Le projet TERRE-ISTEX pour l'identification et l'analyse des terrains d'études dans les corpus ISTEX

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

Eric Kergosien

Journées d'ISTEX - 15 et 16 mars 2018
Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris

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 (MCF), Marta Severo (MCF), Joachim Schöpfel (MCF), Eric Kergosien (MCF)

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

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

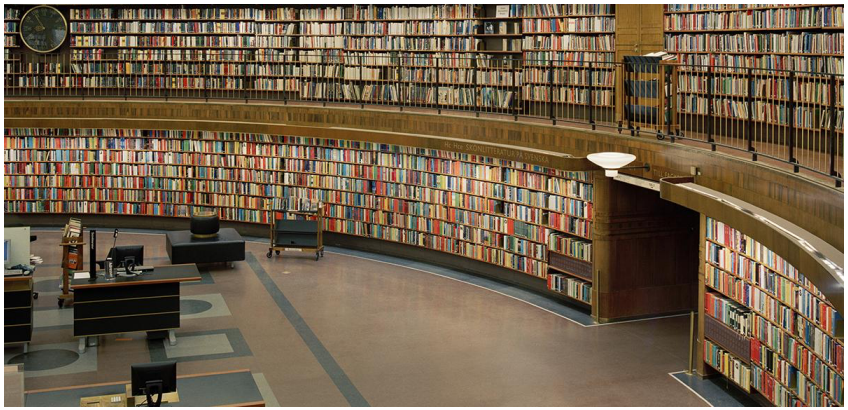
- Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour – LIUPPA, Pau
- Chercheurs impliqués : Marie Noëlle Bessagnet (MCF), Annig Le Parc-Lacayrelle (MCF), Christian Sallaberry (MCF, HDR)

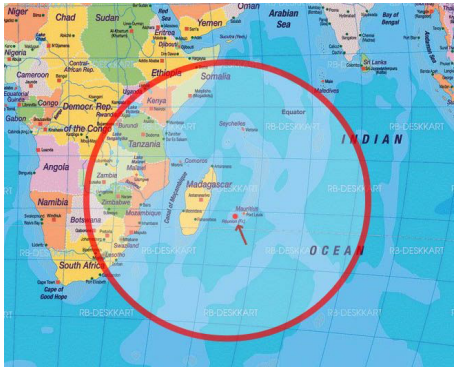
- Atelier National de Reproduction des Thèses (ANRT), Lille
- Chercheurs et personnels impliqués : Joachim Schöpfel (directeur), Rachid Berbache (informaticien, adjoint au directeur), Jérémy Berthe (technicien, chargé de projet).

TERRE-ISTEX

Notre cas d'études général pour le projet « chantier thématiques »

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





ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

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.

  - Cas d'applications :
    - Quel est le territoire d'études associé à la thématique « changement climatique »?
    - Pour les territoires Lac Alaotra (Madagascar) et Fleuve Sénégal (Sénégal), quelles sont les thématiques traitées ?

- Et côté Recherche d'Information :
  - Quels sont les documents qui font mention d'un territoire?
  - De ces documents, quelles sont les périodes et thématiques mentionnées

    → Visualisations spatiales

    → Mobilisation des experts pour l'analyse des résultats

TERRE-ISTEX

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

Exemple pour l'entité spatiale absolue

« ... le fleuve Sénégal ... »

ESA

Outils classiques d'extraction d'entités nommées

La représentation précise

Google

ISTEX
L'excellence documentaire pour tous

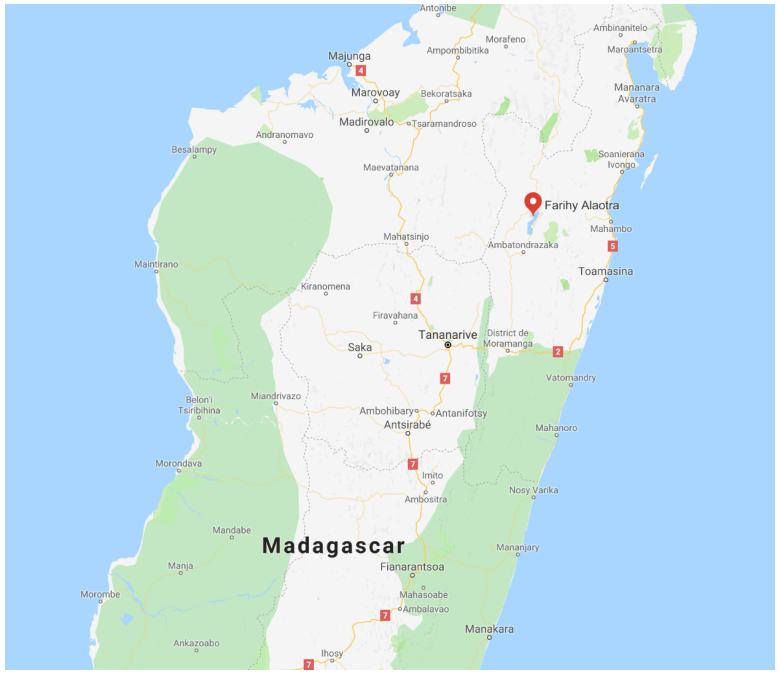
Etudier tout ce qu'il se passe sur un territoire : Usages

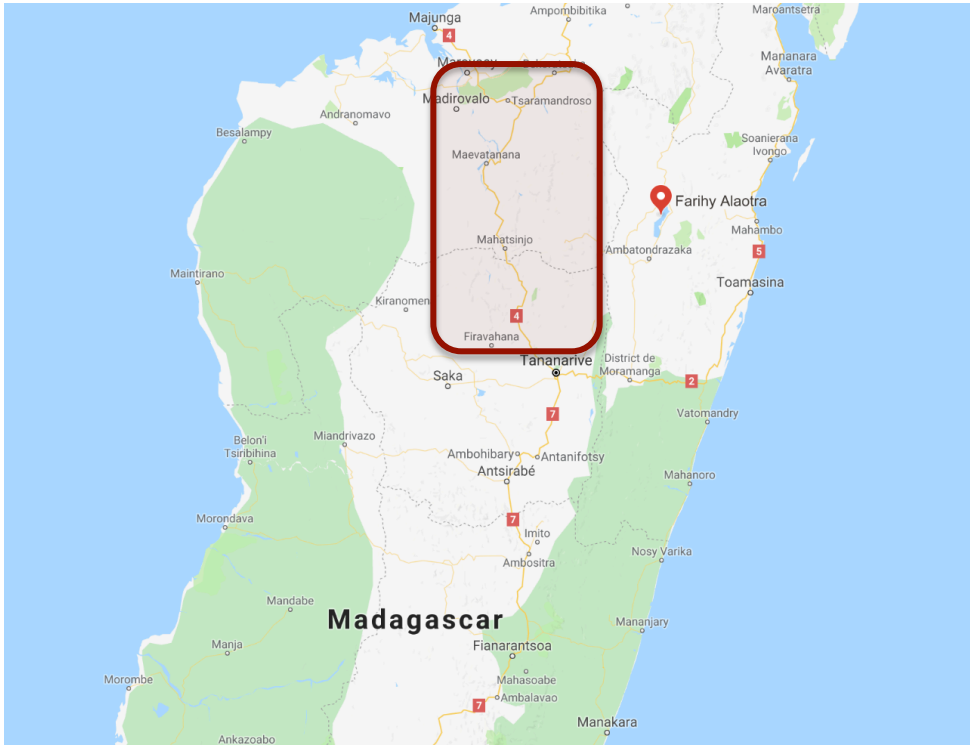
Exemples pour l'entité spatiale relative

« 100 km à l'ouest du lac Alaotra»

Relations :
Distance et orientation

ESA





ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Chaîne de traitements globale du projet

elastic kibana

Documents
Série de publications et thèses

Identification des données pertinentes

Contenus et métadonnées :

- Lieux et coordonnées spatiales

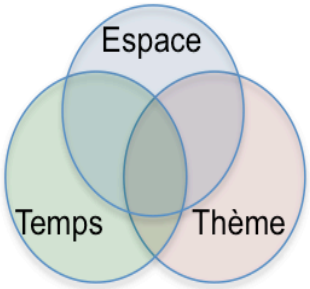
- Dates de publication

- Thématiques et/
- ou disciplines
- Résumés

Validation des données

Indexation

Analyse géographique



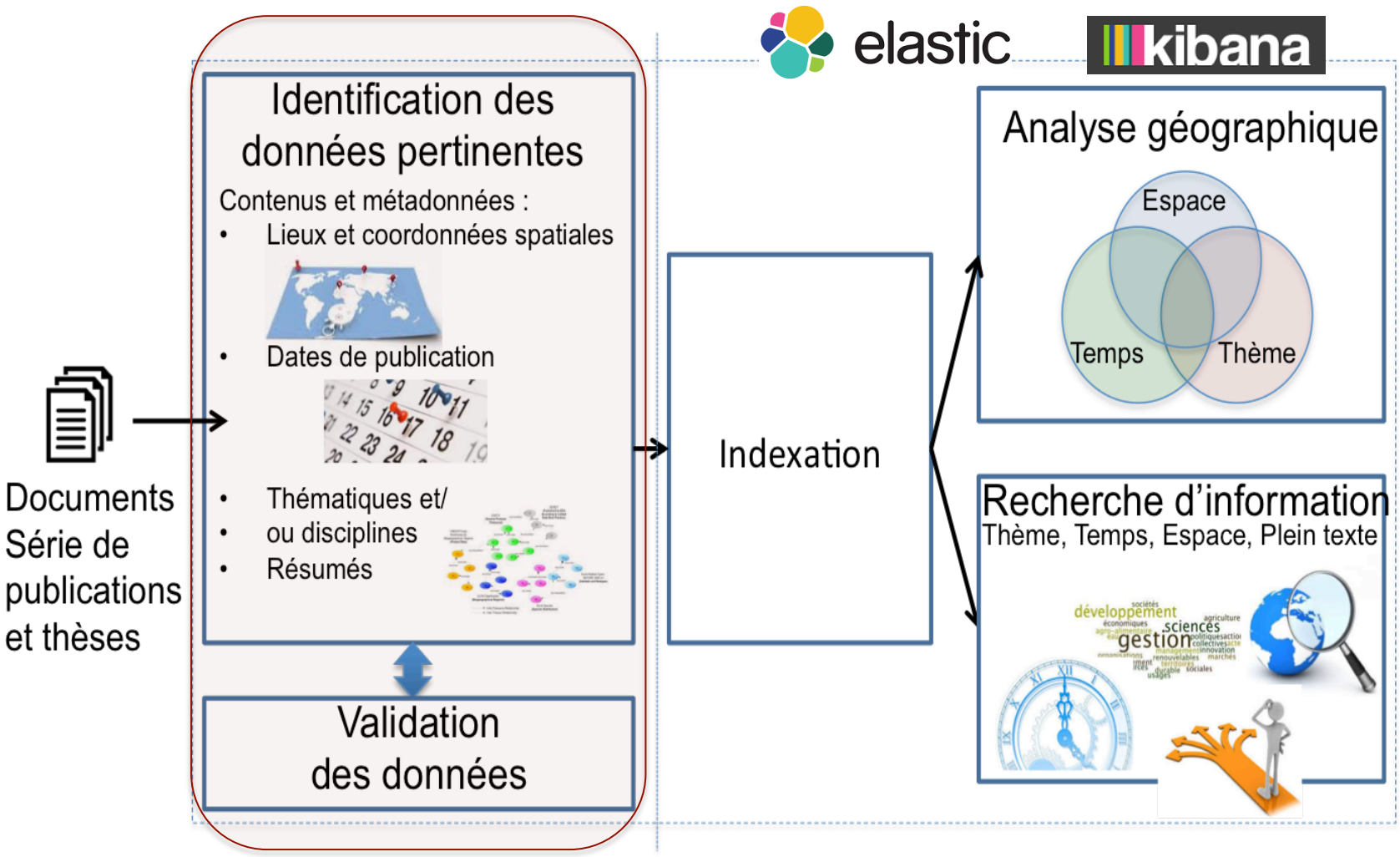
Recherche d'information

Thème, Temps, Espace, Plein texte

ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Actions en cours



- Appui humain pour le projet : Ingénieur d'études recruté (travail de 8 mois puis départ en thèse)

ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Corpus hétérogène multilingue (français – anglais)

Données CIRAD (recherche agronomique pour les pays du sud)

- 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)

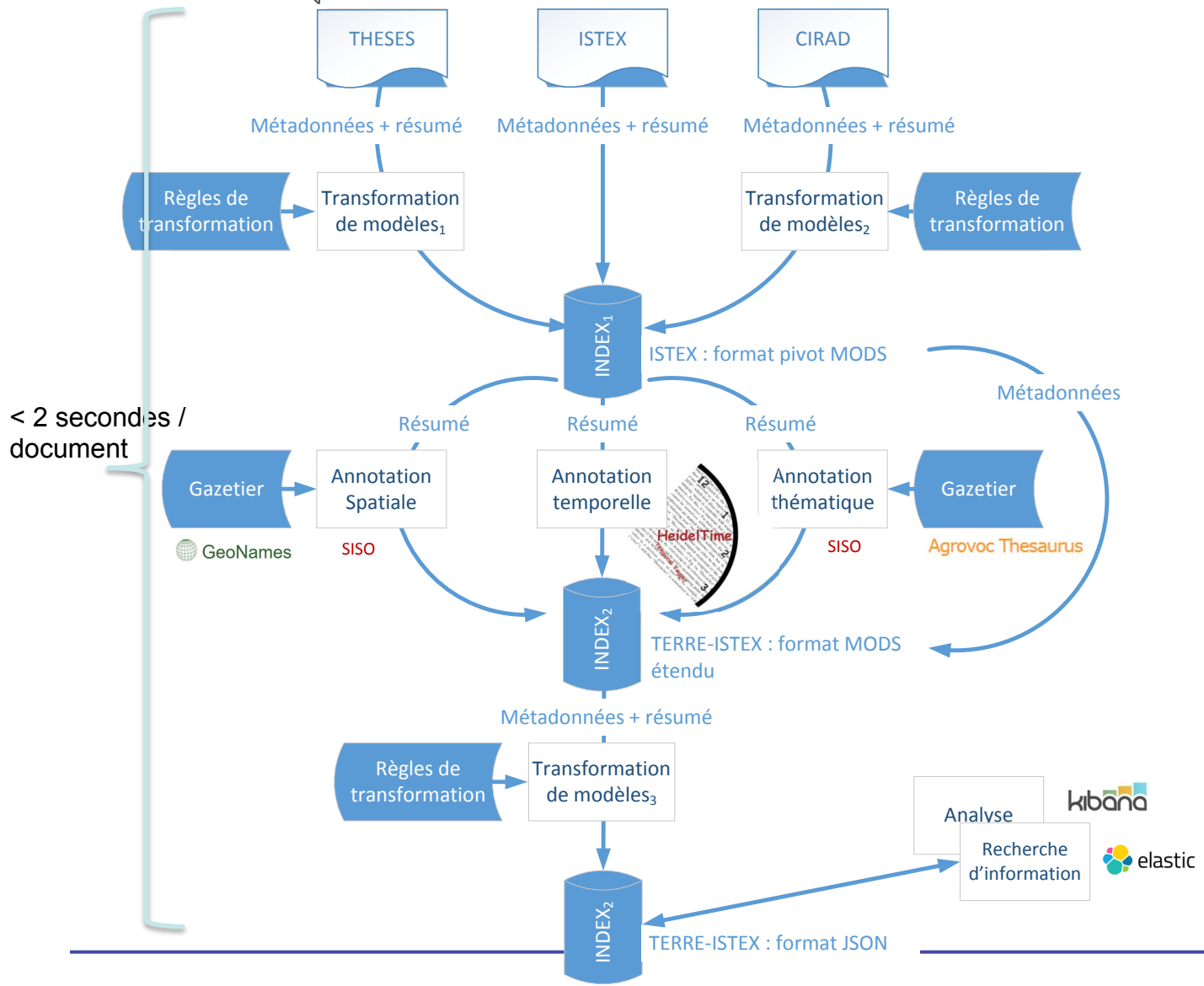
- 200 000 thèses, métadonnées internes, lien SUDOC (ABES).
- 70 000 thèses numérisées
- Notices ABES ? Thésaurus RAMEAU
- Liens avec la thématique du projet :
  - 400 thèses sur la thématique changement climatique (somme thèses.fr et ANRT)

Données ISTEX à partir des requêtes par mots clés suivantes :

- « Climate Change » et « Changement climatique » : 85 800 documents
- « Senegal » et « Sénégal » : 43 293 documents
- « Madagascar » : 41 142 documents

ISTEX
L'excellence documentaire pour tous

Phase 1 : marquage de contenu (Fouille de textes)



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Extraction et localisation des entités spatiales

Documents scientifiques

Gazetteers

Fouille de textes pour
l'extraction d'empreinte spatiale

Empreintes spatiales
(une par document)

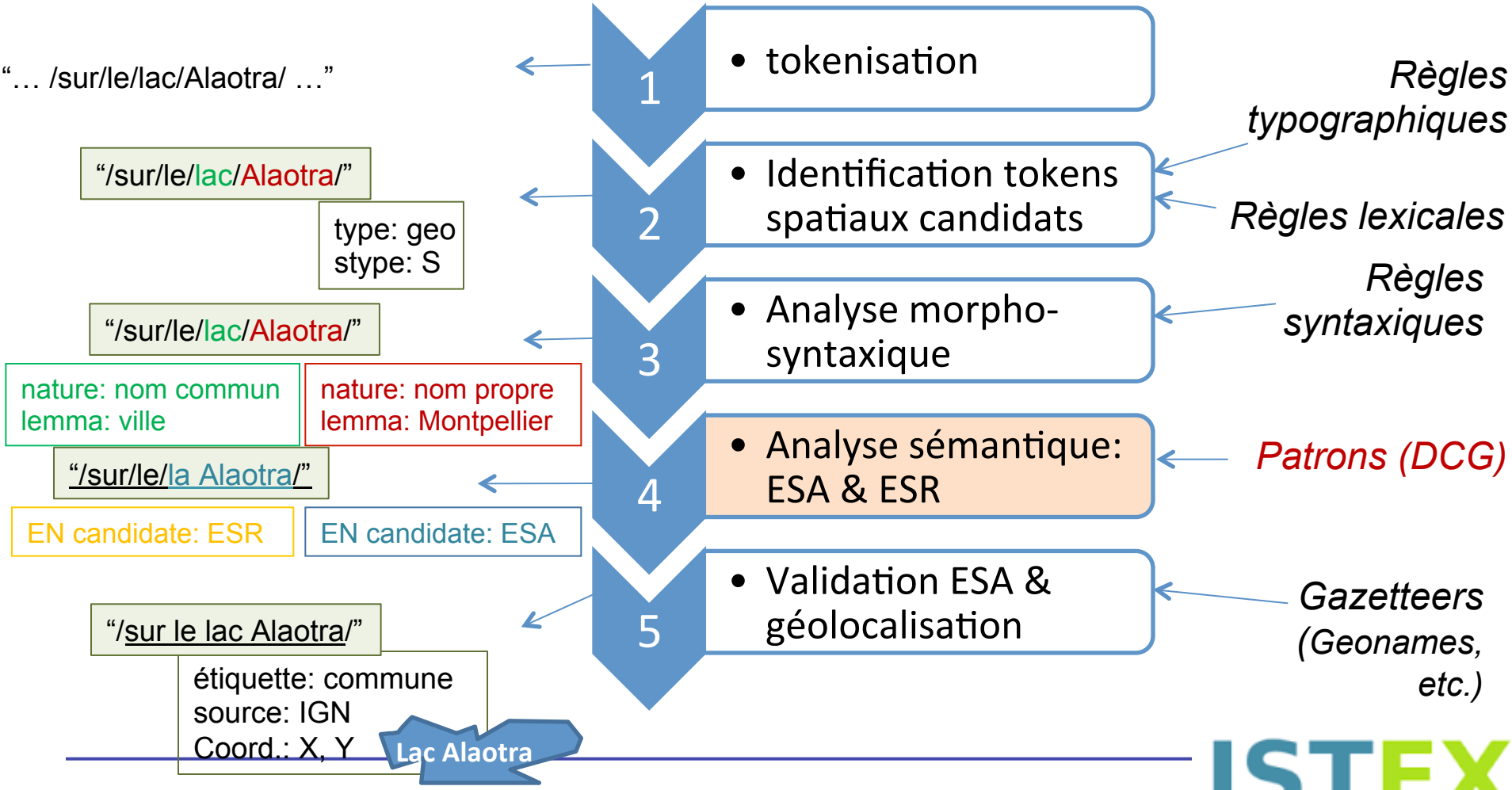
(Tahrat et al., WIMS 2013 ; Kergosien et al., KDIR 2015 ; Zenasni et al., ISMIS 2015)

ISTEX
L'excellence documentaire pour tous

Extraction et localisation des entités spatiales

Patrons linguistiques pour l'extraction d'ES (sur la base des travaux de Lesbegueries et al., 2007)

"… sur le lac Alaotra…"



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Extraction des entités thématiques et temporelles

- 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 de 266 articles scientifiques

Les années **2000** ont été particulièrement animées. Je situe cet événement aux alentours **des années 2000**. Il est né au **début des années 1910**. La chaleur était suffocante à **la fin de l'été 2003**. La dispute a eu lieu bien avant le mois de **mars 1999**. Je crois que la fête a eu lieu à **la fin de l'été**. Nous sommes allés nous coucher tard dans **la soirée** du **mardi**. Il n'a pas cessé de pleuvoir du **01/02/2013** au **21/05/2013**. Le 23, **24** et **25 juin** seront consacrés aux festivités pastorales dans toute la vallée d'Aspe. On dit qu'en **mai**, **juin**, **juillet** et **août**, il ne faut pas manger d'huîtres. **Le début de l'automne** est toujours ensoleillé. Nous partîmes **le 2 juillet 1914**, nous fumes blessés **le 2 août** et rapatriés le 30.

TERRE-ISTEX

Extraction des entités thématiques et temporelles

- 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 : intégration de Agrovoc)
  - Construction d'un monde lexical autour d'entités spatiales (à faire)

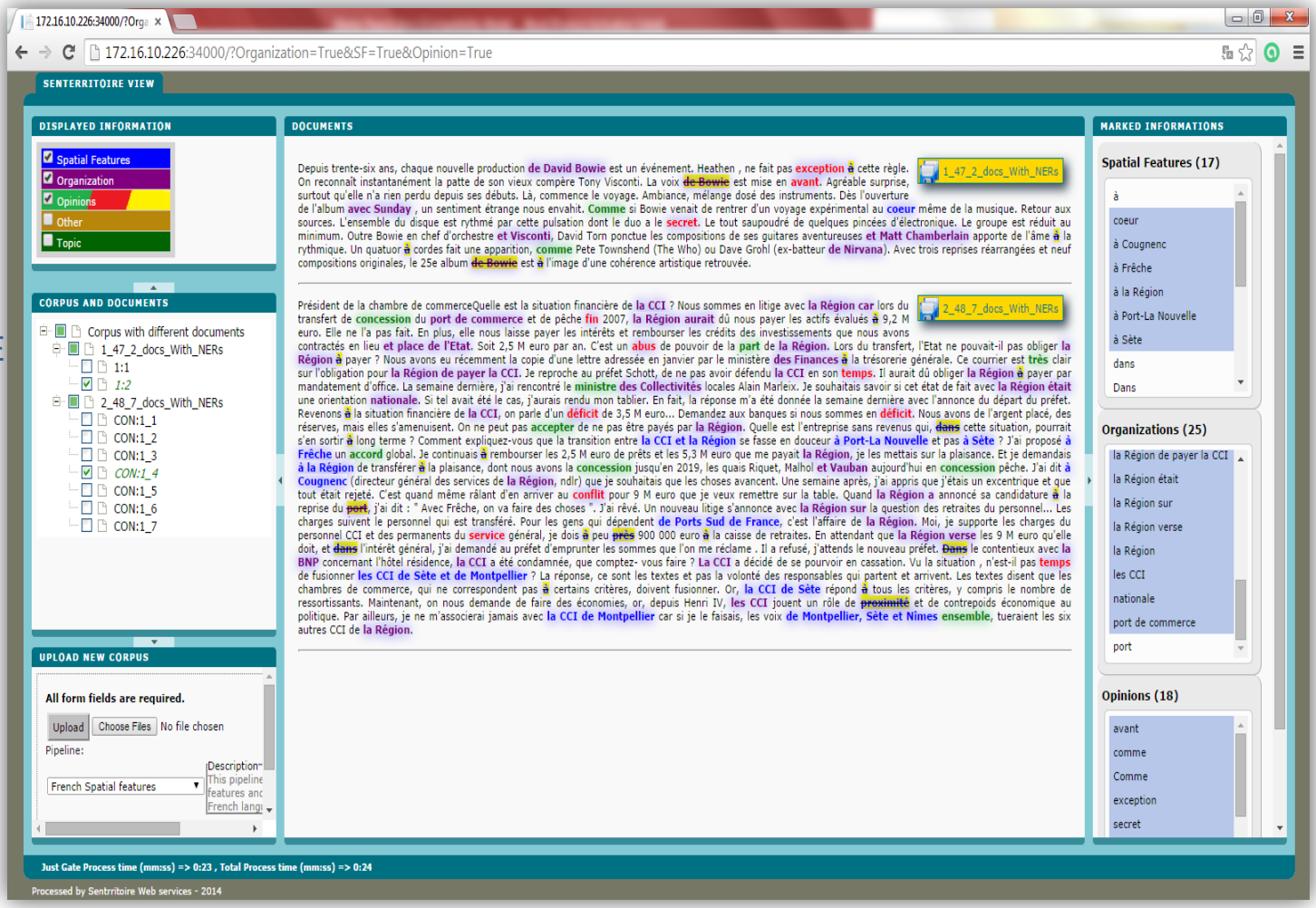
Les grands **programmes** internationaux d'observation des écosystèmes, tels que le Millenium Ecosystem Assessment (Mea), puis Redd (**Réduction** des émissions liées à la déforestation et à la **dégradation des forêts**) et Redd+, préconisent le développement des approches permettant de quantifier et de spatialiser les **services écosystémiques** afin de mettre en oeuvre des pratiques et des politiques de gestion environnementale plus adaptées. La cartographie des **services écosystémiques** apparaît ainsi comme un outil majeur des espaces à forts enjeux environnementaux. Cependant, elle souffre encore de certaines limitations. C'est le cas du stock de carbone dans la biomasse végétale. À l'échelle d'une localité d'**Amazonie** brésilienne de 175 km², cette fonction écologique a été cartographiée avec une résolution spatiale de 30 x 30 m. Afin de quantifier ces **stocks**, des mesures de biomasse arborée et arbustive au sein de 45 " points " et

ISTEX
L'excellence documentaire pour tous

Phase 1 : marquage de contenu
Application Web SISO

Démonstrateur pour l'extraction de contenu et la validation experte (ISWC, octobre 2015, LREC, mai 2018)

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



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Actions à venir

Documents
Série de publications et thèses

Identification des données pertinentes

Contenus et métadonnées :

- Lieux et coordonnées spatiales

- Dates de publication

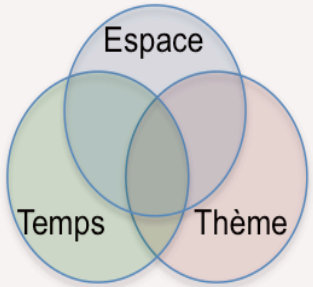
- Thématiques et/
- ou disciplines
- Résumés

Validation des données

elastic kibana

Indexation

Analyse géographique



Recherche d'information

Thème, Temps, Espace, Plein texte

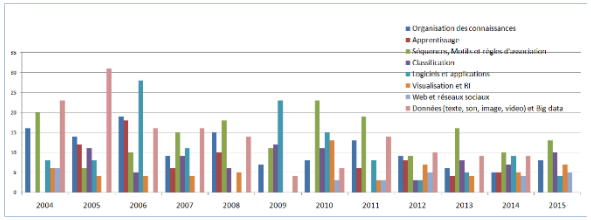
ISTEX
L'excellence documentaire pour tous

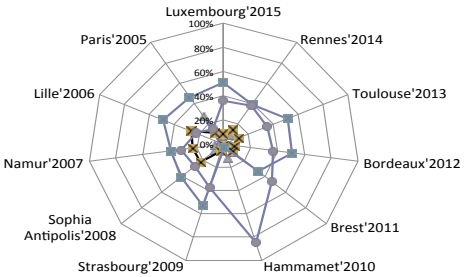
Phase 2 : Prise en main d'Elastic Search et analyse qualitative de données scientifiques

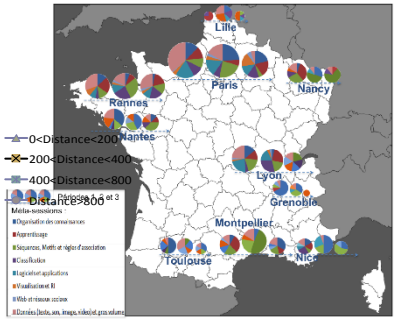
- Analyse géographique de séries de publications : application aux donnés EGC (Kergosien et al., 2017) : indexation, recherche d'information, analyses

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

  - Construction d'un premier index géographique
  - Mise en œuvre d'un moteur de recherche d'information multidimensionnel (Elastic Search)
  - Ensemble d'analyses géographiques disponibles
    http://ekergosien.net/DefiEGC/index.html







Phase 2 : Prise en main d'Elastic Search et analyse qualitative de données scientifiques

ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Et donc?

Bilan intermédiaire

- Modèle de description des données : format MODS enrichi
- Chaînes de traitements linguistiques :
  - Enrichissement des chaînes de marquage des entités spatiales et thématiques pour le passage à l'anglais
  - Prise en compte de la composante temporelle
  - Montée en charge :
    - Intégration des ressources externes (Agrovoc, Geonames),
    - Intégration des 200 000 documents indexés via TERRE-ISTEX dans Elastic Search,
    - Amélioration des modules pour atteindre un temps de traitement de 2 secondes/document,
    - Tests en cours du démonstrateur SISO sur un serveur à Lille : couteux et assez lourd à gérer
- Equipe projet :
  - Consolidation des liens entre les différents acteurs impliqués (publications, évènements organisés en commun, projets à venir en commun)
  - Approche pluridisciplinaire avec des apports des géographes et des SIC.
- Action financée nous permettant de cadrer notre réflexion sur
  - La formalisation des liens entre article scientifique - thèse – données de la recherche
    - Prise en compte de données hétérogènes (thèses, articles scientifiques)
    - Mise en place du projet D4Humanities (coordinateur : J. Schöpfel, http://d4h.meshs.fr)

TERRE-ISTEX

Et donc?

Travaux en cours / perspectives

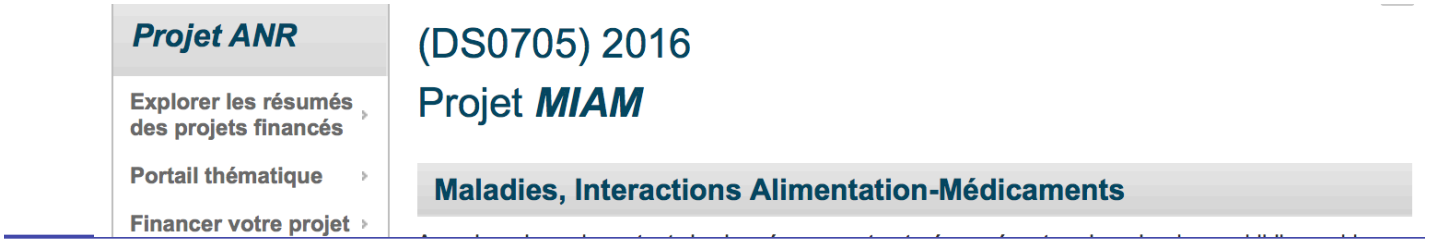
- Mise en place du moteur de recherche géographique dans ElasticSearch : Prochainement sur http://geriico-demo.univ-lille3.fr/siso/

- Analyses quantitatives et qualitatives du cas d'application par des géographes experts du CIRAD
  Appui des membres du projet pour l'utilisation de la couche Kibana.

- Généricité de l'approche sur le thème Library Information Sciences avec les collègues des SIC (GERiiCO) (*Merci Camille et Sabine pour le corpus*)

- Poursuite du projet :

  - Dépôt projet ANR DFG (franco-allemand) : GERiiCO (Lille) , LIUPPA (Pau) , ABES, Universités Oldenburg, Bielefeld et Saarland, etc.
  - Une suite à ISTEX ?....



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Publications

- E. Kergosien, A. Farvardin, M. Teisseire, M.-N. Bessagnet, J. Schöpfel, S. Chaudiron, B. Jacquemin, A. Lacayrelle, M. Roche, C. Sallaberry, J.P. Tonneau. Automatic Identification of Study Fields in Scientific Corpus, In the 11th Edition of its Language Resources and Evaluation Conference (LREC), pp. 4, Japan, may 2018
- E. Kergosien, M. Teisseire, M.-N. Bessagnet, J. Schöpfel, Amin Farvardin, Identification des terrains d'études dans les corpus scientifique, Numéro spécial Document numérique "Analyser la science : les bibliothèques numériques comme objet de recherche », à paraître 2018.
- M.-N. Bessagnet, E. Kergosien, M. Farvardin, A. Le Parc - Lacayrelle et C. Sallaberry, A propos des territoires dans les corpus scientifiques, Atelier sur l'Extraction et la Modélisation de Connaissances à partir de textes scientifiques, 28es Journées francophones d'Ingénierie des Connaissances, Caen (France), juillet 2017
- E. Kergosien, C. Sallaberry, M.-N. Bessagnet, A. Le Parc - Lacayrelle, S. Chaudiron, Using a GIR tool in a Business Intelligence Context: the case of EGC conferences, In *7th. International Conference on Information Systems and Economic Intelligence (SIIE)*, pp. 12, Al Hoceima (Maroc), may, 2017
- A. Le Parc - Lacayrelle, A. Farvardin, TERRE-ISTEX : vers un modèle pour identifier des terrains d'études, In Atelier Valorisation et Analyse des Données de la Recherche (VADOR), conférence Inforsid, Toulouse (France), mai 2017
- J. Schöpfel, E. Kergosien, S. Chaudiron and B. Jacquemin, Dissertations as Data, In *ETD2016 19th International Symposium on Electronic Theses and Dissertations*, 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

ISTEX
L'excellence documentaire pour tous

Communications / Valorisation

Communications

- E. Kergosien, M. Teisseire, M.-N. Bessagnet, J. Schöpfel, Amin Farvardin, Identification des terrains d'études dans les corpus scientifique, In 85e congrès de l'ACFAS, colloque #605 Analyser la science : les bibliothèques numériques comme objet de recherche, Montréal (Canada), Mai 2017
- E. Kergosien, 2017, Identification et analyse des terrains d'études dans les corpus ISTEX, conférencier invité journées Carrefour de l'IST (CARIST 2017), mars 2017, Nancy
- J. Schöpfel, E. Kergosien, H. Prost, « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse, In Atelier Valorisation et Analyse des Données de la Recherche (VADOR), conférence Inforsid, Toulouse (France), mai 2017
- M. Roche, Le projet TERRE-ISTEX, « Two Minutes of Madness » conférence EGC, Grenoble, janvier 2017
- 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
- J. Schöpfel, E. Kergosien. Le projet TERRE-ISTEX pour l'identification et l'analyse des terrains d'études dans les corpus ISTEX, Journée Archives ouvertes et bases de publications : exploration et analyse des sources de données pour la recherche et ses environnements. Paris, mai 2016.
  https://data4ist.sciencesconf.org/program.

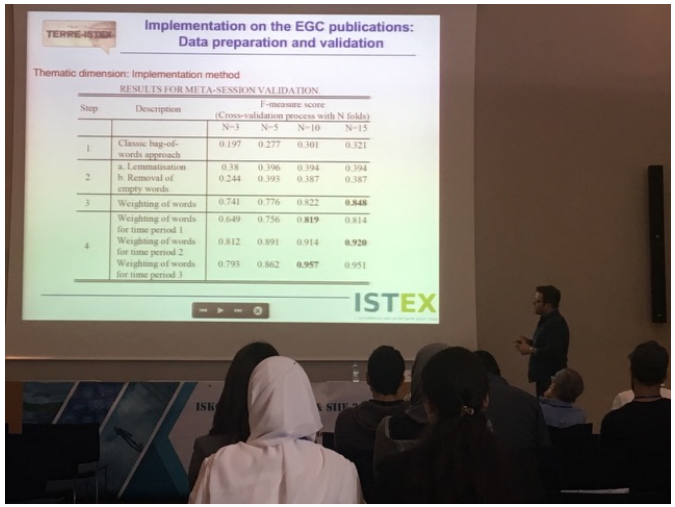
Evènements organisés

- Atelier Valorisation et Analyse des Données de la Recherche (VADOR), conférence Inforsid, Toulouse (France), mai 2017, atelier à venir (CORIA – TALN – RJC 2018)
- Journée d'études « Valorisation et Gestion des données de la recherche », Pau, mars 2017
- 19th International Symposium on Electronic Theses and Dissertations (ETD), Lille, July 2016
- Séminaire doctoral sur les données de la recherche, Université de Lille (2015 – 2018)

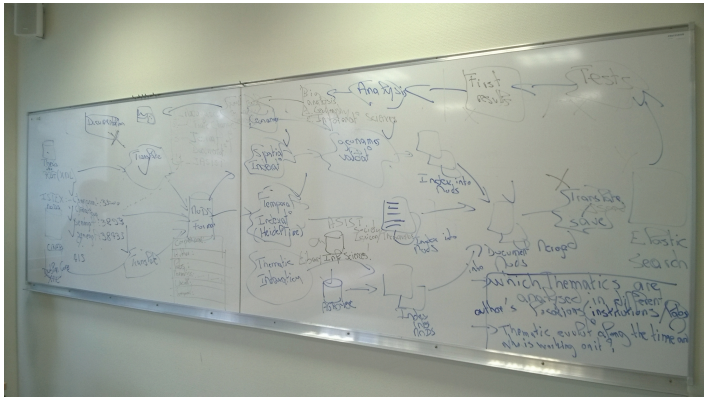
ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Questions



https://terreistex.hypotheses.org



VADOR

ATELIER VADOR - 15 MAI 2018, RENNES. CONFÉRENCES CORIA-TALN-RJC 2018, invité Pascal Cuxac (INIST).

https://vador.sciencesconf.org

ISTEX
L'excellence documentaire pour tous

Extraction et localisation des entités spatiales

- Evaluation
  - 10 documents en français

	ESA, ESR (TERRE-ISTEX)	ESA, ESR (CASEN)
Précision	100%	93%
Rappel	90%	77%
F-Mesure	94,7%	84,2%

  - 10 documents en anglais

	ESA, ESR (TERRE-ISTEX)	ESA, ESR (CASEN)
Précision	90%	94%
Rappel	60%	53,3%
F-Mesure	72,%	68%

  - Travail en cours sur une évaluation de 600 articles (300 francais et 300 anglais)

TERRE-ISTEX

Etudier tout ce qu'il se passe sur un territoire : Temps de traitement

Evaluation de la chaîne complète :

Donnons quelques statistiques d'exécution de notre chaîne de traitements pour un corpus de 8500 documents. Les performances du système sont estimées à :

- Annotation des entités temporelles : 8196 secondes
- Annotation des entités Agrovoc: 1606 s
- Recherche des concepts et concepts liés d'Agrovoc (Ressource Agrovoc offline) : 36 secondes. (En utilisant le web service Agrovoc, ce temps peut être augmenté de 3 à 5 s par corpus)
- Annotation des entités spatiales (français et anglais): 4940 s
- Génération vers le format choisi pour créer les index (JSON) : 55 s

Le processus prend un temps total de 16105 s, soit 1.9 s par document, ce qui est très encourageant.