



TERRE-ISTEX

Identification et analyse des TERRains d'Études dans les corpus ISTEX

Chantiers thématiques d'usage des corpus d'ISTEX 2016 – 2017

Eric Kergosien, Joachim Schöpfel

Nancy, 25 et 26 avril 2016

Gériico

STL
savoirs
langage
extes



LIUPPA



ISTEX
L'excellence documentaire pour tous

Partenaires : équipe pluridisciplinaire



- Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, Université de Lille
- Chercheurs impliqués : Stéphane Chaudiron (Pr.), Bernard Jacquemin (Maître de conférences (MCF)), Marta Severo (MCF), Joachim Schöpfel (MCF), Eric Kergosien (MCF)



- Laboratoire Savoirs, Textes, Langage associé au CNRS
- Chercheurs impliqués : Natalia Grabar



- UMR Territoires, Environnement, Teledetection et Information Spatiale – TETIS, Montpellier, attachement GDR MAGIS
- Chercheurs impliqués : Mathieu Roche, Maguelonne Teisseire, Jean-Philippe Tonneau



- Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour – LIUPPA, Pau
- Chercheurs impliqués : Marie Noëlle Bessagnet (MCF), Annig Le Parc-Lacayrelle (MCF), Christian Sallaberry (MCF, HDR)



- Atelier National de Reproduction des Thèses (ANRT), Lille
- Chercheurs et personnels impliqués : Joachim Schöpfel (directeur), Rachid Berbache (informaticien, adjoint au directeur), Jérémy Berthe (technicien, chargé de projet).

1. Etudier tout ce qu'il se passe sur un territoire sur la thématique changement climatique à partir de données scientifiques hétérogènes (Entrée spatiale)



2. Etudier la genèse des concepts sur la thématique LIS (Library Information Sciences) à partir de données scientifiques hétérogènes (Entrée thématique)
- Annexes : Définir les concepts :
 - Le territoire,
 - Library Information Sciences,
 - ...
 - Période du projet : mars 2016 – juillet 2017
 - Appui humain pour le projet : Ingénieur d'études recruté et financement interne de stages

Chaîne de traitements globale du projet

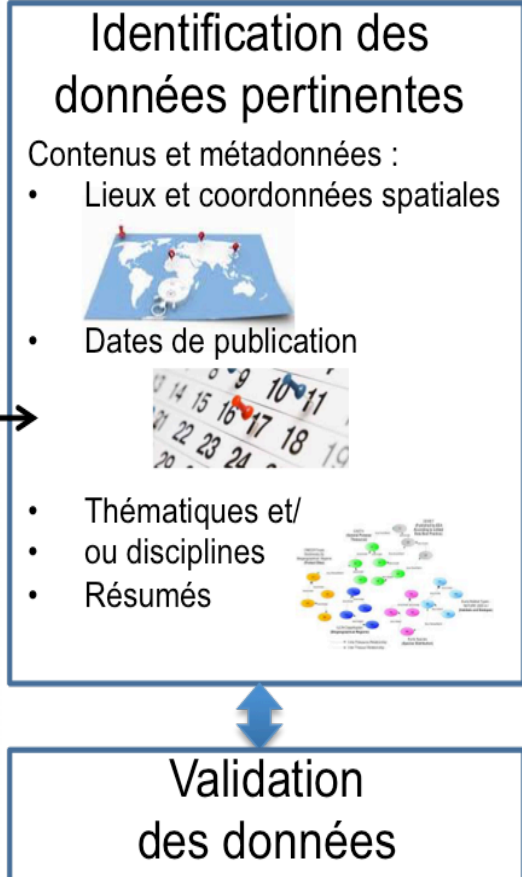


elastic



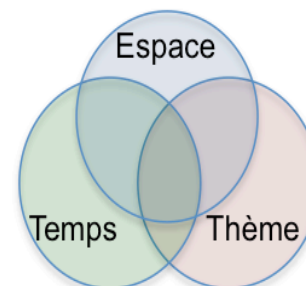
kibana

Documents
Série de
publications
et thèses



Indexation

Analyse géographique



Recherche d'information
Thème, Temps, Espace, Plein texte



- Corpus hétérogène et multilingue (anglais – français)

Données ISTEX

- Emerald : <http://www.licencesnationales.fr/liste-ressource/emerald-revues/>
 - Library management, n° 10 (SCIMAGO) ; Interlending & document supply, n° 197 (SCIMAGO) ; Journal of documentation, n° 201, 1945-2010, (CNU AERES)
- Elsevier : <http://www.licencesnationales.fr/liste-ressource/elsevier-revues/>
 - Information processing & management, n° 1176, de 1975 à 2001 (CNU AERES) ; The International information & library review, n° 1607, 1992-2001 (SCIMAGO) ; Library & information science research, n° 2161 (SCIMAGO) ; International journal of information management, n° 2195 (SCIMAGO)
- Wiley: <http://www.licencesnationales.fr/liste-ressource/wiley-revues/> [revues des origines à 2011]
 - JASIS : n° 192 (CNU AERES) ; JASIST : n° 524 (CNU AERES) ; Proceedings of the American Society for Information Science and Technology: n° 1931 (CNU AERES)
- Corpus à cibler encore pour la thématique Changement climatique
- Vocabulaires contrôlés ciblés :
 - Site web présentant des terminologies de l'INIST : **termsscience**s
 - Société savante qui édite les revues JASIS JASIST : Thésaurus de l'**ASISTT**

Données CIRAD

- Données issues d'Agritrop : archives ouvertes du CIRAD
- 92 000 références et 25 000 documents en texte intégral : publications scientifiques et littérature grise (rapports, etc.)
- Corpus multilingue
- Métadonnées : Titre, auteur, résumé, thématiques indexées à la main via le **thésaurus AGROVOC** et **Agris/Caris de la FAO**. Thèmes Agris : <https://agritrop.cirad.fr/view/subjects/>, métadonnées géographiques gros grains (pays),
- Territoires ciblés pour l'étude : Madagascar et Fleuve Sénégal : corpus encore à filtrer

Données Thèses (ANRT)

- 200 000 thèses, métadonnées internes, lien SUDOC (ABES).
- 70 000 thèses numérisées
- Notices ABES ?. Thésaurus RAMEAU
- Liens avec les thématiques du projet :
 - 300 – 350 thèses sur la thématique LIS (somme thèses.fr et version papier ANRT)
 - 400 thèses sur la thématique changement climatique (somme thèses.fr et ANRT)

Etudier tout ce qu'il se passe sur un territoire : Usages

- Questions :
 - Qu'est ce qu'un **territoire** ?
 - Ensemble d'informations géographiques mises en relation
 - information géographique = entité spatiale + entité thématique + entité temporelle
 - *Exemple : une étude du changement climatique menée dans le sud de Madagascar en 1981.*
 - Quel est le **territoire d'études associé à une thématique**?
 - Territoires d'études : Madagascar versus Lac Alaotra ; Sénégal versus Fleuve Sénégal
 - Pour un territoire, quelles sont les thématiques traitées ?
- **Cas d'application 1 côté Recherche d'Information** : cas concret en cours de définition
 - Quels sont les documents qui font mention d'un territoire?
 - De ces documents, quels sont les auteurs et quelles sont les disciplines?
 - Visualisations spatiales
 - Mobilisation des experts pour l'analyse des résultats

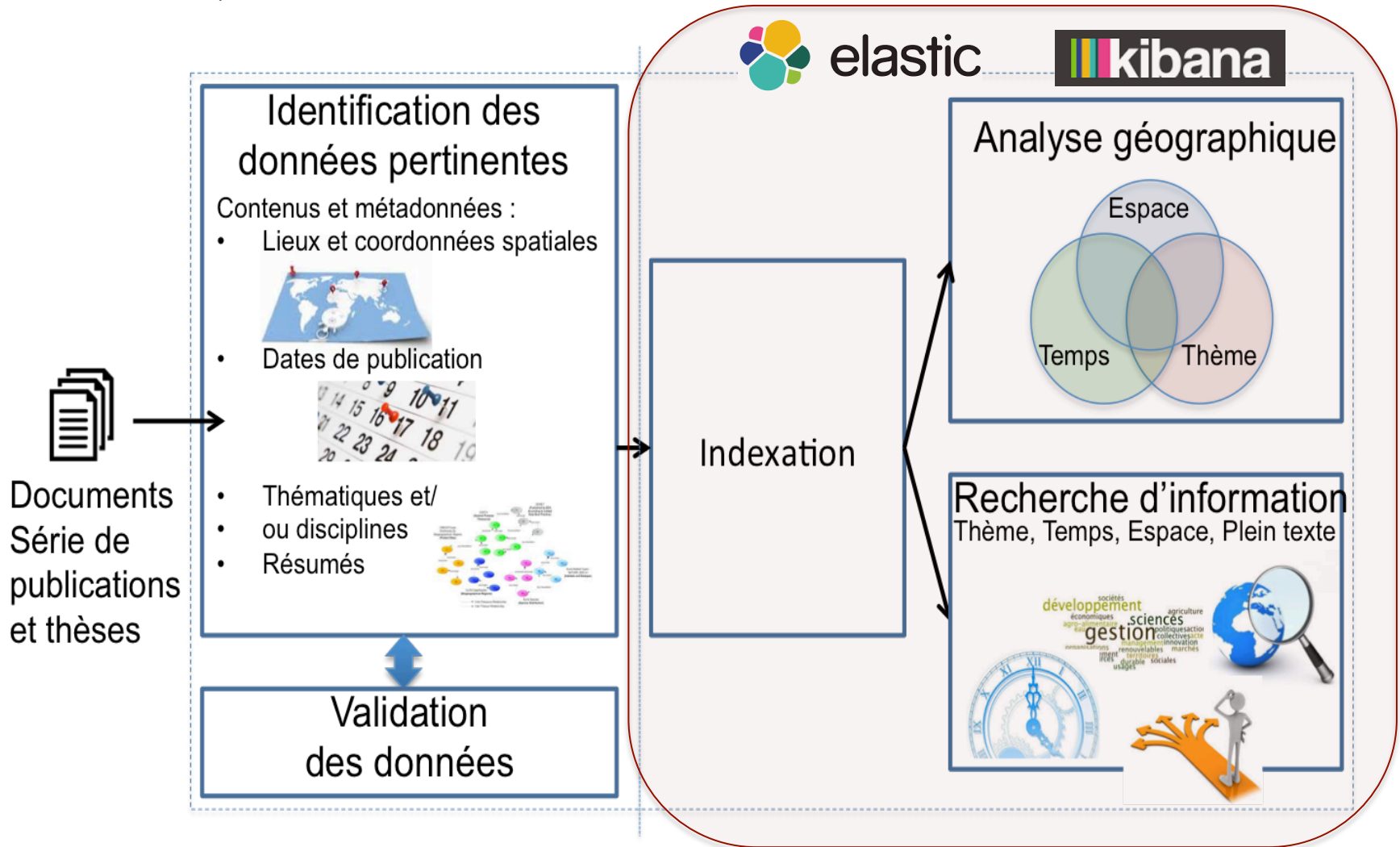
Etudier tout ce qu'il se passe sur un territoire : Méthodologie

- Utilisation de vocabulaires contrôlés du domaine et autres ressources
Thématique Changement climatique : Thésaurus AGROVOC et Agris/Caris de la FAO
Thèses et notices ABES : thésaurus RAMEAU
Bases géographiques : Geonames, Bases de données géographique de TETIS, Base de données de la SAED (polygones des toponymes)...

Approche :

- Phase 1 : Application de l'approche sur les métadonnées et résumés
 - Extraction et localisation des entités spatiales
 - Extraction des thématiques
 - Extraction entités temporelles
- Phase 2 : Application de l'approche sur les contenus des documents.
- Phase 3 : Proposer un moteur de RI géographique sous Elastic Search / Kibana
- Phase 4 : Analyse des résultats (frise chronologique, cartographies spatiales, évaluation sur un corpus annoté, Mobilisation des experts des thématiques étudiés pour validation)

Actions réalisées



Prise en main d'Elastic Search et analyse qualitative de données scientifiques

- Analyse géographique de séries de publications : application aux données EGC (EGC, 2016) : indexation, recherche d'information, analyses

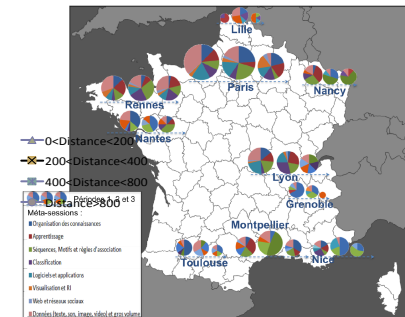
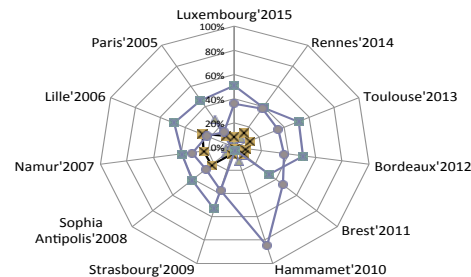
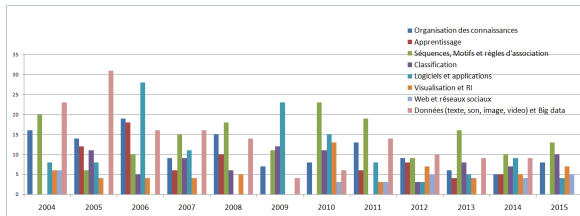
- Construction d'un premier index géographique

Information thématique	Nom ville conférence Noms villes auteurs Noms auteurs Titre article Résumé article Session Domaine
Information spatiale	Coordonnées villes auteurs Coordonnées ville conférence
Information temporelle	Année conférence
Information plein texte	Termes titre article Termes résumé article

- Mise en œuvre d'un moteur de recherche d'information multidimensionnel (Elastic Search)

- Ensemble d'analyses géographiques disponibles

<http://ekergosien.net/DefiEGC/index.html>





elastic

kibana

Documents
Série de
publications
et thèses

Identification des données pertinentes

Contenus et métadonnées :

- Lieux et coordonnées spatiales



- Dates de publication



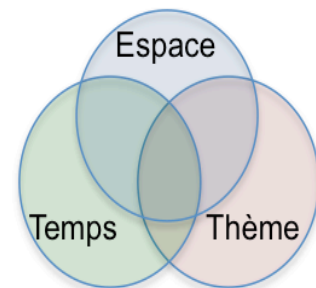
- Thématiques et/ou disciplines
- Résumés



Validation
des données

Indexation

Analyse géographique



Recherche d'information

Thème, Temps, Espace, Plein texte



Extraction et localisation des entités spatiales



Cartes



Documents scientifiques



Gazetteers



Fouille de textes pour
l'extraction d'empreinte spatiale



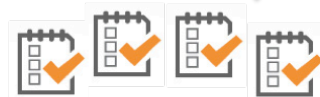
Zone d'études



Mise en relation
Textes / zone d'études



Empreintes spatiales
(une par document)



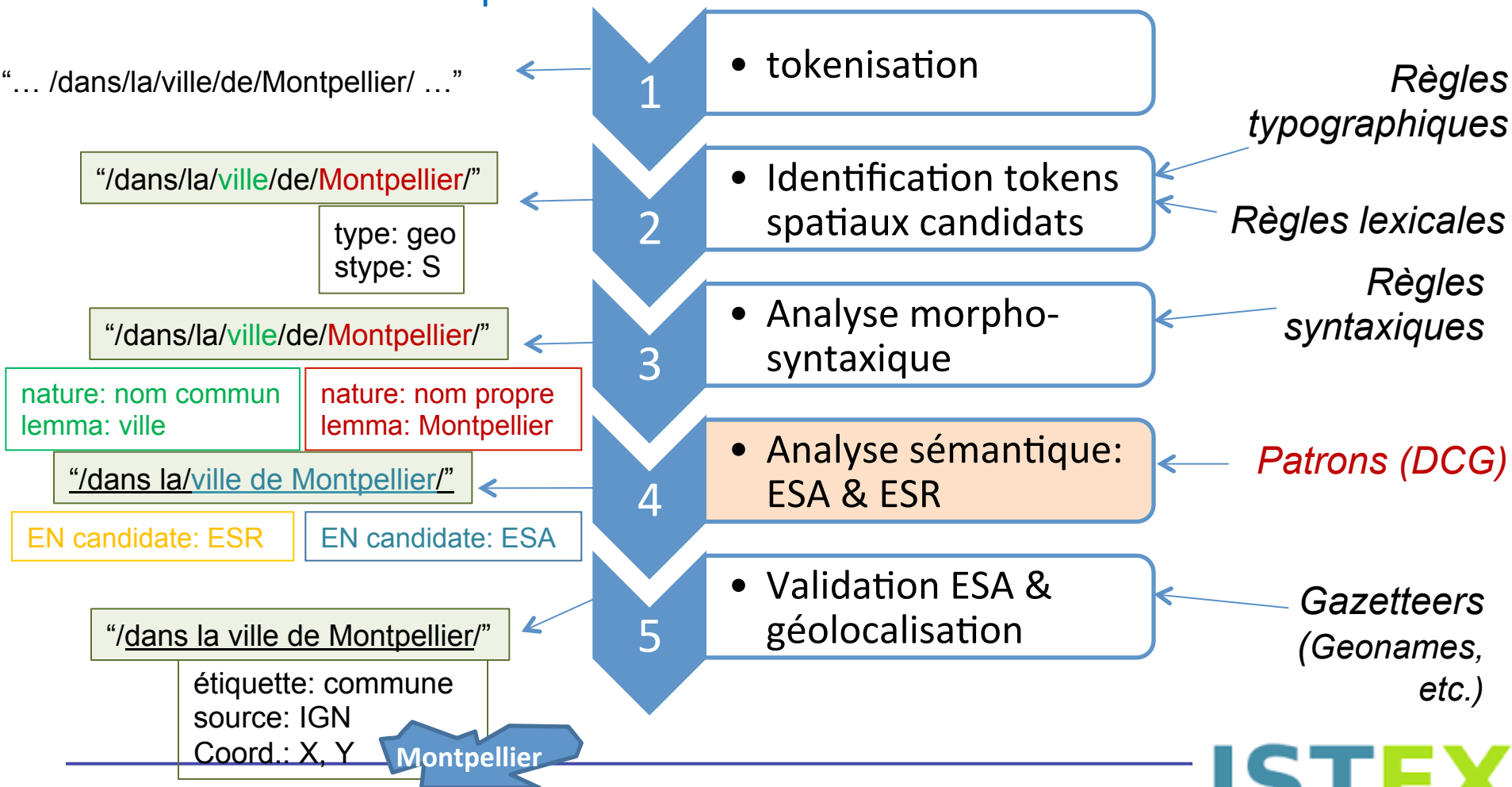
Documents sélectionnés

(Tahrat et al., WIMS 2013 ;
Kergosien et al., KDIR 2015 ;
Zenasni et al., ISMIS 2015)

Extraction et localisation des entités spatiales

Patrons linguistiques pour l'extraction d'ES (Lesbegueries, 2007)

“... dans la ville de Montpellier...”



Extraction et localisation des entités spatiales

1^{ère} Contribution : Ajout de patrons linguistiques pour l'extraction d'ES définis sur la base des travaux de (Lesbegueries, 2007) et le marquage des Ens de type Organisation

- Relation topologique et Indicateurs spatiaux placés après EN de Lieu

Ex : “Montpellier Nord”



- Distribution des relations spatiales et coréférence

Ex : Les environs de Montpellier, Marseille....= les environs de Montpellier & les environs de Marseille.

→ 2ESRs

- Entité spatiale Vs. Entité d'organisation

<Verbe Action, EN > ; <Preposition , EN>

Ex : “Le projet défendu par Montpellier Agglomération ...”

“La France a autorisé un quota de ...”

Limites : Les patrons présentés exploitent un contexte local assez réduit

→ Prendre en compte un contexte plus vaste pour distinguer ES et Organisations



Extraction et localisation des entités spatiales

2^{ème} contribution : approche « sac de mots »

a. Classification supervisée pour construire un modèle de prédiction



Liste de phrases (vérité terrain)



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1. Matrice très creuse

- Mots non vides et non rares (1 occurrence) + Descripteurs TEXT2GEO



ORG
ES
ORG
ES
ES
ES
ORG
ES
ES
ES
Org

2. Classes par type d'EN
La phrase décrit une ES ou une Organisation



b. Calcul proximité vecteurs signatures (SVM) (Joachims (1998))



Indicateurs	Indicateurs actuels ⁴⁰	Scénario tendanciel 2020 ⁴¹	Scénario cible 2020
Croissance potentielle	2 %	1,3 %	2,5 % à 3 %
Espérance de vie ⁴²	80,9 ans	82 ans	83 ans ⁴⁴
Taux de décrochage scolaire ⁴⁵	11,8 %	> 11,8 %	9,5 %
Taux de chômage	9,6 %	9,0 %	4,5 %
Dette publique (% PIB)	83,3 %	Plus de 100 %	Moins de 70 % et vers 60 %
Solde public (% PIB)	- 8,2 %	- 8 %	Equilibré

→ Combinaison avec CASys (Unitex) ?!

- **Extraction d'entités thématiques**

- BioTex (Lossio-Ventura et al., 2015) Extraction de terminologie à partir de textes
 - Approche hybride statistique (combinaison mesure appelée C-value pour mesurer l'association entre les mots composant un terme et différentes pondérations (TF-IDF, Okapi)) et linguistiques (Patrons linguistiques) pour extraire la terminologie à partir de textes libres.
<http://tubo.lirmm.fr/biotex/about.jsp>
 - But de C-value : améliorer l'extraction des termes complexes particulièrement adaptés pour les domaines de spécialité
 - Méthodologie générique qui a été essentiellement appliquée aux domaines scientifiques (biomédical et agronomique)
- Construction d'un monde lexical autour de thématiques : volonté d'intégrer les lexiques de domaine (à faire)
- Construction d'un monde lexical autour d'entités spatiales (à faire)
- Combinaison avec TermSuite ?!

- **Extraction d'entités temporelles**

- Intégration de l'outil multilingue HeidelTime pour l'extraction d'entités temporelles :
 - <https://github.com/HeidelTime/heideltime>
 - Evaluation pertinente sur un autre projet
 - Evaluation en cours sur un sous corpus CIRAD de 266 articles scientifiques

Démonstrateur pour l'extraction de contenu et la validation experte (ISWC, october 2015)

- ✓ Upload de corpus volumineux
- ✓ Chaines de traitements développées sous GATE (ES, thèmes, opinions)
- ✓ Les Experts peuvent corriger le marquage
- ✓ Téléchargement des résultats possibles

En cours :

- Passage à l'anglais
- Evaluation
- Intégration de Biotex et HeidelTime

The screenshot displays the Sentrtoire Web services interface, which is used for document processing and validation. The interface is divided into several sections:

- DISPLAYED INFORMATION:** A sidebar on the left showing a list of displayed information categories: Spatial Features, Organization, Opinions, Other, and Topic. The 'Opinions' category is currently selected.
- CORPUS AND DOCUMENTS:** A section below the displayed information, showing a list of corpus and documents. The '1_47_2_docs_With_NERs' corpus is selected.
- DOCUMENTS:** The main content area displays a list of documents. The first document, '1_47_2_docs_With_NERs', is selected and its content is shown. The text is a French article about David Bowie and the CCI (Chambre de Commerce et d'Industrie).
- MARKED INFORMATION:** A sidebar on the right showing the results of the processing. It includes:
 - Spatial Features (17):** A list of spatial features extracted from the document, such as 'à', 'à Cougnenc', 'à Frêche', 'à la Région', 'à Port-La Nouvelle', 'à Sète', 'dans', and 'Dans'.
 - Organizations (25):** A list of organizations extracted from the document, such as 'la Région de payer la CCI', 'la Région était', 'la Région sur', 'la Région verse', 'la Région', 'les CCI', 'nationale', 'port de commerce', and 'port'.
 - Opinions (18):** A list of opinions extracted from the document, such as 'avant', 'comme', 'Comme', 'exception', and 'secret'.

At the bottom of the interface, there is a status bar showing the processing time: 'Just Gate Process time (mmiss) => 0:23, Total Process time (mmiss) => 0:24'. The interface is processed by Sentrtoire Web services - 2014.

<http://siso.teledetection.fr/>

Publications

- J. Schöpfel, E. Kergosien, S. Chaudiron and B. Jacquemin, Dissertations as Data, In 19th International Symposium on Electronic Theses and Dissertations, To appear, Lille, July 2016
- E. Kergosien, M.-N. Bessagnet, C. Sallaberry, A. Le Parc - Lacayrelle, A. Royer, Analyse géographique de séries de publications : application aux conférences EGC, In Actes de la conférence EGC'2016 (Extraction et Gestion des Connaissances), p.371-382, Reims, 2016

Communications

- E. Kergosien, S. Chaudiron, J. Schöpfel, M. Roche, A. Farvardin, and al., Le projet TERRE-ISTEX pour l'identification et l'analyse des terrains d'études dans les corpus ISTEX, *Séminaire Archives ouvertes et bases de publications : exploration et analyse des sources de données pour la recherche et ses environnements*, Mai 2016, Paris.
- E. Kergosien, M.-N. Bessagnet, C. Sallaberry, A. Le Parc - Lacayrelle, A. Royer, Vers une analyse thématique automatique de séries de publications : application aux articles des conférences EGC, In 84ème conférence de l'ACFAS, Montréal, mai 2016



ETD2016



ETD 2016 "Data and Dissertations"

19th International Symposium on Electronic Theses and Dissertations

11-13 Jul 2016 Lille (France)

<http://etd2016.sciencesconf.org>

Etudier la genèse des concepts sur la thématique LIS : Usages

- Questions :
 - Qu'est ce que la **genèse de termes** ?
 - **Première apparition du terme** dans les articles/thèses à une date donnée
 - Quelles sont les sous-thématiques associées et leur évolution ?
 - **sous-thématique** : ensemble de termes qui revient souvent ensemble dans une fenêtre autour de la sous-thématique
 - Comment ont évoluées dans le temps ces sous-thématiques ?
 - Ces sous-thématiques évoluent-elles selon les **types de documents** ? (thèses vs. Articles)
- **Cas d'application 2 côté Recherche d'Information** : cas concret en cours de définition
 - Quels sont les documents qui font mention d'une sous-thématique?
 - De ces documents, quels sont les auteurs et sur quel territoire (quels espaces) cette sous-thématique a été appliquée?
 - Visualisations sous forme de frise chronologique
 - Analyses des résultats obtenus

Etudier la genèse des concepts sur la thématique LIS : **Méthodologie**

- Utilisation de vocabulaires contrôlés du domaine
Thématiques LIS : Thésaurus de l'ASISTT et termsciences
Thèses et notices ABES : thésaurus RAMEAU

Approche :

- Phase 1 : Application de l'approche sur :
 - les métadonnées : termes provenant des vocabulaires contrôlés utilisés pour décrire le contenu des documents ;
 - Et les résumés : identifier un monde lexical (par extraction libre et/ou contrôlée) qui apparaît "autour" d'un terme étudié
- Phase 2 : Application de l'approche sur les contenus des documents
- Phase 3 : Analyse des données (frise chronologique, Mobilisation des experts (scientifiques experts de la thématique étudiée) pour valider nos résultats)

Expérimentations (2/3) : Approche TAL



exhaustivité

Rappel =

$\frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de réponses correctes existantes}}$

qualité

Précision =

$\frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de toutes les réponses extraites}}$

Corpus
300 phrases

F-mesure = $2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$



Patrons
Pivot



Patrons Text2Geo



Le sens
dépend du
contexte

	ESA	ESR
Rappel	63%	27%
Précision	20%	48%
F-mesure	30%	34%

Mesure	ESA	ESR	ORG
Rappel	94%	66%	35%
Précision	53%	84%	92%
F-Mesure	67%	74%	50%

Expérimentations (3/3) : méthode hybride

- **Ensemble d'apprentissage :**
 - 138 phrases contenant des entités spatiales
 - 134 phrases contenant des ENs Organisation.
- Les évaluations données utilisent le principe de **validation croisée**.

CLASSIFICATION DES PHRASES

SVM

	sp	org
sp	103	35
org	44	90

Taux d'exactitude: **70%**

CLASSIFICATION DES PHRASES AVEC CONTRAINTES

Contrainte Préposition org
(avec, par...)

	sp	org
sp	108	30
org	47	87

Taux d'exactitude: 71,69%

Contrainte Préposition sp et
indicateurs spatiaux
(à, en, sud...)

	sp	org
sp	112	26
org	19	115

Taux d'exactitude: **83,45%**

Les deux contraintes à la fois

	sp	org
sp	113	25
org	19	115

Taux d'exactitude: 83,82%