CybergeoNetworks, une application interactive pour l'analyse géographique et sémantique des publications scientifiques

Clémentine Cottineau, Juste Raimbault, Pierre-Olivier Chasset, Hadrien Commenges, Arnaud Banos, Denise Pumain, Christine Kosmopoulos

CybergeoNetworks, une application interactive pour l'analyse géographique et sémantique des publications scientifiques

Clémentine Cottineau, Juste Raimbault, Pierre-Olivier Chasset, Hadrien Commenges, Arnaud Banos et Denise Pumain

(Chapitre in Bouzeghoub M. et Mosseri R. (dir.) - *Les Big Data à découvert*. Paris, CNRS Editions, 2017, 272-273.)

La croissance exponentielle du nombre des articles publiés et la multiplication des revues ont fait entrer les publications scientifiques dans le règne des données massives. Les bouleversements introduits par les usages d'Internet semblent avoir bien davantage contribué à l'explosion des références qu'ils n'ont simplifié le travail des équipes éditoriales ou facilité l'accès des utilisateurs à la littérature scientifique. Les difficultés de la maîtrise de ces données vont bien au-delà de ce qu'Eugène Garfield avait anticipé en créant en 1964 la base de données de l'ISI (*Institute for Scientific Information*, qui devait devenir le *Web of Science* après rachat par Thomson Reuters). Peu à peu, le bénéfice du travail des « pairs » qui évaluent bénévolement la production scientifique avant publication, principalement par une mise en situation dans les connaissances existantes, a été capté par des sociétés d'édition censées garantir la qualité tout en ponctionnant les bibliothèques universitaires et en dressant des barrières à la diffusion, malgré les injonctions publiques pour un libre accès à la science.

La question de la maîtrise de la littérature reste cruciale pour tout scientifique qui souhaite connaître et dresser un « état de l'art ». Elle devient encore plus difficile pour les sujets qui se situent aux interfaces de plusieurs disciplines, et qui risquent, si l'on s'en tient aux « niches » disciplinaires habituelles, d'être traités de manière partielle en réduisant la portée des solutions que la science peut apporter pour résoudre des problèmes sociaux. Les géographes en sont très conscients car depuis longtemps ils construisent une discipline à la charnière des sciences naturelles et sociales, sur des questions d'habitat et d'urbanisme, d'environnement et de santé, d'aménagement et de développement. Ce n'est peut-être pas un hasard si c'est à partir d'une revue de géographie, *Cybergeo*, qu'un outil est proposé pour aider à améliorer l'exploration d'un univers de publications.

Les progrès réalisés en matière de collecte et d'analyse des données massives permettent aujourd'hui de s'orienter de manière inédite dans les réseaux de publications scientifiques. Nous avons construit une application originale, en libre accès, qui permet d'explorer le contenu du texte et des mots-clés des quelque 900 articles publiés depuis 20 ans par Cybergeo ainsi que l'ensemble des références citées dans ces articles, ou des articles qui les citent, ou qui citent les mêmes références, parmi toutes les autres revues scientifiques accessibles par Google Scholar, ce qui constitue une base de données d'environ 200 000 articles avec leurs centaines de milliers de mots-clés associés. L'objectif est de permettre aux internautes de réaliser eux-mêmes à volonté des cartographies des proximités géographiques et sémantiques entre publications et entre zones géographiques en jouant avec ce corpus.

Une première possibilité offerte par l'application est la réalisation d'une cartographie diachronique représentant, pour une période donnée, les États d'affiliation des auteurs ainsi que les États constituant le sujet de l'article. Lorsque l'on croise cette information avec le profil thématique des articles, on fait apparaître une diversité des centres d'intérêt selon la localisation, ainsi qu'une proximité sémantique entre États étudiés avec des termes similaires. La proximité spatiale et la proximité thématique se recoupent par endroits. Ainsi, par exemple, l'Europe institutionnelle est identifiée comme espace de proximité sémantique forte où le champ lexical des frontières est sur-représenté et celui du risque sous-représenté (figure 1).

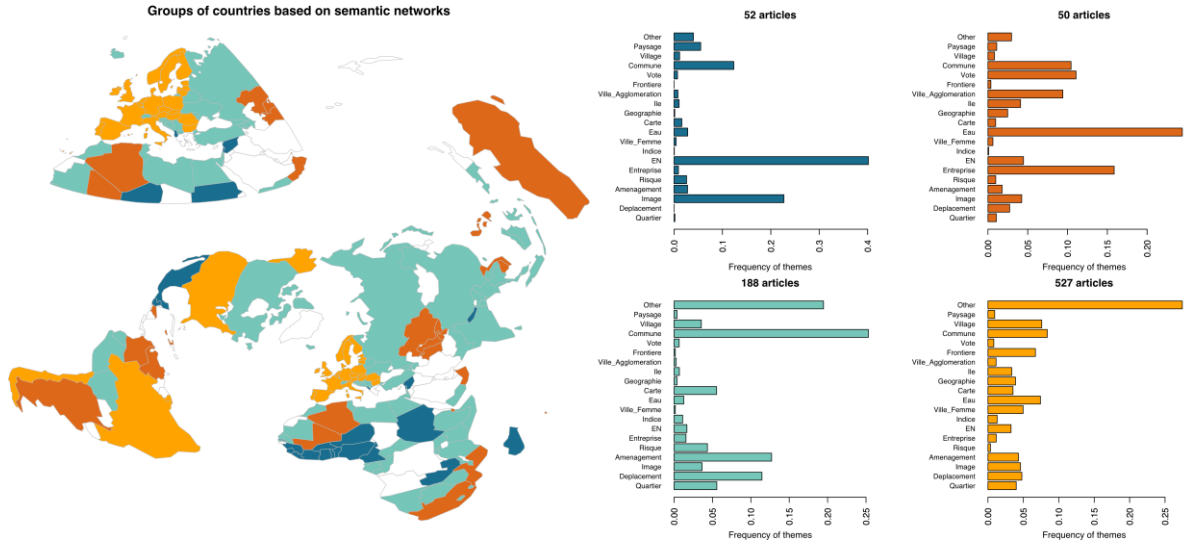

Figure 1: Classification des États étudiés dans Cybergeo en fonction du profil thématique des termes utilisés dans le corps du texte.

Les mots-clés des articles de la revue et des articles cités permettent de connecter les articles qui les utilisent et de constituer des réseaux pondérés en fonction du nombre de ces co-occurrences. La structure de ces réseaux nous renseigne de manière très fine sur les proximités entres des thématiques, réunies dans des « communautés » identifiées par des couleurs dans la figure 2 et que l'on peut explorer à plusieurs niveaux de finesse en zoomant sur le graphe du réseau.

La figure 2 spatialise ce réseau et montre que des disciplines comme la géographie physique et l'économie géographique sont reliées par leur pratique commune de méthodes comme l'analyse spatiale et les statistiques ou le paradigme de la complexité.

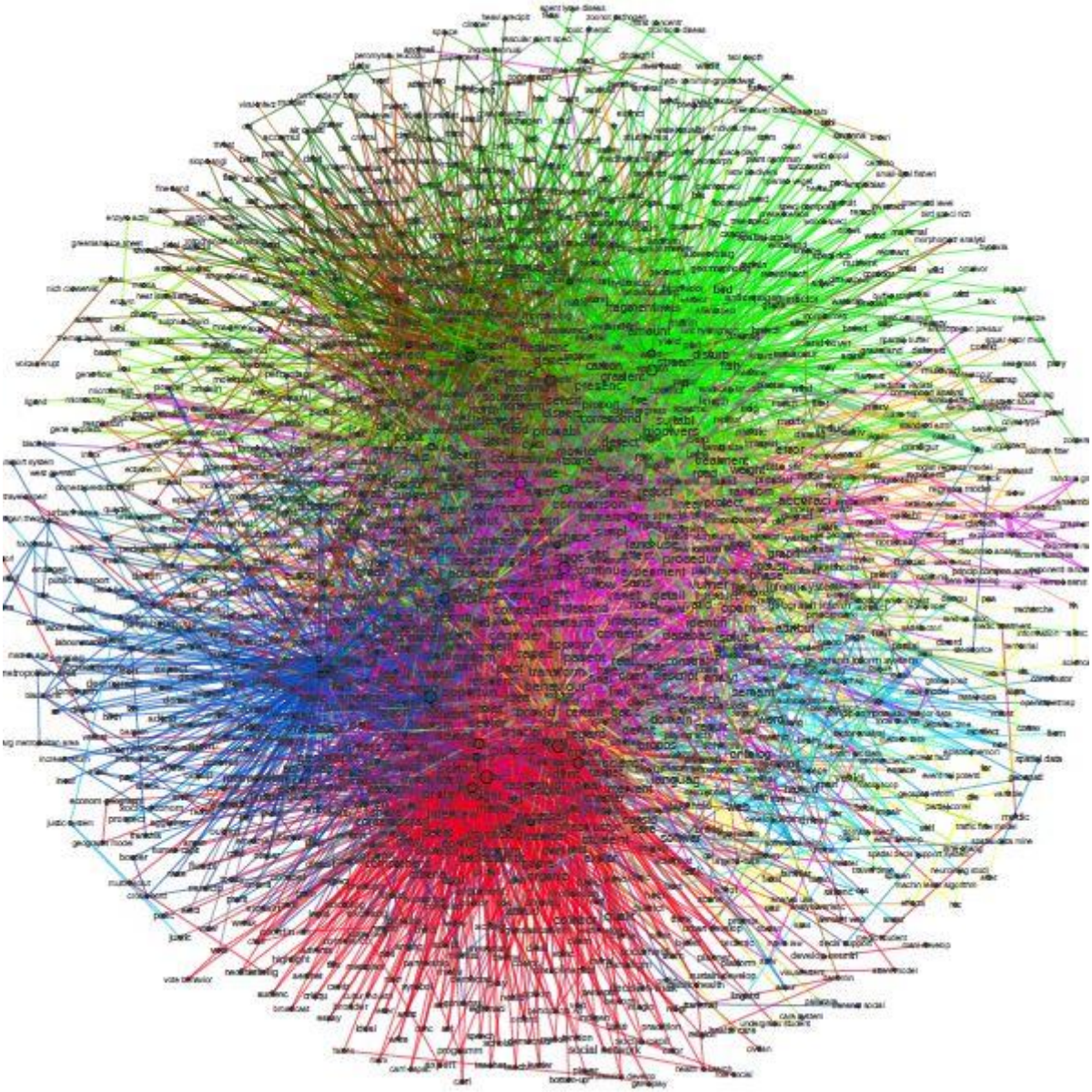

Figure 2 Réseau des mots-clés dans les articles cités dans Cybergeo ou les citant

L'analyse sémantique du contenu intégral des textes permet également de former des nuages de mots qui se regroupent par grands thèmes. On peut faire varier la finesse du regroupement des thèmes et mesurer la fréquence d'apparition des mots dans un thème (figure 3).

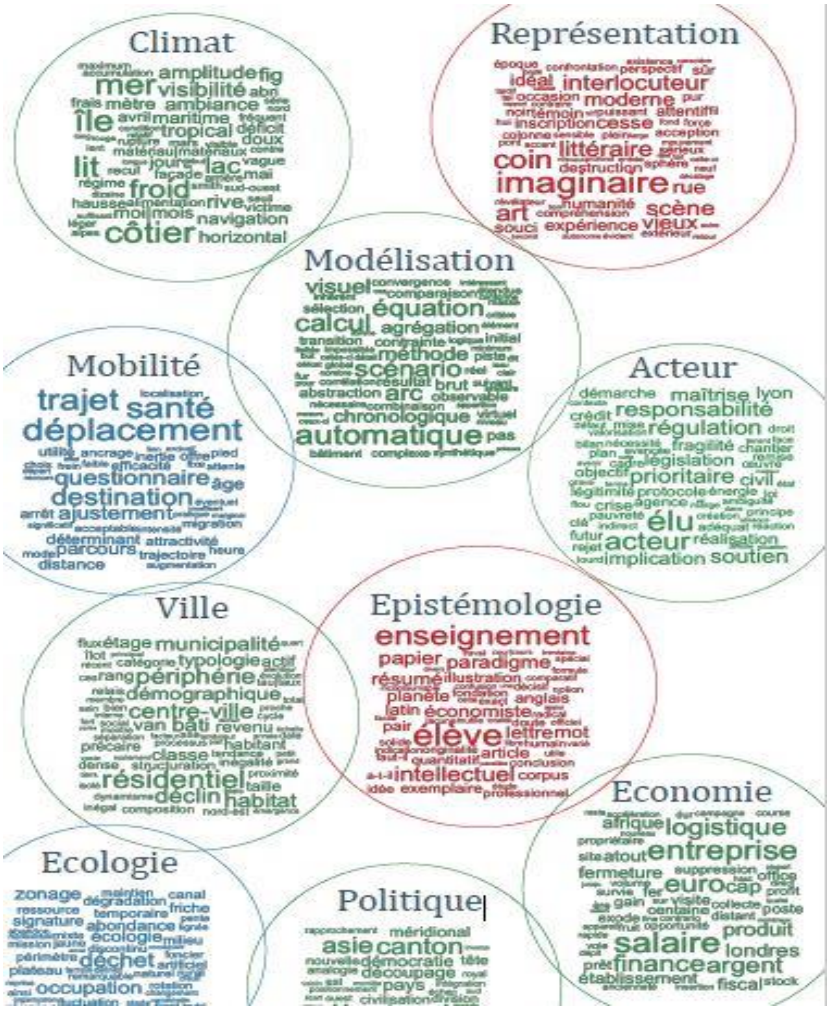

Figure 3 Une dizaine de grands thèmes issus de l'analyse des textes de Cybergeo et mis en nuages de mots. (La couleur du nuage varie selon le nombre de documents porté par chaque thème ; la taille des mots dans les nuages est proportionnelle à la fréquence d'apparition du mot dans le thème du nuage)

L'utilité de cet outil pour les auteurs est d'améliorer leur connaissance des thématiques de la revue, et de situer l'apport potentiel d'un nouvel article dans l'univers de ses références et des disciplines voisines qui peuvent être touchées. Pour l'équipe éditoriale, c'est un instrument de pilotage de la politique de la revue, qui peut choisir de maintenir sa ligne plutôt généraliste ou de se spécialiser davantage. Enfin, l'application pourrait être mise au service d'autres publications en ligne.

Les grandes sociétés d'édition scientifique proposent des outils d'analyse bibliométrique permettant aux chercheurs et aux revues d'optimiser leur investissement pour se situer au mieux sur le marché de la citation. Nous proposons ici un outil libre d'auto-analyse, conçu pour alimenter la réflexivité et l'intégrité de la recherche.

Bibliographie : Chavalarias D. Cointet J.-P. 2011, Phylomemetic patterns in science evolution- The rise and fall of thematic fields. PLoS ONE 8(2): e54847. doi:10.1371/ journal.pone.0054847

Kosmopoulos C., Pumain D., 2007 Citation, Citation, Citation: Bibliometrics, the Web and the Social Sciences and Humanities, *Cybergeo*, 411, 13 p.