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, Maguelonne Teisseire, Marie-Noëlle Bessagnet, Joachim Schöpfel

Colloque « Analyser la science : les bibliothèques numériques comme objet de recherche »
85ème congrès de l'ACFAS
Montréal, 8-9 mai 2017

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)

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

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

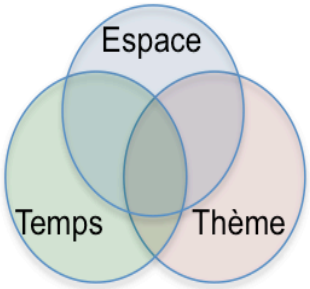
- Dates de publication

- Thématiques et/
- ou disciplines
- Résumés

Validation des données

Indexation

Analyse géographique



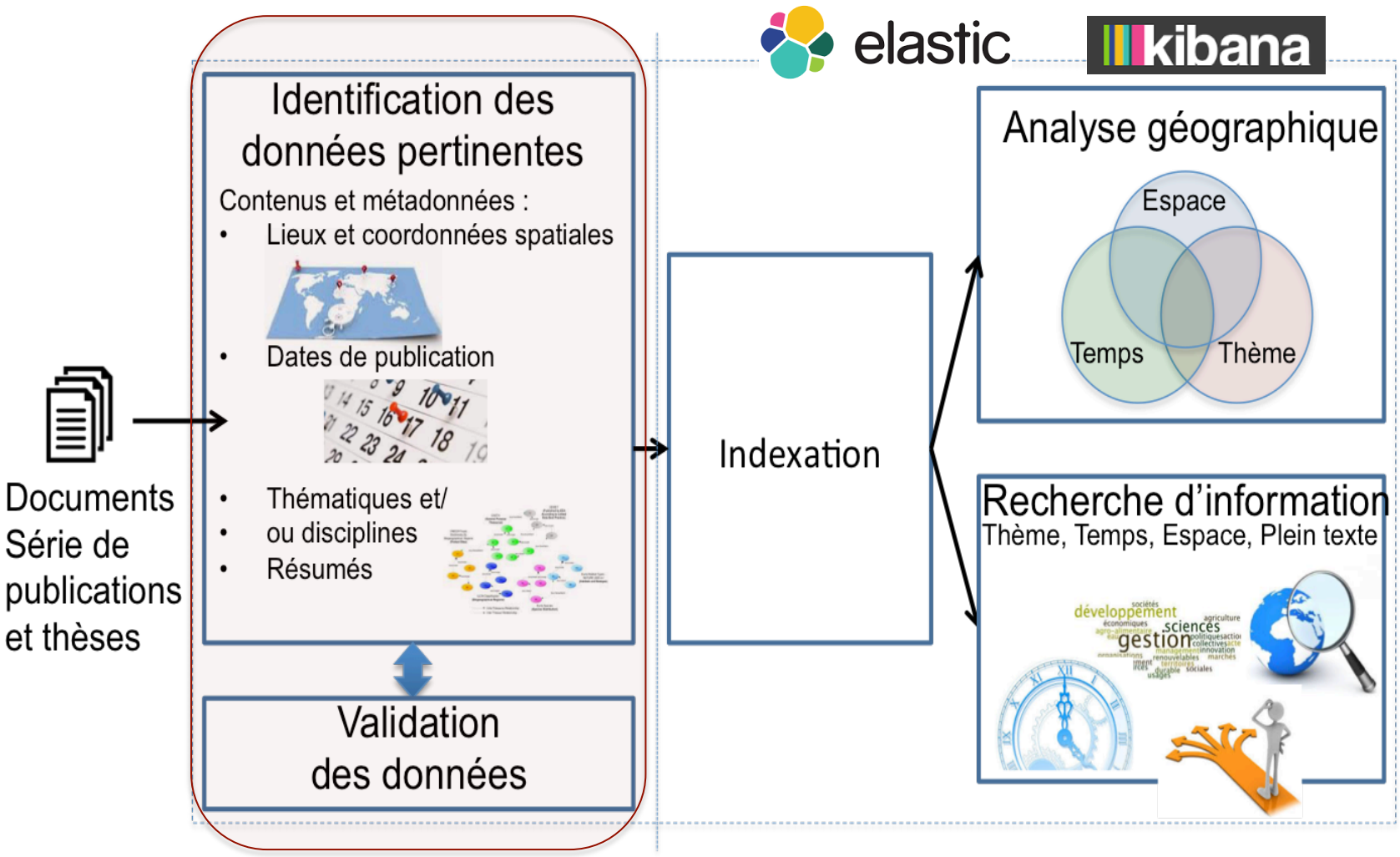
Recherche d'information

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

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.

Corpus hétérogène multilingue (français – anglais)

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

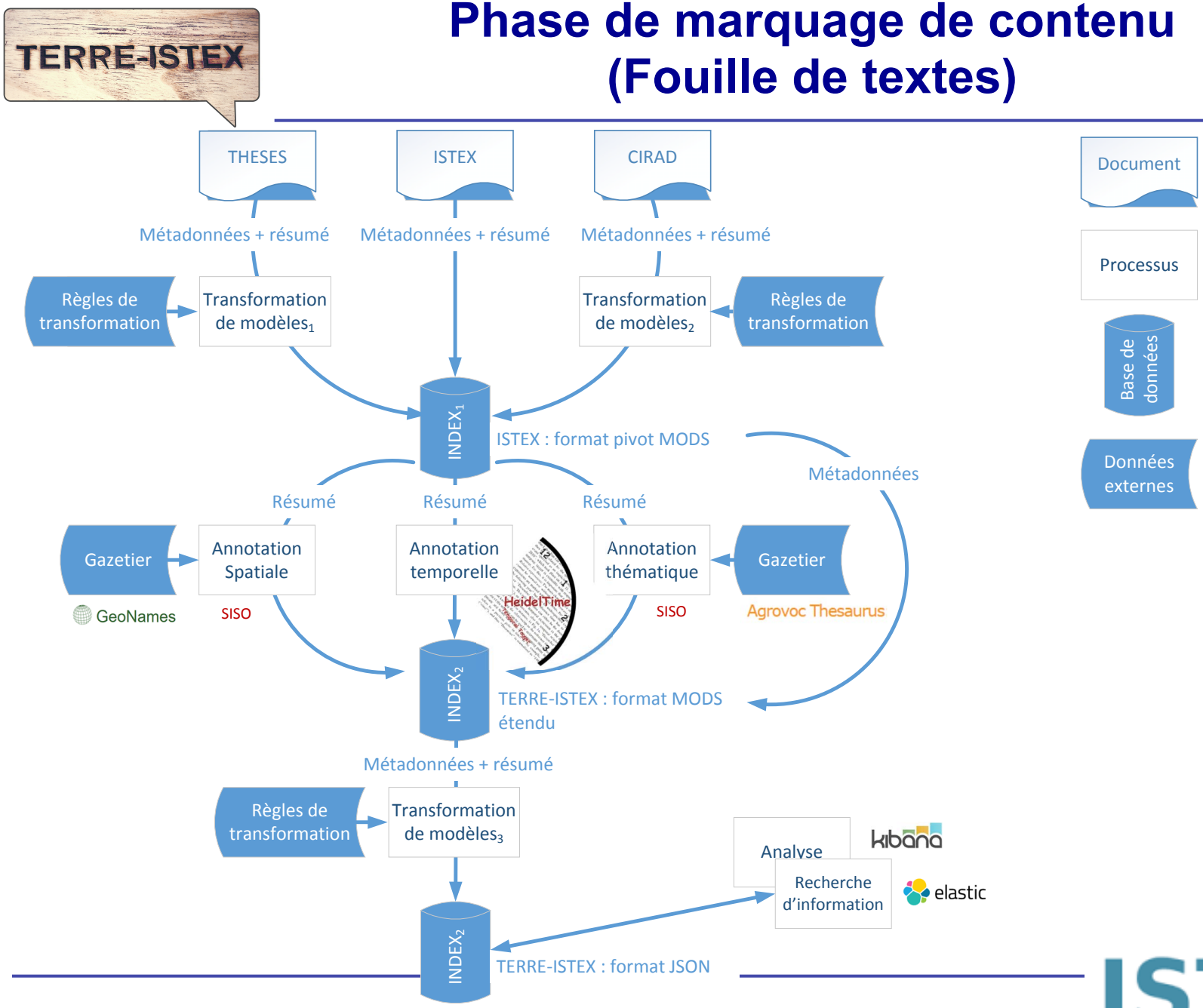
Données Thèses (ANRT)

- 200 000 thèses, métadonnées internes, lien SUDOC (ABES).
- 70 000 thèses numérisées
- Notices ABES ? Thésaurus RAMEAU
- Liens avec 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

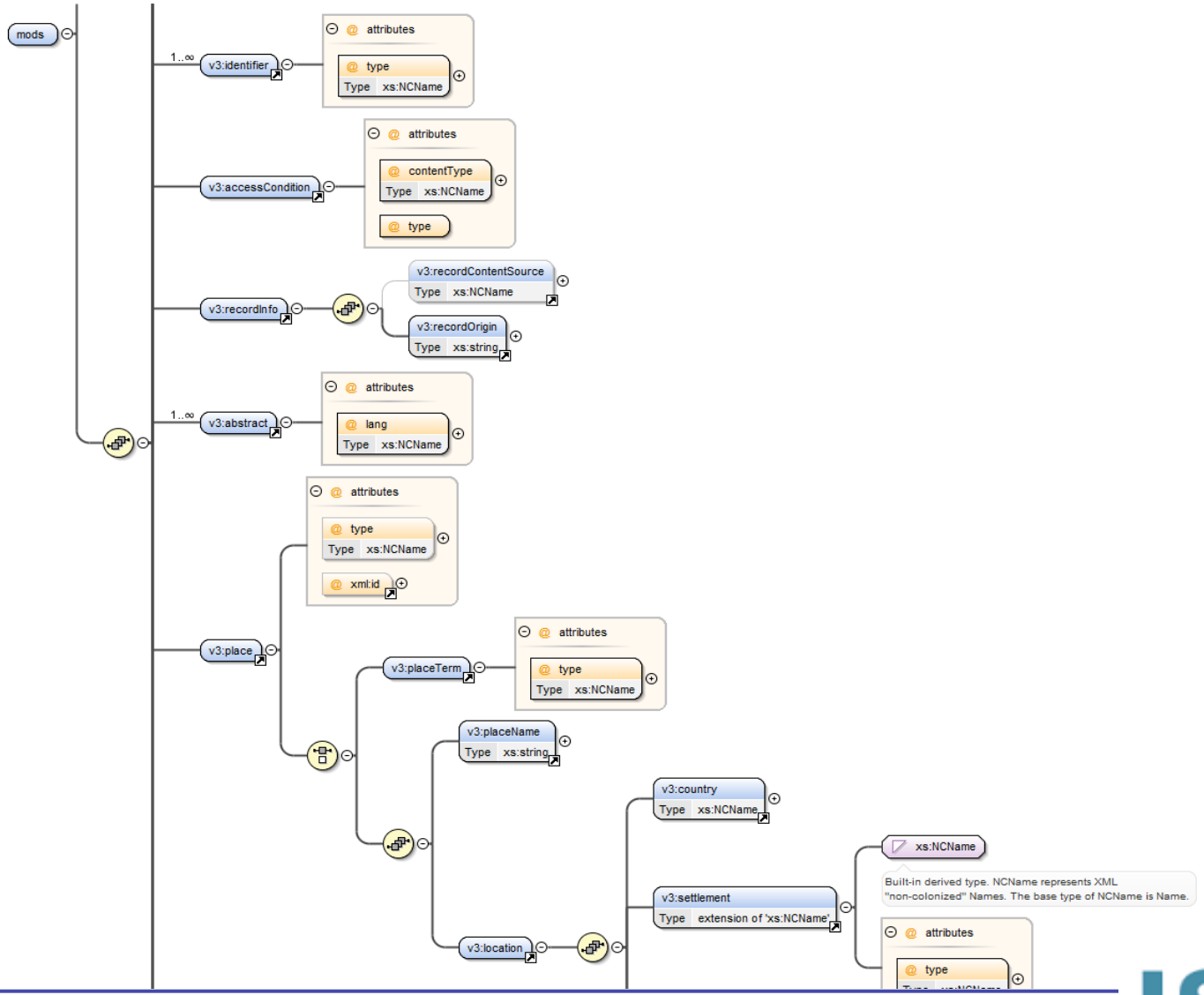
Phase de marquage de contenu (Fouille de textes)



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

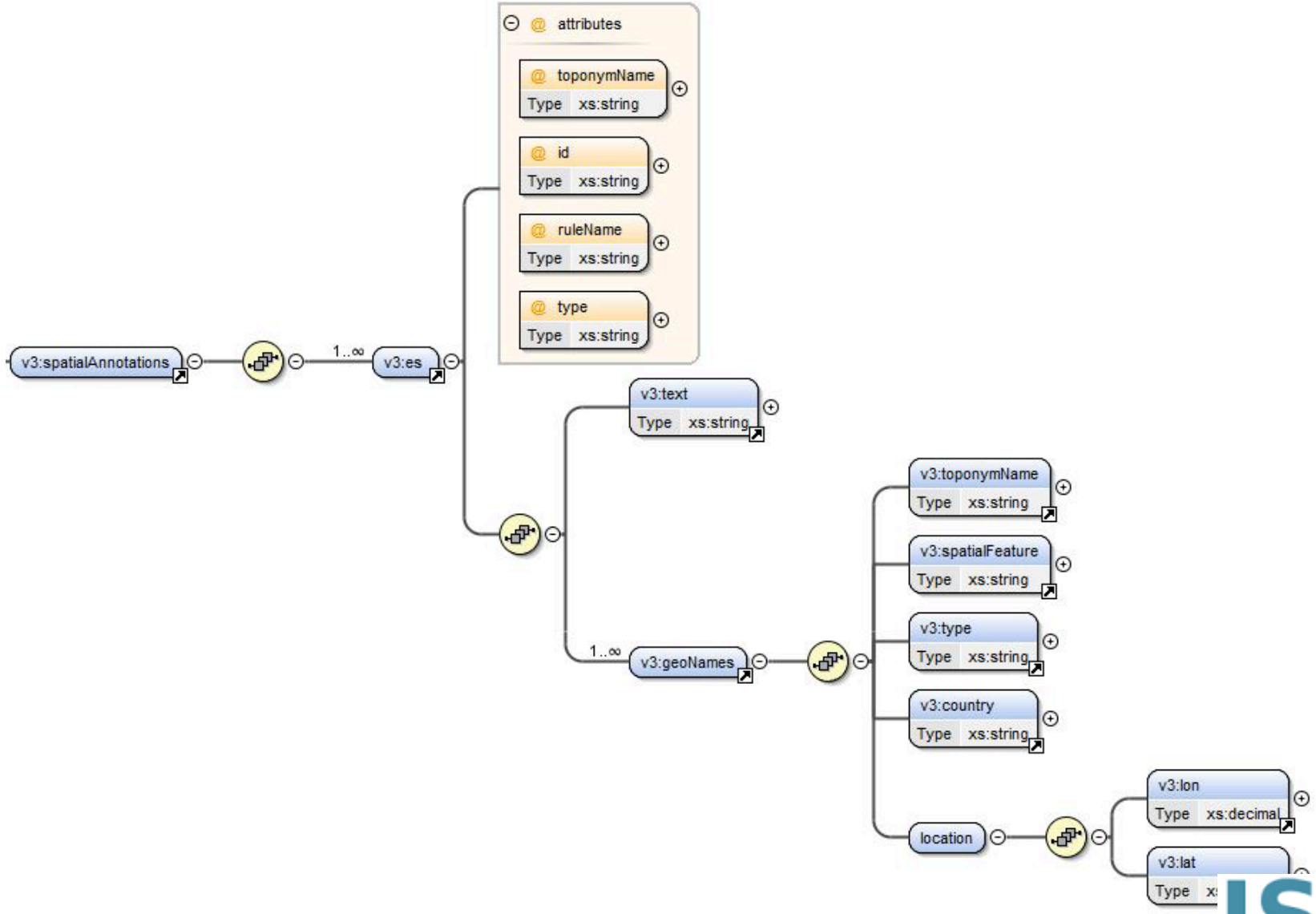
Modélisation des descripteurs, une contribution importante



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

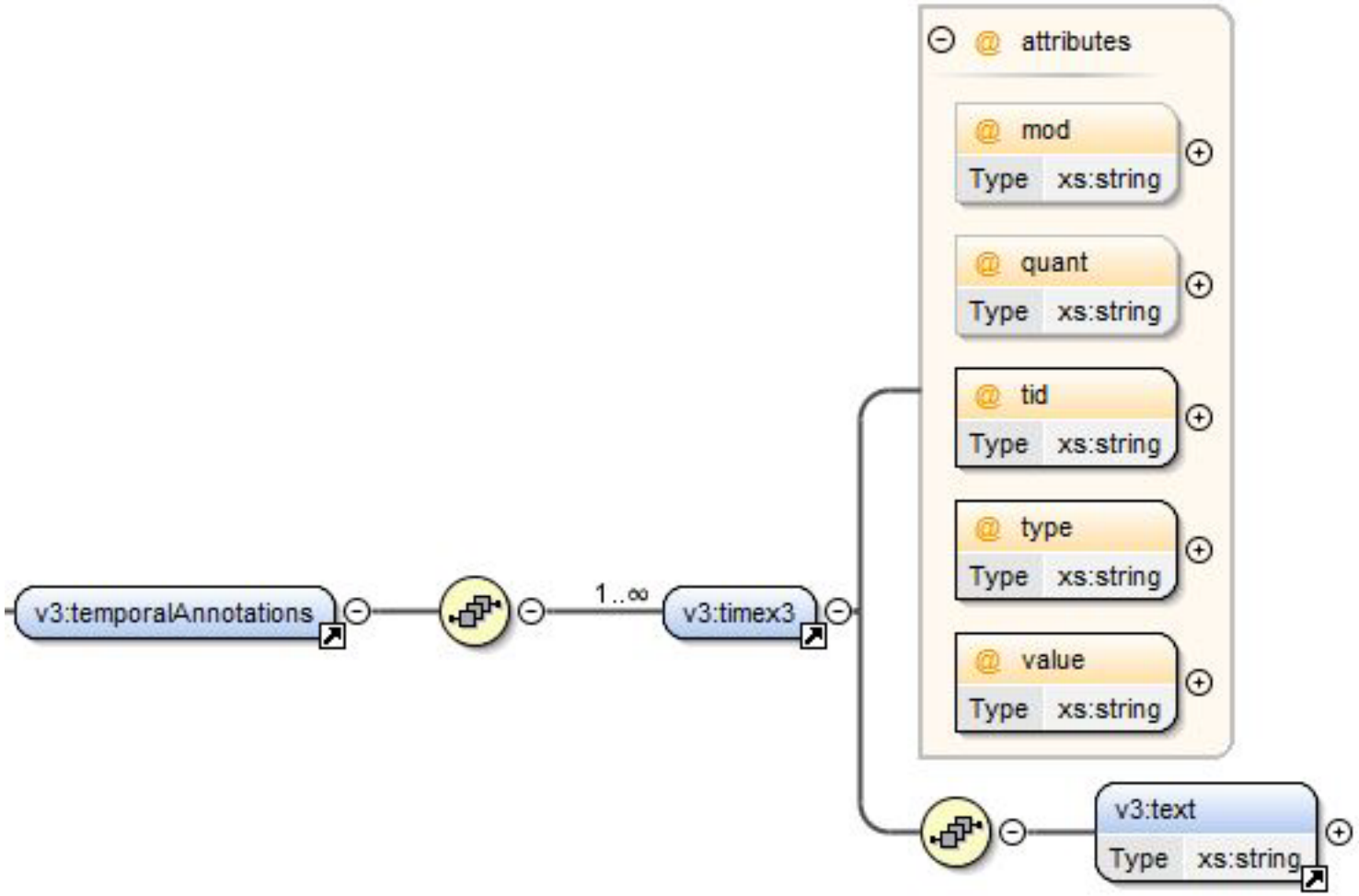
Modélisation du descripteur SpatialAnnotations



ISTEX
L'excellence documentaire pour tous

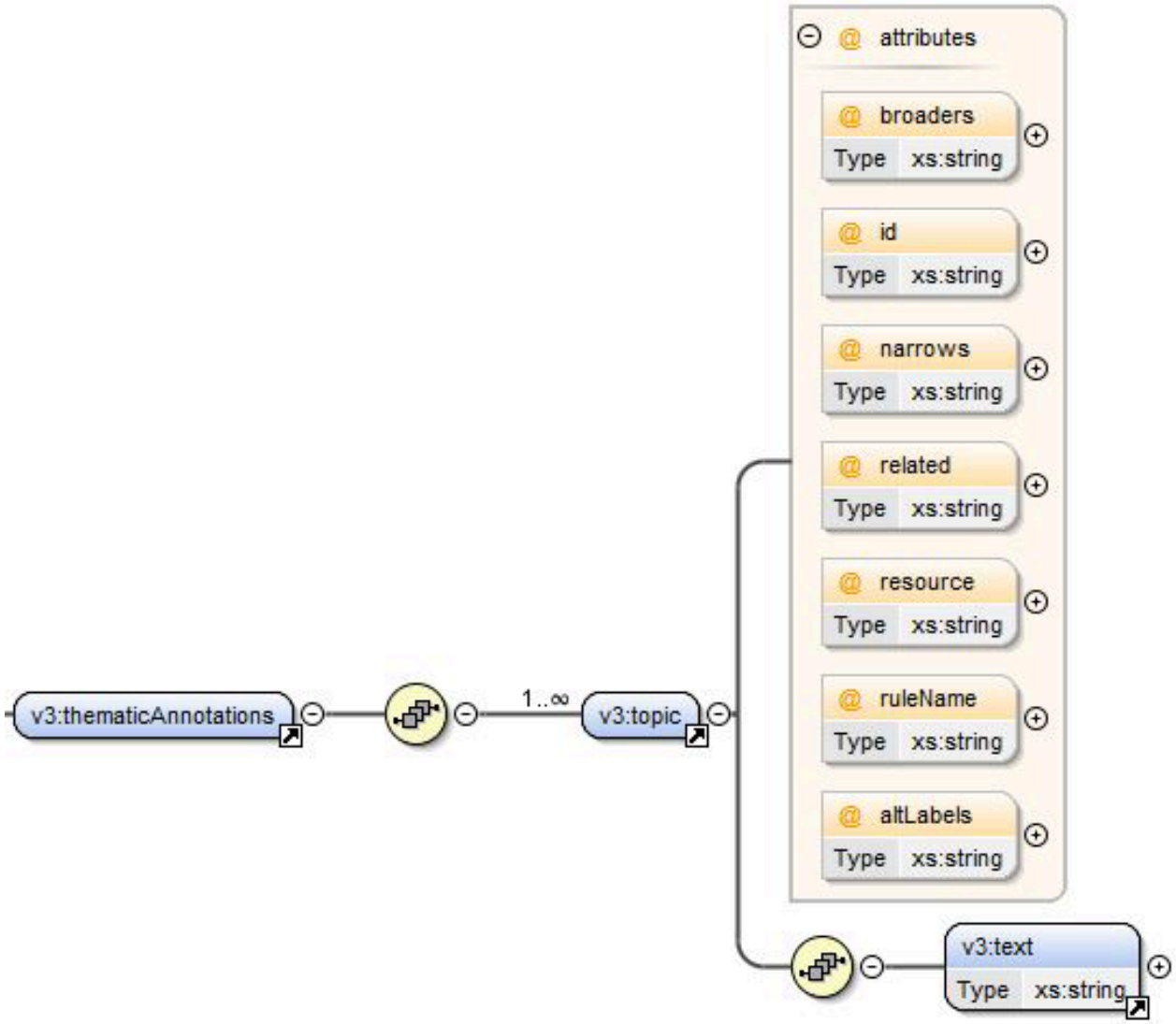
TERRE-ISTEX

Modélisation du descripteur temporalAnnotations



TERRE-ISTEX

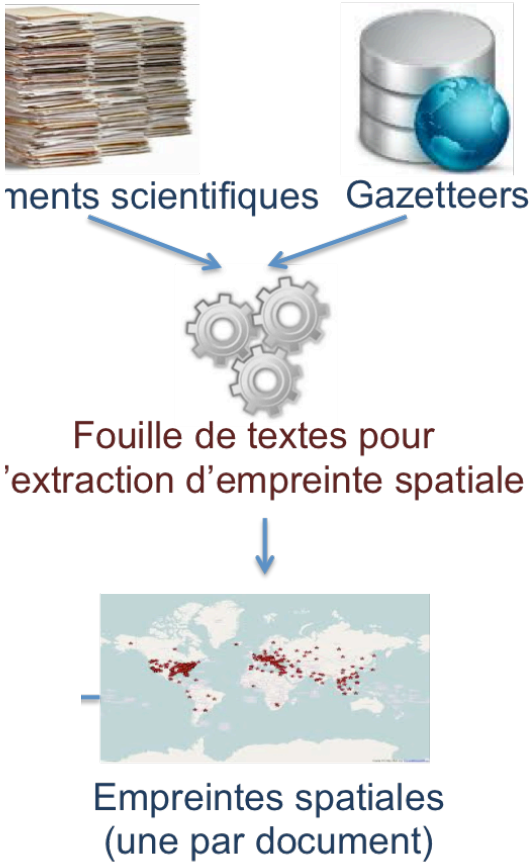
Modélisation du descripteur ThematicAnnotations



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Extraction et localisation des entités spatiales



(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)

"… dans la ville de Montpellier…"

"… /dans/la/ville/de/Montpellier/ …"

1. • tokenisation

Règles typographiques

"/dans/la/ville/de/Montpellier/"

type: geo
stype: S

2. • Identification tokens spatiaux candidats

Règles lexicales

"/dans/la/ville/de/Montpellier/"

nature: nom commun
lemma: ville

nature: nom propre
lemma: Montpellier

3. • Analyse morpho-syntaxique

Règles syntaxiques

"/dans la/ville de Montpellier/"

EN candidate: ESR

EN candidate: ESA

4. • Analyse sémantique: ESA & ESR

Patrons (DCG)

"/dans la ville de Montpellier/"

étiquette: commune
source: IGN
Coord.: X, Y

Montpellier

5. • Validation ESA & géolocalisation

Gazetteers (Geonames, etc.)

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

- 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 ?!

TERRE-ISTEX

Actions en cours

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

- ✓ 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

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

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

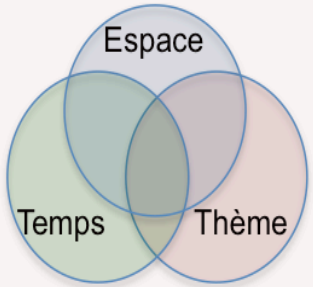
- Thématiques et/
- ou disciplines
- Résumés

Validation des données

elastic kibana

Indexation

Analyse géographique



Recherche d'information

Thème, Temps, Espace, Plein texte

ISTEX
L'excellence documentaire pour tous

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

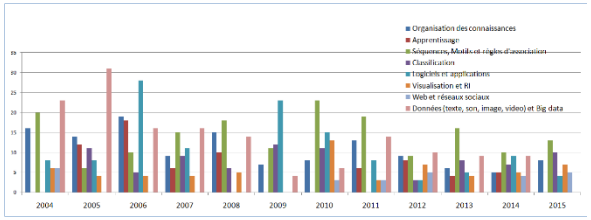
- Analyse géographique de séries de publications : application aux donnés EGC (Kergosien et al., 2017) : indexation, recherche d'information, analyses

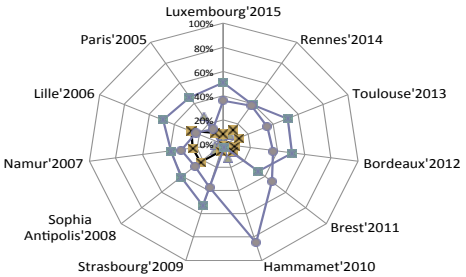
  - Construction d'un premier index géographique

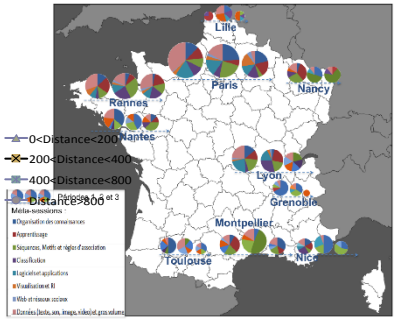
Information thématique	Nom ville conférence Noms villes auteurs Noms auteurs Titre article Résumé article Session Domaine
Information spatiale	Coordonnées villes auteurs Coordonnées ville conférence
Information temporelle	Année conférence
Information plein texte	Termes titre article Termes résumé article

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

  - Ensemble d'analyses géographiques disponibles
    http://ekergosien.net/DefiEGC/index.html

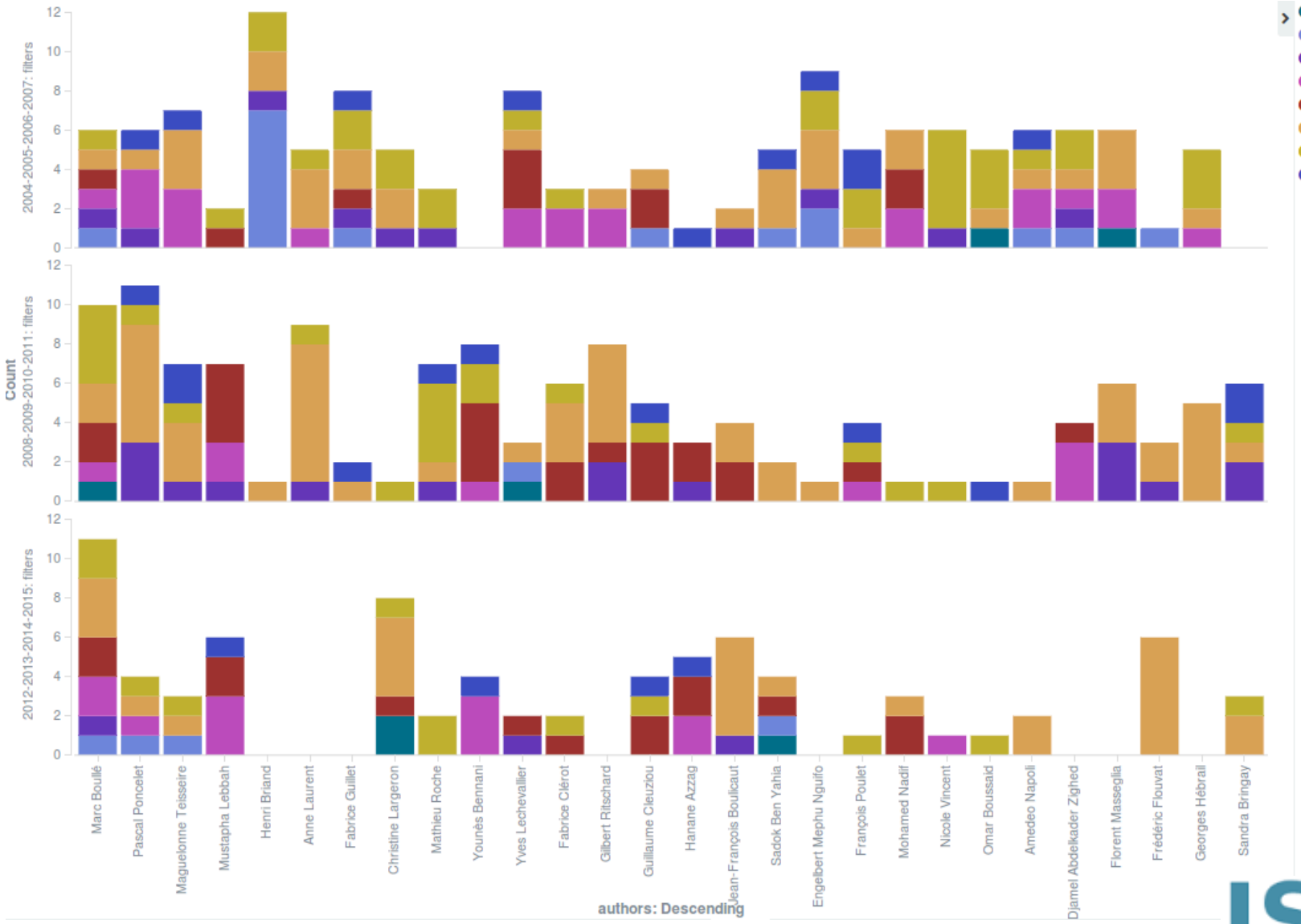






TERRE-ISTEX

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



ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

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

ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Valorisation / Publications

- 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), 2017

- 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, communication à venir (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, Vers une analyse thématique automatique de séries de publications : application aux articles des conférences EGC, In *84ème conférence de l'ACFAS*, Montréal, mai 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

- 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.

ISTEX
L'excellence documentaire pour tous

TERRE-ISTEX

Questions

https://terreistex.hypotheses.org

VADOR

https://vador.sciencesconf.org

D4H

https://www.meshs.fr/page/d4humanities

ISTEX
L'excellence documentaire pour tous