



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

*séminaire technique « Chantiers d'usage » d'ISTEX
6 et 7 juin 2017, INIST Nancy*

Gérialco

STL
savoirs
langage
extes



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



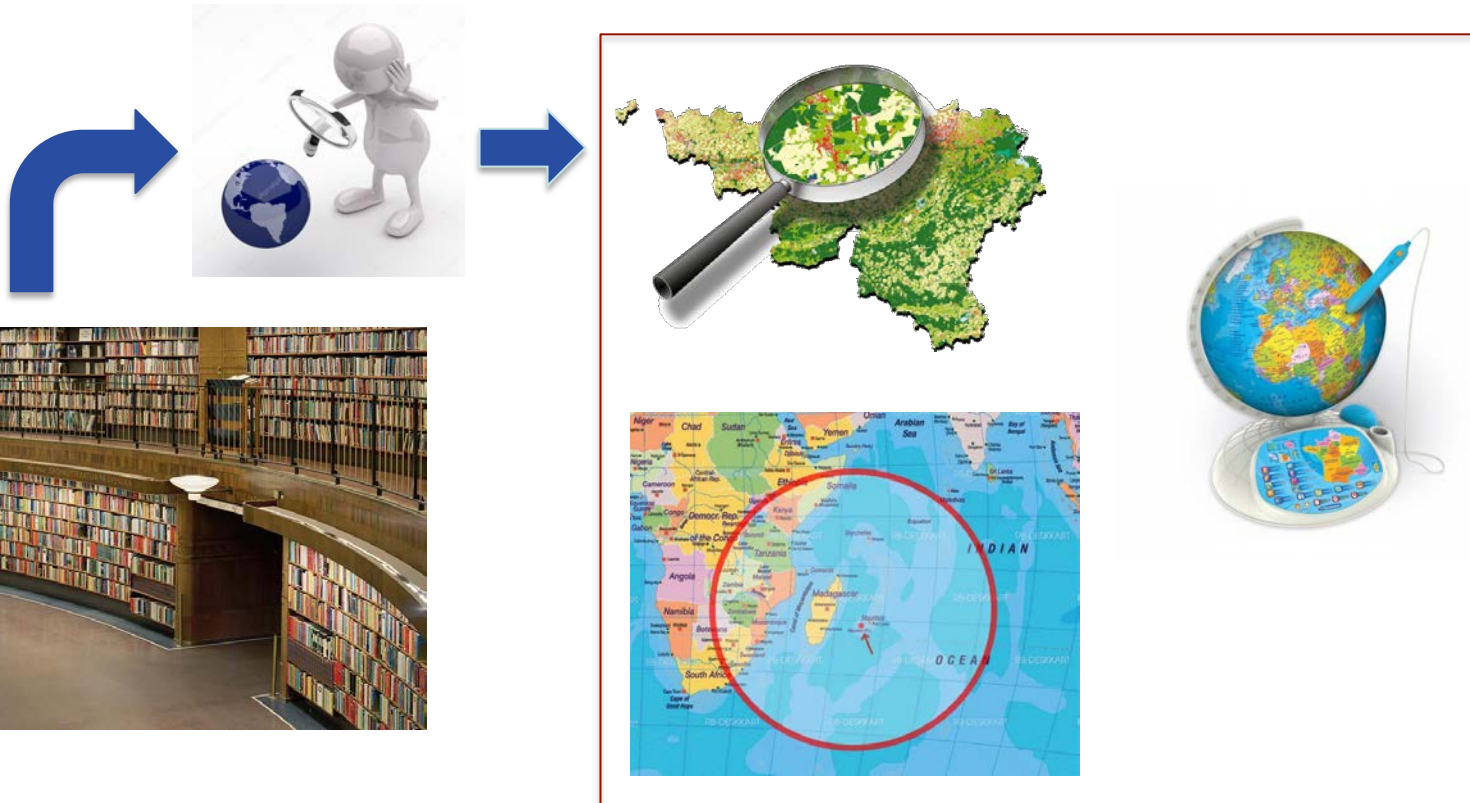
- Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour – LIUPPA, Pau
- Chercheurs impliqués : Marie Noëlle Bessagnet (MCF), Annig Le Parc-Lacayrelle (MCF), Christian Sallaberry (MCF, HDR)



- 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).

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)



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 :

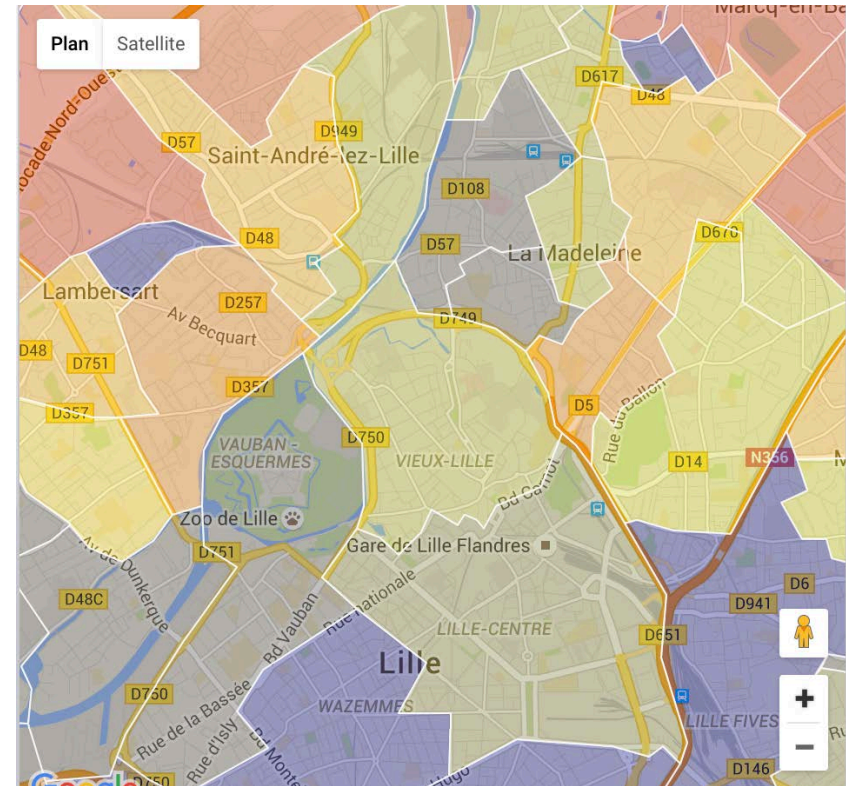
- Quels sont les documents qui font mention d'un territoire?
 - De ces documents, quelles sont les périodes et thématiques mentionnées
 - Visualisations spatiales
 - Mobilisation des experts pour l'analyse des résultats

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

Exemple pour l'entité spatiale absolue

« le quartier Vieux-Lille »

ESA



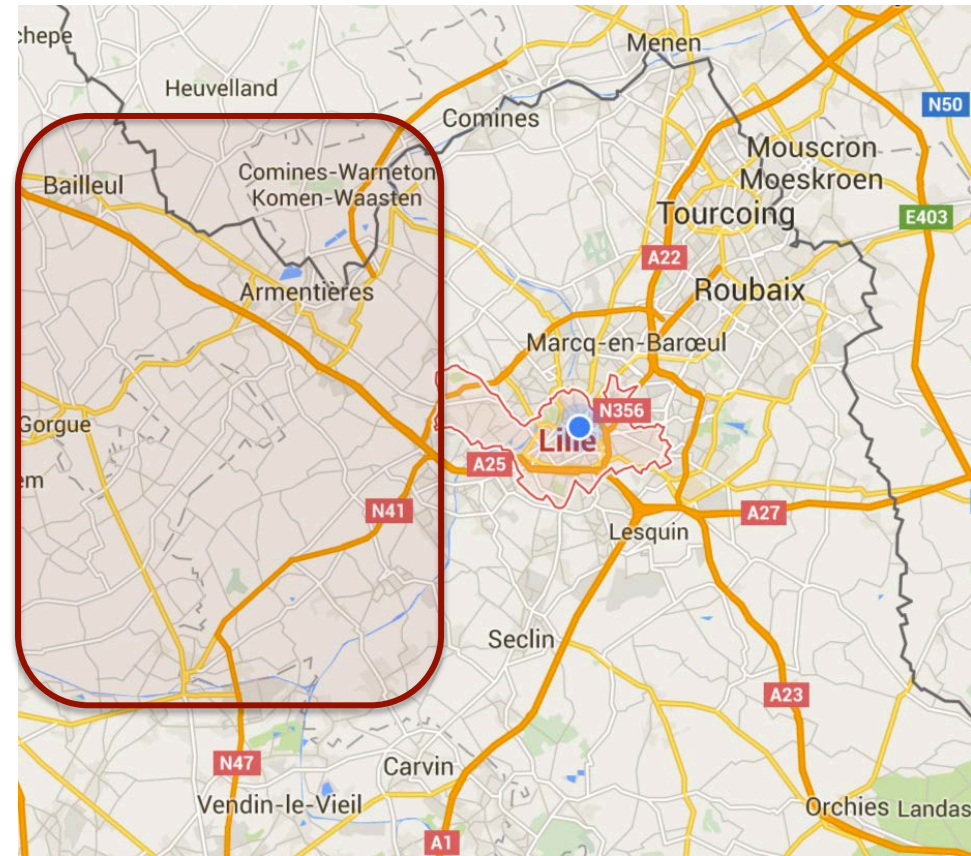
Etudier tout ce qu'il se passe sur un territoire : Usages

Exemples pour l'entité spatiale relative

« à l'ouest de la ville de Lille »

Relation:
orientation

ESA



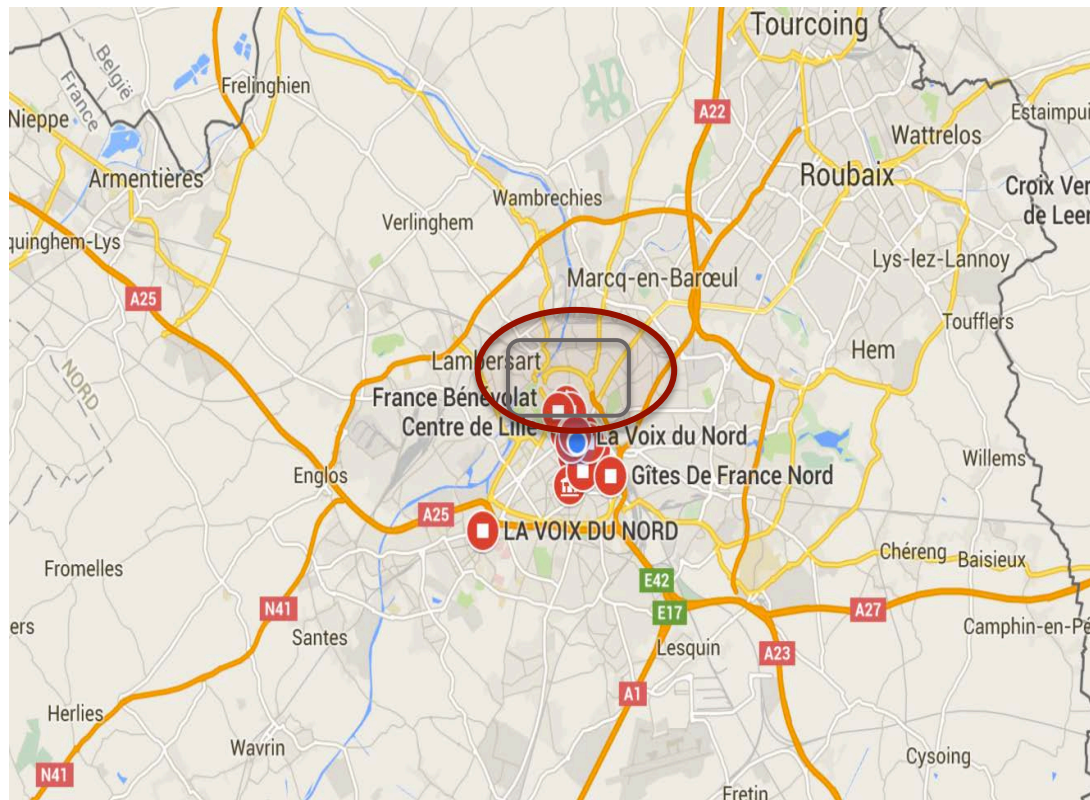
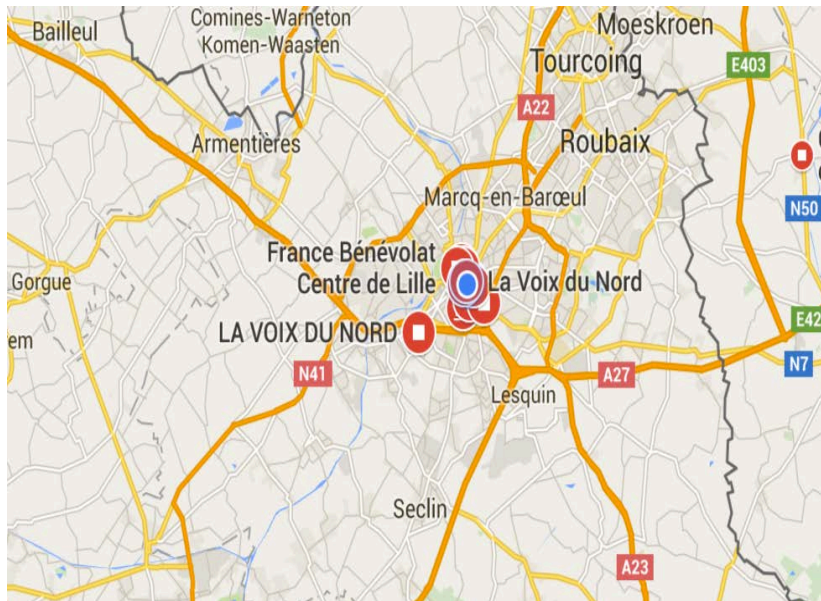
Etudier tout ce qu'il se passe sur un territoire : Usages

Exemples pour l'entité spatiale relative

« dans les environs du nord de Lille »

Relation:
adjacence

ESR





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

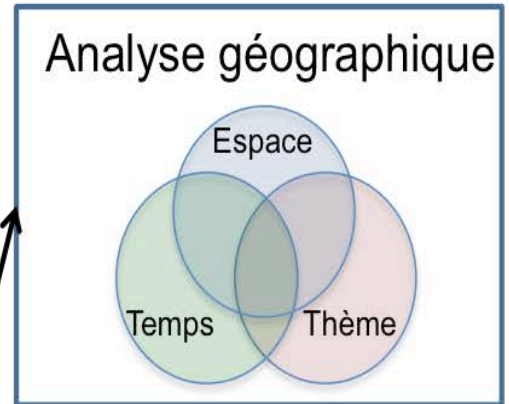
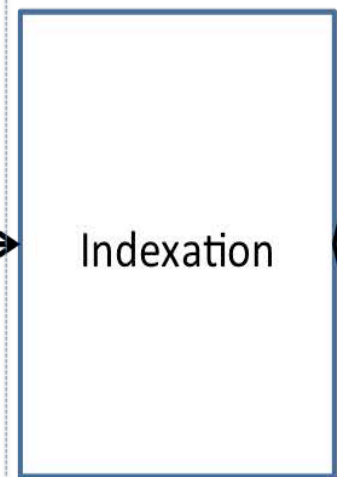
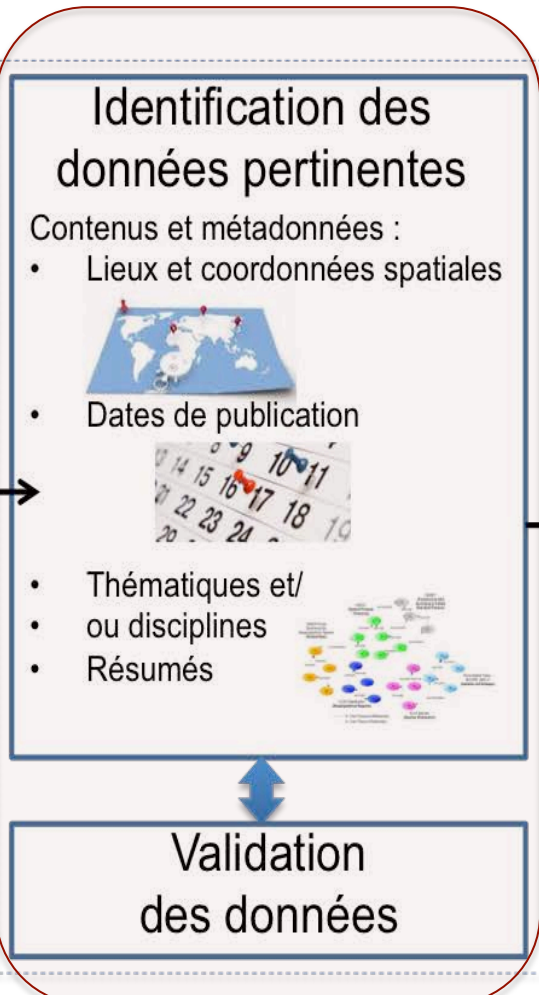
Thème, Temps, Espace, Plein texte



elastic



Documents
Série de publications
et thèses



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

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

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

Données CIRAD

- Données issues d'Agrotrop : 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://agrotrop.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)

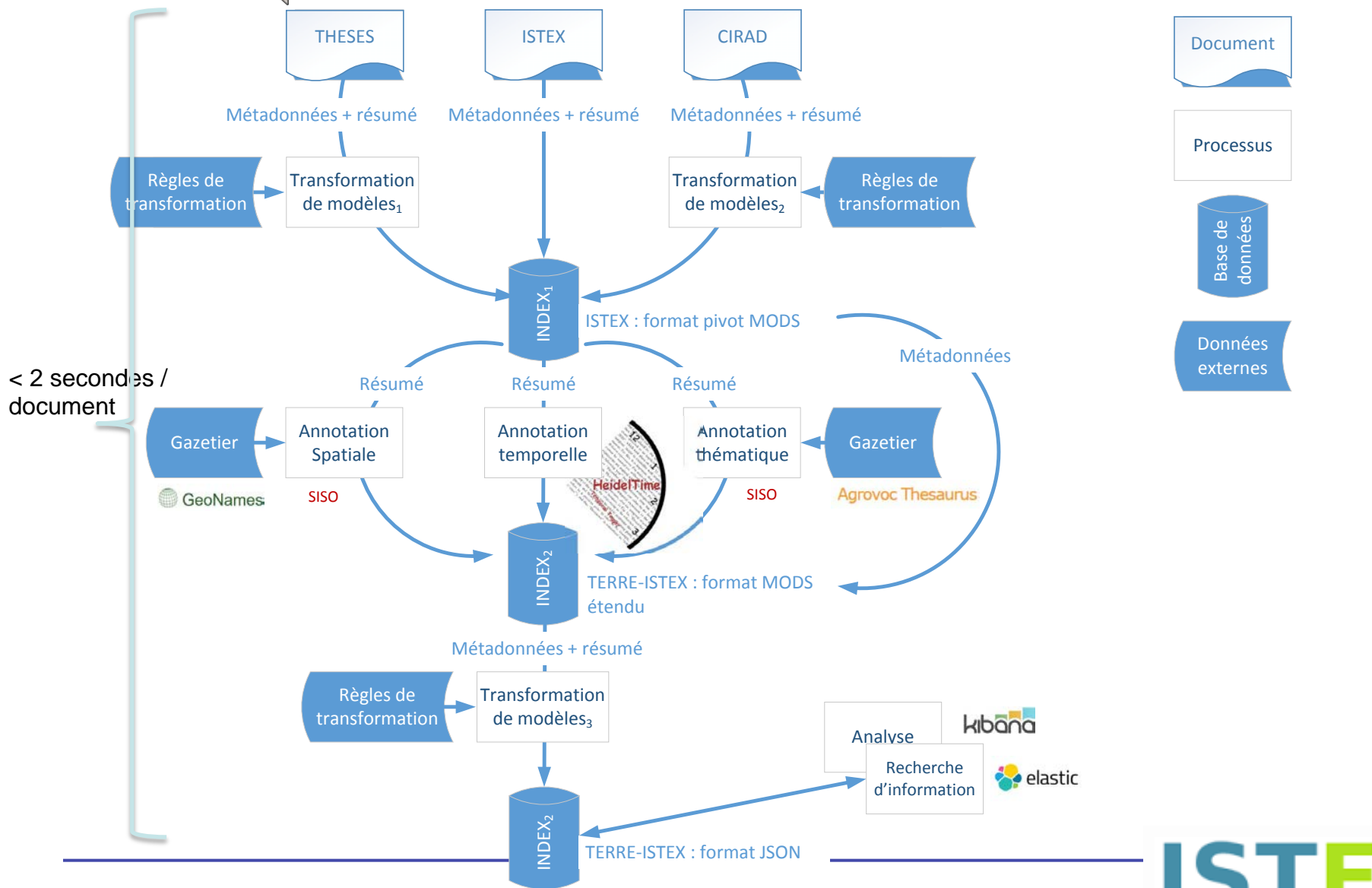
- 200 000 thèses, métadonnées internes, lien SUDOC (ABES).
- 70 000 thèses numérisées
- Notices ABES ? Thésaurus RAMEAU
- Liens avec la thématique du projet :
 - 400 thèses sur la thématique changement climatique (somme thèses.fr et ANRT)

Données ISTEX à partir des requêtes par mots clés suivantes :

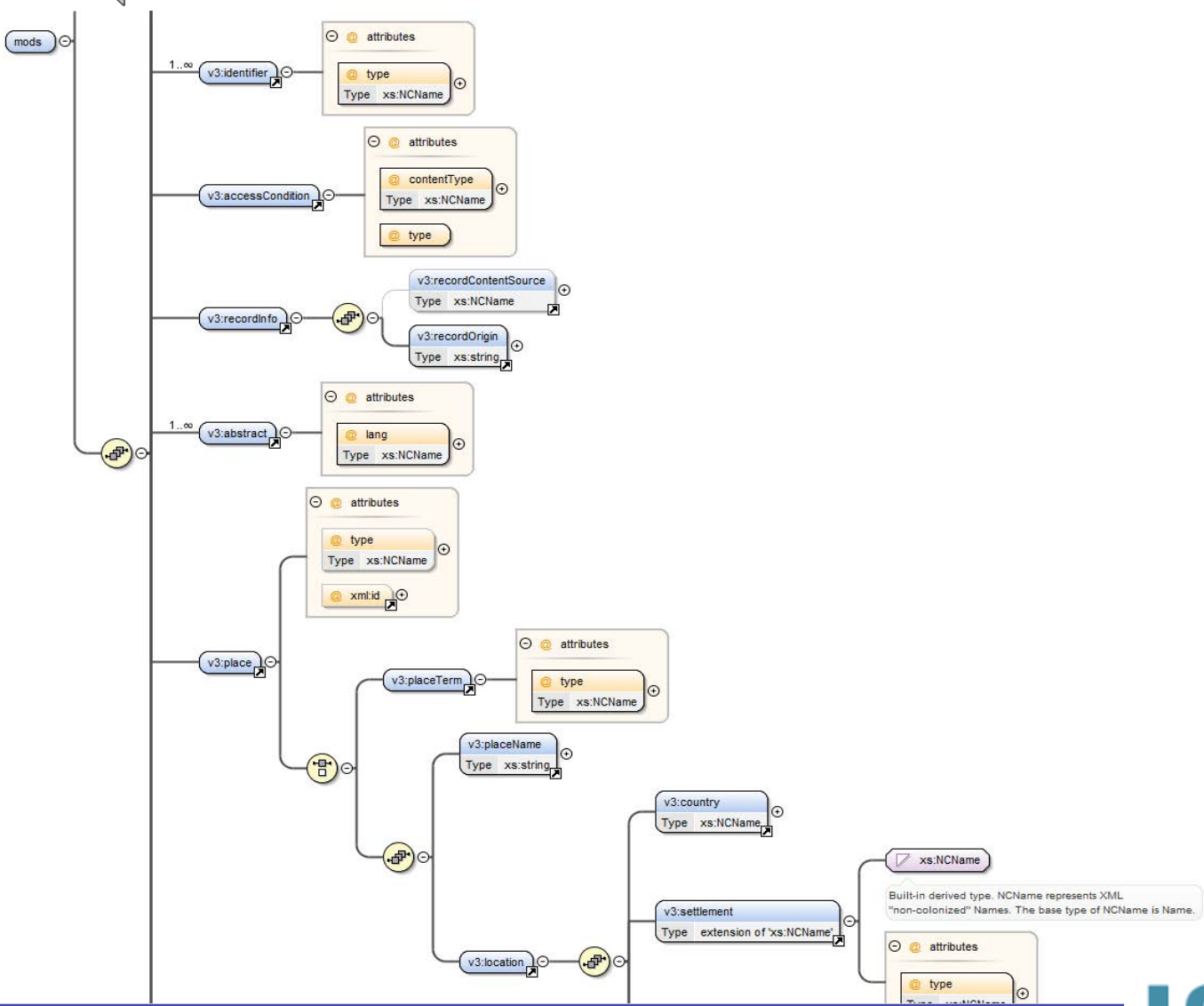
- « Climate Change » et « Changement climatique » : 85 800 documents
- « Senegal » et « Sénégal » : 43 293 documents
- « Madagascar » : 41 142 documents

Phase de marquage de contenu (Fouille de textes)

TERRE-ISTEX

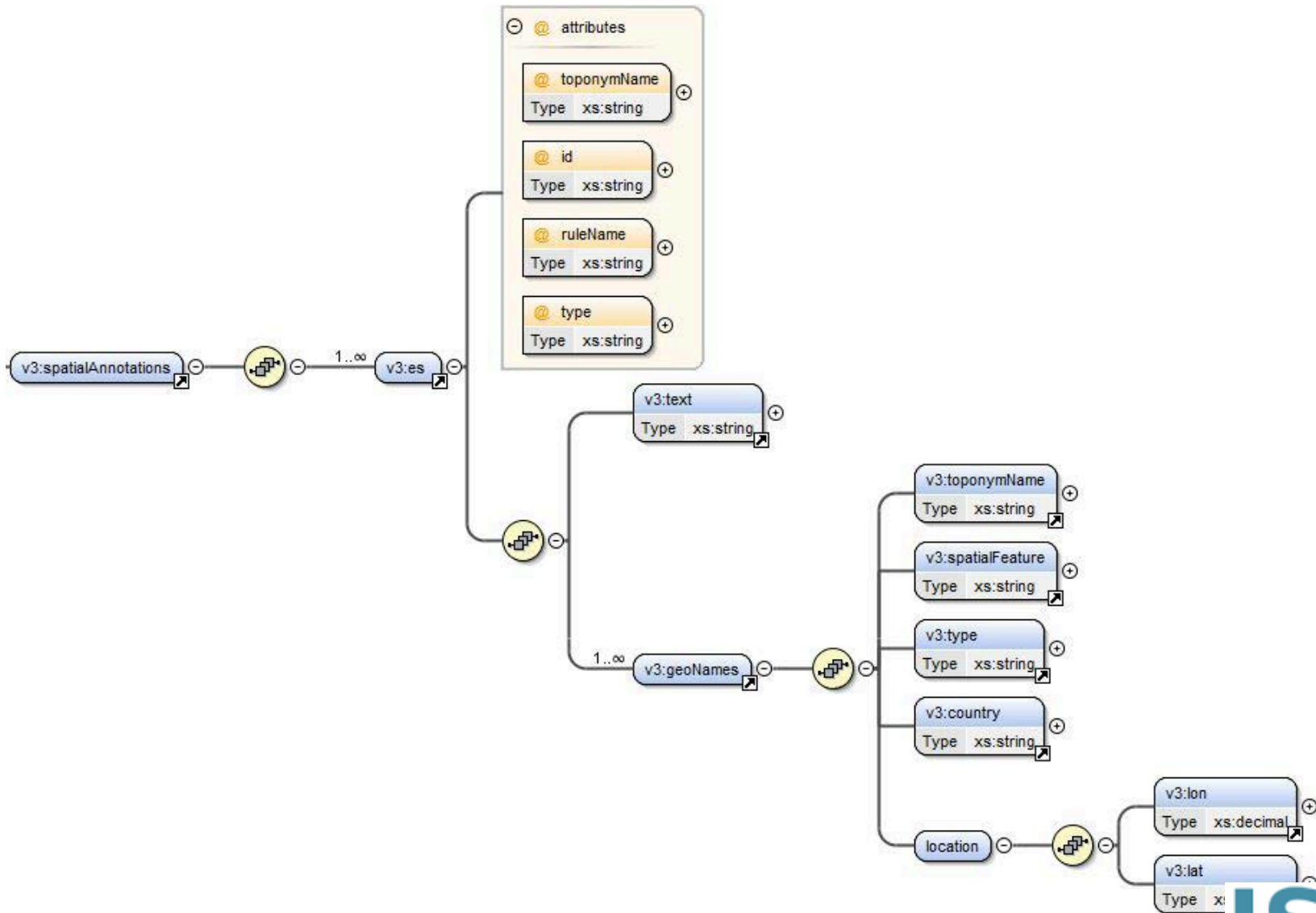


Modélisation des descripteurs, une contribution importante

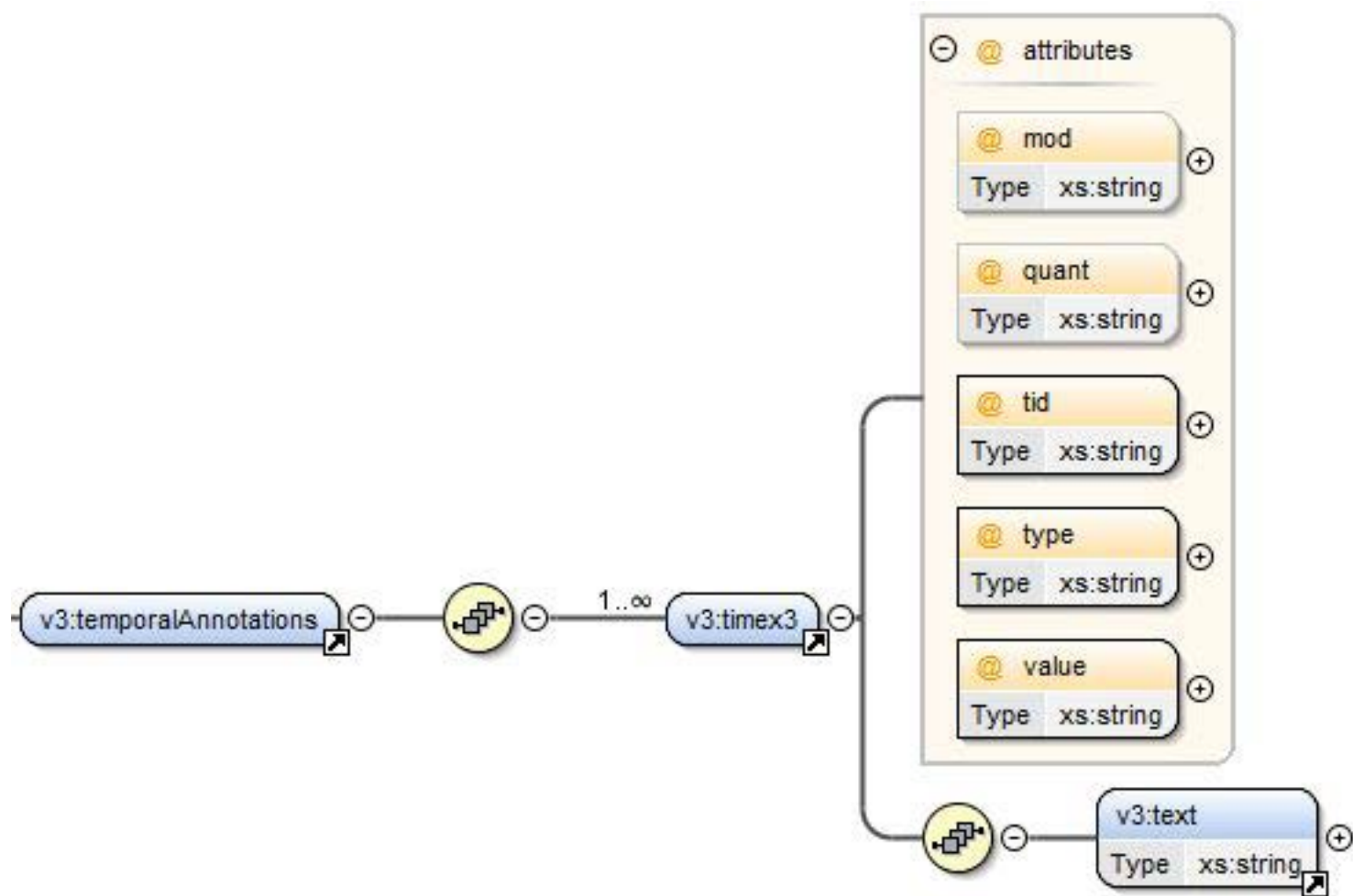


Modélisation du descripteur SpatialAnnotations

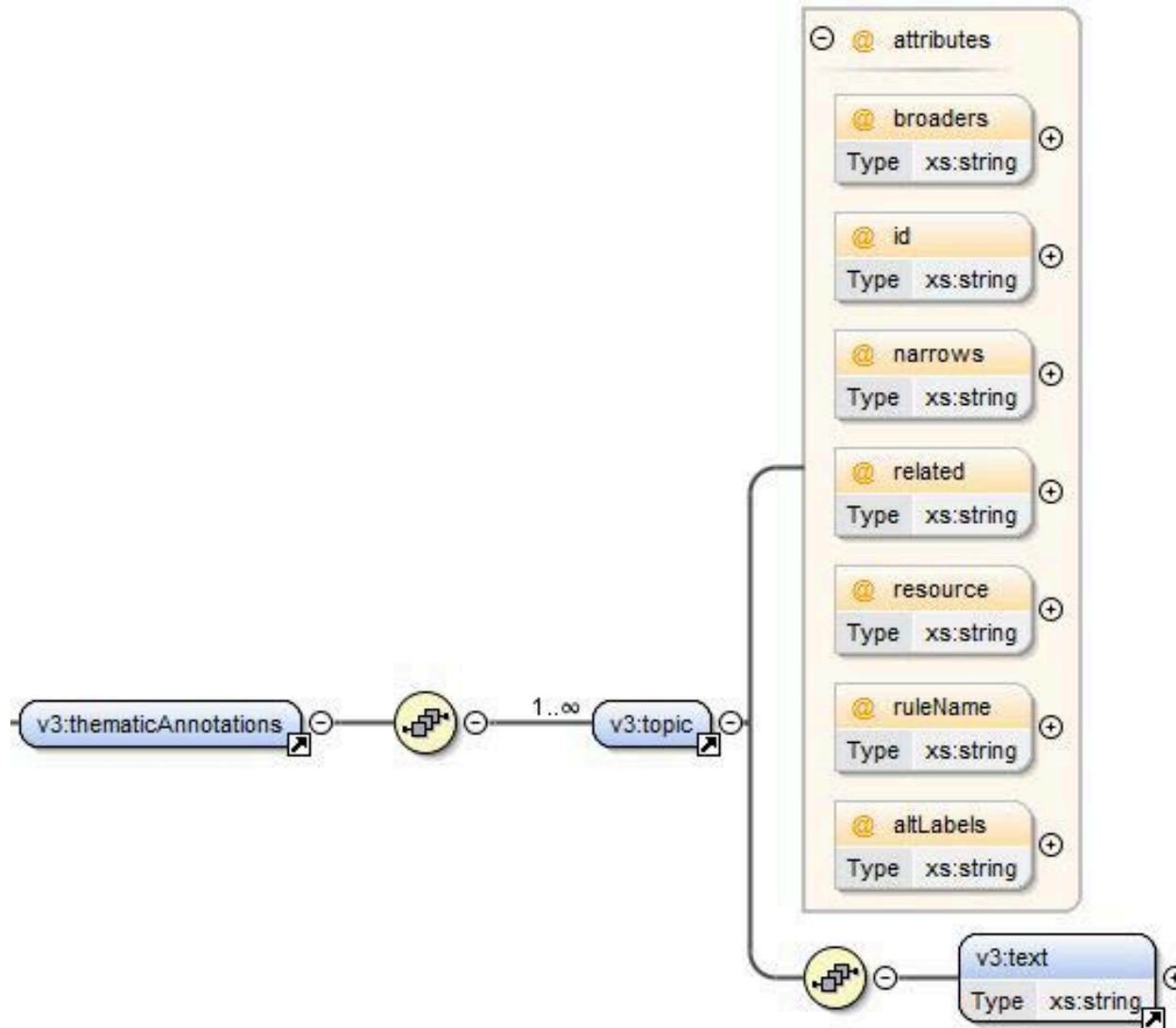
TERRE-ISTEX



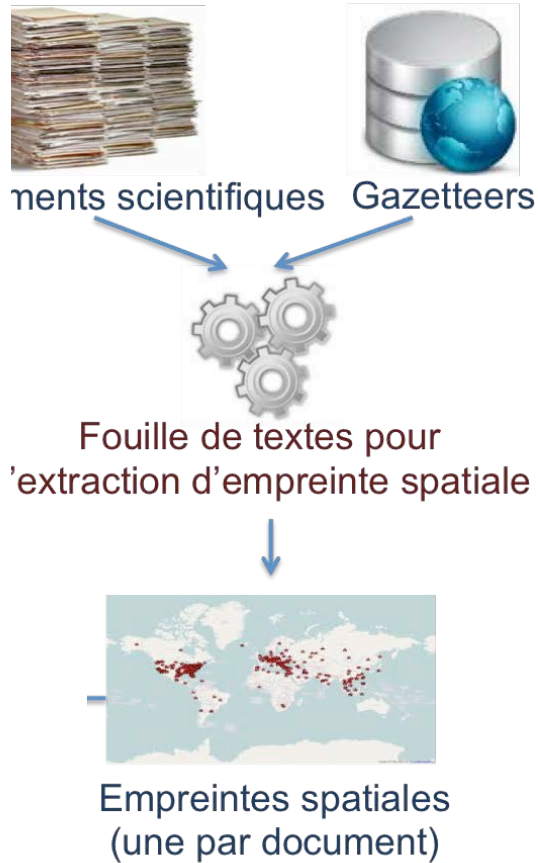
Modélisation du descripteur temporalAnnotations



Modélisation du descripteur ThematicAnnotations



Extraction et localisation des entités spatiales

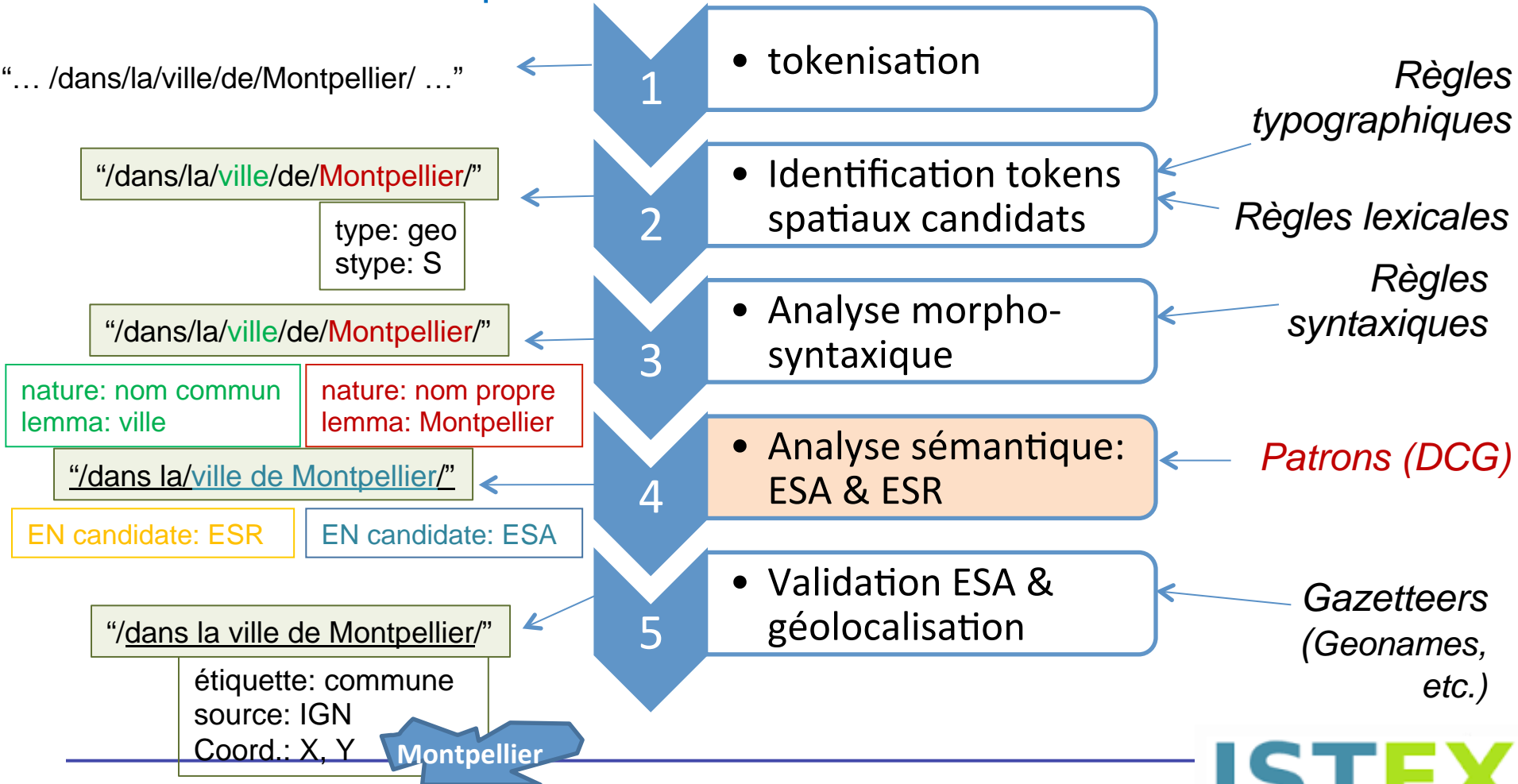


*(Tahrat et al., WIMS 2013 ;
Kergosien et al., KDIR 2015 ;
Zenasni et al., ISMIS 2015)*

Extraction et localisation des entités spatiales

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

“... dans la ville de Montpellier...”



- **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%

- 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.

- **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

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

SENTERRITOIRE VIEW

DISPLAYED INFORMATION

- Spatial Features
- Organization
- Opinions
- Other
- Topic

CORPUS AND DOCUMENTS

- Corpus with different documents
 - 1_47_2_docs_With_NERs
 - 1:1
 - 1:2
 - 2_48_7_docs_With_NERs
 - CON:1_1
 - CON:1_2
 - CON:1_3
 - CON:1_4
 - CON:1_5
 - CON:1_6
 - CON:1_7

DOCUMENTS

Depuis trente-six ans, chaque nouvelle production de David Bowie est un événement. Heathen , ne fait pas exception à cette règle. On reconnaît instantanément la patte de son vieux compère Tony Visconti. La voix de Bowie est mise en avant. Agréable surprise, surtout qu'elle n'a rien perdu depuis ses débuts. Là, commence le voyage. Ambiance, mélange dosé des instruments. Dès l'ouverture de l'album avec Sunday , un sentiment étrange nous envahit. Comme si Bowie venait de rentrer d'un voyage expérimental au cœur même de la musique. Retour aux sources. L'ensemble du disque est rythmé par cette pulsation dont le duo a le secret. Le tout saupoudré de quelques pinces d'électronique. Le groupe est réduit au minimum. Outre Bowie en chef d'orchestre et Visconti, David Torn ponctue les compositions de ses guitares aventureuses et Matt Chamberlain apporte de l'âme à la rythmique. Un quatuor à cordes fait une apparition, comme Pete Townshend (The Who) ou Dave Grohl (ex-batteur de Nirvana). Avec trois reprises réarrangées et neuf compositions originales, le 25e album de Bowie est à l'image d'une cohérence artistique retrouvée.

Président de la chambre de commerce Quelle est la situation financière de la CCI ? Nous sommes en litige avec la Région car lors du transfert de concession du port de commerce et de pêche fin 2007, la Région aurait dû nous payer les actifs évalués à 9,2 M euro. Elle ne l'a pas fait. En plus, elle nous laisse payer les intérêts et rembourser les crédits des investissements que nous avons contractés en lieu et place de l'Etat. Soit 2,5 M euro par an. C'est un abus de pouvoir de la part de la Région. Lors du transfert, l'Etat ne pouvait-il pas obliger la Région à payer ? Nous avons eu récemment la copie d'une lettre adressée en janvier par le ministère des Finances à la trésorerie générale. Ce courrier est très clair sur l'obligation pour la Région de payer la CCI. Je reproche au préfet Schott, de ne pas avoir défendu la CCI en son temps. Il aurait dû obliger la Région à payer par mandatement d'office. La semaine dernière, j'ai rencontré le ministre des Collectivités locales Alain Marleix. Je souhaitais savoir si cet état de fait avec la Région était une orientation nationale. Si tel avait été le cas, j'aurais rendu mon tablier. En fait, la réponse m'a été donnée la semaine dernière avec l'annonce du départ du préfet. Revenons à la situation financière de la CCI, on parle d'un déficit de 3,5 M euro... Demandez aux banques si nous sommes en déficit. Nous avons de l'argent placé, des réserves, mais elles s'amenuisent. On ne peut pas accepter de ne pas être payés par la Région. Quelle est l'entreprise sans revenus qui, dans cette situation, pourrait s'en sortir à long terme ? Comment expliquez-vous que la transition entre la CCI et la Région se fasse en douceur à Port-La Nouvelle et pas à Sète ? J'ai proposé à Frêche un accord global. Je continuais à rembourser les 2,5 M euro de prêts et les 5,3 M euro que me payait la Région, je les mettais sur la plaisance. Et je demandais à la Région de transférer à la plaisance, dont nous avons la concession jusqu'en 2019, les quais Riquet, Malhol et Vauban aujourd'hui en concession pêche. J'ai dit à Cougnenc (directeur général des services de la Région, ndr) que je souhaitais que les choses avancent. Une semaine après, j'ai appris que j'étais un excentrique et que tout était rejeté. C'est quand même révoltant d'en arriver au conflit pour 9 M euro que je veux remettre sur la table. Quand la Région a annoncé sa candidature à la reprise du port, j'ai dit : " Avec Frêche, on va faire des choses ". J'ai rêvé. Un nouveau litige s'annonce avec la Région sur la question des retraites du personnel... Les charges suivent le personnel qui est transféré. Pour les gens qui dépendent de Ports Sud de France, c'est l'affaire de la Région. Moi, je supporte les charges du personnel CCI et des permanents du service général, je dois à peu près 900 000 euro à la caisse de retraites. En attendant que la Région verse les 9 M euro qu'elle doit, et dans l'intérêt général, j'ai demandé au préfet d'emprunter les sommes que l'on me réclame. Il a refusé, j'attends le nouveau préfet. Dans le contentieux avec la BNP concernant l'hôtel résidence, la CCI a été condamnée, que comptez-vous faire ? La CCI a décidé de se pourvoir en cassation. Vu la situation, n'est-il pas temps de fusionner les CCI de Sète et de Montpellier ? La réponse, ce sont les textes et pas la volonté des responsables qui parlent et arrivent. Les textes disent que les chambres de commerce, qui ne correspondent pas à certains critères, doivent fusionner. Or, la CCI de Sète répond à tous les critères, y compris le nombre de ressortissants. Maintenant, on nous demande de faire des économies, or, depuis Henri IV, les CCI jouent un rôle de proximité et de contrepois économique au politique. Par ailleurs, je ne m'associerai jamais avec la CCI de Montpellier car si je le faisais, les voix de Montpellier, Sète et Nîmes ensemble, tueraient les six autres CCI de la Région.

MARKED INFORMATION

Spatial Features (17)

- à
- cœur
- à Cougnenc
- à Frêche
- à la Région
- à Port-La Nouvelle
- à Sète
- dans
- Dans

Organizations (25)

- la Région de payer la CCI
- la Région était
- la Région sur
- la Région verse
- la Région
- les CCI
- nationale
- port de commerce
- port

Opinions (18)

- avant
- comme
- Comme
- exception
- secret

UPLOAD NEW CORPUS

All form fields are required.

Upload No file chosen

Pipeline:

French Spatial features Description: This pipeline features anc French lang

Just Gate Process time (mmiss) => 0:23 , Total Process time (mmiss) => 0:24

Processed by Sentritoire Web services - 2014

Prochainement sur
<http://geriico-demo.univ-lille3.fr/siso/>

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.

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



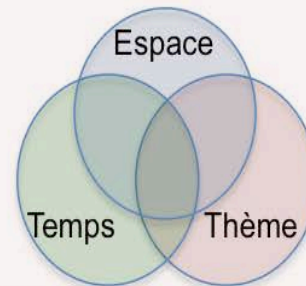
Indexation



elastic

kibana

Analyse géographique



Recherche d'information

Thème, Temps, Espace, Plein texte



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

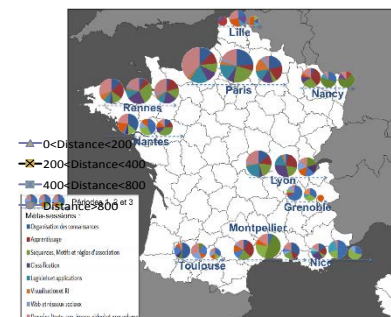
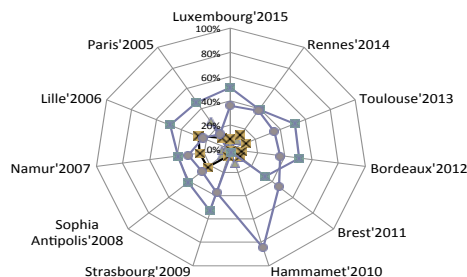
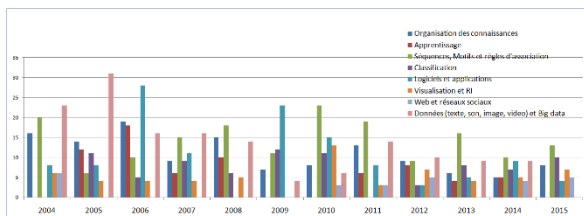
- Construction d'un premier index géographique

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)

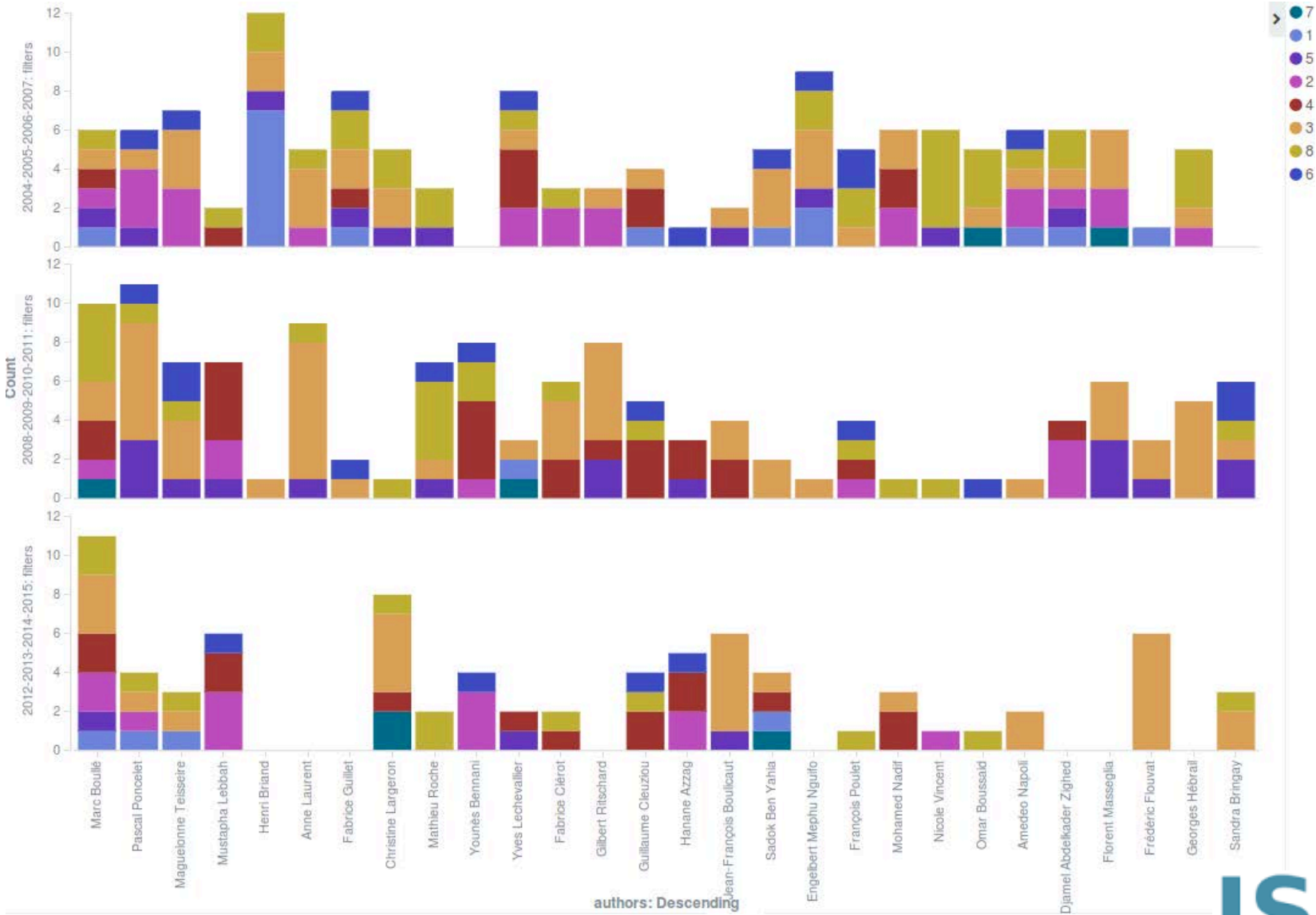
- Ensemble d'analyses géographiques disponibles

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



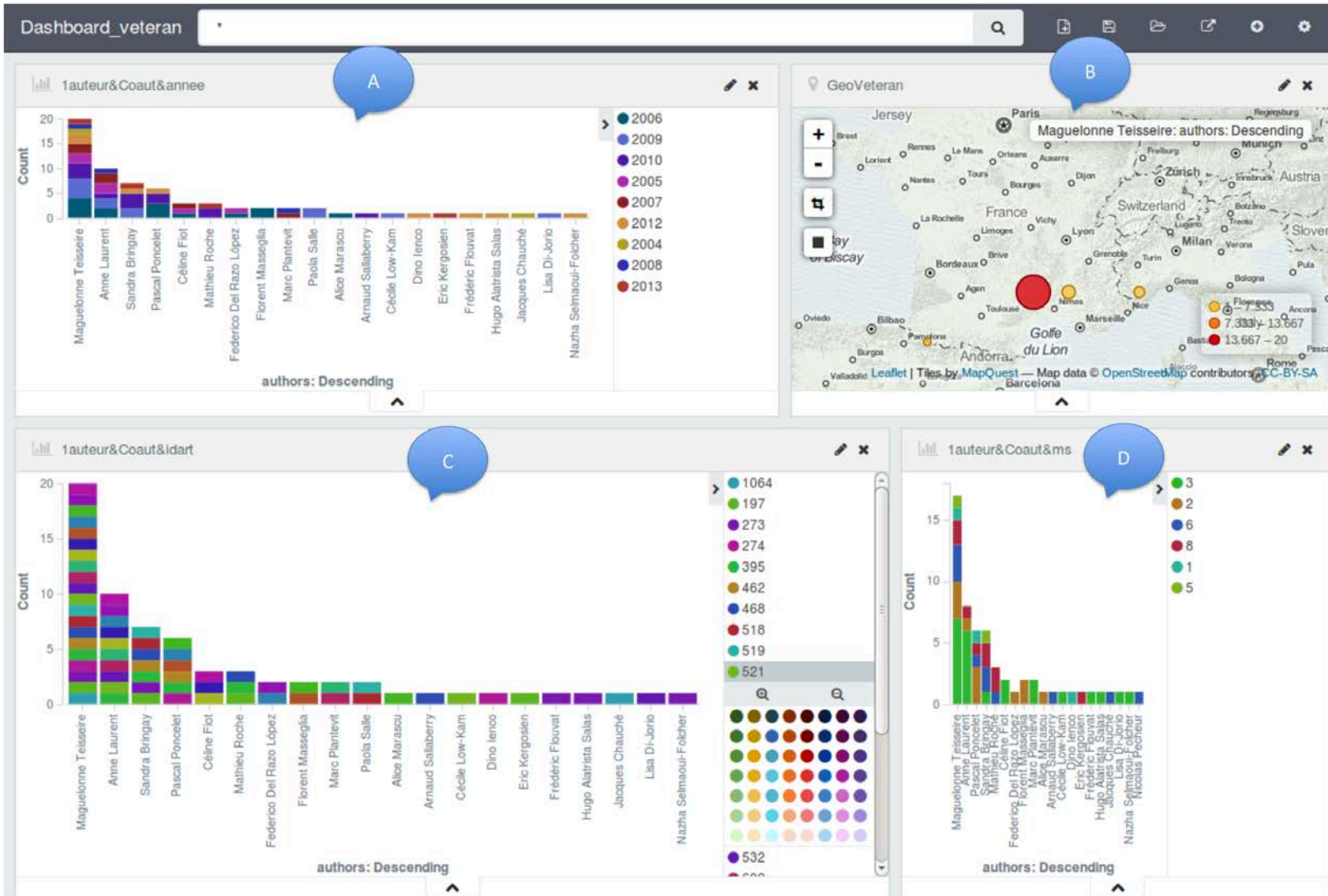


Prise en main d'Elastic Search et analyse qualitative de données scientifiques





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



- 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), à venir juillet 2017
- 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 venir décembre 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 (SIIIE)*, 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

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 ISTEEX, 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 ISTEEX, 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
- 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 (2015 – 2017)

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)
 - Amélioration des modules pour atteindre un temps de traitement de 2 secondes/document,
 - Amélioration du module d'indexation, de correction et d'upload des résultats d'annotation
 - Mise en place d'un serveur pour le démonstrateur SISO et le moteur de recherche ElasticSearch
 - 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 du concept donnée de la recherche
 - 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>)

Travaux en cours / perspectives

- Intégration des données dans ElasticSearch
- 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 (GERiCO) (*Merci Camille et Sabine pour le corpus*)
- Poursuite du projet (dans quel cadre)?!



Questions

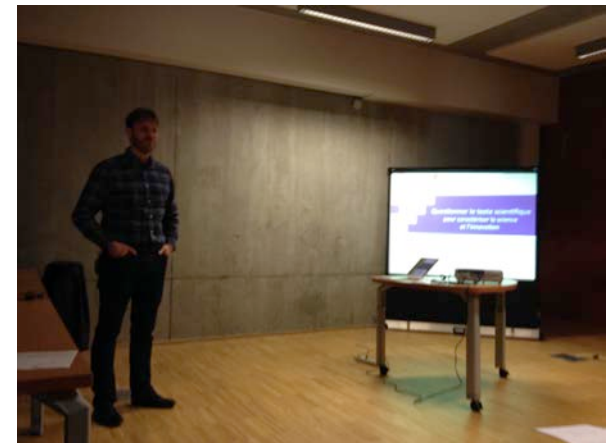
Implementation on the EGC publications:
Data preparation and validation

Thematic dimension: Implementation method

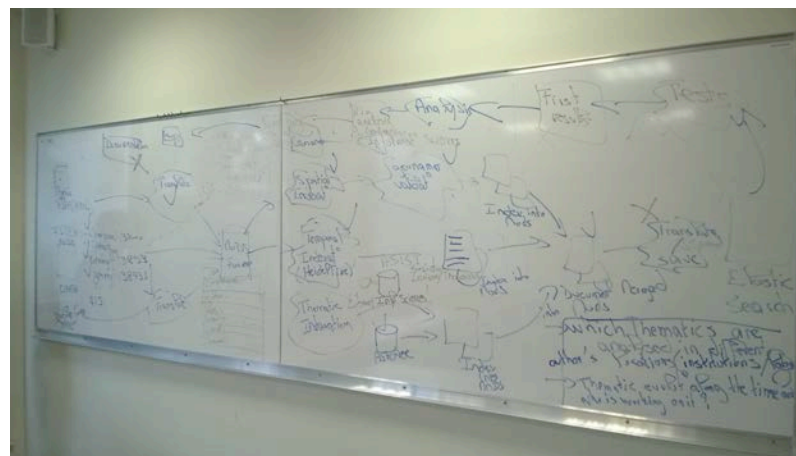
RESULTS FOR META-SESSION VALIDATION

Step	Description	F-measure score (Cross-validation process with N folds)			
		N=3	N=5	N=10	N=15
1	Classic bag-of-words approach	0.197	0.277	0.301	0.321
2	a. Lemmatization	0.38	0.396	0.394	0.394
	b. Removal of empty words	0.244	0.393	0.387	0.387
3	Weighting of words	0.741	0.776	0.822	0.848
4	Weighting of words for time period 1	0.649	0.736	0.819	0.814
	Weighting of words for time period 2	0.812	0.891	0.914	0.920
	Weighting of words for time period 3	0.793	0.862	0.957	0.951

ISTEX



<https://terreistex.hypotheses.org>



<https://vador.sciencesconf.org>