

Big data et information géographique

La nature de l'information scientifique pour les sciences sociales et les humanités a été profondément transformée par le développement des technologies numériques, par la place qu'elles occupent désormais dans tous les compartiments de la vie quotidienne et par les gigantesques flux de données qu'elles génèrent. S'agissant de l'information géographique et de la manière de la mobiliser pour comprendre les espaces contemporains et les processus qui les structurent, la transformation est encore plus massive avec le développement du rôle joué par les systèmes d'information spatialisés et la diversité des formes de représentation numérique qui leur est associée.

La géographie et toutes les sciences sociales qui ont part à l'analyse des interactions espaces/sociétés sont en effet confrontées, comme de nombreuses autres sciences, à l'augmentation vertigineuse de l'information disponible. Le temps semble déjà lointain où il s'agissait seulement de rechercher et de construire l'information à partir d'un nombre limité de bases de données connues. Au *data mining* a fait place le *data deluge*. La géolocalisation et la généralisation de l'usage d'appareils connectés a transformé chaque utilisateur en un collecteur de données muni de plusieurs capteurs potentiels. Les réseaux sociaux génèrent des flux d'information dont l'usage apparaît pertinent pour de nombreux secteurs de la géographie et de l'analyse des territoires : modes de transports, géographie de la santé, géographie économique, flux illégaux de marchandises, inégalités... Face à ces flux de données pour la maîtrise desquels il faut inventer de nouveaux outils, l'enjeu pour l'analyse géographique est d'abord de savoir faire face à la quantité mais surtout de le faire au bénéfice d'une meilleure compréhension du monde qui nous entoure, et d'une meilleure maîtrise par les personnes concernées de la circulation et de l'usage des données.

Face à cet enjeu, les recherches menées dans le champ de l'information géographique et de l'analyse des espaces et des sociétés et dont les contributions réunies dans ce dossier donnent de passionnants exemples, se font en développant des interfaces fécondes autour de la modélisation et de l'analyse et de la représentation des systèmes complexes. Ces recherches interdisciplinaires sont menées souvent en étroite collaboration avec l'informatique et différentes ingénieries, notamment dans le domaine des risques, des mobilités et des dynamiques urbaines. Dans ces dialogues, les spécialistes de l'analyse géographique apportent des savoir-faire et une grande inventivité dans le domaine de la géomatique et de la représentation des données géographiques. La carte et l'atlas restent des outils plus que jamais pertinents à l'ère du *big data* mais ils se déploient dans une toute autre dimension. Des portails ouverts à tous publics permettent par exemple de représenter des phénomènes à une résolution spatiale et à une fréquence temporelle de plus en plus élevée. Au-delà des usagers experts, tels que les enseignants, les gestionnaires de l'espace, les personnes en charge de l'action publique, les agriculteur.trices, les militaires et les forces de l'ordre, l'ensemble des habitants connectés peut enrichir la connaissance de son milieu de vie en se connectant à des portails mis en place par les agences publiques. L'agence pour la protection de l'environnement aux États-Unis a ainsi mis en place un [portail](#) dédié aux inégalités environnemen-

tales qui permet de connaître les propriétés de son espace de vie sous le rapport des nuisances environnementales, afin d'exiger, au besoin, que la législation soit bien respectée. Plus généralement, les *big data* ouvrent l'espoir d'une amélioration de tous les services aux habitants des territoires notamment dans le cadre des villes intelligentes.

Parmi les horizons de recherche ouverts par les données massives, un des premiers est la recherche de dispositifs d'analyse et de modélisation parcimonieux. Modéliser, simuler et visualiser des interactions socio-spatiales fait croître considérablement la quantité des données à analyser. Il en résulte un besoin croissant en moyens technologiques d'analyse et de traitement de ces données qui stimule la créativité des modélisateurs. Il faudra en outre être capable de décrire mais aussi d'expliquer : les *big data* devront être maîtrisées pour mieux comprendre les processus socio-spatiaux. Un autre défi est celui de la collaboration du public dans des études à caractère scientifique. Même si la plupart des *smartphones* contiennent des capteurs utiles pour mesurer de nombreux comportements spatiaux, leur collecte et leur utilisation par les scientifiques dans des actions de *crowdsourcing* posent des questions qui relèvent des relations entre science et société : comment concevoir des protocoles d'acquisition de données qui tiennent compte de la réticence de participants à faire connaître leur localisation malgré un encadrement juridique qui garantit l'anonymat ? Comment engager dans la durée des participants bénévoles au sein d'études longitudinales, notamment avec le risque de l'exploitation d'un travail gratuit ? À quel point faut-il intégrer des non spécialistes dans l'interprétation et la discussion des résultats ?

Les *big data* et leur usage pour la compréhension des interactions espaces-sociétés posent aux disciplines des SHS et à leurs partenaires des défis techniques, méthodologiques, épistémologiques et éthiques passionnants. Les nouveaux horizons ouverts à la recherche obligent à renforcer encore la réflexivité dans la pratique de la recherche, en particulier au sujet du potentiel offert par l'intensification des interactions entre chercheur.e.s et habitants-usagers de l'espace et des territoires pour la production de connaissances, et de nouvelles formes de participation de tou.te.s aux décisions.

Pascal Marty, InSHS

Données massives et information géographique

Timothée Giraud est géomaticien au sein du Réseau interdisciplinaire pour l'aménagement du territoire européen (Riate, UMS2414, CNRS / Université Paris Diderot / Commissariat Général à l'Égalité des Territoires). Directeur du Collège international des sciences du territoire (CIST), Claude Grasland est membre de l'unité Géographie-cités (UMR8504, CNRS / Université Paris 1 Panthéon-Sorbonne / Université Paris Diderot). Ses recherches portent entre autres sur l'aménagement du territoire européen et les dynamiques économiques et démographiques de l'espace mondial. Également membre de l'unité Géographie-cités, Marianne Guérois étudie les dynamiques d'étalement urbain et les interactions entre morphologies urbaines et pratiques de mobilité individuelles, analysées dans une perspective comparative. Au sein du Pôle de recherche pour l'organisation et la diffusion de l'information géographique (Prodig, UMR8586, CNRS / Université Paris 1 Panthéon-Sorbonne / Université Paris Diderot / IRD), Malika Madelin mène des recherches en environnement (climatologie), sur l'information à l'échelle locale. Marta Severo, enfin, s'intéresse aux méthodes numériques pour les sciences sociales et les représentations de l'espace sur Internet, au sein du laboratoire Dicen-IDF (Dispositifs d'Information et de Communication à l'Ère Numérique – Ile-de-France, CNAM / UPEM / Université Paris Ouest Nanterre).

Introduction

Au-delà des effets de modes qui conduisent aujourd'hui à l'emploi obligé de *buzzwords* tels que *Digital Humanities* ou *Big Data* dans tout appel à projet qui se respecte... de nombreux chercheurs en sciences sociales, en informatique et en linguistique ont engagé depuis plusieurs années des recherches approfondies et cumulatives sur l'analyse des données massives issues du Web, des réseaux sociaux et plus généralement de l'ensemble des gisements de données ouvertes.

À la croisée de plusieurs disciplines, les spécialistes des sciences de l'espace¹ et des sciences territoriales² proposent un point de vue original sur l'exploitation de ces nouveaux gisements de données en privilégiant l'étude de leur contenu en termes d'information géographique. Plus précisément, ils développent des outils et des concepts permettant d'analyser à la fois des coordonnées spatiales de localisation de type latitude-longitude et des attributs territoriaux de nature sémantique désignant des lieux aux contours plus imprécis tels que des villes ou des pays.

Sans prétendre épuiser la richesse des travaux menés actuellement dans de très nombreuses équipes de recherche, le présent article souhaite illustrer la diversité des approches à l'aide d'une série d'exemples tirés de travaux du Collège international des sciences du territoire (CIST) et des équipes de recherche qui en sont membres.

Traces numériques des plateformes de l'économie « collaborative » : l'exemple des données AirBnb

Les données *AirBnb* issues des plateformes Internet de location touristique entre particuliers offrent un exemple riche de traces numériques du Web 2.0 détournées en sources d'information géographique pour l'analyse des dynamiques métropolitaines. Les enjeux scientifiques, voire politiques, liés à ces informations sont en effet nombreux et ont déjà suscité plusieurs études portant sur le développement de cette offre d'hébergement, qu'elle appartienne à la sphère de l'économie touristique collaborative ou s'inscrive dans de nouvelles niches de spéculation immobilière³.



Qu'est-ce que le CIST ?

La mission du Collège international des sciences du territoire est de promouvoir, par ses activités, la constitution des sciences territoriales, associant étroitement la théorisation et la pratique, ainsi que l'exploration de l'information territoriale et de son impact sur la vie en société. Il s'appuie pour cela sur une communauté pluridisciplinaire de chercheurs et ingénieurs des 24 équipes de recherche qui le constituent, engagés dans ses 9 axes scientifiques, dont les axes Information territoriale locale et Médias et territoires, plus particulièrement impliqués dans des recherches mobilisant les *Big Data* (projets Géomédia, Grandes métropoles...). Créé en tant que GIS en 2010, le CIST devrait se transformer en Fédération de Recherche en 2017.

[En savoir plus](#)

D'un point de vue méthodologique, l'intérêt suscité par ces données tient tout d'abord à leur exhaustivité (*a priori* tous les biens de ce marché sont renseignés et localisés), à leur haute résolution temporelle (exploitable pour peu qu'un archivage permette d'en garder la mémoire) et spatiale⁴. De plus, dans un contexte où l'accès aux données immobilières localisées à l'adresse s'avère en général très coûteux, les données *AirBnb* sont facilement accessibles et peuvent être obtenues soit directement *via* l'extraction des données de la plateforme, soit par l'intermédiaire du site indépendant *Inside AirBnb* qui facilite le téléchargement d'un grand nombre d'informations pour les centres-villes d'une quarantaine de métropoles dans le monde, avec parfois plusieurs enregistrements temporels. Enfin, ces données se caractérisent par une grande richesse sémantique, aussi bien en termes de description qualitative et quantitative des caractéristiques du bien (type de location, capacité d'accueil, prix de la nuitée, propriétaires multiples ou non...) que de commentaires associés aux locations.

1. Goodchild M. F., Janelle D. G. (dir.). 2004, *Spatially integrated social science*, Oxford University Press.

2. Beckouche P., Grasland C., Guérin-Pace F. & Moisseron J. Y. (dir.). 2012, *Fonder les sciences du territoire*. Karthala.

3. Voir à ce sujet : Gutierrez J., Garcia-Palomares J. C., Romanillos G., Salas Olmedo M. H. 2016, « Airbnb in tourist cities: comparing spatial patterns of hotels and peer-to-peer accommodation », arXiv:1606.07138 ; Quattrone G., Proserpio D., Quercia D., Capra L., Musolesi M., 2016, « Who benefits from the « sharing » economy of AirBnb ? », *Proceedings of the 25th International Conference on World Wide Web* ; [travaux du groupe de recherche Net\(h\)no-graphies](#).

4. Le fait que l'information diffusée soit ponctuelle ne doit cependant pas faire illusion sur la précision absolue de la localisation qui est « floutée » à 150 mètres près pour des raisons de confidentialité. Les motivations de ce floutage font l'objet de plusieurs interprétations : protection des biens ou partage restreint d'une information soumise à des contrôles publics ?

Le projet Grandes métropoles

Le projet *Grandes métropoles* du CIST vise à constituer une plateforme d'échanges autour des défis théoriques et méthodologiques soulevés par le croisement de données locales de plus en plus foisonnantes (ouverture des données publiques, multiplication de données localisées *via* les plateformes Web et les médias sociaux, diffusion de capteurs individuels — de pollution, de température...). Le projet a d'abord effectué un travail préliminaire d'harmonisation des périmètres, des maillages territoriaux et des sources statistiques utilisables pour comparer les métropoles retenues (initialement Paris, Chicago et Mexico). Puis, des ateliers ont été organisés dans le cadre de ce projet, qui permettent d'avancer sur différents fronts, à la fois théoriques et méthodologiques, pour poser *in fine* la question des connaissances originales apportées par cette profusion d'information et de leur interopérabilité.

Dans le cadre du projet *Grandes métropoles* du CIST, plusieurs pistes d'analyse territoriale du phénomène *AirBnb* ont été expérimentées à partir de jeux de données disponibles sur le site *Inside AirBnb* : les informations sur les caractéristiques des biens loués ont ainsi permis d'esquisser un tableau de la diffusion de la présence de ce service dans plusieurs grandes métropoles (Figure 1). Celle-ci, reconstituée à partir de la date du premier commentaire⁵ associé à chaque hébergement, met en valeur la croissance exponentielle de l'offre et l'importance des cycles saisonniers. Dans Paris, l'offre *AirBnb* présente des spécificités très nettes en comparaison d'autres centres de grandes métropoles : en particulier, on y dénombre une part très importante de logements loués en entier (86 %), mais une très faible proportion d'hôtes « multipropriétaires » (7 %). Cette offre se concentre principalement autour de Montmartre et dans le 3^e arrondissement (Figure 2), avec une répartition spatiale des prix des locations similaire à celle d'autres biens immobiliers (centre et ouest de Paris).

	Barcelone	Berlin	Chicago	Londres	New York	Paris	San Francisco
Population (millions d'hab.)	1,6	3,5	2,7	8,4	8,6	2,2	0,9
Superficie (en km ²)	100	890	590	1 570	790	105	120
Offre							
Nombre de logements	17 369	15 373	5 147	49 348	40 227	52 725	8 619
% avec commentaires	82,5 %	79,5 %	80,5 %	70,3 %	77,6 %	72,5 %	74,2 %
% avec disponibilité > 1/2 an	61,3 %	63,1 %	67,8 %	50,3 %	40,1 %	50,6 %	41,4 %
Logement entier	50,4%	60,7%	56,9%	51,2%	49,5%	85,7%	57,6%
Chambre privée	48,4%	38,0%	38,3%	47,3%	47,0%	13,3%	36,9%
Chambre partagée	1,2%	1,3%	4,8%	1,4%	3,4%	1,0%	5,5%
Hôtes							
Nombre	10 112	12 405	3 848	34 678	33 582	44 874	6 705
% avec plusieurs locations	27,0 %	8,3 %	15,6 %	16,0 %	12,0 %	7,4 %	13,7 %
Prix							
Moyen	25,8	23,0	50,2	33,5	58,0	32,6	85,4
Médian	21,0	20,0	41,7	27,5	47,5	28,0	62,5
Date des données (sur <i>InsideAirBnb.com</i>)	12/16	10/15	10/15	10/16	12/16	07/16	07/16

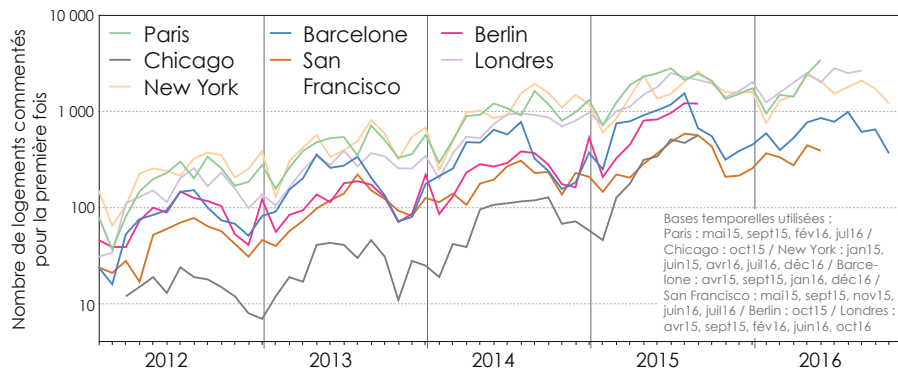
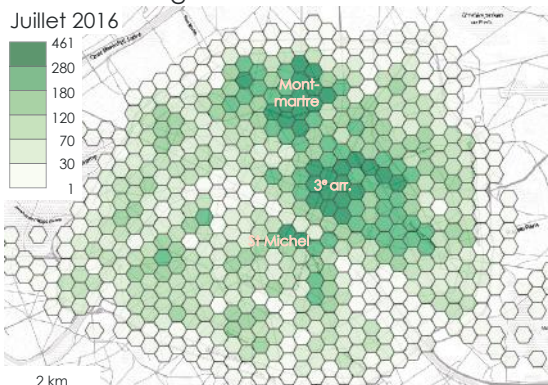


Figure 1 : Principales caractéristiques de l'offre *AirBnb* dans sept métropoles (communes centres) et évolution du nombre de logements commentés pour la première fois, entre 2012 et 2016

Densité de logements AirBnb



Moyenne par logement du prix / personne

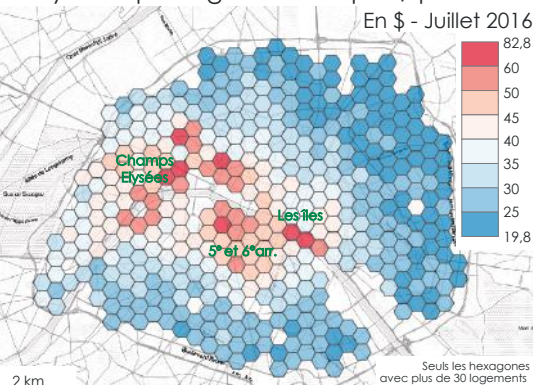


Figure 2 : Densité de l'offre et prix des locations *AirBnb* à Paris, en juillet 2016

Des fils RSS aux événements géomédiatiques

Le projet *ANR Géomédia*, qui vient juste de s'achever, se proposait l'objectif ambitieux de discuter l'existence d'un espace public mondial à travers la collecte et l'analyse d'un grand corpus de nouvelles de presse internationales. L'intérêt pour ce type d'objet était motivé, entre autres, par la disponibilité massive d'un nouveau type de données médiatiques, le fil RSS. Le fil RSS est un fichier XML, mis à disposition librement par un site Web pour diffuser les dernières nouvelles publiées en ligne. Cet outil de diffusion des actualités s'est largement répandu sur les sites Web des journaux de presse écrite qui l'ont adopté pour communiquer en temps réel. Par conséquent, ces données numériques offrent au/à la chercheur/se en sciences sociales une source d'information alternative aux bases de données médiatiques payantes comme Factiva ou Europresse : non seulement les fils RSS sont librement disponibles sur Internet, mais ils peuvent être collectés, archivés et analysés grâce à leur structure standardisée (titre, date, résumé).

Pour étudier les relations internationales au prisme des médias, l'équipe *Géomédia* a stocké le contenu des fils RSS associés aux articles publiés par une centaine de quotidiens dans différents pays. L'analyse a privilégié les actualités internationales (publiées dans le fil RSS international du

journal), c'est-à-dire les informations ayant trait à des événements se produisant en dehors des frontières du pays. L'analyse des actualités collectées dans cette base a permis de déduire deux types d'information portant respectivement sur les flux entre États et sur les événements internationaux.

5. Cet indicateur est discutable, la pratique du commentaire n'étant pas aussi forte d'un pays à l'autre.

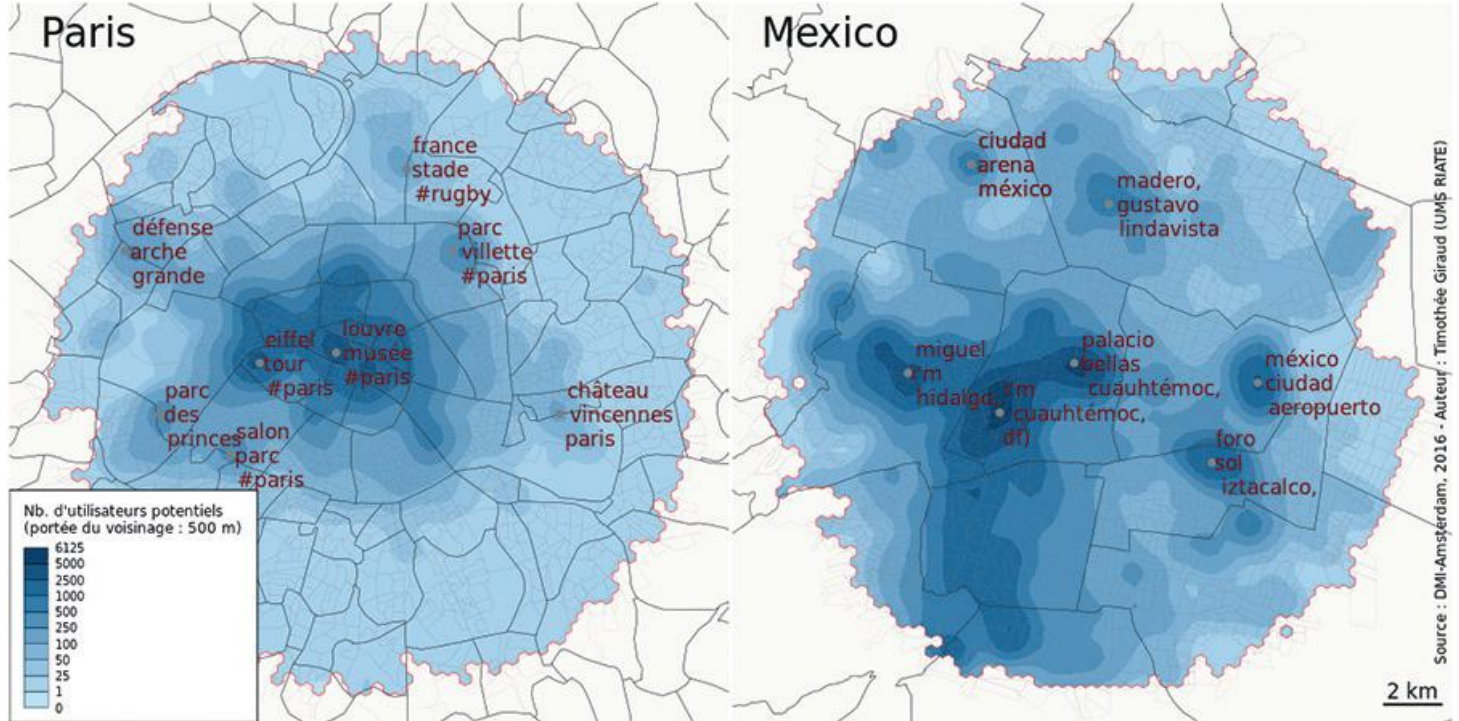


Figure 4 : Densités d'utilisateurs à Paris et Mexico (février 2016)

Sur cette figure, sont mises en valeur dans un périmètre comparable pour Paris et Mexico (cœur central des aires urbaines cumulant 5 millions d'habitants) les zones où l'on trouve le plus grand nombre d'utilisateurs actifs dans un rayon de 500 m. Les pics d'utilisateurs sont associés à des équipements touristiques, culturels ou sportifs majeurs des deux métropoles, comme le révèlent les principaux mots-clés extraits du contenu des messages. Cette méthode d'analyse croisée des localisations et du texte permet aussi, en sens inverse, de détecter les lieux où certains sujets sont abordés.

Une limite importante concerne la représentativité de ces données : les utilisateurs de *Twitter* sont de plus en plus nombreux, mais les *tweets* géolocalisés (qui possèdent des coordonnées de latitude et de longitude) ne représentent que 1 à 2 % du total des *tweets* publiés⁷. On peut de plus faire l'hypothèse que la sous-population qui laisse accessibles ses données de géolocalisation a un profil spécifique par rapport au reste de la population (minorité d'utilisateurs jeunes). Enfin, la majorité des utilisateurs localisables publiée très peu, tandis qu'une minorité (comptant la présence de « robots » programmés qui publient des messages automatiques — spams ou publicités pour la plupart — en grande quantité) s'avère très active. C'est pourquoi il est souvent préférable de raisonner à partir d'un indicateur de densité d'utilisateurs (Figure 4) plutôt que de densité de *tweets*.

Même si ces résultats sont encourageants, la faible représentativité des *tweets* géolocalisés spatialement peut conduire à analyser les informations territoriales éventuellement présentes dans le message même. La tâche n'est pas du tout anodine pour plusieurs raisons. D'abord, *Twitter* n'est pas une plateforme strictement liée au territoire. Au contraire, les échanges sous forme de *tweets* peuvent concerner n'importe quel sujet et, dans la plupart des cas, ils jouent surtout une fonction phatique, c'est-à-dire qu'ils servent pour construire et alimenter une relation. Par ailleurs, ces dernières années, les utilisateurs les plus actifs, qu'ils soient des personnes ou des robots, sont engagés dans une guerre de visibilité et de réputation. Enfin, l'analyse du contenu pose d'importants défis linguistiques parce que les *twittos* peuvent utiliser des *hashtags* ou des abréviations qui ne sont pas faciles à détecter avec des techniques d'analyse textuelle classiques. Tout cela produit un effet de bruit significatif qui rend difficiles l'identification univoque des localisations dans le texte et le repérage des informations qui concernent les territoires.

Malgré ces obstacles, les travaux réalisés dans le cadre du CIST ont permis de développer un certain nombre de techniques qui facilitent l'extraction d'informations utiles pour l'analyse territoriale à partir du contenu des *tweets*. Il est par exemple possible de

conduire l'analyse du contenu qualitatif et/ou quantitatif sur des sous-échantillons de *tweets* localisés dans des endroits-clés pour comprendre la raison d'une densité plus importante d'activité. Une deuxième approche qui a donné des résultats encourageants consiste à sélectionner de manière qualitative un groupe de *twittos* qui peuvent être considérés comme des « influenceurs » du territoire qu'on entend étudier. Enfin, nous avons expérimenté l'intérêt de reproduire la méthodologie développée dans le projet Géomédia pour le fil RSS sur les *tweets*, c'est-à-dire nous avons reconstruit les flux entre pays à partir des pays mentionnés dans les *tweets* géolocalisés dans un certain État.

Conclusion

À l'heure où l'information circule plus rapidement que jamais, il n'est sans doute pas inutile de plaider résolument en faveur d'une science lente qui prend le temps de construire et analyser précisément les données massives à la lumière de concepts et d'hypothèses explicites. Cette construction sera d'autant plus performante à long terme qu'elle aura pris le temps du dialogue entre plusieurs disciplines apportant des éclairages complémentaires : géographie, informatique, sciences de la communication, linguistique...

contact&info
 ► Marion Gentilhomme,
 CIST
marion.gentilhomme@gis-cist.fr

7. Gerlitz C. & Rieder B. 2013, *Mining One Percent of Twitter: Collections, Baselines, Sampling*, in *MIC Journal*, 16(2) ; Severo M., Giraud T. & Pecout H. 2015, *Twitter data for urban policy making: an analysis on four European cities*, in *Handbook of Twitter for Research*, EMLYON Press.

la lettre de l'InSHS

- ▶ **Directeur de la publication** Patrice Bourdelais
- ▶ **Directrice de la rédaction** Marie Gaille
- ▶ **Responsable éditoriale** Armelle Leclerc armelle.leclerc@cnrs-dir.fr
- ▶ **Conception graphique** Sandrine Clérisse & Bruno Roulet, Secteur de l'imprimé PMA
- ▶ **Graphisme Bandeau** Valérie Pierre, direction de la Communication CNRS
- ▶ **Crédits images Bandeau**
© Photothèque du CNRS / Hervé Théry, Émilie Maj, Caroline Rose, Kaksonen
- ▶ **Pour consulter la lettre en ligne**
www.cnrs.fr/inshs/Lettres-information-INSHS/lettres-informationINSHS.htm
- ▶ **S'abonner / se désabonner**
- ▶ **Pour accéder aux autres actualités de l'INSHS**
www.cnrs.fr/inshs

Institut des sciences humaines et sociales CNRS

• 3 rue Michel-Ange 75794 Paris cedex 16 •

ISSN : 2272-0243