Proposition de sujet de Thèse :

Titre :	
La Géographie à l'heure des données massives (Big Data) et de l'informatique en nuage (Cloud Computing) ou comment la Géographie ne peut échapper aux nouveaux paradigmes de l'information.	
Mots clés :	Big Data, Semantic Data, Cloud Computing, Services WEB, Modélisation mathématique, Information géographique.
Directeur :	Dominique LAFFLY (dominique.laffly@univ-tlse2.fr)
Codirecteurs :	Yannick Le Nir et Astrid Jourdan, EISTI (École Internationale des Sciences du Traitement de l'Information)
Axes :	Axe 2 et Axe 3.

Le contexte

L'information géographique est essentielle à la recherche en sciences environnementales. Depuis les premiers satellites géostationnaires météorologiques (début des années 60) puis ceux à défilement d'observation de la Terre en orbite polaire (1972) c'est un véritable arsenal de plusieurs centaines de capteurs optiques et de *radar* qui acquièrent régulièrement voire en continue de l'information à des résolutions spatiales et géographiques variées. Outre ces images de télédétection, les bases de données éco-démo-socio-épidémio-géographiques se multiplient et s'affinent quotidiennement. On croule littéralement sous les données alors qu'auparavant on en manquait, "*With geospatial technology, we are no longer rich in data but poor in information*" (G. Renzhong, http://www.geospatialworld .net/). On touche ici la problématique généralisée des données massives (*Big Data)*.

Seule l'informatique est à même de stocker et gérer toutes ces informations. Selon les disciplines, les outils informatiques utilisés reposent sur des configurations simples à la distribution parallèle des calculs (pratiquée de longues dates quand ceux-ci s'avèrent très consommateurs des ressources des machines). C'est le domaine limité des *HPC* (*High Performance Computing* – les supers calculateurs, grille de calculs…) qui se réduit

principalement à quelques disciplines des sciences environnementales (météorologie, géologie structurale, océanographie), et reste inaccessible à la majorité des chercheurs, enseignants-chercheurs et étudiants. Les coûts d'acquisition et de maintenance, la rigidité des infrastructures et la limitation des accès des *HPC* expliquent principalement cet état de fait, tout comme le manque d'ingénieurs et techniciens susceptibles de les maîtriser au sein des laboratoires et des universités. Pour ceux qui y ont accès, les coûts engendrés sont aussi un élément qui limite les investigations scientifiques, le quota d'heures allouées pouvant très vite être atteint (tout particulièrement en configuration « grille de calculs », cf . le projet GéoBigData, Traitement d'images géographiques réparti et ordonnancement, porté par GEODE au CALMIP – Calcul en Midi-Pyrénées). Techniquement et sans entrer dans les détails on peut dire que la technologie *HPC* s'avère rapide mais très contraignante.

De plus, la distribution parallèle des calculs ne répond pas à tous les besoins inhérents à l'information géographique et particulièrement ceux de la mise à disponibilité des données sur l'internet à l'image de ce que proposa Google en 2004 en rachetant le logiciel commercial *Earth Viewer* (*Keyhole Inc*.) pour devenir *Google Earth* aujourd'hui téléchargé plus de 1 milliard de fois. C'est le *cloud computing* – rendu populaire par les sociétés Amazon et Google au début des années 2000 – qui offre l'accès aux « fontaines web ». Très schématiquement, le *cloud computing* permet d'instancier des machines virtuelles accessibles en réseau dans un schéma privé ou publique. C'est l'émergence du *cloud computing* qui permet de répondre à la contrainte généralisée du *Big Data* (pour résumer, quand il y a trop de données pour être traitées avec des solutions informatiques classiques). Plus lent mais beaucoup moins contraignant, moins coûteux que la solution HPC, le *cloud computing* peut se voir comme une structure pyramidale composée de trois niveaux :

- *Infrastructure as a Service (IaaS)* : permet de créer un ensemble de machines, soit pour simuler un *HPC*, soit pour créer d'autres environnements ;
- *Plateforme as a Service (PaaS)* : fournit des machines prêts à l'emploi ;
- *Software as a Service (SaaS)* : logiciel en ligne utilisable directement.

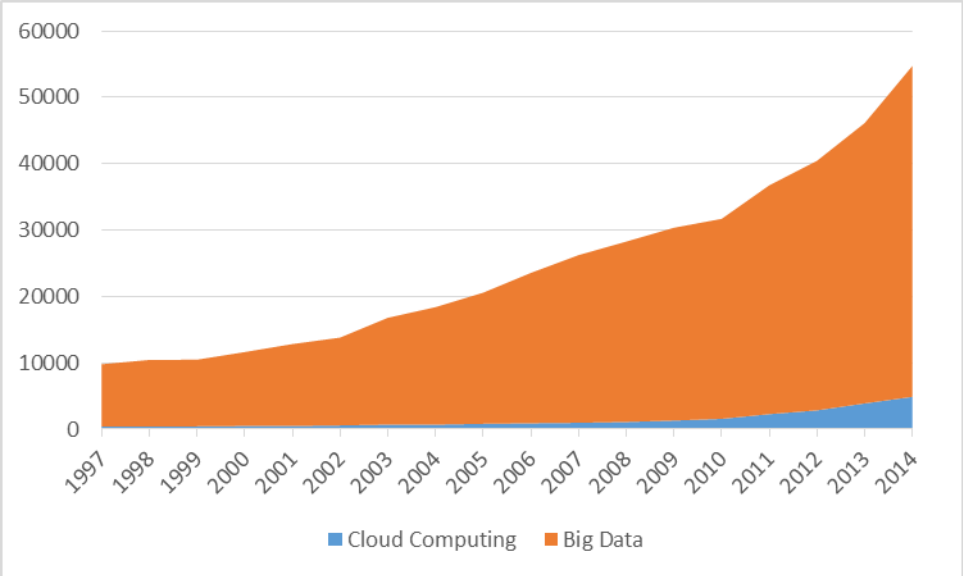

Figure 1 : évolution de l'occurrence des termes « *Big Data* » et « *Cloud Computing* » depuis 1997 dans Science Direct.

Très présent dans nos quotidiens – que ce soit par les offres commerciales de stockage et sauvegarde ou par les logiciels que l'on utilise via l'internet – le *cloud computing* l'est beaucoup moins dans le domaine des sciences environnementales. C'est un paradigme trop récent pour qu'il soit déployé à la hauteur de ses potentialités. Cependant, les séminaires, colloques et numéros spéciaux de revues scientifiques témoignent de la demande croissante de la communauté scientifique et universitaire (figure 1). Dans la même veine, les fournisseurs de logiciels de géomatique – payants ou libres de droit – proposent des solutions de plus en plus nombreuses fondées sur le *cloud computing* et ce depuis plusieurs années déjà (ArcGIS, QGIS, ERDAS, ENVI, R, MatLab…).

Le projet de thèse

La géographie ne peut ignorer ces nouveaux paradigmes que sont le *Big Data* et le *Cloud Computing* qui bouleversent littéralement les conceptions classiques aujourd'hui dépassées que nous avons de « l'informatique » et « l'information géographique ». La géographie, même si ce n'est pas encore visible pour tous les géographes, est aussi embarquée dans ce mouvement inexorable qui représente l'enjeu majeur du monde contemporain de l'information : comment, d'une part, stocker des données de plus en plus massives et diversifiées qui sont actualisées de plus en plus rapidement et, d'autre part, comment identifier ces données et y trouver un sens plus rapidement qu'elles ne s'accumulent ?

Du point de vue des données, de nombreux verrous offrent actuellement des sujets d'études passionnant en prise directe avec les problématiques géomatiques. À titre d'exemple, le stockage de l'ensemble des données qu'il est possible de récolter en flux quasi continu nécessite l'utilisation de formalismes spécifiques que l'on regroupe souvent sous l'appellation *NoSQL*. La découverte et l'acquisition de ces données est également un enjeu majeur qu'il est important de reconsidérer à l'heure des données connectées. Citons, par exemple, les données sémantiques qu'il est possible de récolter dynamiquement à travers une description dans un langage proche du langage naturel (gestion des synonymes, du multilingue, taxonomie des termes, ...) ainsi que les données ouvertes mise à disposition par un nombre d'acteurs en perpétuelle croissance (*NASA EOSDIS, applying semantic web technologies to earth science* ; *ESA, Easy Semantic Approximation with Explicit Semantic Analysis* ; *CNES, activities on semantic search* ; INSEE, espace RDF[1]…). Le volume, la diversité et le flux de ces données offrent, certes, de nouvelles perspectives de traitement et d'analyse, mais dépend de notre capacité à maîtriser des métriques aussi variées que leur normalisation, leur corrélation ou leur véracité. Dans cette thèse nous porterons un intérêt tout particulier à l'analyse des techniques mathématiques permettant d'extraire ces

[1] « *Resource Description Framework (RDF) est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions. Développé par le W3C, RDF est le langage de base du Web sémantique. L'une des syntaxes (ou sérialisations) de ce langage est RDF/XML. D'autres sérialisations de RDF sont apparues ensuite, cherchant à rendre la lecture plus compréhensible ; c'est le cas par exemple de Notation3 »*(Wikipédia). L'INSEE publie expérimentalement des données conformes aux standards du web sémantique, c'est-à-dire formatées en RDF et interrogeables par le langage de requête SPARQL.

métriques afin de proposer des modèles à la fois pertinents et efficaces utilisant des données géographiques au sein d'un environnement de type *cloud*.

L'objectif est double :

- mettre en œuvre des méthodes descriptives permettant de décrire et mieux comprendre cette masse de données (Analyses multivariées : Composantes Principales, Correspondances, Correspondances Multiples, Canoniques, Procrustéennes… ; *clustering* : K-means, Random Forest Trees, Hierachical...), notamment l'analyse de corrélations/liens entre les variables et leur regroupement afin de détecter les redondances et de diminuer la dimension du problème.
- définir des modèles prédictifs à l'aide de techniques d'apprentissage (régressions, réseaux de neurones, Bayes et Dempster-Schaffer, Monte-Carlo-Markovian-Chain...), permettant de prévoir le comportement futur des phénomènes observés depuis l'ensemble des données collectées.

Les méthodes statistiques ou de *machine learning*[2] pour atteindre ces objectifs sont connues depuis des décennies, mais leur utilisation dans le contexte des données massives nécessite des adaptations. Un des principaux problèmes vient des matrices de très grandes tailles générées par les données. Elles ont la particularité d'avoir un nombre très important de variables (10^3 à 10^8 colonnes) par rapport au nombre d'observations et elles nécessitent d'être stockées de façon répartie sur la nébuleuse du web. L'enjeu est alors de diminuer la taille de ces matrices avec comme contrainte un accès très limité aux données (3 à 5 lectures maximum). Pour atteindre cet objectif, on s'intéressera plus particulièrement aux méthodes fondées sur les matrices aléatoires récemment proposées par Halto *et al.* (2011) et Witten and Candès (2013)[3].

En s'appuyant sur des exemples thématiques de la discipline et pour lesquels des projets scientifiques sont en cours à GEODE – croissance urbaine et pollution (Hanoï), impact du changement climatique (Spitsberg, Pyrénées) pour n'en citer que deux – la thèse aura à cœur de développer et d'intégrer les outils du *big data* et du *cloud computing* dans le champ disciplinaire de la Géographie.

Projets en cours ou à venir en relation avec le sujet :

2105-2016 **CSAIDE**, *Compressive Sensing* appliquée à l'imagerie et aux dynamiques environnementales. PEPS INS2I/INSMI 2015, Fondements et Applications de la

[2] CSAIDE - *Compressive Sensing* appliquée à l'imagerie et aux dynamiques environnementales. PEPS INS2I/INSMI 2015, Fondements et Applications de la Science des Données (FaSciDo) Porteur : Dorothée NORMAND-CYROT (LSS-CNRS UMR 8506), Partenaires : J.-P. Barbot (ENSEA), A. Jourdan et Y. Le Nir (EISTI), D. Laffly (UT2 – GEODE). Une partie du financement obtenu (14 000 euros) viendra en appui de la thèse.

[3] R. Witten and E. J. Candès. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. To appear in **Algorithmica**. N. Halko, P.G. Martinsson, J. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." **SIAM Review**, 53(2), 2011, pp. 217-288

Science des Données (FaSciDo) Porteur : Dorothée NORMAND-CYROT (LSS-CNRS UMR 8506), Partenaires : J.-P. Barbot (ENSEA), A. Jourdan et Y. Le Nir (EISTI), D. Laffly (UT2 – GEODE).

2014-2016 **SOCRATE** – Soil Organic Carbon Research in Arctic Environement. Projet international (République de Corée, Norvège, France) financé par le KOPRI (Korea Polar Research Institute).

2015-2018 **TORUS** : Toward Open Resources Upon Services – Cloud Computing of Environmental Data
Soumission (fév. 2015) à ERASMUS+ Capacity Building (1 million d'€, 3 ans)
Réponse en juin 2015
Ce projet fédère l'université de Ferrara (Italie), la Vrije Universiteit Brussel (Belgique), l'université de Toulouse 2 et l'École Internationale des Sciences du Traitement de l'Information – EISTI (France), les universités de Nong Lam de Ho Chi Minh et Nationale du Vietnam ainsi que l'Institut de l'Agriculture et de l'Environnement à Hanoï (Vietnam), l'Asian Institute of Technology et l'université de Walailak (Thaïlande) avec plus de 40 chercheurs et enseignants-chercheurs.
Responsable : D. Laffly

2015-2017 ***Super_SOCRATE*** *: Analysing application of superspectral remote sensing sensors for Soil Organic Carbon Research in ArcTic Environment.*
Soumission (fév. 2015) au programme VENµS (*Vegetation and Environment monitoring on a New MicroSatellite*)
Accès gratuit aux images (1/jour pendant 2 années)
Réponse premier semestre 2015
Responsable : D. Laffly
Partenaires : Y. Le NIR (EISTI), L. Nilsen (Univ. of Tromsø), Y. J. Jung (KOPRI)

2015-2017 ***UPAC*** *: Urban Air Pollution and Cloud Computing*
Soumission (fév. 2015) au programme VENµS (*Vegetation and Environment monitoring on a New MicroSatellite*)
Accès gratuit aux images (1/jour pendant 2 années)
Réponse premier semestre 2015
Responsable : Nguyen Thi Nhat Thanh (Vietnam National University)
Partenaire : D. Laffly (UT2), Y. Le NIR (EISTI)

Profil du candidat

Le/La candidat(e) sera titulaire d'un master 2 en sciences environnementales où la géomatique est maîtrisée à un haut niveau (géographie, écologie, agronomie, sciences environnementales…), en mathématique appliquée, en informatique ou sera titulaire d'un diplôme d'ingénieur des grandes écoles.

Le/La candidat(e) doit maîtriser parfaitement la géomatique d'une manière générale et les langages de programmation impératifs, objets et fonctionnels. Une parfaite maîtrise de l'analyse de données, des statistiques, de l'algèbre linéaire et des techniques de « *machine learning* » est également nécessaire. Une connaissance du paradigme de la programmation distribuée (acteurs, *map/reduce*,…) serait un atout supplémentaire.

Maîtrise de l'anglais exigée.