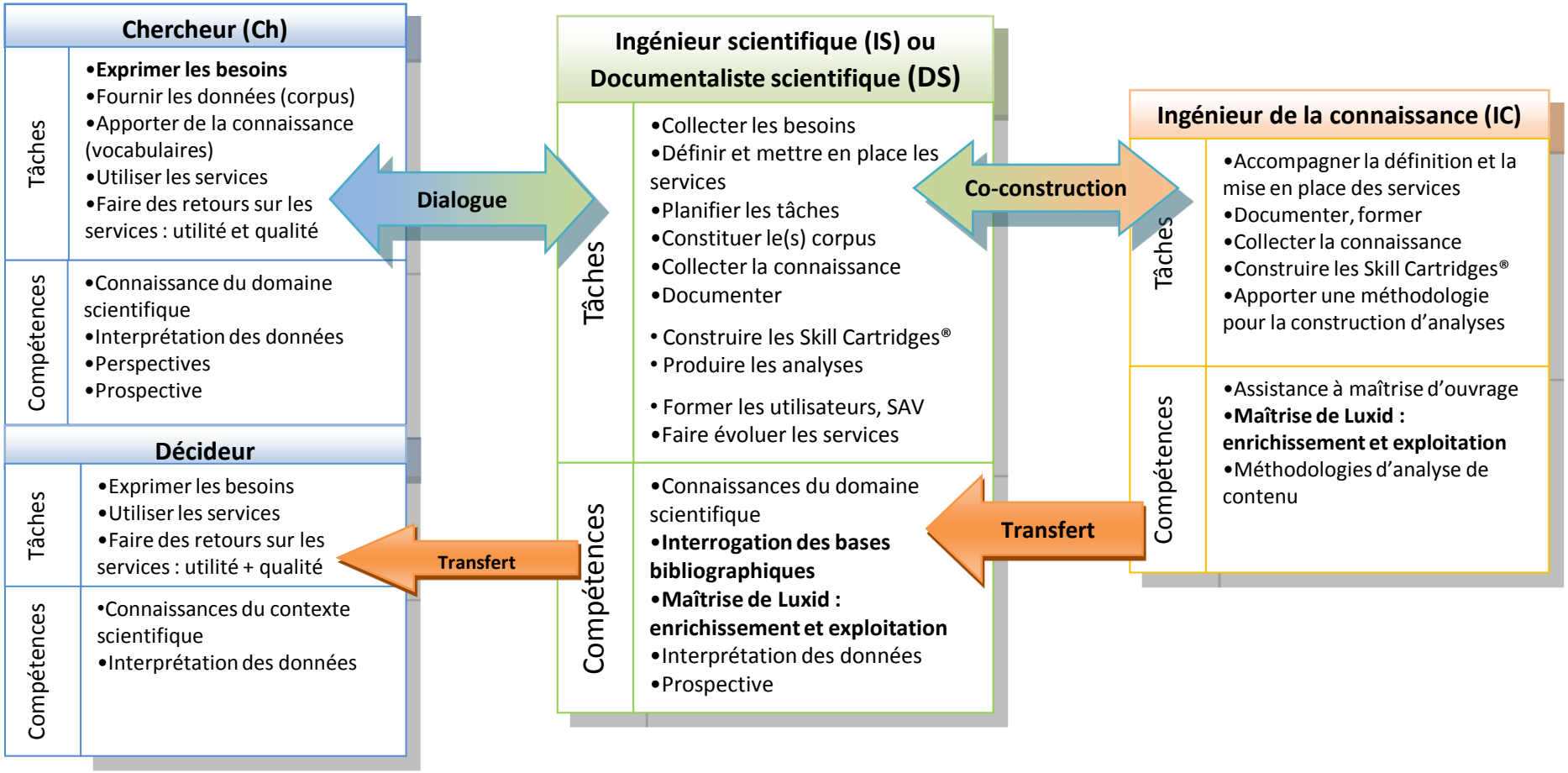
_03

Moyens

Des métiers en interaction

Compétences

Des métiers en interaction



Visuel emprunté à TEMIS

Coûts de la solution Luxid

•Licence perpétuelle: 40 000 €
• 30 000 €/Gb/an (1 Gb = 200 000 pages full text)
•20% /an maintenance et support

Annotation Factory

3 cartouches incluses

Skill Cartridge® Library

Content Enrichment Studio

Inclus

Information Analytics

Licence perpétuelle (non nominative):

•Professionnel de l'information: 2500€

•analyste: 1250€

•lecteur: de 50€ (+200 licences) à 250€ (de 1 à 10 licences)

•20% /an maintenance et support

_04

Services envisagés

1. Service d'analyses autour de la base BergeRicrochGMLibrary

Forces

Base de connaissance experte, spécialisée et complète
Collaboration entre scientifiques et IC DV-IST
Analyses produites par des experts scientifiques
Granularité ajustable aux besoins
Maturité des technologies envisagées
Communauté Luxid

Faiblesses

Reconstitution du corpus historique sur les OGMs
Besoin d'accompagnement des utilisateurs
Compétences pour la construction de ressources
Paternité multiple des données
Tentation d'une approche non critique des résultats
Besoin d'un ou + chercheur référent pour assurer la veille de concepts, de problèmes et enrichir le vocabulaire OGM

Opportunités

Service attendu par la communauté
Réutilisation des ressources développées sur le prototype
Capacités d'apprentissage de Luxid pour les labels de la base d'origine
Analyse stratégique du domaine
Valorisation de connaissances/ressources

Menaces

Produits concurrents
Usage détourné des services
Moyens humains insuffisants pour maintenir le service
Contraintes de la PI pour des services externes

2. Service d'analyses multi-thématiques

Forces

Multidisciplinarité
Compétences IC à la DV-IST
Capitalisation des ressources construites
Granularité ajustable aux besoins
Maturité des technologies envisagées
Communauté Luxid

Faiblesses

Besoin d'accompagnement des utilisateurs
Compétences pour la construction de ressources
Tentation d'une approche non critique des résultats
Propriété intellectuelle des ressources exploitées
Besoin d'un ou + chercheur référent pour assurer la veille de concepts, de problèmes et enrichir le vocabulaire OGM

Opportunités

Service attendu par la communauté
Analyse stratégique des thématiques Inra (prospective, pilotage de la recherche, positionnement)
Développement de la recherche in silico sur le texte
Valorisation de connaissances/ressources
Diversification des compétences IST

Menaces

Produits concurrents
Usage détourné des services
Moyens humains insuffisants pour maintenir le service
Défiance des usagers

Focus sur 2 insectes non étudiés

Zabre et Sunn Pest

Bonne connaissance de la problématique

Quelle mobilité géographique de la peste ?

Faut-il envisager une arrivée de la peste en France ?

Rapport entre enjeux agronomiques et efforts de recherche

Exemples du Zabre et Sunn Pest

Observation des enjeux agronomiques

On distingue deux groupes de pays en ce qui concerne la problématique agronomique

- Les pays d'Europe de l'Ouest et du Nord pour qui ces insectes ne représentent pas une problématique (France, Allemagne, UK…)
- Les pays plus à l'Est ou au Sud qui eux les considèrent comme problématiques (Roumanie, Bulgarie …)

Observation en termes de publication

Très peu de publications. Leurs origines géographique sont en accord avec les enjeux agronomiques : majoritairement ou exclusivement issues de la Roumanie (Zabre).
Ils sont étudiés principalement par des pays d'Europe de l'est, ce qui permet de s'interroger sur le devenir de ces pestes agricoles

Axes de recherche actuels

Zabre

Il existe un programme roumain d'amélioration variétale par une approche conventionnelle (MAS) pour répondre à la problématique du Zabre mais les publications trouvées ne traitent que des préconisations en matière de traitement par pesticides des grains

Sunn Pest

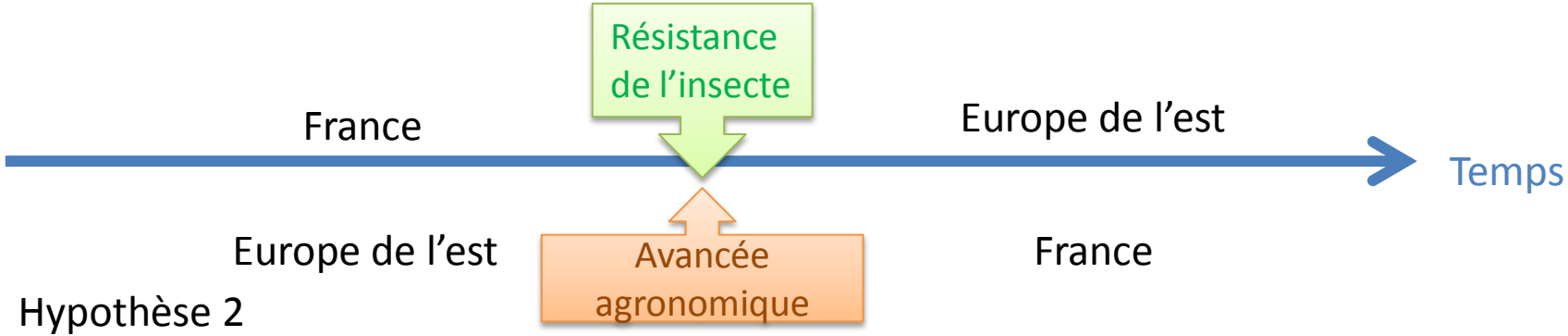
Observation de la résistance des variétés résistantes, de l'impact de l'insecte sur la récolte et de l'insecte lui-même.

Aide à la prospective et au suivi des projets de recherche

Aide à la prospective :
exemples du Zabre et Sunn Pest

- Hypothèse 1

  La peste a évolué et a contourné des modes de lutte qui ne sont plus suffisants en Europe de l'Est mais fonctionnent encore en France par exemple ?



- Hypothèse 2

  Les pays pour lesquels elles ne représentent pas une problématique ont-ils eu accès à une avancée technologique ?

Lire précisément un corpus de documents plus petits

Requête

Entités

AND OR

pest1 OR pest2 OR pest3 OR crop1

pest1 AND pest2 AND pest3 AND crop1

Requête souhaitée pour un couple

(pest1 OR pest2 OR pest 3) AND crop1

Valeur ajoutée de Luxid

Traitement d'un corpus de grande taille

Aide lexicale pour l'identification des couples

Outil **'Guessed species'** permettant d'identifier des coquilles et des espèces voisines.

Exploration du corpus « non sélectionné », avec la peste mais pas la plante.

Possibilité d'analyse par plante pour observer si la peste est étudiée sur des problématiques transposables (d'où l'intérêt d'un corpus global sur la maladie)

Développements souhaités

Mise en relation systématique du nom latin avec le nom vernaculaire anglais

Optimisation 'Guessed species'

Attention

Certains résultats proposés par Luxid ont du être vérifiés **manuellement** : l'observation experte en connaissant bien le sujet conduit à demander des améliorations.

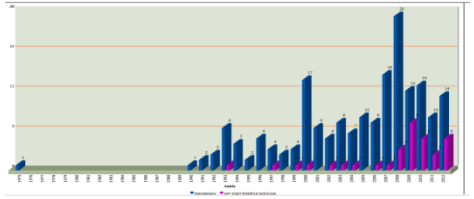
Opérateurs booléens non associables AND ou OR - *contournable mais plus lent et plus compliqué*

Pas de graphique en 3D possible par Luxid

Entité GM

Sous-corpus GM
(couples + autres)

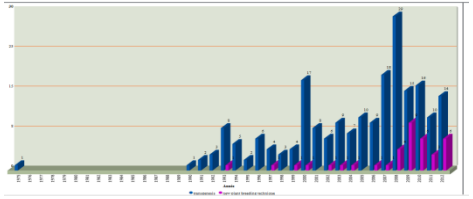
Analyse temporelle



Entité peste
+ plante

Sous-corpus
couple

44 sous-
corpus
couples



Observations sur les NTPB

Sur le corpus sélectionné par Luxid traitant des NTPB on trouve :

- des articles de réflexion sur la méthode

- un article singulier et au positionnement particulier

Attention lecture attentive de la notice par une personne connaissant le sujet.

RNAi : Analyse Luxid par Pays et plantes concernées

Filtrer / Location ▶ / ▼ Cultivated plant	Asia	Europe	North America
Allium	1		
Brassica		1	
Glycine			1
Hordeum		4	3
Nicotiana tabacum	2	1	1
Oryza			1
Solanum	4	3	6
Triticum	1	2	4
Zea		1	4

> Exclure les doublons
> Sous-corpus 29 documents

Les lire tous pour connaitre les conclusions et les résultats

8 - 12 - 20