



HAL
open science

Infodoc Express : Introduction au text mining

Mathieu Andro, Sophie Aubin

► **To cite this version:**

Mathieu Andro, Sophie Aubin. Infodoc Express : Introduction au text mining. Infodoc Express (Introduction au text mining), 2016. hal-02793404

HAL Id: hal-02793404

<https://hal.inrae.fr/hal-02793404v1>

Submitted on 5 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.

Infodoc Express

Introduction au text mining

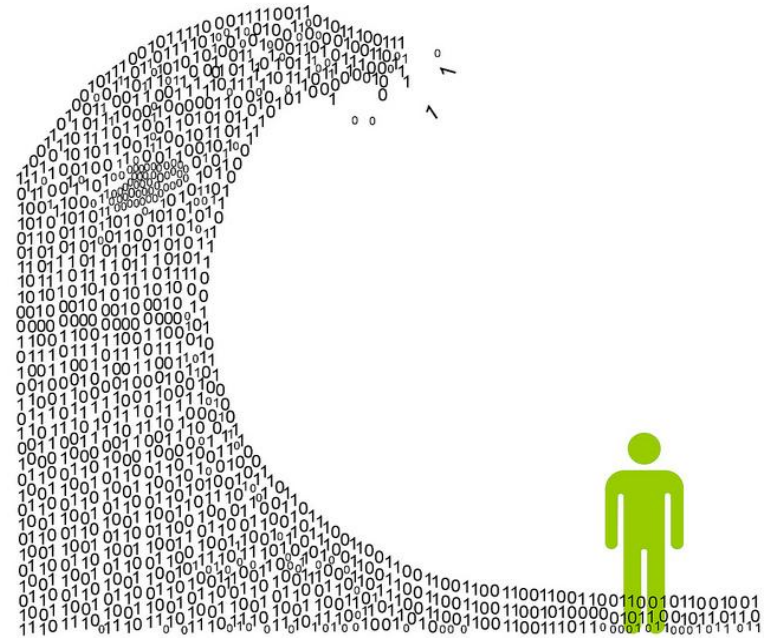
mathieu.andro@versailles.inra.fr
sophie.aubin@versailles.inra.fr



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...

dans des masses de documents...

et **identifier les relations** entre ces mots

et **extraire des données** de ces documents



Flickr, Kate Ter Haar, CC BY 2.0

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

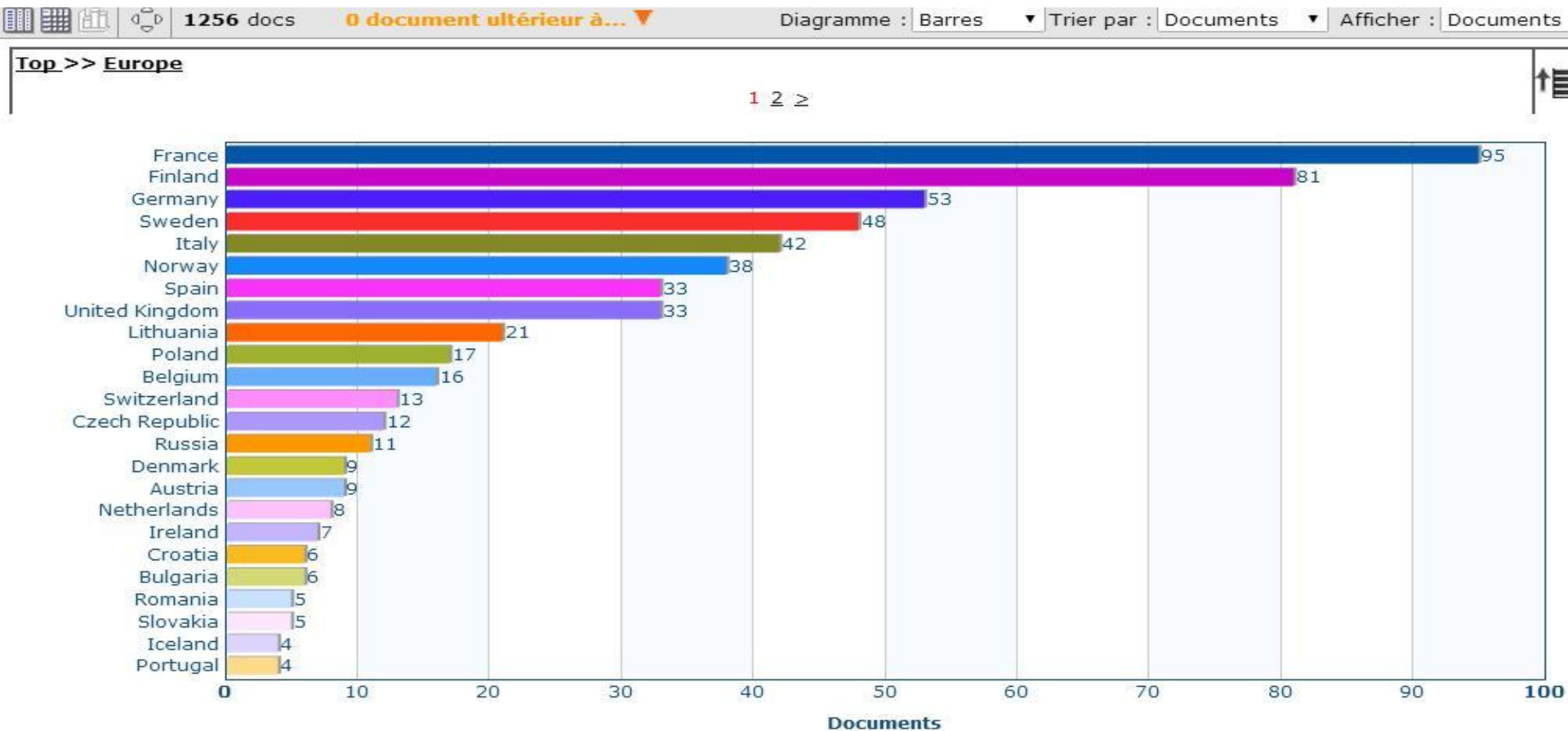
Objectif : état de l'art sur la recherche sur la phénologie des arbres forestiers.

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 :

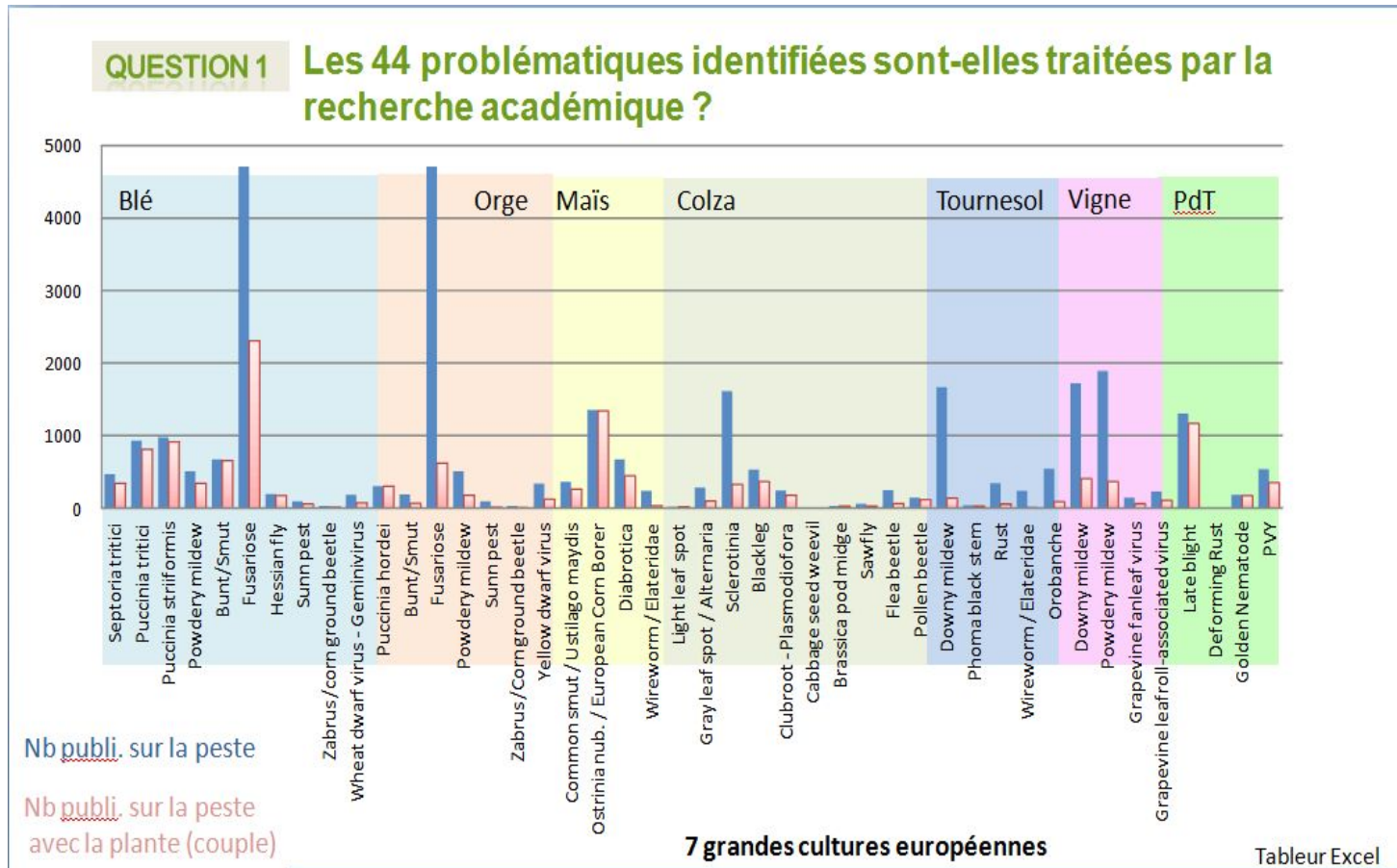
- produire des analyses plus rapidement et avec moins d'efforts (changement d'échelle)
- faire évoluer facilement la question scientifique et le périmètre
- formaliser l'expertise pour la réutiliser

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



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

Projet Pest & Crops



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

Projet GIS Biocontrôle

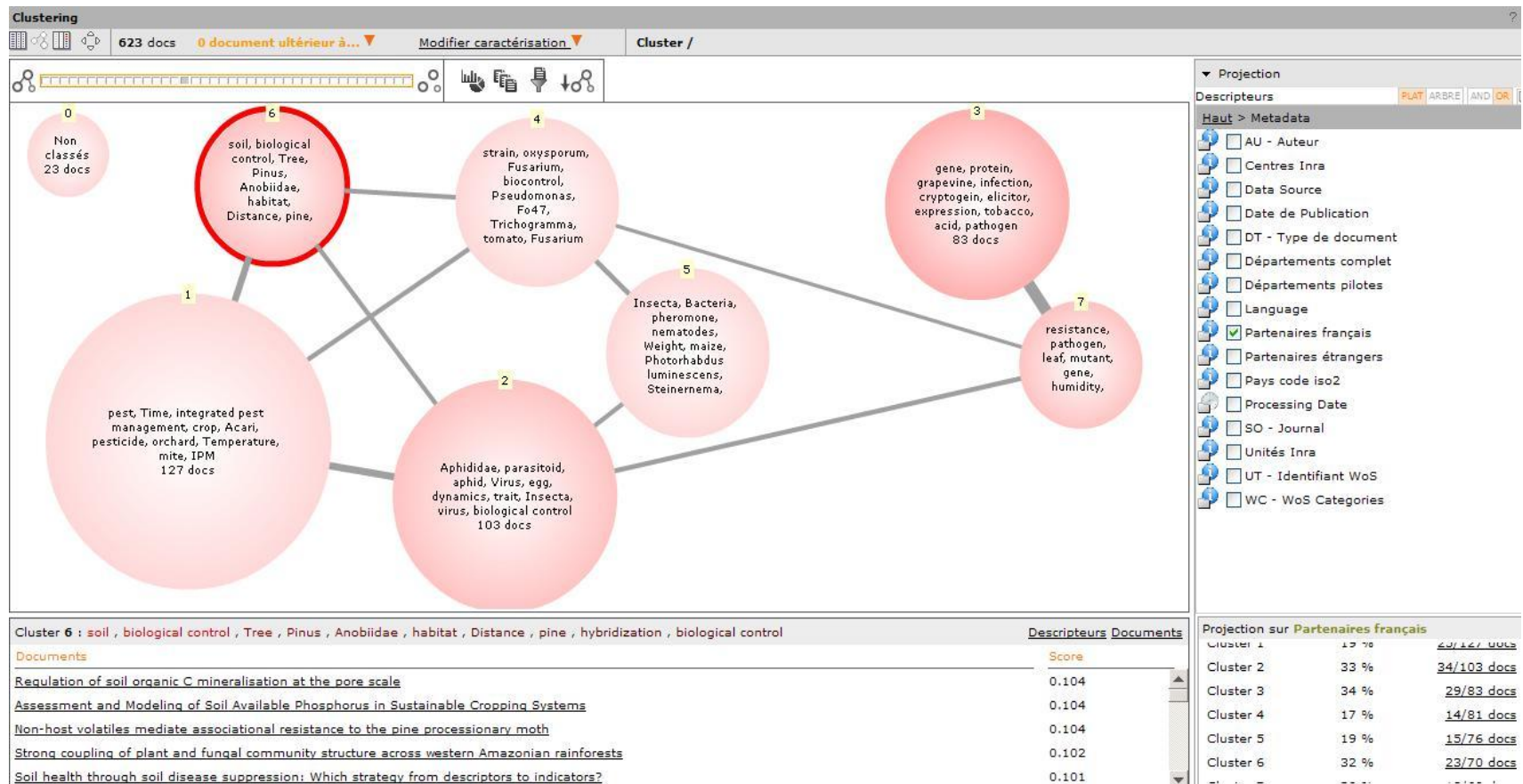
Objectifs : identifier des partenaires publics et privés pour un GIS, proposer un service mutualisé aux membres du GIS.

Technologies originales mobilisées : 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 :

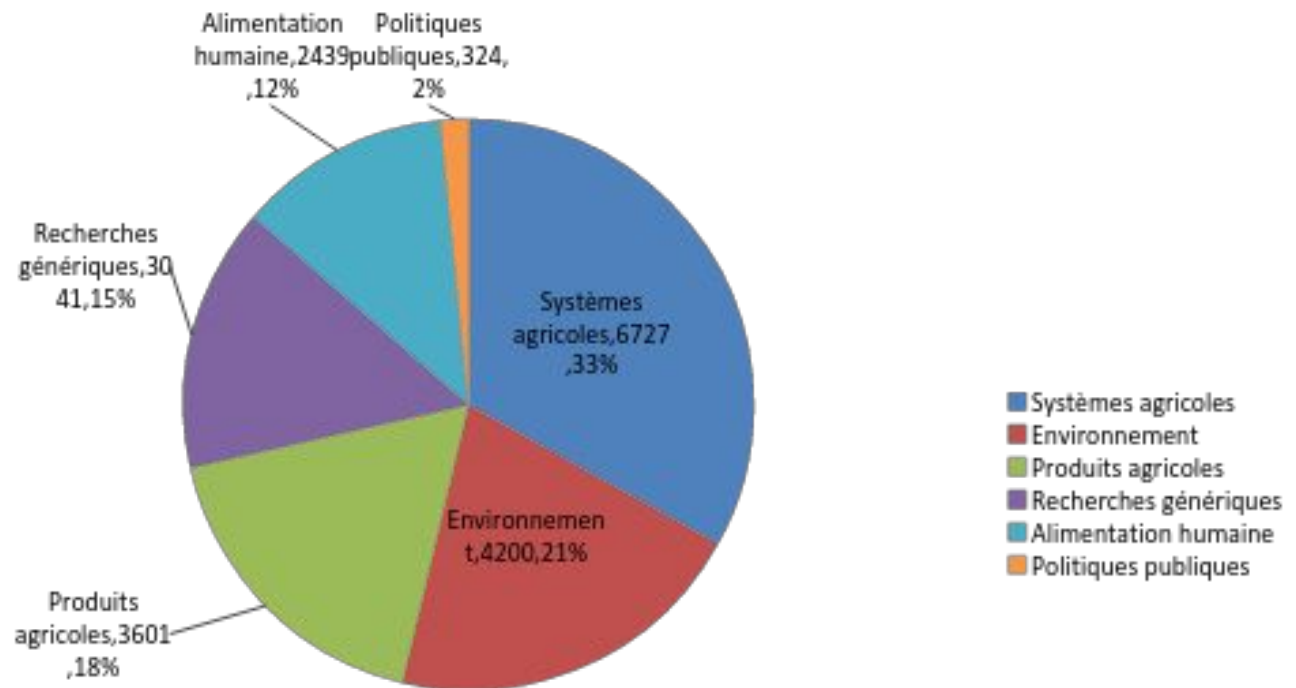
- Apprentissage par analyse de l'empreinte sémantique de corpus existants classés par un humain.
- Complémentaire à bibliométrie



pixabay.com, ejaugsburg, CC0

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

Analyses stratégiques



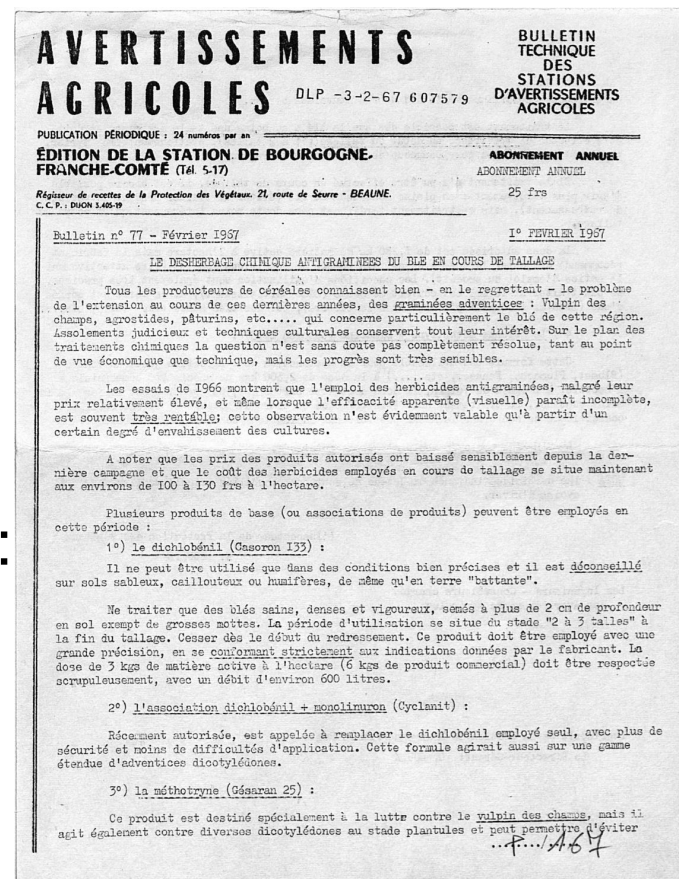
Répartition estimée des publications Inra par axes stratégiques (2001-2007) d'après CorText

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



Exemple de tapuscrit numérisé

Avertissements agricoles

The screenshot displays the Vespa Mining web interface. At the top, there is a navigation bar with a map of France icon, the text 'Vespa Mining', and links for 'Mon profil' and 'Déconnexion'. Below this is a search panel with several filters: 'Plante' (blé), 'Maladie' (rouille brune du blé), 'Ravageur', 'Date de début' (02/11/1945), and 'Date de fin' (28/07/2011). A 'Recherche Textuelle' field and a 'LANCER LA RECHERCHE' button are also present.

The main content area is divided into three sections:

- Les Bulletins**: A sidebar with 'Grandes Cultures' and a list of 'Années' from 1977 to 2005.
- Table of Results**: A table with columns 'Nom', 'Région', and 'Date'. It lists bulletin entries such as 'aa_tc_midi_pyrenees_1977_010' from 'MIDI-PYRÉNÉES' on '01/04/1977'.
- Map**: A map of France titled 'Bulletins citant une relation entre blé et rouille brune du blé du 02/11/1945 au 28/07/2011'. The map shows bulletin counts by region: 8 in the north, 38 in the west, 117 in the center, and 85 in the south.

At the bottom of the interface, there are logos for 'Contact', 'INRA SCIENCE & IMPACT', 'CORTEXT', 'Réseau PIC', and the French flag.

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.

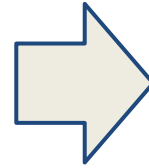


Flickr, Marc, Pescadería – Fish Shop, Madrid
HDR, CC BY-NC-SA 2.0

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

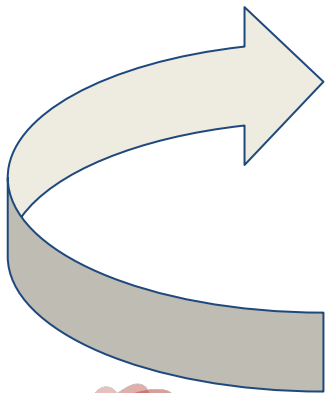
Quelques solutions de text mining

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

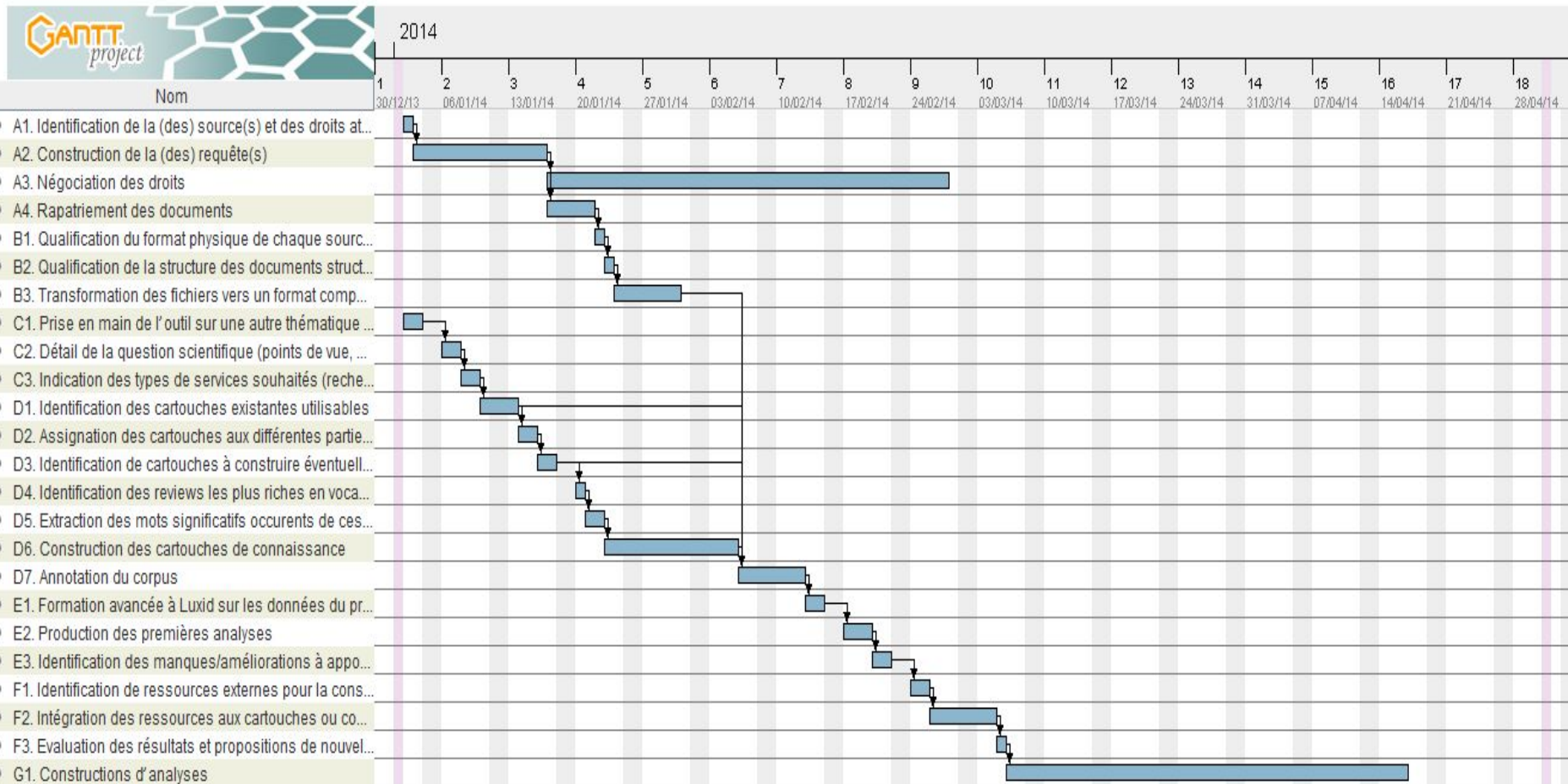
- Approches complémentaires
- Large éventail de fonctionnalités, capacité : volumes, formats, rapidité
- Solution éprouvée notamment par les grands éditeurs (Elsevier, Nature Publishing Group, Thomson Reuters...)
- Supports de formation, communauté
- Outils de construction de ressources
- Réutilisation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

