Détermination de l'emprise géographique d'une carte à partir d'un texte, Geoffrey Brun[1], [1]IGN/COGIT

L'« emprise géographique » d'une carte correspond à la surface comprenant la totalité des données thématiques de la carte. La méthode proposée dans le cadre de cette thèse consiste à déterminer une emprise géographique « pertinente » à partir d'un texte, c'est-à-dire prenant en compte la localisation des informations relatives au sujet principal du texte. La méthode mise au point repose sur l'analyse des noms de lieu présents dans le texte en fonction de leur nombre d'occurrences et de leur localisation, et s'articule en trois étapes successives.

Dans un premier temps, les noms de lieu du texte sont identifiés. Ceux-ci peuvent s'avérer complexes à identifier, selon qu'ils contiennent des concepts géographiques et des indirections : *Madrid*, *la Chine méridionale*, *l'est de la Russie*, *l'archipel des Kouriles* ou encore *l'ouest du plateau de Sibérie orientale*. Ces noms de lieu sont extraits sous la forme d'entités nommées spatiales et d'organisation : nous jugeons pertinent, par exemple, de toujours exploiter le nom de lieu *Paris*, que celui-ci fasse référence à la ville ou à la municipalité de manière indifférenciée dans le texte. Cette étape d'extraction repose sur des dictionnaires et des patrons morpho-syntaxiques modélisés via le logiciel Unitex.

La seconde étape consiste à lever les ambiguïtés résidant sur ces noms de lieu – lesquels ne sont que de simples chaînes de caractères – en leur associant un toponyme. *Vienne* peut par exemple faire référence à une ville en France ou en Autriche, et *Niger* à un fleuve ou un pays. Cette étape de désambiguïsation s'appuie sur le gazetier GeoNames : les noms de lieu non ambigus, c'est-à-dire ne possédant qu'un seul toponyme possible dans le gazetier, sont utilisés afin de trouver le meilleur toponyme possible pour chaque nom de lieu ambigu.

Enfin, une emprise géographique est générée à partir des toponymes identifiés. Cependant, tous les noms de lieu d'un texte ne sont pas toujours utiles à la création d'une emprise géographique illustrant ce texte. Celui-ci peut par exemple traiter de puits de pétrole en Irak et citer néanmoins quelques noms de lieu extérieurs à ce pays, comme *Washington* et *Etats-Unis*. Afin de prendre en compte cette difficulté, la méthode proposée consiste à regrouper les toponymes dans des clusters (algorithme DBSCAN) en se fondant sur leurs coordonnées géographiques. Selon la proximité spatiale des clusters et les nombres d'occurrences des toponymes qui les composent, certains clusters sont ensuite sélectionnés pour générer une première emprise. Afin de générer une seconde emprise plus précise, les toponymes des clusters sélectionnés sont appariés à des entités géographiques de type polygonal provenant de la base de données spatiales Natural Earth.