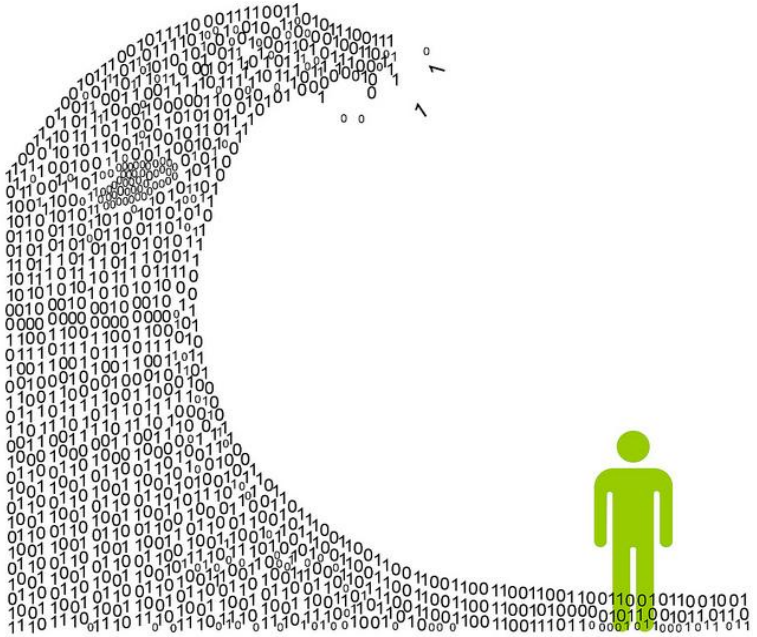
Infodoc Express
Introduction au text mining

mathieu.andro@versailles.inra.fr
sophie.aubin@versailles.inra.fr

Durée : 35 minutes
19/01/2016

Qu'est-ce qui a rendu nécessaire le text mining ?

- Infobésité, 40 % des recherches d'informations infructueuses
  (> 11 h / semaine / salarié)
- Limites de la recherche par mots
- Coûts importants de la production de métadonnées… mais coûts pour la production de vocabulaires
- et de la production d' observations, de connaissances.

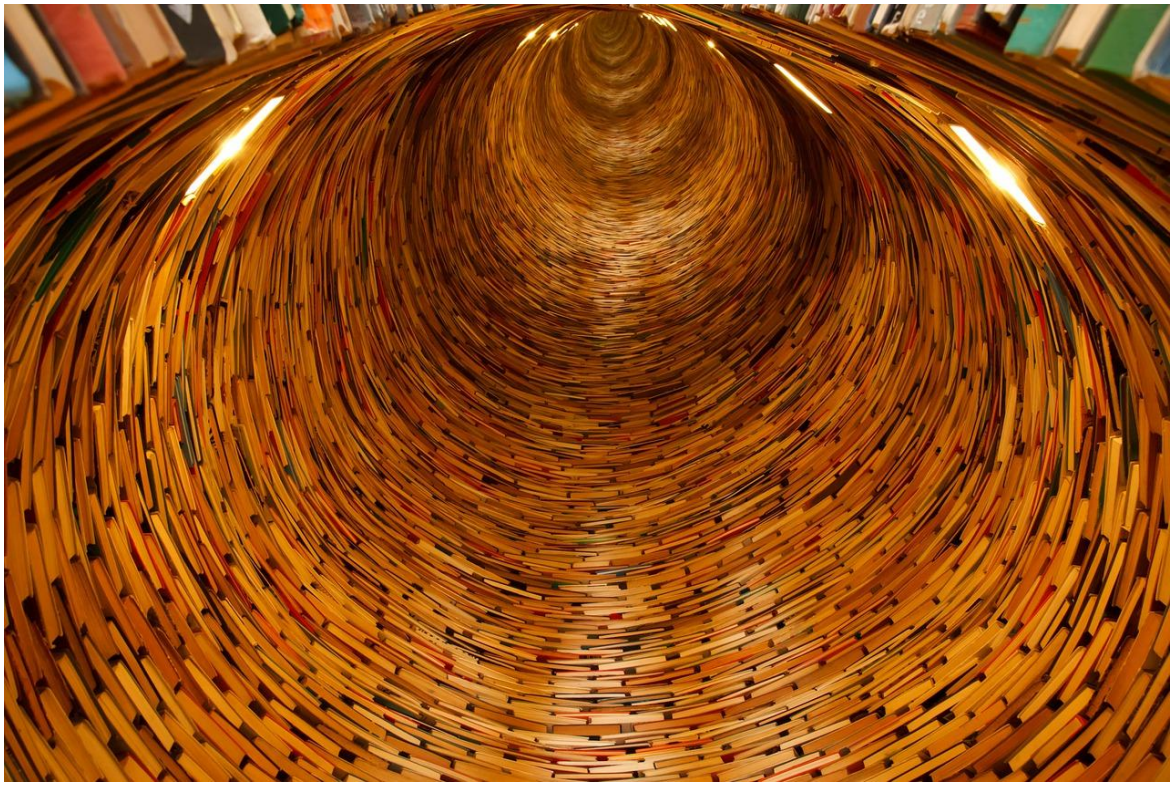


Flickr, Mark Smiciklas, CC BY-NC 2.0

Le text mining, c'est quoi ?

Au lieu de lire quelques documents…

exploiter des montagnes de documents.



Pixabay, Tunnel de Livres, CCO Public Domain

Le text mining, c'est quoi ?

Au lieu de rechercher un mot dans un document...

rechercher des dictionnaires de mots :

- espèces
- lieux géographiques
- institutions
- vocabulaires métier, etc...

Flickr, Kate Ter Haar, CC BY 2.0

dans des masses de documents...

et **identifier les relations** entre ces mots

et **extraire des données** de ces documents

Le text mining, ça sert à quoi ?

Structurer les documents pour :

- **Améliorer les résultats d'un moteur de recherche en élargissant (synonymes) et en restreignant (cf "ray")**
- **Filtrer l'information, la structurer, l'indexer, la classer**
- **Analyser le contenu textuel**
  - Qui fait quoi ? Avec qui ? Comment évolue un sujet ?
  - Analyse du sentiment et du discours (cf socio, veille)
  - Découvrir de l'information, clustering
- **Convertir des textes en données et peupler une base de données par extractions automatiques d'informations**

Quelques autres exemples

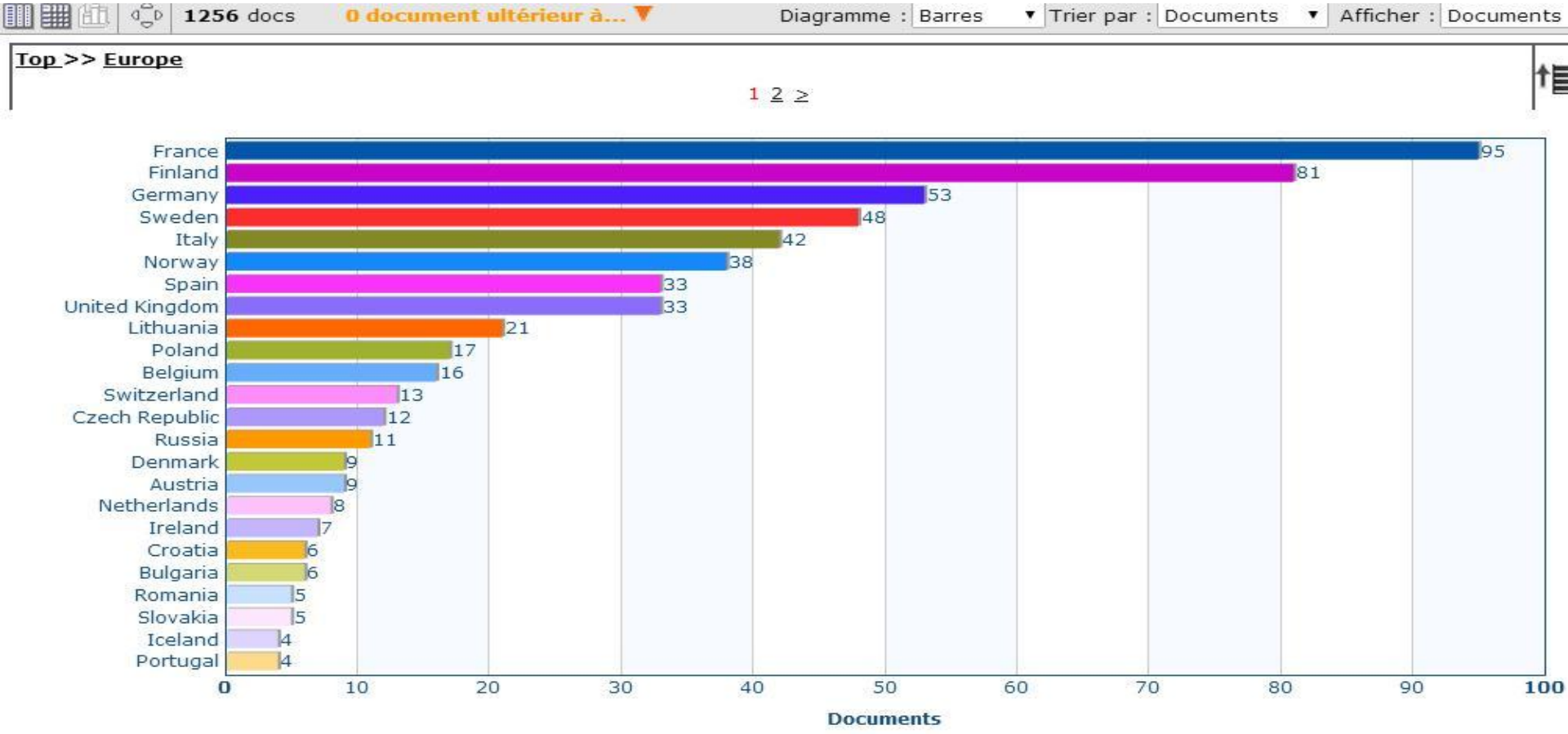
Projet phénologie

Objectif : état de l'art sur la recherche sur la phénologie des arbres forestiers.

Technologies originales mobilisées : analyses en entonnoirs avec analyses de sous corpus.

Wikimedia Commons, Antony.sorrento, CC-BY-SA-3.0

Projet phénologie



En Europe, qui travaille sur le développement des bourgeons des arbres ?

Projet Pest & Crops

Objectifs : identifier les relations entre les cultures et leurs menaces, les techniques utilisées

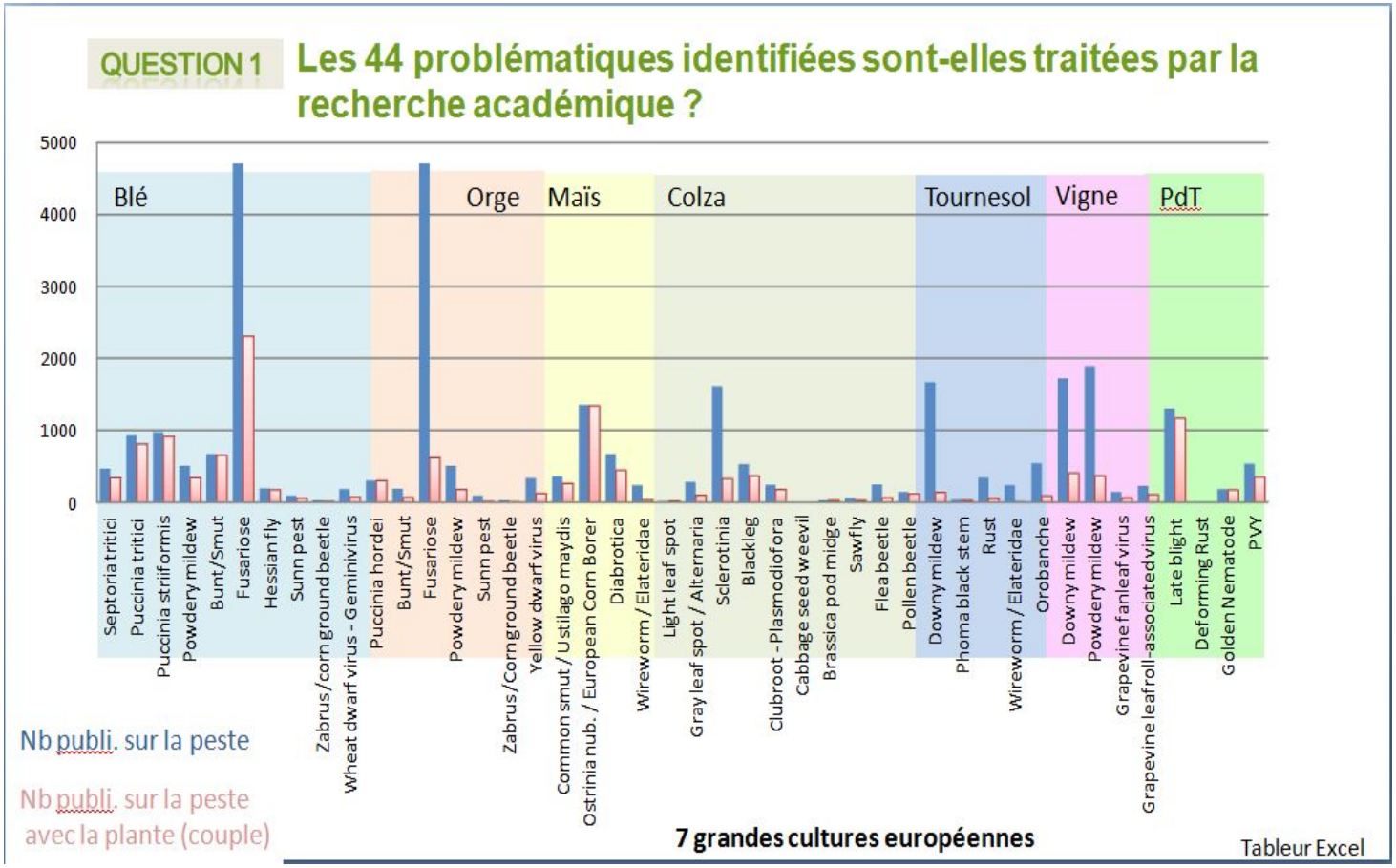
Technologies originales mobilisées :

- produire des analyses plus rapidement et avec moins d' efforts (changement d'échelle)
- faire évoluer facilement la question scientifique et le périmètre
- formaliser l'expertise pour la réutiliser

Commons Wikimedia, Agricultural Research Service, the research agency of the United States Department of Agriculture, Public Domain

En Europe, qui travaille sur le développement des bourgeons des arbres ?

Projet Pest & Crops



Etat des lieux de la recherche scientifique mondiale sur des couples plante-maladie

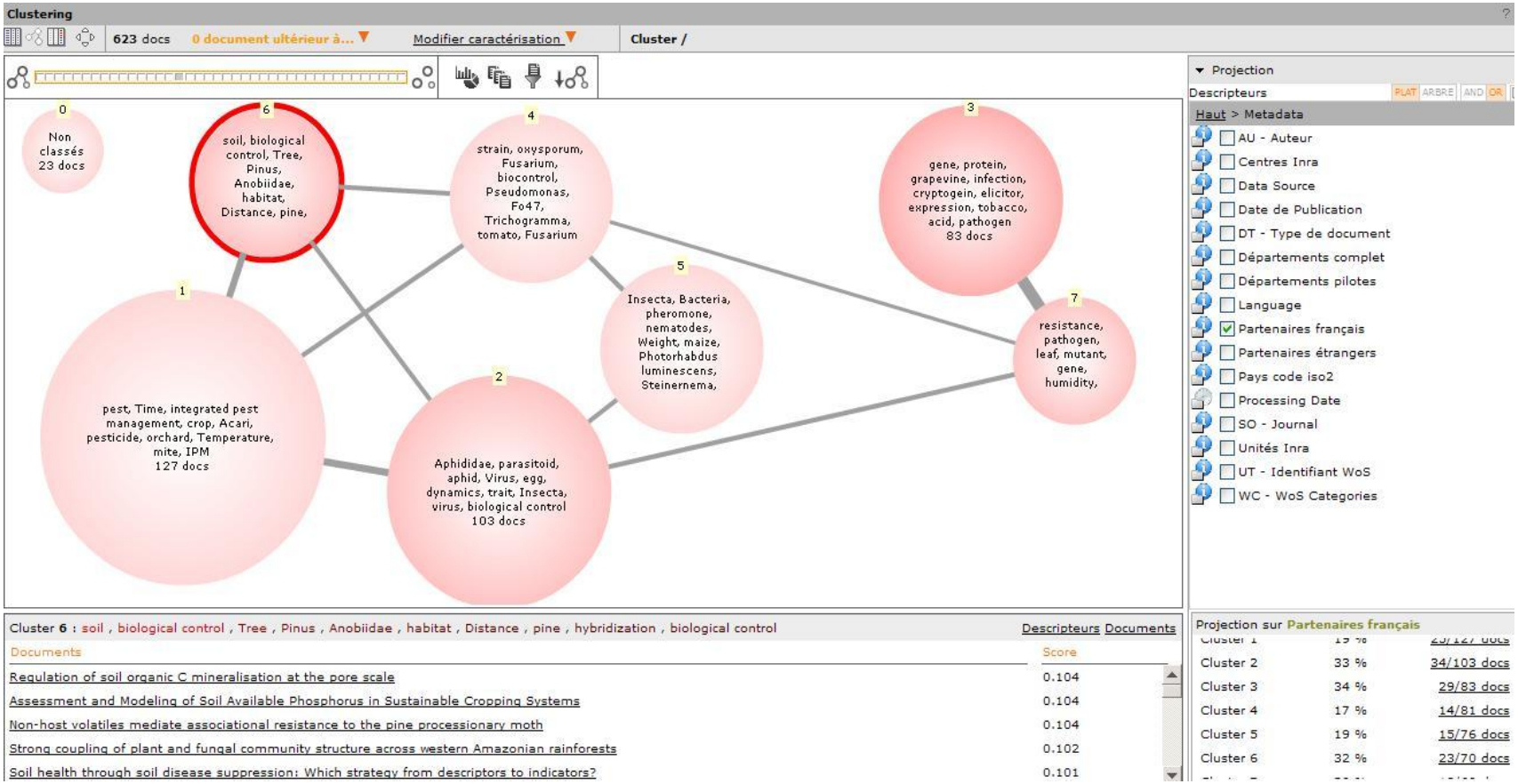
Projet GIS Biocontrôle

Objectifs : identifier des partenaires publics et privés pour un GIS, proposer un service mutualisé aux membres du GIS.

Technologies originales mobilisée : Clusterisation pour définir un périmètre sans a priori

Commons Wikimedia, Jon Sullivan, CCO

Biocontrôle



Clusterisation : identification automatique de groupes thématiques

Analyses stratégiques

Objectif : identifier les publications relevant de tel ou tel axe stratégique de l'Inra.

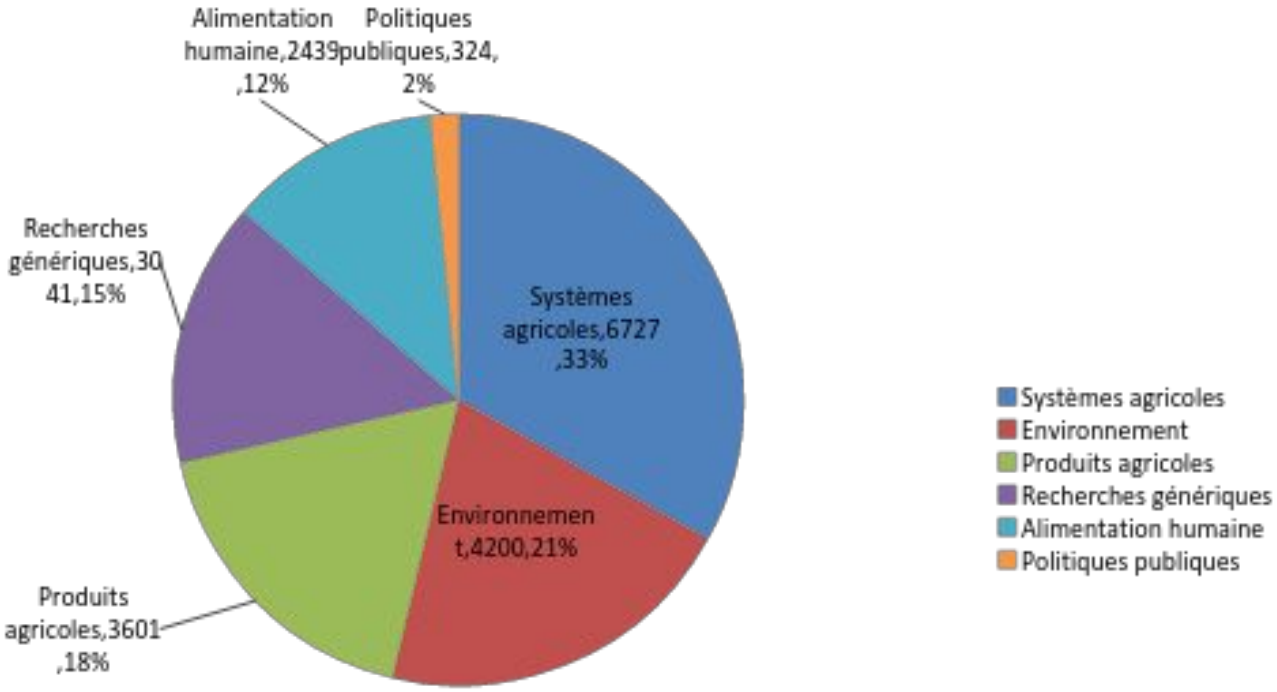
Technologies originales mobilisées :

- Apprentissage par analyse de l' empreinte sémantique de corpus existants classés par un humain.
- Complémentaire à bibliométrie

pixabay.com, ejaugsburg, CC0

✚ Existence d'un corpus de publications Inra 1999-2014 permettant de produire des analyses très rapidement.

Analyses stratégiques



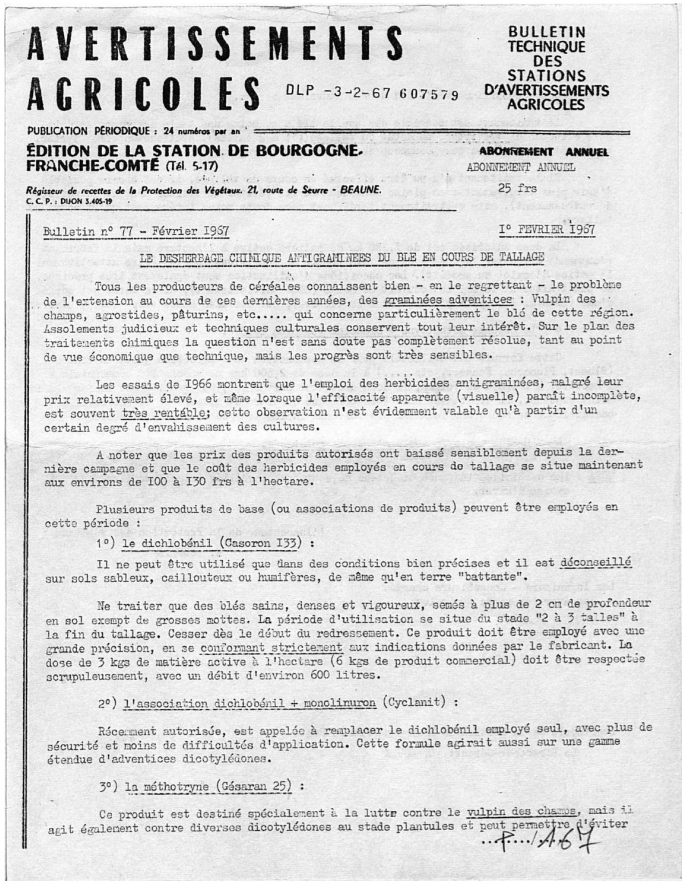
Répartition estimée des publications Inra par axes stratégiques (2001-2007) d'après CorTexT

Projet avertissements agricoles

Objectif : extraire dans les avertissements agricoles des données sur les cultures, leurs attaques et les conditions de ces attaques pour les modéliser.

Techniques originales mobilisées :

- Texte intégral et non notices
- Numérisation, océrisation, correction de l'OCR



AVERTISSEMENTS AGRICOLES

DLP -3-2-67 607579

BULLETIN TECHNIQUE DES STATIONS D'AVERTISSEMENTS AGRICOLES

PUBLICATION PÉRIODIQUE : 24 numéros par an

ÉDITION DE LA STATION DE BOURGOGNE-FRANCHE-COMTÉ (Tél. 5-17)

ABONNEMENT ANNUEL

ABONNEMENT ANNUEL 25 frs

Régisseur de recettes de la Protection des Végétaux. 21, route de Seurre - BEAUNE. C.C.P. : DIJON 3.405-19

Bulletin n° 77 - Février 1967 — 1° FEVRIER 1967

LE DESHERBAGE CHIMIQUE ANTIGRAMINEES DU BLE EN COURS DE TALLAGE

Tous les producteurs de céréales connaissent bien - en le regrettant - le problème de l'extension au cours de ces dernières années, des graminées adventices : Vulpin des champs, agrostides, pâturins, etc..... qui concerne particulièrement le blé de cette région. Assolements judicieux et techniques culturales conservent tout leur intérêt. Sur le plan des traitements chimiques la question n'est sans doute pas complètement résolue, tant au point de vue économique que technique, mais les progrès sont très sensibles.

Les essais de 1966 montrent que l'emploi des herbicides antigraminées, malgré leur prix relativement élevé, et même lorsque l'efficacité apparente (visuelle) paraît incomplète, est souvent très rentable; cette observation n'est évidemment valable qu'à partir d'un certain degré d'envahissement des cultures.

A noter que les prix des produits autorisés ont baissé sensiblement depuis la dernière campagne et que le coût des herbicides employés en cours de tallage se situe maintenant aux environs de 100 à 130 frs à l'hectare.

Plusieurs produits de base (ou associations de produits) peuvent être employés en cette période :

1°) le dichlobénil (Casoron 133) :

Il ne peut être utilisé que dans des conditions bien précises et il est déconseillé sur sols sableux, caillouteux ou humifères, de même qu'en terre "battante".

Ne traiter que des blés sains, denses et vigoureux, semés à plus de 2 cm de profondeur en sol exempt de grosses mottes. La période d'utilisation se situe du stade "2 à 3 talles" à la fin du tallage. Cesser dès le début du redressement. Ce produit doit être employé avec une grande précision, en se conformant strictement aux indications données par le fabricant. La dose de 3 kgs de matière active à l'hectare (6 kgs de produit commercial) doit être respectée scrupuleusement, avec un débit d'environ 600 litres.

2°) l'association dichlobénil + monolinuron (Cyclanit) :

Récemment autorisée, est appelée à remplacer le dichlobénil employé seul, avec plus de sécurité et moins de difficultés d'application. Cette formule agirait aussi sur une gamme étendue d'adventices dicotylédones.

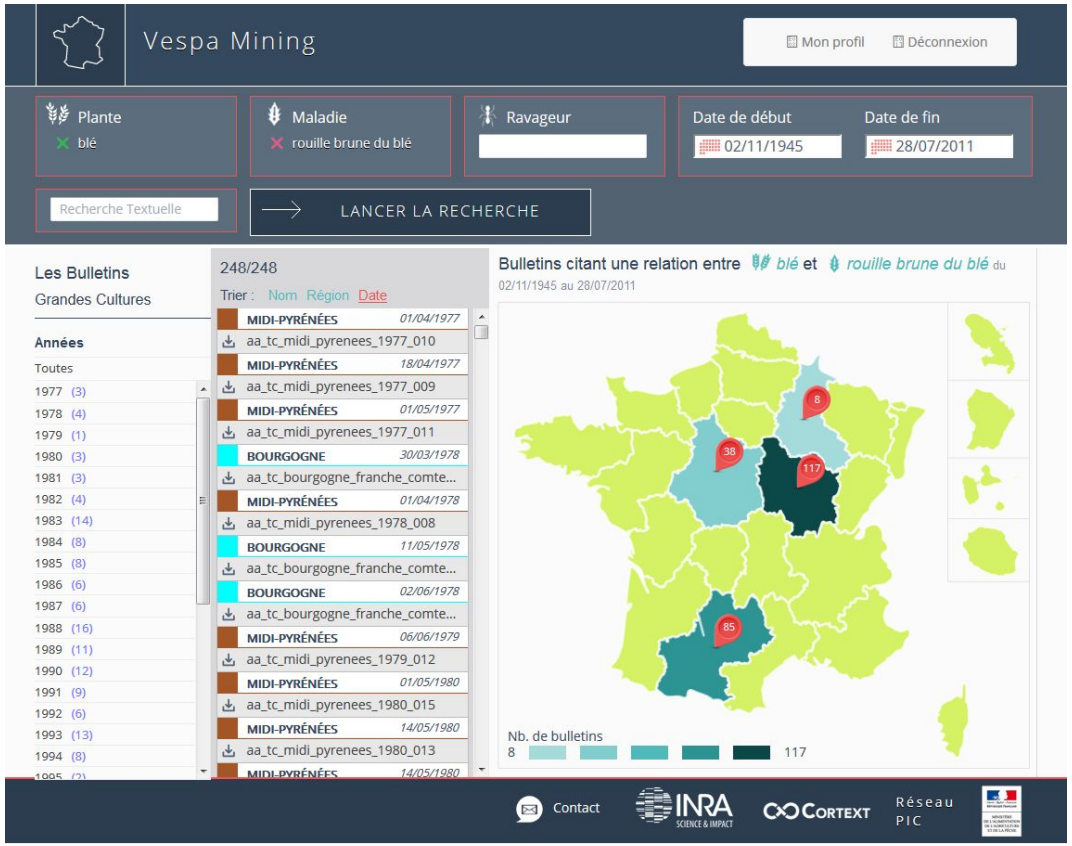
3°) la méthotryne (Gésaran 25) :

Ce produit est destiné spécialement à la lutte contre le vulpin des champs, mais il agit également contre diverses dicotylédones au stade plantules et peut permettre d'éviter

...F.../467

Exemple de tapuscrit numérisé

Avertissements agricoles



Plateforme Vespa Mining

Exemple, la domestication des poissons

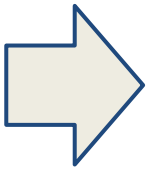
- Créer une base de données avec les traits de reproduction et d' alimentation des poissons extraits des textes.
- Identifier les caractères communs aux espèces d' aquaculture.
- Découvrir d'autres espèces à domestiquer.

Flickr, Marc, Pescadería – Fish Shop, Madrid HDR, CC BY-NC-SA 2.0

Transformer les textes en connaissances

"Sardines spawn in a much wider temperature range (13-25°C) than anchovy (11.5-16.5°C)."

(LLUCH-BELDA ET AL.: SARDINE AND ANCHOVY SPAWNING AS RELATED TO TEMPERATURE AND UPWELLING CalCOFI Rep., Vol. 32,1991)



Poisson : *Sardina pilchardus*
Trait de reproduction
Température de frai min. (C°) *: 13*
Température de frai max.(C°) *: 25*

Poisson : *Engraulidae*
Trait de reproduction
Température de frai min. (C°) : *11.5*
Température de frai max.(C°) : *16.5*

Comment nous travaillons

Quelques solutions de text mining

Commerciales	Académiques
Luxid (TEMIS)	Nooj / INTEX
SemioLab (Noopsis)	Cortext
Syllabs Text Mining	Alvis NLP/ML
IBM SPSS	UIMA
Antelope (Proxem)	OpenNLP
	GATE

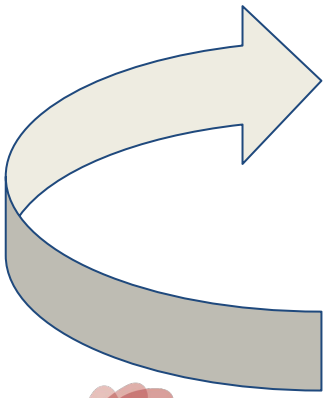
Luxid

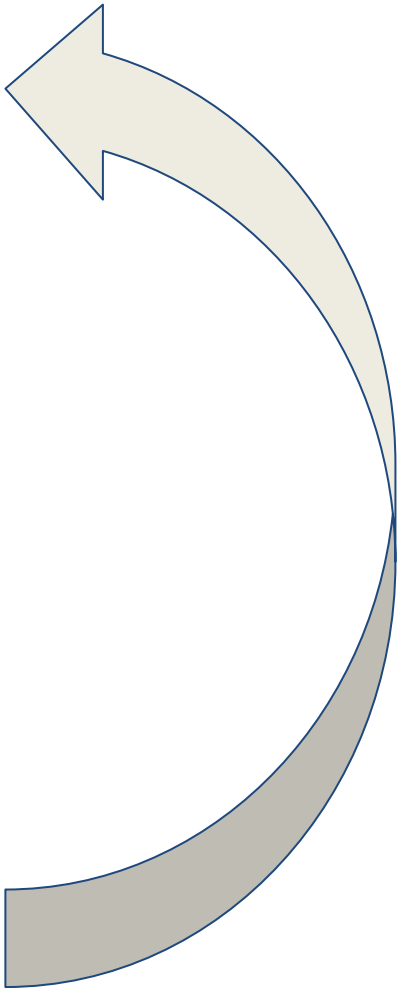
TEMIS
STRUCTURE THE UNSTRUCTURED

- Approches complémentaires
- Large éventail de fonctionnalités, capacité : volumes, formats, rapidité
- Solution éprouvée notamment par les grands éditeurs (Elsevier, Nature Publishing Group, Thomson Reuters...)
- Supports de formation, communauté
- Outils de construction de ressources
- Réutilisation et capitalisation de ressources

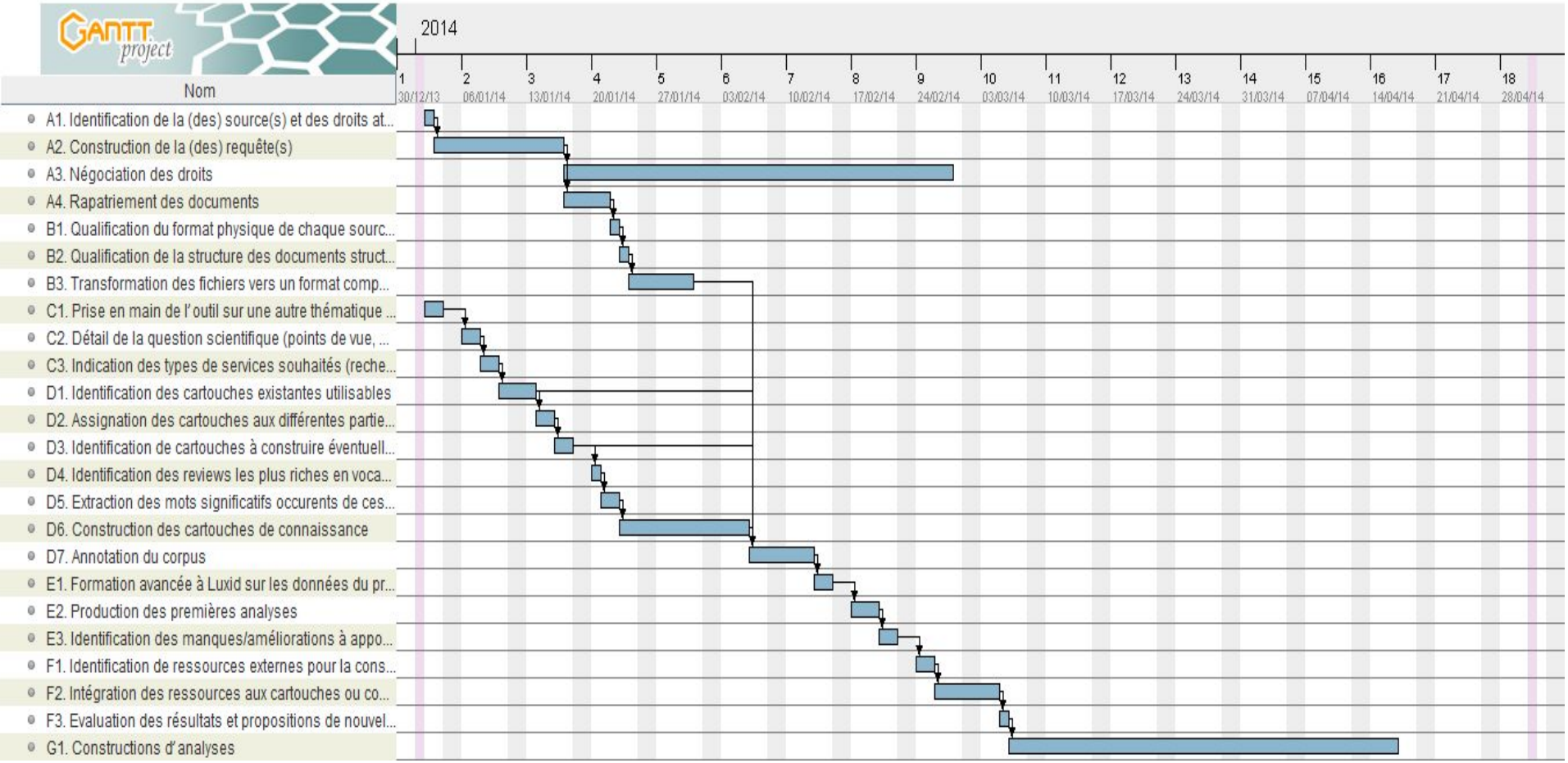
Grandes étapes d'un projet text mining

1. Expression des besoins et des objectifs poursuivis (analyses, sources). Quels titres de diagrammes ?
2. Acquisition du corpus (notices, textes, numérisation, OCR, veille…)
3. Développement des vocabulaires d'annotation
4. Production d'analyses et extractions





Planning type



Offre de services IST en text mining

- Traitement de tous types de documents (numérique ou papier)
- Constitution de corpus avec identification des sources, conseils juridiques, récupération et mise au format des documents
- Fourniture ou aide à la construction de ressources d' annotation
- Paramétrage de la chaîne d'annotation
- Mise à disposition d'une interface personnalisée pour produire les analyses
- Formations
- Conduite de projet

Analyses sur le corpus Inra

https://text-mining.inra.fr/Luxid

login : demo

mot de passe : demo1

- Service permanent
- Corpus Inra (Web of Science 1999-2014) constitué par les équipes bibliométrie
- Analyses sur les sociétés, les institutions, les unités, les départements, les pays, les taxons, les revues, les sujets...

Merci de votre attention

Diaporama : http://tinyurl.com/md6gckl
Contact : mathieu.andro@versailles.inra.fr