Soft data

Essai d'une nouvelle définition des données pour les études territoriales

Marta Severo, Alberto Romele

Introduction

L'étiquette de «*soft data*» que nous utilisons dans le titre de ce chapitre constitue un essai provisoire de réponse à une sensation d'insatisfaction terminologique émergeant des travaux de l'axe scientifique Médias et territoire du Collège international des sciences du territoire (CIST, Paris). Ces dernières années, la recherche et l'administration publique se trouvent confrontées à un phénomène qui semble pouvoir renouveler la façon de gérer et d'étudier le territoire. Nous nous référons au déluge de nouveaux types de données générées par les technologies numériques, notamment par l'Internet, qui se proposent comme nouvelle source d'information sur le territoire [Hey et Trefethen, 2003].

Ces données représentent en même temps une réalité, un désir et un besoin. Une réalité, car une des caractéristiques fondamentales des technologies numériques est sûrement leur pouvoir de générer des traces. Toute action qui les traverse laisse des traces qui peuvent être facilement récoltées et traitées. Un désir, car ces données, grâce à leur valeur de trace de l'action sociale, promettent de répondre à un désir généralisé des politiques publiques. Aujourd'hui le décideur public sent la nécessité de trouver de nouvelles sources de données concernant la vie collective, des informations disponibles en temps réel et produites avec des dynamiques *bottom-up* qui puissent l'aider à rendre son action plus efficace. Dans ce sens, il suffit de penser au nombre de collectivités territoriales qui se sont lancées dans des projets de *big* et d'*open data*. Enfin, ces données représentent également un besoin pour des méthodes et des outils plus adaptés à leur traitement.

Si les projets qui incluent des données provenant du Web dans l'étude de la société se sont multipliés ces dernières années, nous pouvons noter comment les auteurs les plus enthousiastes envers les données numériques se trouvent aujourd'hui à faire un pas en arrière et à assumer des positions bien plus prudentes, voire parfois pessimistes. Prenons le cas de David Lazer qui en 2009, dans un article collectif [Lazer *et al.*, 2009], arrive à la définition de la science sociale computationnelle (*Computational social science*). Il observe : « Les nouvelles technologies, tels que la surveillance vidéo, l'email, les badges nominatifs "intelligents", offrent une image instant par instant des interactions sur de longues périodes, en fournissant des informations sur la structure et le contenu des relations ». Il continue : « Internet offre une canal entièrement différent pour comprendre ce que les gens sont en

train de dire, et comment ils se connectent» et il conclut : «En bref, une science sociale computationnelle est en train d'émerger, qui tire profit de la capacité de collecter et analyser les données avec une ampleur, une profondeur et une échelle sans précédents». Le même David Lazer, dans un article plus récent [Lazer *et al.*, 2014], a une attitude bien plus modérée envers les données numériques et souligne la nécessité de prendre des précautions dans ce type d'études, notamment quand on parle de *big data*. Il est intéressant que l'auteur attaque notamment l'outil de Google étudiant l'évolution de la grippe[1], qui peut être sans aucun doute considéré comme un parmi les symboles de cette nouvelle vague des méthodes numériques pour les sciences sociales [Ginsberg *et al.*, 2009].

De même, Venturini *et al.* [2014] identifient trois malentendus à l'usage des méthodes numériques en sciences sociales et précisent toute une série de précautions à l'usage des traces numériques dans ce type d'études qui contredisent paradoxalement certains travaux antérieurs des même auteurs au Médialab de Sciences Po [Venturini et Latour, 2012 ; Latour *et al.*, 2013]. Ce mouvement vers l'arrière est certainement représentatif d'un besoin d'une réflexion plus approfondie sur ces données et d'une définition de méthodes plus adéquates aux différents contextes d'analyse. Noortje Marres a été une des pionnières dans cette direction en mettant en avant plusieurs limites des outils qu'elle-même avait utilisés précédemment [Marres et Weltevrede, 2013]. En se concentrant sur la question de l'intervention des technologies numériques dans les sciences sociales, l'autrice développe le concept de «redistribution des méthodes» [Marres, 2012]. A travers cette expression, reprise des STS [Latour, 1988 ; Rheinberger, 1997 ; Whatmore, 2009], elle souligne comment la numérisation rend nécessaire la participation de nombreux acteurs à la définition des méthodes utilisées dans la recherche : «Une approche redistributive à la recherche sociale redéfinit les méthodes comme impliquant la combinaison et la coordination de compétences différentes : classification, conception visuelle, analyse automatisée, et ainsi de suite. Derrière les débats sur la faillibilité des données générées par les sujets de recherche et le "désordre" des contenus en ligne auto-indexés, se trouve un débat sur la redistribution des méthodes entre chercheurs, dispositifs, informations et utilisateurs dans les environnements en ligne» [Marres, 2012 : 161]. D'ailleurs, nous trouvons un autre exemple de cette attitude dans l'article de Marres et Gerlitz publié ci-dessus.

Sans vouloir prendre position entre optimistes et pessimistes [Woogar, 2002], cet article veut s'interroger sur ce besoin définitoire et contribuer à la réflexion générale sur ces méthodes et ces données. L'objectif de ce texte est notamment de contribuer à la prise de distance par rapport à la pratique pour revenir vers une réflexion théorique sur l'usage des données numériques dans la recherche et dans les politiques publiques. Nous n'avons pas la prétention ici de nous confronter aux usages des *data* dans tout terrain de manière indistincte, mais au contraire

1 http://www.google.org/flutrends.

notre réflexion sera basée sur des expériences d'usage de données dans les études territoriales [Delaney, 2005]. En introduisant la catégorie de *soft data*, nous ne voulons pas théoriser l'existence d'un type particulier de données. Les *soft data* ne sont pas une alternative aux *hard data*, aux *big data* ou aux *open data*. Notre intention est plutôt de suggérer une manière différente de regarder les données disponibles sur Internet. A notre avis, la notion de *soft data* est plus inclusive, car elle arrive à rendre compte de certaines données qui ne sont pas facilement catégorisables dans les catégories déjà existantes. Cela est le cas par exemple des tweets, souvent utilisés dans des analyses de phénomènes territoriaux [Wilken, 2014 ; Romele et Severo, 2014] ou les *checks-in* Facebook qui promettent de nous faire découvrir de nouvelles géographies relationnelles [Vienne *et al.*, 2014]. Ces données, qui sûrement ne sont pas des *open data*[2]*, peuvent fournir des informations intéressantes même si elles n'ont pas non plus toujours les caractéristiques des big data.*

L'analyse sera structurée en trois parties. Dans une première partie, nous reprendrons les définitions principales proposées ces dernières années pour parler des données du Web. Nous poserons l'accent en particulier sur les termes de *data*, *big data* et d'*open data*, ces deux dernières étant les plus utilisées dans le contexte de l'analyse territoriale. Dans une deuxième partie, nous analyserons plus précisément les caractéristiques des données employées dans ce type d'études. En premier lieu, nous observerons que leur succès est dû principalement à l'insatisfaction générée par celles qu'on appellera *hard data*, c'est-à-dire les données créées par les fournisseurs des données traditionnels (comme Eurostat ou l'Insee) généralement employées dans les études territoriales. En deuxième lieu, nous poserons l'accent sur la caractéristique principale des données employées pour étudier l'espace, la géolocalisation, mais nous chercherons à montrer comme l'accent sur une telle caractéristique a rendu difficile l'identification du nouvel apport des données du Web dans les études territoriales. Cela nous portera à notre troisième partie, dans laquelle nous proposerons une nouvelle définition de *soft data*. Dans cette dernière partie, nous préciserons ainsi la nouveauté de notre définition en rapport aux usages faits précédemment du terme *soft data* et déclinerons les caractéristiques de ces données qui montrent l'imperfection des définitions déjà existantes et justifient la nécessité d'une nouvelle définition.

De nouvelles définitions pour de nouvelles données

Data

Avant d'introduire la nouvelle catégorie de *soft data*, dans cette partie nous allons rendre compte des catégories déjà existantes. Avant de s'orienter sur les adjectifs utilisés – « *big* », « *open* », etc. – il est nécessaire de préciser le sens du terme central « *data* ».

2 À propos des enjeux légaux liés à l'usage de Twitter, voir Beurskens [2014].

Récemment, Venturini *et al.* [2014 : 3] ont suggéré une distinction méthodologique entre traces numériques et données numériques : « "traces numériques" se réfère à toute collection de bit stocké dans la mémoire d'un dispositif numérique (un ordinateur, dans la plupart des cas) comme un résultat de l'implémentation délibérée du système de traçabilité. Les "données numériques" sont par contre la collection organisée d'informations, produite à partir des traces numériques à travers le travail du chercheur qui les sélectionne, les nettoie et les exploite dans un étude spécifique ». Cette distinction est très intéressante, cependant elle n'a aucune valeur du point de vue étymologique, historique ou théorique.

Etymologiquement, le terme « *data* » est le pluriel de *datum*, terme latin qui dérive du verbe *dare*, signifiant « donner ». Du point de vue étymologique donc, « *data* » indique qu'il s'agit d'un élément donné ou accordé ou, encore mieux, qui se donne ou s'accorde à quelqu'un en l'état. Les data ne sont pas travaillées par un individu – par son regard, sa conscience, son intellect, etc. – ou par un groupe d'individus, mais se donnent à cet individu ou à ce groupe dans leur évidence. De ce point de vue, les données contrastent avec les faits, car *factum* est le participe passé du verbe latin *facere* qui veut dire « faire », « exécuter », « accomplir ».

Historiquement, c'est précisément le caractère d'évidence qui est au cœur du mot « *data* » et c'est aussi la raison de l'adoption de ce mot par l'anglais[3]. Jusqu'au XVII^e^ siècle, le terme était tout simplement absent en anglais. La première occurrence signalée par l'*Oxford English Dictionary* date de 1646, dans un traité théologique qui parle d'un « amas (*heap*) de données » [Rosenberg, 2013 : 18]. En général, pendant tout le XVII^e^ siècle le mot a été utilisé dans le sens technique qu'il avait chez Euclide, pour indiquer des quantités données dans des problèmes mathématiques, contrairement aux *quaesita*, qui étaient des quantités *cherchées*. En théologie, la parole signifiait les vérités scripturaires données par Dieu et donc incontestables. En philosophie, les *data* étaient les principes qui ne pouvaient pas être disputés, à cause de leur auto-évidence ou par convention. Au cours du XVIII^e^ siècle, le terme devient plutôt commun en anglais, au delà des domaines spécifiques des mathématiques, de la théologie et de la philosophie, mais il assume aussi un sens nouveau, presque inverse. Les *data* ne sont plus les bases indiscutées d'une argumentation, mais les faits mis en évidence suite à une expérience : « Il est devenu habituel de penser aux données comme le résultat d'une investigation plutôt que sa prémisse » [Rosenberg 2013 : 33]. Il est intéressant d'observer comment même dans cette nouvelle acception, qui d'ailleurs est proche du sens actuel du mot, la prétention d'immédiateté reste intacte. Bien que résultants d'un processus d'investigation, les données se révèlent comme quelque chose qui était déjà là et qu'il s'agissait tout simplement de dévoiler.

3 Dans ce contexte, bien que nous utiliserons quelque fois le français « données », c'est le terme anglais « *data* » qui nous intéresse.

D'ailleurs, ce n'est pas sans raison que le philosophe Martin Heidegger traduisait le terme *aletheia*, vérité en grec, par « dévoilement ».

Or, si nous approchons le terme « *data* » d'un point de vue plus théorique, nous assistons à une sorte de renversement de perspective. Les données semblent être non pas quelque chose de *donné*, mais plutôt quelque chose de *capturé :* « Dans l'usage général, *data* se réfère aux éléments qui sont *pris ;* extraits à travers observations, computations, expériences et enregistrements […]. Techniquement donc, ce que nous comprenons comme *data* sont en réalité *capta* […] ; les unités des données qui on été sélectionnées et récoltées à partir de la somme de toutes données potentielles » [Kitchin, 2014a : chapitre 1, paragraphe *What are data?*]. Au fond, les sciences « dures » comme les sciences « molles » n'ont jamais à voir avec un référent qui se donne. Ce que Latour et Woolgar soulignaient à propos de la construction de faits dans leur *Laboratory Life*, nous pouvons le dire aussi des données : « Un fait devient tel quand il perd toute qualification temporelle et devient incorporé dans un grand corps de connaissance fondé sur d'autres » [Latour et Woolgar, 1986 : 106]. Les faits ne sont rien d'autre que des métaphores mortes, « mythologie blanche », pour utiliser l'expression par laquelle Derrida indiquait le processus à travers lequel la culture occidentale finissait par prendre « sa propre mythologie, l'indo-européenne, son *logos*, c'est-à-dire le *mythos* de son idiome, pour la forme universelle de ce qu'il doit vouloir encore appeler la Raison » [Derrida, 1971 : 254].

Le philosophe de l'information Luciano Floridi refuse les interprétations informationnelle et computationnelle des *data*, précisément à cause de leur tendance à comprendre les données à partir de quelque chose d'autre – l'information dans le premier cas, le support numérique dans le deuxième. Les *data* ont plutôt une nature « diaphorique », dans le sens que le terme « données » indique tout simplement une manque d'uniformité – *diaphora* en grecque signifie « différence ». Les *data* sont alors neutres d'un point de vue taxonomique, car elles sont des entités relationnelles, typologiques, dans la mesure où on peut distinguer entre données primaires, secondaires, *metadata*, etc. [Floridi, 2008 : 7-10]. Rosenberg nous semble à ce propos encore plus clair, en affirmant que les *data* n'ont ni une nature ontologique, qui serait celle des faits, ni épistémologique, celle des évidences, mais une nature rhétorique : « Les faits sont ontologiques, l'évidence est épistémologique, *data* est réthorique. […] L'existence d'une donnée est indépendante de toute considération sur la vérité ontologique correspondante. Quand un fait est prouvé être faux, il cesse d'être un fait. Une données fausse est quand même une donnée » [Rosenberg, 2013 : 18]. Un regard théorique sur les *data* révèle en somme leur nature « molle » du point de vue ontologique. En reprenant la fameuse distinction de Frege, nous pourrions dire que, considérées par elles-mêmes, les données ont un « sens » (*Sinn*) mais n'ont pas de « référence » (*Bedeutung*), au moins immédiate. Cette dernière dépend plutôt de qui les utilise, dans certains contextes et selon des intérêts spécifiques.

Big data

Selon Steve Lohr [Lohr, 2013], la première trace du terme « *big data* » remonte à un article d'Erik Larson, publié en 1989 dans *Harper Magazine* puis republié par le *Washington Post*. D'habitude, on reconnaît l'origine du terme dans la présentation de John Mashey, déjà chercheur en chef à la Silicon Graphics, intitulée « Big data and the Next Wave Infrastress » [Mashey, 1998], dans laquelle l'auteur se démontre conscient du phénomène. Dans le contexte académique, la première occurrence du terme remonte à l'ouvrage *Predictive Data Mining: A Practical Guide* [Weiss et Indurkhia, 1998] où, dans les toutes premières lignes de la préface, les deux auteurs affirment que « à l'époque d'Internet, des intranet et des entrepôts de données, les paradigmes fondamentaux de l'analyse classique de données sont mûr pour le changement. […] De très grandes collections de données […] sont aujourd'hui compilées dans des entrepôts de données centralisés, permettant aux analystes d'utiliser des méthodes puissantes pour examiner les données d'une manière plus détaillée. En théorie, les "*big data*" peuvent conduire à des conclusions beaucoup plus fortes pour les applications de fouille de données, mais en pratique beaucoup de problèmes surgissent » [Weiss et Indurkhia, 1998 : xi]. En 2003, Diebold publie un article intitulé « Big data Dynamics Factor Models for Macroeconomic Measurement and Forecasting ». Comme l'affirme l'auteur lui-même dans une intervention qui suit, il s'agit de la première référence au terme « *big data* » dans les domaines de la statistique, de l'économétrie, etc. [Diebold, 2012]. Néanmoins, il paraît que le vrai succès du terme arrive seulement en 2008. Dans un contexte de vulgarisation, Anderson publie son fameux article « The end of theory: the Data Deluge Makes the Scientific Method Obsolete », en annonçant l'entrée dans l'âge du petabyte : « Les petabyte nous permettent de dire : "la corrélation est suffisante". Nous pouvons arrêter de chercher les modèles. Nous pouvons analyser les données sans hypothèses sur ce qu'elles devraient montrer. Nous pouvons jeter les chiffres dans les plus grandes clusters d'ordinateurs que le monde ait jamais vu et laisser les algorithmes statistiques trouver des motifs (*patterns*) que la science ne peut pas trouver ». Dans le contexte scientifique, Bryant, Kats et Lazowska, membres du *Computer Community Consortium*, ont dédouané le terme avec leur article intitulé « Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society ».

La définition la plus classique des *big data* est celle des 3V (volume, vélocité et variété) selon laquelle les *big data* sont d'énormes bases de données, produites en temps quasi réel, structurées ou non-structurées, et souvent temporellement et spatialement référencées. La littérature émergente décrit les *big data* comme étant (1) exhaustives, parce qu'elles ne se contentent pas d'un échantillon mais veulent être représentatives de la totalité de leur objet de recherche ; (2) détaillées en résolution ; (3) relationnelles, en ayant des aspects en commun qui permettent l'entrecroisement entre différentes bases de données ; (4) flexibles, et donc facilement extensibles et scalaires [Kitchin, 2014a : chapitre 6, *Introduction*]. A bien y voir donc, l'adjectif « *big* » est en quelque manière trompeur, car les *big data* sont caractérisées par bien plus que

leur volume : « En effet, certaines collections de données "*small*" peuvent être très grandes en dimension, comme les recensements nationaux qui visent également à être exhaustifs et à avoir une forte résolution et relationnalité. Néanmoins, les collections de données concernant les recensements manquent de vélocité […], de variété […], et de flexibilité » [Kitchin, 2014b : 2]. Or, la question des *big data* a été envisagée selon différentes approches, techniques et méthodologiques, économiques et sociales, éthiques, politiques et même philosophico-existentielles – voir l'article de de Mul ci-dessus. Toutefois, c'est la perspective épistémologique qui a intéressé le plus les interprètes et qui a polarisé le débat.

Surtout dans la première vague d'études sur les grandes bases de données, les experts et les commentateurs ont eu tendance à reconnaître dans les *big data* un vrai tournant épistémologique pour les sciences de la nature et les sciences humaines et sociales. C'est le cas d'un auteur comme Lazer, que nous avons cité dans l'introduction. Dans Mayer-Schönberger et Cukier [2013], ouvrage à la limite entre scientificité et vulgarisation, nous trouvons plusieurs exemples du succès du traitement des *big data*, comme le Google Flu Trends, l'algorithme d'Amazon pour la recommandation de produits, la création de Farecast par Oren Etzioni ou le cas étrange des Pop-Tarts chez Walmart aux Etats-Unis. A propos de ce dernier, les auteurs affirment : « Nous n'avons plus forcément besoin d'une hypothèse substantive valide à propos du phénomène pour commencer à comprendre notre monde. […] Nous n'avons pas besoin de nous soucier des goûts culinaires des clients de Walmart. […] A la place de l'approche orientée hypothèse, nous peuvent utiliser une approche orientée données » [Mayer-Schönberger et Cukier, 2013 : 55]. Dans le premier livre de la *Métaphysique* [981b 5-7], Aristote dit que « ce n'est pas l'habileté pratique qui rend, à nos yeux, les chefs plus sages, c'est parce qu'ils possèdent la théorie et connaissent les causes » et il définit la philosophie comme la science des causes primaires. Dans cette approche aux *big data*, la recherche de la cause est clairement supprimée en faveur de la simple constatation de corrélations, en voulant transformer ainsi un des paradigmes fondateurs de la science occidentale.

La littérature plus récente a montré les limites et même les dangers de cet enthousiasme initial. Dans leur « Six Provocations for Big data, boyd et Crawford, en prenant position contre Chris Anderson et en citant Berry – qui à son tour se réfère implicitement au « deuxième » Heidegger [boyd et Crawford, 2011 : 8] –, soulignent : « Significativement, le rejet généralisé de toutes les autres théories et disciplines proposé par Anderson révèle une allusion arrogante dans plusieurs débats concernant les *big data* où toutes les autres formes d'analyse peuvent être mises de côté […]. A la place de la philosophie – que Kant voyait comme la base rationelle de toutes institutions – "la computationnalité pourrait être comprise comme une ontothéologie, créant une nouvelle 'époque' ontologique comme une nouvelle constellation historique d'intelligibilité" ». L'histoire des sciences guidées par les données (*data-driven*) a commencé bien avant les grandes bases de données numériques, et « la perception d'une "surcharge d'informations" a émergé à

plusieurs reprises depuis la Renaissance, au début de la période moderne et dans la modernité et à chaque fois des technologies spécifiques ont été inventées pour traiter la surcharge considérée» [Strasser, 2012 : 85]. D'ailleurs, les attentes relatives au traitement automatique de grandes bases des données ont été constantes au moins à partir de la diffusion des cartes perforées utilisées pour le recensement de 1880 aux Etats-Unis [Driscoll, 2012]. Comme nous l'avons dit dans la partie introductive, même les enthousiastes comme Lazer, Latour et leurs collaborateurs sont aujourd'hui plus prudents par rapport aux possibilités relatives au traitement des données numériques pour les sciences sociales.

Certains aspects traditionnels de la recherche, comme le regard du chercheur, l'interprétation, la formulation des hypothèses, et parfois même l'attitude critique, ont été réadmis en tant que moments incontournables du parcours de traitement de grandes bases de données. Toutes les étapes de l'analyse des *big data* – *data selection*, *pre-processing*, *data reduction and projection* et *data enrichment* – et plus précisément l'ensemble des techniques qui permet le recueil et l'analyse des *data*, sont le résultat de choix, possibilités, stratégies et limites qui méritent d'être attentivement questionnées. Ce qui semblait ramener les sciences de la nature à leur esprit d'exactitude et rapprocher les sciences humaines et sociales de ce même modèle, se révèle ainsi dans son caractère «mythologique» [boyd et Crawford, 2012 : 663] : les *big data* «ne parlent pas» par elles-mêmes et la marge de manœuvre du chercheur ou du décideur publique reste importante.

Open data

Selon Chignard [2013], le terme *open data* est apparu pour la première fois en 1995 dans *On the full and open Exchange of scientific Data*, une publication du Committee on Geophysical and Environmental Data du National Research Council des Etats-Unis. En réalité, l'idée d'appliquer les principes des biens communs au domaine de la connaissance remonterait au moins aux travaux des années 1940 de Robert King Merton, parmi les pères de la sociologie des sciences. Dans son «The normative Structure in Science», il définit quatre normes éthiques qui régissent les comportements scientifiques : universalisme, «communisme» (*communism*), désintéressement et scepticisme organisé. Le «communisme» en sciences consiste «in the non technical and extended sense of common ownership of goods […]. The substantive findings of science are a product of social collaboration and are assigned to the community» [Merton, 1973 : 273]. Internet et les nouvelles technologies ont donné un nouveau crédit à cette idée. Notoirement, la naissance de la Toile est strictement liée à la culture du partage des connaissances [Castells, 2001 : 40]. Dans ses recherches aux limites de la description et de la prescription, Yochai Benkler [Benkler, 2006 : 31] affirme qu'«une confluence particulière de changements techniques et économiques est maintenant en train d'altérer la manière dont nous produisons et échangeons les informations, la connaissance et la culture». Du point de vue économique, les biens informationnels sont dits «non-rivaux» : «Nous

considérons un bien comme non-rival quand sa consommation de la part d'une personne ne le rend pas moins disponible pour la consommation de la part d'un autre» [Benkler, 2006 : 36]. De surcroît, les biens informationnels sont des biens non-rivaux de nature particulière car, à la différence des biens publics, par exemple, quand la consommation augmente ils ne s'épuisent pas, mais s'enrichissent.

Selon le site http://opengovdata.org, l'histoire du mouvement pour l'ouverture des données commence en août 2005, avec «*the open definition*» selon laquelle «la connaissance est ouverte si n'importe qui est libre d'y accéder, de l'utiliser, la modifier et la partager – soumis, au plus, à des mesures qui en préservent la provenance et l'ouverture». Entre temps, de l'autre côté de l'Atlantique, *The Guardian* lançait en mars 2006 l'initiative «*free our data*»[4]. En 2007, un groupe d'intellectuels et activistes d'Internet se sont rencontrés à Sebastopol en Californie pour trouver une définition de *public open data* et la faire adopter aux candidats pour les élections présidentielles aux Etats-Unis. Les participants se sont mis d'accord sur huit principes, selon lesquels les *open data* doivent être : (1) complètes ; (2) primaires (3) opportunes (*timely*) ; (4) automatiquement exploitables ; (5) accessibles ; (6) non-discriminatoires : «les données sont disponibles pour tous, sans besoin d'enregistrement» ; (7) non-propriétaires ; (8) *licence-free*[5]. En janvier 2009, Barack Obama signe une note «On Transparency and Open Government», dans laquelle les trois principes de la transparence, de la participation et de la collaboration sont établis. A partir de 2010, une série d'institutions nationales et internationales dans le monde entier commencent à ouvrir des bases de données jusque-là inaccessibles.

Or, sous certains aspects le mouvement des *open data* s'oppose à celui des *big data*. En effet, le traitement des *big data* implique des investissements importants au niveau économique et des compétences que très peu d'institutions sont aujourd'hui en mesure de proposer. Sous cet aspect, le *big data* ressemblent plutôt à des formes classiques de traitement de données, telles que les archives d'Etat, qui sont fermées par nature [Derrida, 1995]. Le mouvement des *open data* veut changer cette situation. Pour certains, l'ouverture des données concerne exclusivement les données elles-mêmes, tandis que pour d'autres l'ouverture est un concept plus vaste, qui concerne aussi leur utilisation, analyse et distribution [Kitchin, 2014a : chapitre 3, *Open Data*]. En outre, certains auteurs [Gurstein, 2013] ont affirmé qu'il faut considérer les données ouvertes non pas comme un produit mais comme un service, *i.e.* comme une relation entre le fournisseur et l'utilisateur. Bien évidemment, en cherchant à démocratiser l'accès et l'utilisation des bases des données, le mouvement des open data est orienté vers des questions de nature éthique, politique, économique et sociale. La plupart des arguments en faveur des *open data* se joue précisément en cette direction : les *open data* servent à évaluer l'activité d'institutions publiques et non-gouvernementales, elles encouragent une

4 http://www.theguardian.com/technology/free-our-data.

5 https://public.resource.org/8_principles.html.

participation active à la chose publique, elles poussent les institutions elles-mêmes à améliorer leur productivité, etc. [Kitchin, 2014a : chapitre 3, *The Case for Open Data*]. La majorité des critiques va aussi en ce sens : les open data alimentent une idéologie néo-libérale et la monétarisation des services publiques, elles ne sont pas en soi un processus de démocratisation mais peuvent conduire aussi à l'exercice d'un pouvoir disciplinaire et finalement elles risquent, comme le dit Gurstein [2011], de « donner du pouvoir à ceux qui en ont déjà (*empowering the empowered*) ».

Toutefois, pour arriver à préciser la différence entre les *open data* et notre proposition de *soft data*, il importe de considérer ce mouvement du point de vue épistémologique. Considérons pour un instant les huit principes des *open data* dont nous avons dressé la liste auparavant. Les quatre premiers principes ne sont rien d'autre qu'une reprise des caractéristiques des *big data*. Les quatre autres principes sont par contre ce qui distingue les *open data* par rapport à toute autre approche vis-à-vis des données. Leur mise en pratique par les administrations et les institutions est encore loin d'être réalisée de manière extensive. Parfois, à cause de décisions politiques et stratégiques ; d'autres fois, à cause des difficultés – manque d'expertise, de financements, de confrontation entre les différentes réalités, etc. – liées à une série d'opérations nécessaires pour rendre ces données vraiment ouvertes, comme la mise en forme et la standardisation. Le fait que les données des administrations soient de plus en plus disponibles en ligne ne signifie pas en effet qu'elles peuvent être considérées comme open. Sans doute, la poursuite d'un accès plus immédiat aux données représente une tâche qui mérite d'être poursuivie. Toutefois, la nouvelle catégorie de *soft data* voudrait répondre à la situation actuelle. Dans les pratiques actuelles, comme sans doute dans celles du futur proche, les institutions publiques et privées qui veulent utiliser les données en ligne ne peuvent pas attendre l'ouverture totale des données, mais doivent être capables de travailler à partir d'un amas de données qui peuvent être ouvertes, semi-ouvertes ou non-ouvertes.

Les données pour les études territoriales

Insatisfaction des *hard data*

Traditionnellement, la décision publique liée à la gestion du territoire est basée sur la collecte et l'analyse de ce qui peut être qualifié comme données « hard », à savoir les statistiques officielles et plus généralement les données produites par l'administration publique à différents niveaux (local, national, international). Ces données sont soigneusement harmonisées et stockées dans des bases de données, soumises à divers contrôles, complétées par l'estimation de valeurs manquantes et de métadonnées. Ces données représentent une valeur ajoutée exceptionnelle pour les personnes intéressées par la politique territoriale. Elles garantissent une structure claire, une qualité standard et un niveau de fiabilité de l'information qui

en font une base solide pour des statistiques géographiques utiles à l'action des décideurs publics [Burt *et al.*, 2009 : 18].

Néanmoins, ces dernières années, les décideurs publics ont révélé certaines lacunes ou des frustrations importantes liées à ces données qui peuvent être résumées par trois éléments principaux :

1) Le trop long délai de publication. Les données officielles font l'objet d'un processus technique et parfois politique d'harmonisation et de validation. Ce long processus passe par des validations réitératives des données qui peuvent prendre longtemps. Par exemple Fassmann *et al.* [2009 : 39 ; 115] observent que pour l'étude des migrations la période entre deux recensements ainsi que la période d'attente jusqu'à la disponibilité des données pour les chercheurs et politiciens sont trop longues.

2) La couverture insuffisante de certains sujets d'intérêt pour la cohésion territoriale comme l'attractivité des lieux, les sentiments des citoyens, la perception des actions des décideurs publics. Ces sujets ne sont pas faciles à représenter avec des données territoriales. Ils sont généralement abordés par de grandes enquêtes, mais souvent le lien avec l'espace est très faible. Nous pouvons avoir par exemple des enquêtes à l'échelle nationale, mais les données obtenues sont difficilement transposables à des échelles plus précises comme la ville.

3) La définition *top-down* des données. En effet, la qualité de ces données est garantie par l'application d'une méthodologie de production que seuls des professionnels peuvent mettre en place et, comme nous l'avons déjà observé, c'est exactement cette qualité et ce contrôle qui rendent ces données intéressantes pour les politiques publiques. Pourtant, « les technologies de l'information sont devenues des instruments qui permettent aux résidents des villes de participer à la renégotiation et redéfinition des espaces urbains » [Unsworth *et al.*, 2014]. Par conséquent, les administrations publiques sentent de plus en plus la nécessité d'insérer dans leurs démarches d'analyse territoriale des données participatives, ouvertes et élaborées par les citoyens, les entreprises, les collectivités locales et régionales [Guermond, 2011]. L'approche *bottom-up* [Frasera *et al.*, 2006] pour la définition des données d'intérêt est une dimension qui ne peut plus être ignorée par les décideurs publics. La diffusion de l'approche *bottom-up* est notamment liée au développement des systèmes d'information géographique pour la participation publique (PPGIS). Comme Sieber [2006 : 503] a observé, « PPGIS fournit une approche unique pour engager le public dans la prise de décisions à travers son objectif d'incorporer la connaissance locale, d'intégrer et de contextualiser des informations spatiales complexes, de permettre aux participants d'interagir dynamiquement avec les données (*input*), d'analyser les alternatives et de valoriser les individus et les groupes ». Michael Frank Goodchild [2007] a évoqué également

la naissance d'un *crowdsourcing* développé à partir d'initiatives individuelles où le citoyen devient capteur de phénomènes territoriaux.

Aucune de ces critiques n'était très importante il y a dix ans. Tant que des données officielles étaient la principale source d'information pour les décideurs et les citoyens, les gens étaient susceptibles d'accepter un certain retard dans le processus de suivi des territoires. Cependant, l'ordre du jour de la cohésion territoriale a été fortement modifié par la croissance exponentielle de l'information disponible sur Internet. Un grand nombre d'informations concernant le développement territorial est maintenant disponible sur le Web, en introduisant une concurrence claire pour les producteurs classiques de données. Ce phénomène va bien au-delà de la néogéographie et du géoweb[6] [Haklay *et al.*, 2008 ; Elwood, 2010]. L'apport du Web 2.0 à la cartographie constitue certainement l'interface la plus évidente de la rencontre entre Web et territoire, mais cet article veut souligner la présence d'autres données disponibles sur la Toile, qui correspondent à une vision de cartographie 2.0 [Mericskay et Roche, 2011] plus vaste que celle généralement implicite dans le concept de géoweb, et qui promettent de transformer les études territoriales.

La traçabilité de la vie des territoires

Comme nous l'avons déjà noté à plusieurs reprises, les nouvelles technologies, et notamment Internet, ont radicalement changé plusieurs secteurs de la société. Internet « est le premier moyen de communication moderne qui étend sa portée en décentralisant la structure capitale de production et distribution de l'information, de la culture et de la connaissance. Beaucoup du capital physique qui intègre la plus grande part de l'intelligence dans le réseau est largement diffusée et possédée par les utilisateurs finaux » [Benkler, 2006 : 29].

Ce qui rend ce changement particulièrement intéressant est le fait qu'il affecte à la fois la société elle-même et la façon de l'étudier et de la gérer. En effet, la communication numérique a secoué les conditions de la recherche et de la politique, en multipliant la disponibilité de traces de phénomènes collectifs. L'avantage des médias électroniques est que toutes les interactions qui les traversent laissent des traces numériques qui peuvent être facilement enregistrées, massivement stockées, puis récupérées et analysées. Ainsi, les médias numériques offrent de nouvelles bases de données énormes qui peuvent être utilisées pour améliorer l'analyse des phénomènes sociaux et, par conséquent, le processus de prise de décision qui leur est lié [Rogers, 2013].

6 Le terme « géoweb » se réfère en général à un ensemble de technologies géospatiales et aux informations géographiques disponibles sur le Web [Herring, 1994], tels que Google Earth et MapQuest, où des outils basés sur la localisation, des données et contenus géospatiaux peuvent être générés et partagés par toute personne ayant un connexion Internet.

Les traces numériques ne sont pas seulement produites de façon automatique par les technologies numériques, mais aujourd'hui, nous avons aussi de grandes quantités de données provenant de nouveaux fournisseurs de données tels que des membres de réseaux sociaux en ligne et des utilisateurs des plates-formes de partage de contenu. Dans le contexte du Web 2.0, le succès des médias sociaux n'est plus en doute et leurs taux de diffusion ont atteint des niveaux sans précédent. Des centaines de millions d'utilisateurs sont inscrits à ces médias. Ils échangent via des forums, des blogs et ils maintiennent des pages Facebook, ils racontent leurs dernières pensées, humeurs ou activités en quelques mots. Le développement d'appareils mobiles tels que les smartphones ou tablettes a favorisé l'émergence de ces nouvelles pratiques. En conséquence, les utilisateurs de médias sociaux laissent des traces de leurs activités en ligne et hors ligne qui peuvent devenir de nouvelles sources d'information, que dans la suite nous appellerons *soft*, extrêmement utiles pour des études territoriales et pour les politiques publiques.

Le besoin d'un nouveau concept est dû, par ailleurs, à la nécessité d'aller au-delà du phénomène de l'information géographique volontaire [Gooldchild, 2007] qui a occupé de manière quasi exclusive l'attention des chercheurs qui se sont intéressés à la rencontre entre géographie et Web. Sans oublier le rôle joué par le citoyen comme capteur géographique dans des plates-formes comme Wikimapia ou OpenStreetMap, il est aujourd'hui essentiel de poser notre attention sur d'autres données disponibles sur le Web qui n'ont pas toujours les caractéristiques de l'information géographique volontaire. Prenons l'exemple de check-in Facebook, c'est-à-dire quand un membre de réseau social partage sa position avec ses amis. Comme l'ont noté Vienne *et al.* [2014], ces déclarations permettent des définir de nouvelles géographies de proximité, mais ce n'est pas pour autant qu'on peut utiliser l'étiquette de information géographique volontaire. En effet, si l'action du check-in est sûrement une action volontaire, la motivation qui est à la base est difficilement comparable à celle à l'œuvre dans des applications de VIG dans des contextes de crise [Roche *et al.*, 2013] ou dans d'autres contextes comme ceux décrits par la *Citizen science* [Hand, 2010]. Alors, comme cet exemple le rend clair, de nouvelles traces numériques sont disponibles aujourd'hui sur le Web qui peuvent fournir une nouvelle information, plus fraîche et participative, et répondre aux désirs que les *hard data* ne pouvaient plus satisfaire. L'objectif de la suite de cet article est de chercher à en produire une définition qui puisse mettre en avant la nouveauté de cette source de données et son impact potentiel sur les études territoriales.

Centralité et limite de la géolocalisation

L'usage des traces numériques pour l'étude des phénomènes collectifs est en train de se diffuser rapidement dans tous les champs de la société. Prenons par exemple tous les terrains d'application des *big data*, résumés efficacement par le

rapport de Bulger *et al.* [2014] pour le Oxford Internet Institute. Si l'on considère le terrain des études territoriales, il faut noter d'abord que ce type d'études exige une caractéristique précise aux données : la présence d'une information géographique. Une information, pour être utile dans une analyse territoriale, doit être liée à l'espace, à une échelle spécifique, encore mieux à un point spécifique identifiable par des coordonnées géographiques. Les potentialités du numérique ont sûrement accru ce besoin. Paéz et Scott [2005 : 53], en faisant une courte histoire du rôle des nouvelles technologies dans les études territoriales, observent : « une récente expression de la longue tradition dans l'analyse urbaine qui consiste à adopter rapidement les développements technologiques peut être observée dans l'adoption des systèmes d'information géographique ». Des outils comme les GPS ont rendu finalement possible de générer de données géo-référencées en grande quantité. Cette possibilité technique est certainement très précieuse pour les études territoriales qui peuvent aujourd'hui s'appuyer sur des informations très précises concernant les territoires qu'elles analysent. Cependant, il faut noter des risques importants liés à cette nouvelle centralité de la géolocalisation.

En premier lieu, comme nous l'avons noté dans un article précédent [Romele et Severo, 2014], l'action de géolocalisation n'est pas neutre, elle cache toute une série de choix liés à la plate-forme technique qui génère les coordonnées géographiques et à la personne qui déclare sa position. En effet, il faut prendre en compte les « intentions de l'utilisateur » ainsi que le fait que l'action de géolocalisation sur une plate-forme du Web 2.0 peut être une action volontaire mais est plus souvent le résultat d'un artefact technologique. Comme on l'a déjà noté, cela signifie que ces traces numériques, apparemment fiables et objectives, doivent être problématisées et approfondies. Notamment, l'impact des outils employés pour récolter et analyser ces informations géo-taggées ne doit pas être sous-évalué. L'analyse des outils de *scraping* proposée par Marres et Weltevrede [2013] constitue un exemple optimal de l'attitude qu'il faut tenir vers les méthodes numériques employées pour étudier ces traces du social.

En deuxième lieu, l'importance et la puissance de la géolocalisation, à notre avis, ont porté à négliger d'autres caractéristiques de ces nouvelles données qui méritent l'attention du chercheur. Dans cette direction, il faut citer la réflexion de Crampton *et al.* [2013] qui soulignent la nécessité d'aller au-delà du géotag dans l'analyse des *big data* pour les études territoriales. Les auteurs proposent cinq solutions alternatives à l'information géographique classique pour identifier les lieux dans les discours sur Internet. Ces cinq extensions des coordonnées géographiques sont : « (1) aller au-delà du média social qui est explicitement géographique ; (2) aller au-delà des spatialités du "ici et maintenant" ; (3) aller au-delà de ce qui est proche ; (4) aller au-delà de l'humain vers des données produites par des robots et des systèmes automatisés et (5) aller au-delà du géoweb même, en exploitant ces sources contre les donnés accessoires, tels que les bulletins d'informations et les données de recensements. Nous voyons ces extensions

des méthodologies existantes comme fournissant le potentiel pour dépasser les limitations actuelles de l'analyse du géoweb» [Crampton *et al.*, 2013 : 2].

Par ailleurs, Mark Graham et Mattew Zook soulignent que «les lieux sont de plus en plus définis par des couches denses et complexes de représentation qui sont créées, accédées, et filtrées à travers les technologies numériques et souvent à travers des lignes opaques d'algorithmes codés» [Graham et Zook, 2013 : 77]. Ils continuent en expliquant que «ces dimensions numériques de lieux sont fragmentées en plusieurs axes comme le site, le langage et les réseaux sociaux, avec des représentations également éclatées en fonction d'ensembles uniques d'individus» [*op. cit.* : 78]. En ligne avec telle vision théorique, les auteurs définissent de nouvelles géographies en étudiant la distribution des langues dans les médias sociaux. En deux mots, les auteurs concluent : «outre à dévoiler où, nous cherchons de comprendre quoi, pourquoi et qui» [*op. cit.* : 95].

Ces études nous aident à poser l'accent sur la variété et la richesse des informations, que les données numériques peuvent offrir aux chercheurs et décideurs publics qui s'occupent de gestion du territoire à toute échelle. Les coordonnées géographiques sont certainement des données précieuses et puissantes qui permettent d'avoir une vision ponctuelle de l'espace physique, même si cela ne doit pas devenir une limite dans l'exploitation du potentiel qu'offre le numérique pour étudier l'espace social et les phénomènes collectifs. C'est pour cela que dans le prochain paragraphe nous essayerons de proposer une nouvelle définition de données numériques pour les études territoriales basée sur un système plus articulé des caractéristiques.

Le besoin d'une définition alternative : les *soft data*

Le terme *soft data* n'est sans doute pas nouveau. Selon Cole [1983], il trouve ses racines dans la hiérarchisation des sciences proposée par Comte au XIX^{e} siècle. En épistémologie des sciences, la distinction est introduite par Bertrand Russel en 1914. Dans son *Our knowledge of the external world as a field for scientific method in philosophy*, le philosophe définit d'abord les *data* comme étant «des objets (*matters*) de connaissance commune, vagues, complexes et inexacts comme l'est toujours la connaissance commune, mais néanmoins commandant de quelque manière notre assentiment comme sur le tout et dans certaines interprétations quasi certainement vraies» [Russel, 1914 : 65]. En d'autres termes, les données sont des connaissances préthéoriques, communément acceptées et généralement vraies, bien qu'encore confuses. Dans le livre posthume *De la Certitude*, Wittgenstein appellera ces formes de connaissance «images du monde» (*Weltbildern*). A la différence de Wittgenstein, selon qui les visions du monde ne peuvent pas être soumises à des jugements de vérité et fausseté, Russel utilise la vérité empirique et de la logique pour ordonner les données sur une échelle qui va du *hard* au *soft :* «Par "*hard*" *data* je veux dire celles qui résistent à l'influence dissolvante de la réflexion critique, et par "*soft*" *data* celles

qui, sous l'opération de ce procès, deviennent à nos esprits plus ou moins douteuses. Les plus dures des *hard data* sont de deux espèces : les faits particuliers des sens et les vérités générales de la logique » [Russel, 1914 : 70-71]. Déjà dans cette première définition, la frontière entre *hard data* et *soft data* n'est pas claire. De surcroît, les deux notions sont données par défaut par rapport à un modèle qui, comme l'a bien démontré le Wittgenstein des *Recherches philosophiques*, est déjà problématique en soi.

À partir de cette distinction entre données dures et molles, les interprètes sont arrivés à parler de *hard sciences* et *soft sciences*. Or, il y a au moins trois manières de discerner ces deux types de sciences. La première, la plus immédiate, est à partir de leurs objets respectifs. Toutefois, comme nous venons de le dire, l'idée selon laquelle les objets des mathématiques, de la logique ou des sciences tels que la physique ou la biologie sont des référents durs a été défiée par les considérations du « deuxième » Wittgenstein. La deuxième est à partir des différentes méthodologies au sens large, *i.e.* les techniques et technologies que chaque science utilise. Le philosophe Hans-Georg Gadamer, dans son œuvre majeure *Vérité et Méthode*, distinguait par exemple entre les disciplines de la vérité, comme l'art, l'histoire et la philosophie, et les sciences de la méthode. La validité de cette distinction a été néanmoins niée par le succès des sciences structurales du langage, qui s'efforçaient d'utiliser une méthodologie rigoureuse dans des contextes normalement considérés comme mous comme la littérature. En outre, la perspective gadamérienne ne fournissait pas de réponse pour les cas limites de certaines sciences à l'époque déjà florissantes, comme la sociologie et la psychanalyse, trop « dures » pour être considérées comme disciplines de la vérité mais encore trop « molles » pour être accueillies parmi les sciences de la méthode.

Loin d'être dépassées, ces deux manières simplistes de discriminer entre sciences dures et sciences molles sont encore bien présentes dans la littérature. En 1986, le mathématicien Serge Lang accusait le candidat à la *National Academy of Sciences*, Samuel Huntington, d'utiliser des pseudo-mathématiques dans ses analyses en sciences politiques : « Comment Huntington mesure-t-il des choses comme la frustration sociale ? Possède-t-il un compteur de frustration sociale ? Je conteste à l'académie de certifier comme science de simples opinions politiques » [Lang, 1988]. Comme l'observe Diamond [1987], « la question que Lang soulève est centrale à toute science, *hard* ou *soft*. Elle pourrait être définie comme la nécessité d'"opérationnaliser" un concept. […] Malheureusement, opérationnaliser se prête au ridicule en sciences sociales, parce que les concepts étudiés tendent à être des concepts familiers sur lesquels nous tous imaginons être des experts ». Les distinctions entre données dures et molles « orientées objet et méthode » sont aujourd'hui communes dans le domaine des affaires [Stawarski et Phillips, 2008 : 108], de la géostatistique [Zhang *et al.*, 2008 ; Lu *et al.*, 2010], des sciences de l'information [Pravia *et al.*, 2008 ; Prentice et Shapiro, 2011], et d'autres encore.

La troisième manière de distinguer entre *hard* et *soft sciences* est celle développée par la sociologie de la connaissance scientifique, à ne pas confondre avec la sociologie des sciences [Collins, 1983]. En s'appuyant sur les travaux liminaires de Robert King Merton sur la reconnaissance et la récompense dans le domaine des sciences, Norman W. Storer [1967] affirmait que la distinction entre sciences dures et sciences molles ne peut s'appuyer ni sur une question d'objets visés, ni sur l'idée qu'une *hard science* comme la physique nécessite une concentration majeure, plus d'heures de travail et d'exercices en laboratoire qu'une *soft science* telle que la sociologie[7]. La différence repose plutôt sur la difficulté qu'ont les sciences dures à contribuer de manière significative à la discipline et donc aussi sur le risque assumé par le chercheur à chaque contribution : « "*Hardness*" en ce sens […] suggère aussi un degré de difficulté impliqué dans la tentative de contribuer au sujet et ainsi un degré de risque qu'un scientifique prend quand il propose une contribution » [*op. cit.* : 79]. La *hardness* des sciences a des effets directs sur les relations sociales entre scientifiques : « si nous concluons qu'il y a plus de risque impliqué dans la contribution en sciences dures qu'en sciences molles parce que les collègue peuvent plus facilement identifier toute faiblesse dans le travail de quelqu'un, il pourrait être que ce quelqu'un se sent moins "proche" de ces collègues en termes de chaleur et confiance » [*op. cit.* : 79]. L'hypothèse était que l'impersonnalité des sciences dures pouvait être mesurée « en cherchant si, en citant le travail d'autres scientifiques, l'auteur d'un rapport de recherche a utilisé le prénom ou seulement les initiales » [*op. cit.* : 80]. Les résultats des recherches conduites sur un échantillon de publications dans dix domaines scientifiques confirmèrent cette hypothèse. Sur la même ligne s'insère le travail de Solla Price dans *Citation Measures of Hard Science, Soft Science, Technology, and Non Science* (1969), qui avance l'hypothèse selon laquelle les *hard sciences* se caractérisent par l'immédiateté des citations, c'est-à-dire par le fait que les références dans un article tendent à ne pas dépasser les cinq années précédentes. Pour cette raison, plus haute est la quantité de vieilles citations, plus probable est que l'article en question appartienne au domaine des sciences molles.

De cette histoire nous pouvons retenir au moins deux éléments. Premièrement, la difficulté générale à tracer une ligne nette entre *hard sciences* et *soft sciences*. D'ailleurs, plusieurs auteurs ont distingué entre une attitude « molle » et une attitude « dure » dans la même discipline. La différence entre philosophie continentale et analytique

7 Similairement, Soler [2000 : 24] affirme que « l'opposition sciences dures/sciences molles n'est pas à placer sur le même plan que [les autres classifications des sciences], dans la mesure où elle repose essentiellement sur un jugement de valeur : parler de sciences "molles" est évidemment péjoratif ». Dans l'éditorial de *Nature* 487/271 (19 juillet 2012), on trouve une défense des sciences sociales en ces termes : « Une partie de la responsabilité doit se trouver dans la pratique d'étiqueter les sciences sociales comme étant *soft*, que trop rapidement se traduit comme signifiant vagues et stupides (*soft-headed*). Dans la mesure où ils ont à avoir avec des systèmes hautement complexes, capables de s'adapter et non réglé de manière rigoureuse, les sciences sociales sont parmi les disciplines les plus difficiles, méthodologiquement et intellectuellement ».

repose précisément sur l'idée que cette dernière utilise une méthodologie plus solide et a des objets d'investigation bien définis, notamment la logique. Dans la sociologie aussi on a essayé de distinguer entre *hard-data sociology* et *soft-data sociology* [Eriksson, 1978]. Face à cette difficulté, les interprètes ont ainsi dû développer des méthodologies astucieuses comme celles que nous venons de voir. Deuxièmement, nous pouvons retenir quelque chose du fait que les dernières définitions dont nous avons parlé ont la particularité d'être intra- et inter-textuelles, mais de ne pas aller au-delà de la production scientifique comme production de textes scientifiques. Une fois écartée la possibilité de distinguer entre sciences dures et molles à partir de leurs objets et de leurs techniques, il reste une distinction toute interne aux manières respectives de produire des textes. Par conséquence, en utilisant la terminologie de la linguistique, nous pouvons dire que la distinction entre sciences dures et molles est ici de nature synchronique et non diachronique. Mais alors toutes les sciences finissent par perdre au moins une partie de leur hardness. Les travaux de Bruno Latour sur la construction des objets scientifiques (Pasteurisation) nous mènent précisément dans cette direction.

Or, nous choisissons ici l'étiquette *soft data* pour trois raisons. Premièrement, précisément parce que nous croyons qu'en général les données ont une nature plus molles que ce qui a leur été attribué dans les sciences. Comme nous l'avons dit dans le deuxième chapitre, les *data* ne sont ni de nature ontologique, comme les faits, ni de nature épistémologique, comme les évidences, mais de nature rhétorique ou «diaphorique». Selon cette perspective, nous pouvons dire que si «*raw data is an oxymoron»* [Gitelman, 2013], alors «*soft data* est un pléonasme». Dans leur réflexion, Gerardi et Turner [Gerardi et Turner, 2002] cherchent à construire des bases théoriques solides pour la recherche qualitative en sciences sociales. Dans *Real man don't collect soft data*, ils affirment justement que la distinction entre *hard* et *soft* «néglige le substrat de méthodes, perceptions, capacités et attentes implicites qui imprègnent toute investigation, quelle que soit son organisation systématique, et surestime l'importance des chiffres et des mesures» [*op. cit.:* 34].

Deuxièmement, nous choisissons ici l'étiquette *soft data* parce que nous pensons que ceci est d'autant plus vrai dans le cas des données numériques. Si dans l'opposition traditionnelle le terme *soft* sert à souligner l'absence de rigueur et de structure des données dans les sciences humaines et sociales, ici nous volons poser l'accent sur le fait que c'est exactement cette forme moins structurée qui rend intéressante l'information contenue dans ces données. L'étiquette «*soft data*» permet de marquer une différence par rapport à la prétention de *hardness* qui se cache derrière les définitions de *big data* et d'*open data*. Encore une fois, comme nous l'avons dit en introduisant ces termes dans la deuxième partie, la littérature a récemment mis à dure épreuve l'idée selon laquelle les grandes bases de données représentent une nouveauté épistémologique, une automatisation absolue et sans faille de tout processus de connaissance. De surcroît, nous avons montré comment les données

ouvertes ne font rien d'autre, du point de vue épistémologique, que reprendre ce modèle d'exactitude présumée.

Troisièmement, c'est dans la pratique concrète de l'utilisation des données numériques pour l'étude des territoires que nous voyons un avantage à l'utilisation du terme *soft data*. Dans ce contexte, les *soft data* peuvent être définies comme des données disponibles sur Internet[8], généralement non contrôlées par une administration. Elles sont constituées principalement – mais non seulement – par les nouveaux types de données issues du Web 2.0 (Facebook, Twitter, fils RSS, etc.) qui s'offrent au décideur public comme une source originale et riche d'informations sur les phénomènes sociaux qui ont lieu dans un territoire. Elles se caractérisent aussi par le fait d'être visible sur Internet et donc potentiellement accessibles et récoltables. En ce sens, les *soft data* présentent des avantages par rapport à d'autres sources de données : (1) un délai plus court de publication utile pour l'action publique. Un exemple classique de cette réactivité est donné par l'enregistrement des tremblements de terre par les médias sociaux comme Twitter. De nombreux chercheurs [Sakaki *et al.*, 2010] ont démontré que les utilisateurs de réseaux sociaux peuvent être considérés comme des capteurs, capables de localiser les événements catastrophiques en temps réel et de suivre leur développement ; (2) la couverture de nouveaux sujets d'intérêt comme les modes de déplacement dans les zones urbaines, la pauvreté et l'exclusion sociale, les sentiments des citoyens envers les politiques publiques [Weller *et al.*, 2014] ; 3) l'élaboration *bottom-up*, comme le montre l'exemple d'*Open Street Map*, qui offre une alternative aux cartes officielles produites par les instituts géographiques, mais encore plus toutes les données qui peuvent être récoltées à partir des média sociaux. Ces données participatives peuvent également être utilisées à des fins non prévues par leur créateur pour créer une information sur mesure utile au décideur public [Severo *et al.*, 2015].

Regarder les données du Web comme étant « molles » plutôt que « grandes » ou « ouvertes » a aussi ses avantages. Par rapport à l'étiquette *big data*, parler de *soft data* nous permet de souligner que les données numériques peuvent être bien utiles pour les politiques publiques même si elles n'ont pas le volume, la vitesse et la variété des grandes bases de données. Dans une époque de *big data*, les *small data* ont encore une grande valeur : « les *small data* peuvent se focaliser sur des cas spécifiques et raconter des histoires individuelles, nuancées et contextuelles. Les études à partir des *small data* visent à trouver de l'or en travaillant dans des mines étroites, tandis que les études à partir des *big data* visent à extraire des pépites en creusant à ciel ouvert, en ramassant et criblant de grandes étendues

8 Souvent l'étiquette « données non structurées » est utilisées pour identifier ces données disponibles sur le Web. Cependant, nous trouvons cette étiquette restrictive pour deux raisons. En premier lieu, les données disponibles sur Internet ne sont pas toujours non structurées, mais si on y retrouvent pas les structures classiques des données pour la recherche, elles sont souvent très codifiées et enrichies de métadonnées. En deuxième lieu, le fait de caractériser ces données simplement pour leur forme, ne permet pas de capturer toutes les spécificités liées à leur provenance et leur signification pour les SHS.

de terrain» [Kitchin, 2014b : 4]. Par rapport aux *open data*, un regard orienté *soft* permet d'accueillir toutes les données qui ne sont pas libres de droit. Les données sur Internet sont souvent produites par des sujets, individuels ou collectifs, privés ou publics, qui en gardent la propriété. *Last but not least*, le terme *soft data* offre un abri à toutes les données numériques qui ne sont pas construites avec la rigueur ontologique et épistémologique exigée par les *big data* et les *open data*.

Conclusion

Cet article visait à réfléchir sur l'usage des données numériques dans les études territoriales. Dans la partie introductive, nous avons montré comment les auteurs qui ont contribué le plus à développer cet enthousiasme pour les données numériques se trouvent aujourd'hui à faire un pas en arrière et à assumer des positions bien plus prudentes, voire pessimistes. Comme nous avons argumenté dans la deuxième partie, c'est la nature même des données, et des données numériques en particulier, qui nous impose une attitude différente. En ce qui concerne les premières, nous avons dit qu'elles n'ont pas une nature ontologique «dure», car leur référence dépend en grande partie des intentions – de la pragmatique, pour utiliser le langage de Peirce – de leur utilisateur. A propos des données numériques, nous avons travaillé les catégories de *big data* et *open data* avec l'intention de déconstruire leur prétention à une épistémologie et méthodologie *hard*. La littérature sur les grandes bases de données admet de plus en plus le regard (critique) du chercheur là où auparavant elle voyait un procès automatisé, anonyme et objectif. Si du point de vue socio-économique et socio-politique les données ouvertes représentent une possible alternative aux *big data*, du point de vue de leur fonctionnement elles reprennent le désir de complétude, primauté, opportunité et exploitation automatique des *big data*. Dans la troisième partie, nous avons orienté notre critique vers les traces numériques pour l'étude du territoire. A ce propos, le cas des données géoréférencées s'est démontré paradigmatique. Premièrement, parce qu'en dépit de son importance indiscutable, l'action de géolocalisation n'est pas neutre, elle cache toute une série de choix liés à la plate-forme technique qui génère les coordonnées géographiques et à la personne qui déclare sa position. Deuxièmement, car les chercheurs ont souvent surestimé la puissance de la géolocalisation à représenter de manière fidèle la réalité sociale, en négligeant ainsi d'autres caractéristiques de ces nouvelles données.

L'attitude prudente et critique démontrée jusqu'ici ne s'est pourtant pas transformée en pessimisme. Dans la quatrième partie nous avons en effet jeté les bases pour une nouvelle définition qui peut rendre compte des faiblesses attestées, en transformant ces faiblesses en une ressource potentielle. La catégorie de *soft data* ne représente pas une alternative aux *big data* ou aux *open data*, mais il s'agit plutôt d'une étiquette plus inclusive. Au cours de cette partie, nous avons

d'abord montré comment la réserve générale envers les donnés *soft* ne tient pas. La distinction entre sciences « dures » et « molles » ne peut se baser ni sur leurs objets ni sur leurs méthodologies respectives. De surcroît, la discrimination inter- et intra- textuelle proposée par la sociologie de la connaissance scientifique ne fait que faire perdre un peu de *hardness* à toutes les sciences. Le terme de « *soft data* » est préférable non seulement pour nommer les données scientifiques en général, mais aussi les données numériques et les traces numériques pour l'étude du territoire en particulier. Il nous permet surtout d'accueillir les *data* qui n'ont pas la voluminosité, la vitesse et la variété des grandes bases de données ni la forme déjà structurée et libre de droit des données ouvertes. C'est précisément avec ces données de nature mixte que les institutions publiques doivent aujourd'hui apprendre à travailler. En conclusion, cette proposition de définition est une invitation aux chercheurs et aux décideurs publics à prendre conscience de la spécificité des données du Web pour l'étude du territoire et de la nécessité d'identifier des procédures et des techniques adéquates pour gérer certaines des problématiques mentionnées jusqu'ici, comme la question de la propriété des données, de leur hétérogénéité et de leur lien complexe avec l'espace.

Références bibliographiques

[Anderson, 2008] Anderson, C., « The end of theory », *Wired magazine*, 16(7), 16-17.

[Benkler, 2006] Benkler, Y., 2006, *The wealth of Networks. How social Production transforms Markets and Freedom*, New Haven et London : Yale University Press.

[Beurskens, 2014] Beurskens, M., 2014, « Legal Questions of Twitter Research », In Weller, K., Bruns, A. et Burgess, J. E. (dir.), *Twitter and society*, Peter Lang, pp 123-133.

[Bizer *et al.*, 2009] Bizer, C., Heath, T. et Berners-Lee, T., 2009, « Linked Data: The Story so Far », http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf.

[boyd et Crawford, 2011] boyd, D. et Crawford, K., « Six Provocations for Big data », *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, http://ssrn.com/abstract=1926431.

[boyd et Crawford, 2012] boyd, D. et Crawford, K., 2012, « Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon », *Information, communication & society*, 15(5), 662-679.

[Bulger *et al.*, 2014] Bulger, M., Taylor, G. et Schroeder, R., 2014, *Data-Driven Business Models: Challenges and Opportunities of Big data*, Oxford Internet Institute.

[Burt *et al.*, 2009] Burt, J. E., Barber, G. M. et Rigby, D. L., 2009, *Elementary Statistics for Geographer*, New York : Guilford Press.

[Castells, 2001] Castells, M., 2001, *The Internet galaxy: Reflections on the Internet, business, and society*, Oxford : Oxford University Press.

[Chignard, 2013] Chignard, S., 2013, « A brief History of Open Data », *ParisTech Review*, http://www.paristechreview.com/2013/03/29/brief-history-open-data.

[Cole, 1983] Cole, S., 1983, « The Hierarchy of the Sciences? » *American Journal of Sociology*, 89, 111-139.

[Collins, 1983] Collins, H. M., 1983, « The Sociology of Scientific Knowledge: Studies of Contemporary Science », *Annual Review of Sociology*, 9, 265-285.

[Crampton *et al.*, 2013] Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W. et Zook, M., 2013, « Beyond the geotag: situating "big data" and leveraging the potential of the geoweb », *Cartography and Geographic Information Science*, 40(2), 130-139.

[Delaney, 2005] Delaney, D., 2005, *Territory: A Short Introduction*, Malden, MA : Blackwell.

[Derrida, 1995] Derrida, J., 1995, *Mal d'archive*, Paris : Galilée.

[Derrida, 1971] Derrida, J., 1971, « La mythologie blanche », *Marges de la philosophie*, Paris : Editions de Minuit.

[Diamond, 1987] Diamond, J., 1987, « Soft sciences are often harder than hard sciences », *Discover*, août.

[Diebold, 2012] Diebold, F. X., 2012, « A personal Perspective on the Origin(s) and Development of "Big data": the Phenomenon, the Term, and the Discipline », Working paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843.

[Driscoll, 2012] Driscol, K., 2012, « From Punched Cards to "Big data": A Social History of Database Populism », *communication* + 1, 1(4), http://scholarworks.umass.edu/cpo/vol1/iss1/4.

[Elwood, 2010] Elwood, S., 2010, « Geographic information science: Emerging research on the societal implications of the geospatial web », *Progress in Human Geography*, 34(3), 349-357.

[Eriksson, 1978] Eriksson, I., 1978, « Soft-Data Sociology », *Acta Sociologica*, 21(2), 103-124.

[Fassmann *et al.*, 2009] Fassmann, H., Reeger, U. et Sievers, W., 2009, *Statistics and Reality: Concepts and Measurements of Migration in Europe*, Amsterdam : Amsterdam University Press.

[Floridi, 2008] Floridi, L., 2008, « Data », In Darity, W. A. (dir), *International Encyclopedia of the Social Sciences*, Detroit : Macmillan.

[Frasera *et al.*, 2006] Frasera, E. D.G., Dougilla, A. J., Mabeeb, W. E., Reeda, M. et McAlpinec, P., 2006, « Bottom up and top down: Analysis of participatory processes for sustainability indicator identification as a pathway to community empowerment and sustainable environmental management », *Journal of Environmental Management*, 78, 114-127.

[Gerardi et Turner, 2002] Gerardi, S. et Turner, B., 2002, « Real men don't collect soft data », In Huberman, A. Michael et Miles, Matthew B., *The qualitative researcher's companion*, Thousand Oaks : Sage, 81-100.

[Gitelman, 2013] Gitelman, L., 2013, *Raw Data is an oxymoron*, Cambridge, MA : MIT Press.

[Ginsberg *et al.*, 2009] Ginsberg, J., Mohebbi, Matthew H., Patel, R. S., Brammer, L., Smolinski, M. S. et Brilliant, L., 2009, « Detecting influenza epidemics using search engine query data », *Nature*, 457 (7232), 1012-4.

[Goodchild, 2007] Goodchild, M. F., 2007, « Citizens as sensors: the world of volunteered geography », *GeoJournal*, 69(4), 211-221.

[Goodchild *et al.*, 2007] Goodchild, M. F., Fu, P. et Rich, P., 2007, «Sharing geographic information: An assessment of the Geospatial One-Stop», *Annals of the Association of American Geographers*, 97(2), 250-266.

[Graham et Zook, 2013] Graham, M. et Zook, M., 2013, «Augmented realities and uneven geographies: exploring the geolinguistic contours of the web», *Environment and Planning A*, 45(1), 77-99.

[Guermond, 2011] Guermond, Y., 2011, «Les banques de données géographiques régionales e de la révolution du libre accès à la participation citoyenne», *L'Espace géographique*, 2(40), 97-102.

[Gurstein, 2013] Gurstein, M. B., 2013, «Empowering the Empowered or effective Data Use for Everyone?», *First Monday*, 16(2).

[Haklay *et al.*, 2008] Haklay, M., Singleton, A. et Parker, C., 2008, «Web mapping 2.0: The neogeography of the GeoWeb», *Geography Compass*, 22(6), 2011-2039.

[Hand, 2010] Hand, E. (2010). «Citizen science: People power», *Nature*, 466, 7307, 685-687.

[Herring, 1994] Herring, C., 1994, «An architecture of cyberspace: Spatialization of the Internet», *US : Army Construction Engineering Research Laboratory*.

[Hey et Trefethen, 2003] Hey, A. J. G. et Trefethen, A. E., 2003, «The Data Deluge: An e-Science Perspective», In Berman, F., Fox, G., et Hey, A. J. G. (éd.), *Grid Computing: Making the Global Infrastructure a Reality*, John Wiley and Sons, pp 1-17.

[Kitchin, 2014a] Kitchin, R., 2014, The Data Revolution: Big data, Open Data, Data Infrastructures and their Consequences, London : Sage, Kindle Edition.

[Kitchin, 2014b] Kitchin, R., 2014, «Big data, new Epistemologies and paradigm shift», *Big data & Society*, 1(1), 1-12.

[Kitchin et Lauriault, 2014] Kitchin, R. et Lauriault, T. P., 2014, «Small Data in the Era of Big data», *GeoJournal*, 1-13.

[Lang, 1998] Lang, S., 1998, «Academia, Journalism, and Politics: A Case Study: The Huntington Case», *Challenges*, New York : Springer Science & Business Media, 1-222.

[Latour *et al.*, 2013] Latour, B., Jensen, P., Venturini, T., Grauwin, S. et Boullier, D., 2013, «Le tout est toujours plus petit que les parties. Une expérimentation numérique des monades de Gabriel Tarde», *Réseaux*, 31(177), 199-233.

[Latour, 1993] Latour, B., 1993, *The pasteurization of France*, Harvard University Press.

[Latour et Woolgar, 1986] Latour, B. et Woolgar, S. (1986), *Laboratory Life L The Construction of scientific Facts*, Princeton : Princeton University Press.

[Lazer *et al.*, 2014] Lazer, D. M., Kennedy, R., King, G. et Vespignani, A., 2014, «Big data. The parable of Google Flu: Traps in big data analysis», *Science*, 343(6176), 1203-1205.

[Lazer *et al.*, 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. et Van Alstyne, M., 2009, «Computational social science», *Science*, 323(5915), 721-3.

[Lohr, 2013] Lohr, S., 2013, *The Origins of Big data: An etymological detective Story*, http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story.

[Marres et Weltevrede, 2013] Marres, N. et Weltevrede, E., 2013, «Scraping the social?», *Journal of Cultural Economy*, 6(3), 313-335.

[Marres, 2012] Marres, N., 2012, «The redistribution of methods: on intervention in digital social research, broadly conceived», *The sociological review*, 60(S1), 139-165.

[Mashey, 1998] Mashey, J., 1998, «Big data and the Next Wave Infrastress», *Computer Science Division Seminar*, Berkeley : University of California.

[Mayer-Schönberger et Cukier, 2013] Mayer-Schönberger, V. et Cukier, K., 2013, *Big data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/ Mariner Books [réimpression].

[Mericskay et Roche, 2011] Mericskay, B. et Roche, S., 2011, «Cartographie 2.0 : le grand public, producteur de contenus et de savoirs géographiques avec le web 2.0», *Cybergeo/ European Journal of Geography*, 552, http://cybergeo.revues.org/24710.

[Merton, 1973] Merton, R. K., 1973, «The normative Structure of Science», In *The Sociology of Science. Theoretical and empirical Investigations*, Chicago : University of Chicago Press.

[Nature, 2012] «A different agenda», Editorial, *Nature*, 487(7407), 271.

[Paéz et Scott, 2005] Paéz, A. et Scott, D. M., 2005, «Spatial statistics for urban analysis: a review of techniques with examples», *GeoJournal*, 61(1), 53-67.

[Pravia *et al.*, 2008] Pravia, M. A., Prasanth, R. K., Arambel, P. O., Sidner, C. et Chong, C. Y. 2008, «Generation of a fundamental data set for hard/soft information fusion», *Proceedings of the 11th International Conference on IEEE.*

[Prentice et Shapiro, 2011] Prentice, M. et Shapiro, S. C., 2011, «Using propositional graphs for soft information fusion», Information Fusion (FUSION), *Proceedings of the 14th International Conference on IEEE.*

[Rheinberger, 1997] Rheinberger, H.-J., 1997, *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*, Stanford, CA : Stanford University Press.

[Roche *et al.*, 2013] Roche, S., Propeck-Zimmermann, E. et Mericskay, B., 2013, «GeoWeb and crisis management: issues and perspectives of volunteered geographic information», *GeoJournal*, 78(1), 21-40.

[Rogers, 2013] Rogers, R., 2013, *Digital methods*, Cambridge, MA : MIT press.

[Romele et Severo, 2014] Romele, A. et Severo, M., 2014, «Une approche philosophique de la ville numérique : méthodes numériques et géolocalisation», In Carmes, M. et Noyer, J.-M. (dir.), *Devenirs urbains*, Paris : Presses des Mines, pp 205-225.

[Rosenberg, 2013] Rosenberg, D., 2013, «Data before Facts», In Gitelman, L. (dir.), *Raw Data is an oxymoron*, Cambridge, MA : MIT Press, pp 15-25.

[Russell, 1914] Russell, B., 1914, Our knowledge of the external world as a field for scientific method in philosophy, Open Court.

[Sakaki *et al.*, 2010] Sakaki, T., Okazaki, M. et Matsuo, Y., 2010, «Earthquake shakes Twitter users: real-time event detection by social sensors», *Proceedings of the 19th international conference on World wide web.*

[Severo *et al.*, 2015] Severo, M., Giraud T. et Pecout, H., 2015, «Twitter data for urban policy making: an analysis on four European cities», *Twitter for Research conference*, Lyon.

[Sieber, 2006] Sieber, R., 2006, « Public participation geographic information systems: A literature review and framework », *Annals of the Association of American Geographers*, 96(3), 491-507.

[Soler, 2000] Soler, L., 2000, *Introduction à l'épistémologie*, Paris : Ellipses.

[Stawarski et Phillips, 2008] Stawarski, C. et Phillips, P. P., 2008, *Data collection: Planning for and collecting all types of data*, Oxford : John Wiley and Sons.

[Storer, 1967] Storer, N. W., 1967, « The hard Sciences and the Soft: Some Sociological Observations », *Bulletin of the Medical Library Association*, 55, 75-84.

[Strasser, 2012] Strasser, B. J., 2012, « Data-driven sciences: from wonder cabinets to electronic databases », *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43, 85-87.

[Unsworth *et al.*, 2014] Unsworth, K., Forte, A. et Dilworth, R., 2014, « Urban Informatics: The Role of Citizen Participation in Policy Making », *Journal of Urban Technology*, 21(4), 1-5.

[Venturini *et al.*, 2014] Venturini, T., Baya Laffite, N., Cointet, J.-P., Gray, I., Zabban, V. et De Pryck, K., 2014, « Three maps and three misunderstandings: A digital mapping of climate diplomacy », *Big data & Society*, 1, 2.

[Venturini et Latour, 2014] Venturini, T. et Latour, B., 2010, « The social fabric: Digital traces and quali-quantitative methods », In *Proceedings of Futur en Seine 2009*, Paris : Editions Futur en Seine, 87-101.

[Vienne *et al.*, 2014] Vienne, F., Douay, N., Le Goix, R. et Severo, M., 2014, « Lieux et hauts lieux des densités intermédiaires : une analyse par les réseaux sociaux numériques », *ASRDLF 2014*.

[Weiss et Indurkhya, 1998] Weiss, S. et Indurkhya, N., 1998, *Predictive Data Mining: A Practical Guide*, San Francisco : Morgan Kaufmann Publishers.

[Weller *et al.*, 2014] Weller, K., Bruns, A., Burgess, J., Mahrt, M. et Puschmann, C. (éd.), 2014, *Twitter and society*, New York : Peter Lang.

[Whatmore, 2009] Whatmore, S. J., 2009, « Mapping knowledge controversies: science, democracy and the redistribution of expertise », *Progress in Human Geography*, 33(5), 587-598.

[Wilken, 2014] Wilken, R., 2014, « Twitter and Geographical Location », In Weller, K., Bruns, A., Burgess, J., Mahrt, M. et Puschmann, C. (éd.), *Twitter and society*, New York : Peter Lang.

[Woolgar, 2002] Woolgar, S. (dir), 2002, *Virtual society? Technology, cyberbole, reality*, Oxford : Oxford University Press.

[Zhang *et al.*, 2008] Zhang, T., Lu, D. et Li, D., 2008, « A statistical information reconstruction method of images based on multiple-point geostatistics integrating soft data with hard data », *Computer Science and Computational Technology, 2008, ISCSCT'08*, International Symposium on IEEE, pp 573-578.