TERRE-ISTEX : vers un modèle pour identifier des terrains d'études

Annig Le Parc - Lacayrelle [1], **Amin Farvardin** [2]

1. *Université de Pau et des Pays de l'Adour*
   LIUPPA
   64013 Pau cedex, France
   annig.lacayrelle@univ-pau.fr
2. *Université Paris-Dauphine*
   LAMSADE
   75775 Paris cedex 16, France
   amin.farvardin@teledetection.fr

RÉSUMÉ. *Cet article présente les premiers travaux réalisés dans le projet* TERRE-ISTEX *dont les objectifs sont, d'une part, d'identifier les fronts de recherche en relation avec les territoires et, d'autre part, d'offrir un outil de recherche d'information multidimensionnelle.*

ABSTRACT. *This article presents the first works carried out in the* TERRE-ISTEX *project whose objectives are, firstly, to identify research fronts in relation to the territories and, on the other hand, to offer a multidimensional information research tool.*

MOTS-CLÉS : *Information géographique, documents hétérogènes et multi-langues*

KEYWORDS: *Geographical information, multi-lingual and heterogeneous documents*

1. Introduction

La disponibilité accrue des ressources numériques scientifiques, à travers notamment les plateformes de revues (Revues.org), des répertoires d'archives ouvertes (HAL), des entrepôts de thèses électroniques (Theses.fr), des services de fédération de contenus (Isidore), des répertoires de données de la recherche (Nakala) et des bibliothèques numériques (ISTEX) offre de nouvelles et de nombreuses opportunités d'usage. Les historiens et les sociologues des sciences peuvent ainsi analyser la genèse des disciplines, les conditions et les facteurs d'émergence des concepts par les communautés scientifiques, leurs modalités de circulation et d'appropriation par d'autres communautés. Il est également possible d'identifier l'évolution des fronts de recherche, les croisements disciplinaires ainsi que les modalités concrètes de recherche dans la mesure où ce gigantesque corpus scientifique rend compte des terrains, des méthodes et des cadres théoriques mobilisés.

L'objectif de ce papier est de présenter le projet de recherche TERRE-ISTEX [1] qui s'inscrit dans cette dernière perspective. Ce projet fait partie des travaux initiés par le projet ISTEX [2] dont l'objectif est de créer des services de recherche d'information innovants pour accéder à un ensemble de ressources numériques selon différents critères. Mais, au-delà de la seule analyse des fronts de recherche et de l'évolution des tendances dans une optique infométrique, l'objectif de TERRE-ISTEX est d'identifier les territoires au sens géographique du terme. Par "territoire", nous entendons un ensemble d'informations géographiques associant des informations spatiales, temporelles et thématiques sur lesquelles ont porté des études. Bien que de nombreux travaux en scientométrie présentent des méthodes pour analyser des communautés à partir de publications scientifiques (que ce soit en revues ou en conférences) (Cavero *et al.*, 2014) (Cabanac *et al.*, 2015), il n'existe pas à notre connaissance de travaux proposant une analyse géographique d'un corpus de publications, i.e. combinant les dimensions spatiale, temporelle et thématique. Ainsi, à partir d'un corpus de documents scientifiques hétérogènes et multi-langues, le projet TERRE-ISTEX vise d'une part à (1) identifier les lieux qui ont fait l'objet d'études empiriques et dont rendent compte les publications issues du corpus, d'autre part (2) à identifier les approches (méthodes et concepts) mobilisées pour la réalisation de ces études et enfin (3) à situer temporellement les périodes au cours desquelles les études ont été menées. En croisant ces trois dimensions (spatiale, temporelle et thématique), il sera ainsi possible de comprendre quelles recherches ont été menées sur quels territoires, selon quelles approches et à quel moment. L'intérêt de ses travaux est ainsi de compléter les approches classiques de veille qui articulent très rarement ces trois dimensions en se focalisant tout d'abord sur la thématique "changement climatique".

Ce projet comporte 3 aspects :

1. https://terreistex.hypotheses.org/
2. http://www.istex.fr/le-projet/

– le marquage et l'indexation précise des informations spatiales, temporelles et des thématiques traitées dans les documents du corpus ;

– l'articulation de ces 3 axes (spatial, temporel et thématique) pour l'exploration des publications et l'analyse des fronts de recherche ;

– la conception et le développement d'un moteur de recherche multidimensionel.

Ce papier s'intéresse plus particulièrement au premier aspect du projet. La section 2 présente la démarche générale utilisée dans le projet. Le corpus est décrit en section 3. La section 4 aborde la mise en oeuvre de la démarche, et la section 5 décrit le modèle de données TERRE-ISTEX. La section 6 conclut cette présentation.

2. Démarche générale

La démarche mise en oeuvre est décrite figure 1. Elle est générique car indépendante de tout corpus de publication. Nous l'avons d'ailleurs déjà utilisé dans le cadre d'une étude sur les publications EGC (Kergosien *et al.*, 2016).

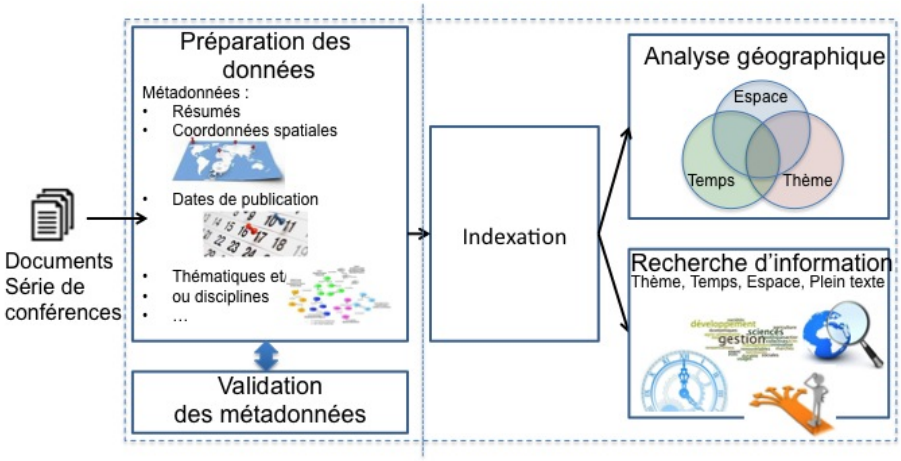

Figure 1. Chaîne de traitement générique pour l'analyse de corpus de publications.

La première étape consiste à extraire, en utilisant une approche TAL, les informations spatiales, temporelles et thématiques contenus dans les documents du corpus et dans leurs méta-données. Ces informations sont ensuites validées en nous appuyant respectivement sur des outils de géocodage, de fouille de textes et sur une base calendaire. Une deuxième étape concerne l'indexation des méta-données et des informations extraites dans un moteur de recherche afin de les exploiter, lors d'une troisième étape, dans des stratégies d'analyse et de recherche d'information combinant des critères spatiaux, temporels et thématiques. Il est à noter que les résumés et les contenus des documents sont eux aussi indexés afin de permettre la recherche "plein-texte".

3. Le corpus du projet TERRE-ISTEX

Le corpus regroupe des publications ayant trait à la thématique "changement climatique sur les territoires du Sénégal et de Madagascar" et provenant des plateformes ISTEX [3] et Agritrop [4] (archive ouverte du CIRAD [5]), ainsi que des thèses de l'ANRT [6] et de theses.fr [7]. Nous avons choisi ce cas d'étude car nous disposions d'un grand nombre de documents sur le sujet, et que les analyses qui en découleraient ont un intérêt pour des membres du projet. Les documents provenant de la plateforme ISTEX (environ 170000 documents) ont été obtenus en faisant des requêtes avec les mots-clés suivant : "climate change", "changement climatique", "Senegal", "Sénégal", "Madagascar". Les documents provenant d'Agritrop ciblent des études traitant de Madagascar et du fleuve Sénégal. Enfin, les 400 thèses provenant de l'ANRT traitent du changement climatique. Chaque document possède, en plus de son contenu, des méta-données et un résumé. Selon la provenance du document, les méta-données sont soit au format MODS [8] (ISTEX), soit un format XML inspiré du Dublin Core (CIRAD), soit en RDF (thèses ANRT). Le corpus est multi-langue : certains documents sont en français et d'autres en anglais, mais on peut également trouver des documents utilisant les deux langues (par exemple, ils comportent un résumé en français et un résumé en anglais). Nous sommes donc face à un ensemble de documents multi-langues et hétérogènes.

4. Mise en oeuvre de la démarche

Dans un premier temps, nous avons choisi d'appliquer notre approche sur les méta-données et les résumés de chaque document. Dans un second temps, nous l'appliquerons en plus sur leur contenu. Comme nous venons de le voir, le corpus est multi-langue (documents en français et en anglais) et hétérogènes (différents formats pour les méta-données). La figure 2 décrit la chaine de traitement mise en oeuvre dans le projet TERRE-ISTEX.

Pour pallier l'hétérogénéité de format des méta-données, nous avons défini un modèle de données TERRE-ISTEX basé sur le format MODS (voir section 5). Des règles de transformation entre modèles ont donc été écrites. Une fois toutes les méta-données au format MODS, une annotation des entités spatiales, temporelles et thématiques contenues dans les résumés est réalisée. Les méthodes d'annotation pour les entités spatiales et thématiques sont basées sur celles de l'outil web SISO (Farvardin *et al.*, 2015). Les entités spatiales (ES) à annoter peuvent être de deux types (Sallaberry *et al.*, 2007) : les ES absolues (ESA) (références directes à un espace géo-localisable,

3. http://www.istex.fr/category/plateforme/
4. https://agritrop.cirad.fr/
5. http://www.cirad.fr/
6. http://www.anrt.asso.fr/
7. http://www.theses.fr/
8. http://www.bnf.fr/fr/professionnels/f_mods/s.mods_presentation.html

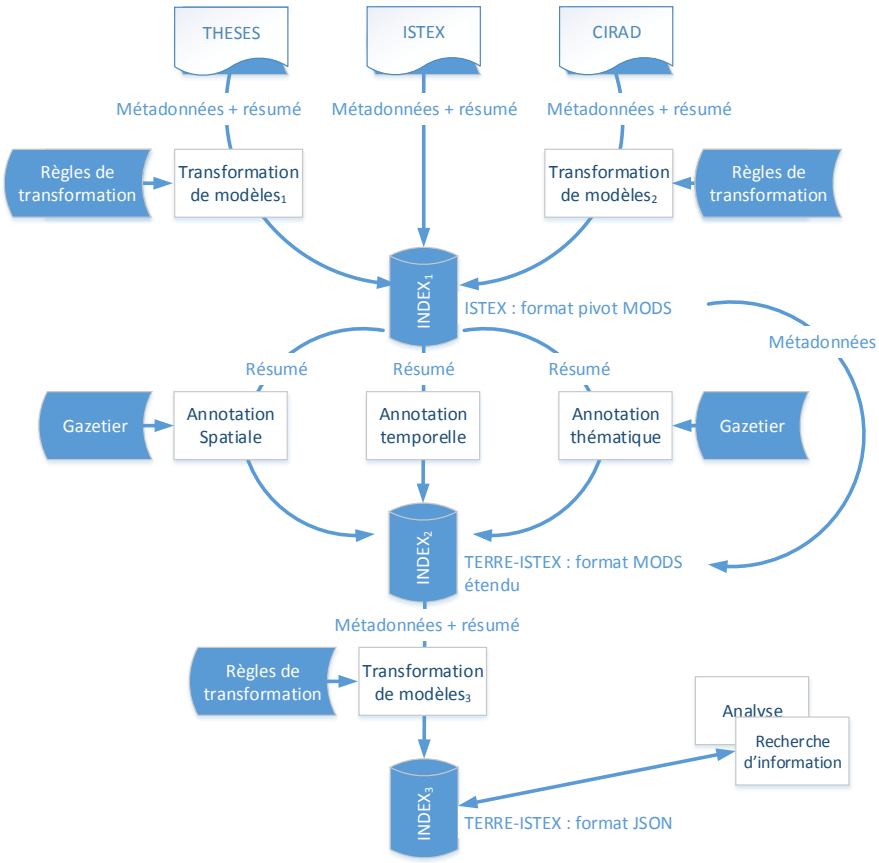

Figure 2. Mise en oeuvre de la chaine de traitement

par exemple "la ville de Paris") et les ES relatives (ESR) (définies à l'aide d'au moins une ESA et d'indicateurs spatiaux d'ordre topologique, par exemple, "près de Paris"). Une fois les ES identifiées, la chaine de traitement calcule l'empreinte spatiale correspondant à chacune d'entre elles (en utilisant la ressource Geonames). Les entités thématiques à annoter étant liées, dans notre cas, au domaine du changement climatique, nous utilisons la ressource Agrovoc (Rajbhandari, Keizer, 2012). Cette approche est générique car le changement de domaine d'étude nécessite juste l'utilisation de la ressource appropriée. En ce qui concerne les entités temporelles, nous annotons uniquement les entités calendaires (dates et périodes) en utilisant HeidelTime (Strötgen, Gertz, 2013).

Ces annotations viennent ensuite compléter les méta-données décrivant chaque document. Ces méta-données sont ensuite indexées sous ElasticSearch [9] ce qui implique une transformation du format MODS-étendu au format json (utilisé par ElasticSearch). Nous avons choisi ElasticSearch car c'est un moteur de recherche basé sur la librairie Lucene, qui permet la recherche plein-texte, la recherche structurée, la gestion des données spatiales et offre des outils d'analyse de données tels que Kibana [10]. Les index ElasticSearch ainsi créés pourront donc être utilisés pour produire les analyses et permettre la recherche d'information multidimensionnelle.

5. Le modèle de données TERRE-ISTEX

Le modèle de données TERRE-ISTEX étend le format MODS afin de lui permettre de décrire les informations spatiales, temporelles et thématiques extraites des documents et de leurs méta-données. Le choix de MODS a été guidé par le fait que MODS est le format utilisé sur la plateforme ISTEX, qu'il est approprié à la description de tous les types de documents et de tous les supports (numériques ou non), qu'il est plus riche que le Dublin Core et plus proches des modèles de structuration des informations bibliographiques utilisées par les bibliothèques.

Ainsi, nous avons rajouté trois balises à un document MODS :

- `<spatialAnnotations>`,
- `<temporalAnnotations>`,
- `<thematicAnnotations>`.

La balise `<spatialAnnotations>` contient un ensemble d'entités spatiales (balise `<es>`), avec pour chacune d'elle, le texte annoté (balise `<text>`) ainsi que son empreinte spatiale obtenue en interrogeant la ressource Geonames. La DTD correspondante est donnée figure 3.

La balise `<temporalAnnotations>` contient un ensemble d'entités temporelles décrites par les balises `<timex3>` provenant d'Heildeltime complété par le texte annoté (balise `<text>`). La DTD correspondante est donnée figure 4.

Enfin, la balise `<thematicAnnotations>` contient l'ensemble des thèmes abordés dans le résumé (balise `<topic>`), avec pour chacun d'eux des informations provenant de la ressource Agrovoc complété le texte annoté (balise `<text>`). La DTD correspondante est donnée figure 5.

6. Conclusion

Dans cet article, nous avons présenté la chaine de traitement mise en oeuvre dans le projet TERRE-ISTEX pour marquer et indexer les informations spatiales, tempo-

9. http://www.elastic.co/fr/
10. http://www.elastic.co/fr/products/kibana/

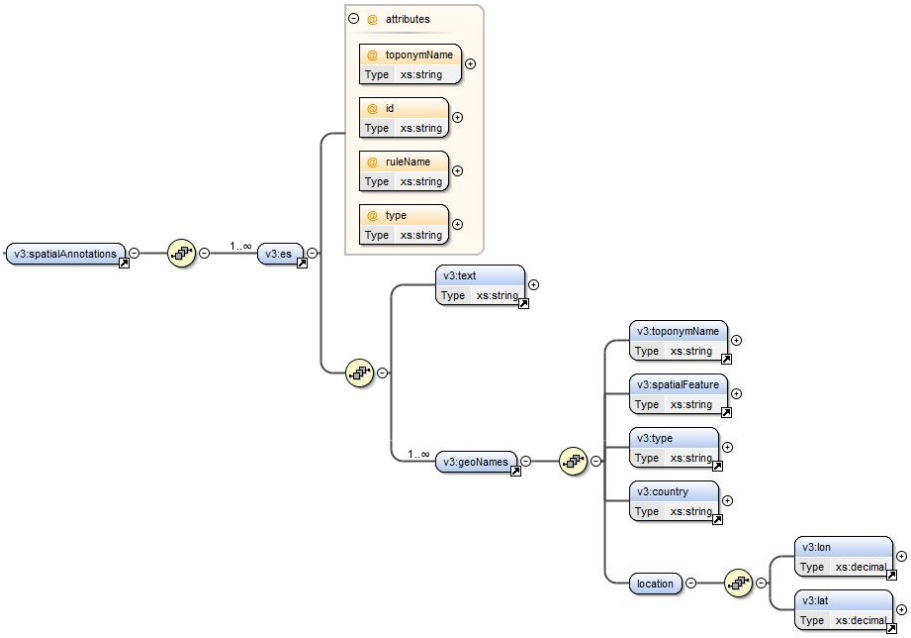

Figure 3. DTD décrivant la balise `<spatialAnnotations>`

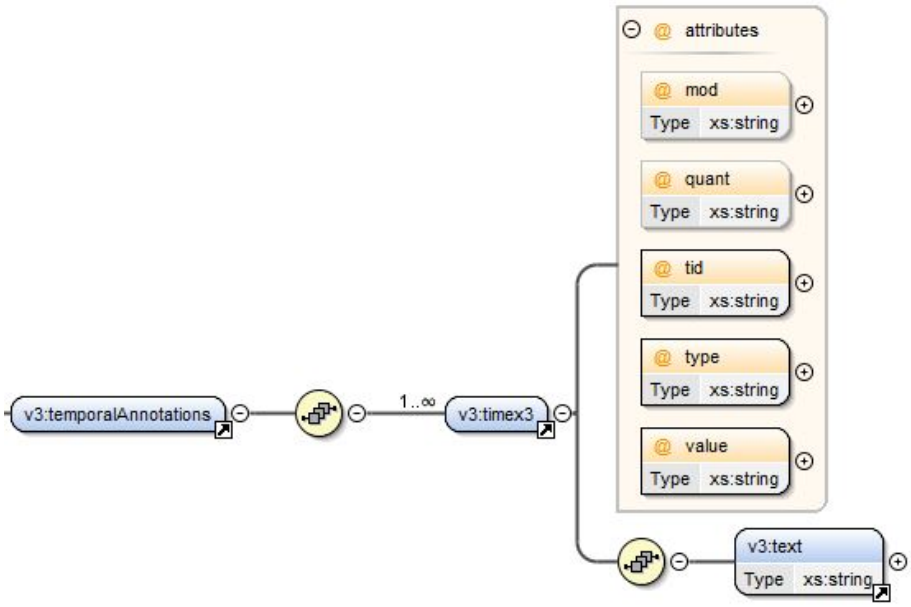

Figure 4. DTD décrivant la balise `<spatialAnnotations>`

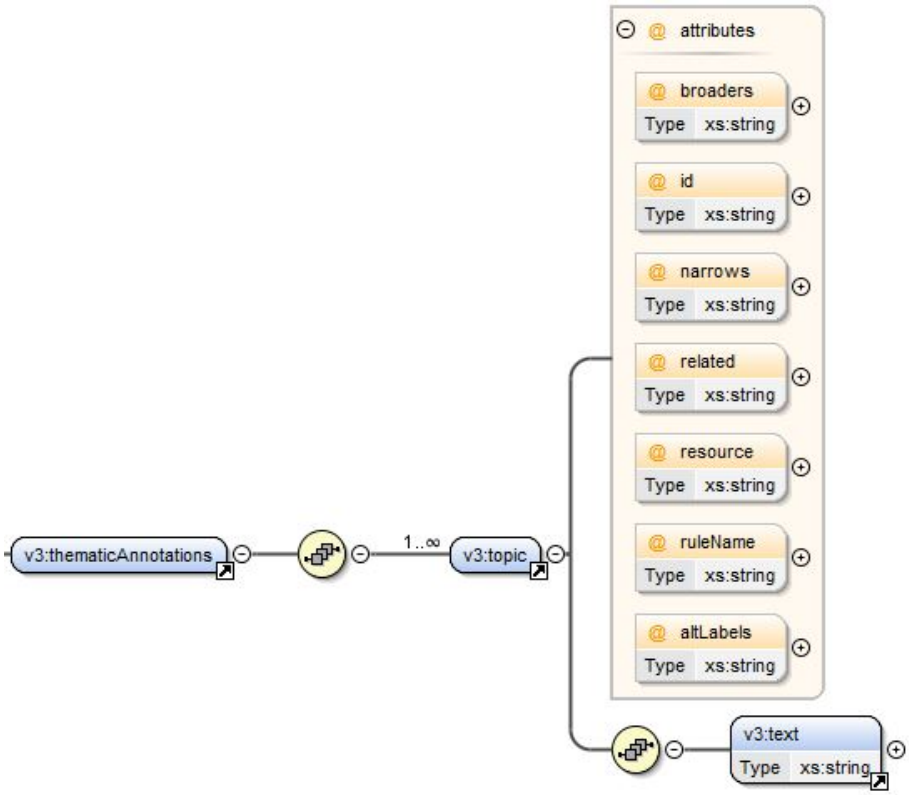

Figure 5. DTD décrivant la balise `<thematicAnnotations>`

relles et thématiques contenues dans un corpus de documents hétérogènes et multilangues. Nous avons choisi comme format pivot MODS, et nous l'avons étendu en ajoutant des descripteurs permettant de décrire les informations annotées. Actuellement, cette chaine est appliquée uniquement sur les méta-données et les résumés des documents. La phase d'annotation est terminée (une première évaluation peut-être trouvée dans (Bessagnet *et al.*, 2017)). Nous disposons donc de chaque document au format MODS-étendu. Nous travaillons actuellement sur la transformation JSON et sur la construction de l'index ElasticSearch. Cet index va ensuite être utilisé, d'une part, pour mettre en oeuvre des analyses permettant d'identifier les fronts de recherche en relation avec les territoires d'études, et d'autre part, pour offrir un outil de recherche d'information multidimensionnelle.

Bibliographie

Bessagnet M.-N., Kergosien E., Farvardin A., Le Parc-Lacayrelle A., Sallaberry C. (2017). A propos des territoires dans les corpus scientifiques. In *Atelier emc^{Sci} 2017 (en cours de soumission).*

Cabanac G., Hubert G., Milard B. (2015). Academic careers in computer science: continuance and transience of lifetime co-authorships. *Scientometrics*, vol. 102, n° 1, p. 135–150. Consulté sur http://dx.doi.org/10.1007/s11192-014-1426-0

Cavero J., Vela B., Cáceres P. (2014). Computer science research: more production, less productivity. *Scientometrics*, vol. 98, n° 3, p. 2103-2111. Consulté sur http://dx.doi.org/10.1007/s11192-013-1178-2

Farvardin A., Kergosien E., Roche M., Teisseire M. (2015). A webtool for analyzing land-use planning documents. In *14th international semantic web conference (demonstration track).*

Kergosien E., Bessagnet M.-N., Sallaberry C., Le Parc-Lacayrelle A., Royer A. (2016). Analyse géographique de séries de publications : application aux conférences egc. In *Conférence internationale francophe sur l'extraction et la gestion des connaissances (egc 2016).*

Rajbhandari S., Keizer J. (2012). The agrovoc concept scheme: A walkthrough. *Journal of Integrative Agriculture*, vol. 11, n° 5.

Sallaberry C., Baziz M., Lesbegueries J., Gaio M. (2007). Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - évaluation. In *Coria*, p. 53-64.

Strötgen J., Gertz M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, vol. 47, n° 2, p. 269–298.