La troisième voie entre données et méthodes : une approche par la visualisation des données SHS

Marta Severo

La troisième voie entre données et méthodes : une approche par la visualisation des données SHS

Marta Severo, maitre de conférence, laboratoire Geriico, Université de Lille 3

Déjà en 1986, Michel Callon *et al* avaient deviné l'impact que l'arrivée d'une nouvelle génération de données qualitatives (à l'époque il s'agissait des données offertes par les grandes bases bibliométriques) pouvait avoir sur les SHS, en signant la fondation des STS. Dans les années suivantes, le déluge de données numériques et son impact sur l'étude du social a confirmé ces premières intuitions.[1] En effet, la diffusion du numérique et notamment l'explosion d'Internet ont changé profondément la manière de gérer et d'étudier la société (Benkler, 2006 ; Venturini & Latour, 2010). Au niveau épistémologique, l'intérêt des technologies numériques réside dans le fait que toute activité qui les traverse laisse des traces qui peuvent être stockées et analysées comme traces du social (Lazer *et al* 2009). Pour définir ces données, on utilise souvent le terme « big data ». Sans entrer dans le détail des différentes définitions et dans l'importante littérature qui a été développée (boyd, 2012), ce qui nous intéresse est de considérer l'impact des *big data* sur le monde professionnel, académique et éducatif, comme nouvelle source d'information sur la société (Ginsberg *et al* 2009) et de valeur économique. Mon projet de recherche a pour objectif d'approfondir du point de vue théorique et de tester du point de vue empirique, tant dans la recherche que dans l'enseignement, l'emploi de ces données comme moteur d'innovation des SHS.

Mon projet s'intéresse principalement aux grandes bases de données natives du Web qui s'offrent comme alternatives aux « hard data » produits par les fournisseurs de données traditionnels. L'avantage de ces données est qu'elles sont facilement accessibles, disponibles en temps réel et extrêmement riches. Pour la première fois, les SHS ont à leur disposition des données qui, au égard à leur quantité et traçabilité, sont comparables aux données des sciences dures, (Latour, 2009). Cependant, ces données posent nombreux problèmes juridiques (*i.e.* le droit d'auteur), éthiques (*i.e.* la protection de la vie privée) et méthodologiques (*i.e.* la représentativité des données). C'est pour cela que leur emploi demande l'approfondissement de toute

[1] La cartographie des controverses en est un exemple (voir FORCCAST http://forccast.hypotheses.org/).

une série de questions théoriques et de définir de nouvelles méthodes de collecte, de traitement et d'analyse. Il convient d'identifier deux types d'approches à ces données :

1) Faire « parler » les données (*exploratory data analysis*). Même si le Web met à disposition ces grandes masses de données, il est souvent délicat d'en tirer des informations intéressantes et de déterminer la méthode adéquate pour traiter ces données. Une approche possible consiste à laisser « parler » les données sans prédéfinir des hypothèses. Il est possible de collecter des corpus de tweets (Severo & Zuolo, 2012), des requêtes Google (Venturini et al, 2013), des actualités (Giraud et Severo, à paraître), ou des citations académiques (Chavalarias & Cointet 2013) et ensuite de les analyser à travers les méthodes numériques.[2] Si la puissance de ces méthodes du point de vue exploratoire est indiscutable, aujourd'hui il est nécessaire de développer un regard critique (Romele & Severo, 2014 ; Marres, 2012). Il faut produire un protocole de recherche qui en définit avec précision les limites et qui combine leur emploi avec la deuxième approche que je présenterai ci-après.

2) Interroger les données (*confirmatory data analysis*). Une démarche exploratoire doit être complétée par des méthodologies codifiées capables de répondre à des questions de recherche spécifiques. Des **méthodes quali-quantitatives** (Giraud et Severo, 2014) peuvent faire parler les données mais également fournir des réponses claires et fiables. En outre, l'application des méthodes numériques doit être accompagnée par une étude ethnographique des outils utilisés pour générer et traiter les données (Marres & Weltevrede, 2013).

Dans ce contexte, la **visualisation des données** joue un rôle primordial. Dans Severo et Venturini (à paraître), je m'attache à démontrer que la cartographie du web ne doit pas être simplement exploitée comme métrique mais surtout comme technique en même temps exploratoire et confirmatoire (Tukey, 1977) afin d'analyser les dynamiques d'un phénomène social sur la base de sa présence sur le web. Cela illustre comment la visualisation peut être la troisième voie entre les deux approches (Latour *et al* 2012). Si, d'une part, aujourd'hui les outils de *dataviz* nous permettent de visualiser la complexité du social, d'autre part, une représentation efficace peut fournir un effet de *zoom* et *dezoom* nécessaire pour arriver à des résultats précis sur un échantillon ciblé (Cardon, 2014). De plus, aujourd'hui la visualisation de données

[2] Une nouvelle groupe de méthodes née pour faire parler les données du Web (Rogers, 2013).

devient la meilleure solution pour croiser plusieurs dimensions de données : spatiale, temporelle, thématique, médiatique (Severo et al 2012). Une telle visualisation a d'ailleurs également trouvé d'importantes applications en entreprise, principalement comme outil d'aide à la décision.

De toute façon, il faut admettre que l'usage de la visualisation des données comme technique d'analyse est encore rare à cause de craintes théoriques mais surtout de la nécessité de développer de nouveaux algorithmes de traitement et visualisation efficaces (Manovich, 2011). C'est sur ce point particulier que je souhaite concentrer mon projet. Même si le thème de la visualisation de données a pris de l'envergure ces dernières années dans le secteur SHS, elle reste souvent un aspect latéral. La visualisation est fréquemment introduite dans un projet comme outil de restitution finale de données plutôt que comme outil d'analyse. Je propose de construire un réseau qui aura pour ambition de s'interroger sur comment la visualisation de données devrait évoluer du point de vue théorique et pratique pour devenir un outil d'analyse de données du web exploratoire et confirmatoire.

Je propose deux terrains où cette question pourrait analysée :

1) **Le rapport entre droit à l'oubli et devoir de mémoire**. Aujourd'hui, le droit à l'oubli est un des thèmes les plus débattus au niveau international (www.google.com/advisorycouncil) dans le cadre du débat plus large sur la protection de la vie privée. Le Conseil d'Etat a récemment proposé de le reconnaître comme un droit fondamental du numérique. Bien que la crainte pour le droit à l'oubli soit aujourd'hui liée au numérique, il ne faut pas oublier l'importance du débat existant concernant la sauvegarde des cultures orales. Si la mémoire de ce patrimoine est considérée comme un devoir indiscutable (UNESCO, 2003), il y a depuis toujours des controverses qui émergent sur la nécessité et la manière de préserver cette mémoire. Cet objet d'étude, à travers l'analyse de l'énorme quantité de données à disposition et de son évolution spatio-temporelle et thématique, peut être un terrain idéal pour développer une méthodologie de visualisation capable de saisir les différents points de vue et acteurs du débat.

2) **Données de la ville 2.0 entre décideur public et citoyen**. Les *big data* sur la ville (Boullier, 2010 ; http://senseable.mit.edu/) sont principalement les open data qui mettent à disposition les acteurs de la ville et les traces qui produisent les citoyens. Les deux sont beaucoup utilisées par les entreprises pour construire leur offre et commencent à être exploitées pour les politiques publiques pour connaître l'impact de leur action et les besoins des citoyens. Dans le projet ESPON Big data, on

développe des techniques pour évaluer le *city branding* et les principaux thèmes liés à une ville à partir des RSS des journaux et de tweets. Dans ce contexte, la visualisation doit se confronter à deux défis : (1) le dialogue avec les techniques et les standards de visualisation développés par la cartographie géographique depuis des années ; (2) la question de l'authenticité des données citoyennes (cf. la diffusion d'un usage activiste des traces liées à la ville par les mouvements citoyens ou par le journalisme de données).

Bibliographie

Benkler Y., *The Wealth of Networks: How Social Production Transforms Markets and Freedom,* Yale University Press, 2006.

boyd d. & Crawford K. CRITICAL QUESTIONS FOR BIG DATA, *Information, Communication & Society*, 15(5), 662-679, 2012.

Callon M., Law J., Rip A., *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, Macmillan, 1986.

Cardon D., Zoomer ou dézoomer ? les enjeux politiques des données ouvertes, *Digital Studies*, B. Stiegler (ed), Fyp éditions, 2014.

Chavalarias D, Cointet J-P. Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields. *PLoS ONE* 8(2), 2013

Giraud T. & Severo M., Le périple d'Edward Snowden : analyse quali-quantitative d'un événement médiatique international, *Netcom*, à paraître.

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

Latour, B. Tarde's idea of quantification, *The Social after Gabriel Tarde: Debates and Assessments*, ed. M. Candea, Routledge, London, pp. 145–162, 2009.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. "The Whole is Always Smaller Than Its Parts" A Digital Test of Gabriel Tarde's Monads. *British Journal of Sociology*, 63(4), 591–615, 2012.

Lazer, David, Pentland, Alex, Adamic, Lada , Aral, Sinan, Barabási, Albert-László, Brewer, Devon, Christakis, Nicholas, Contractor, Noshir , Fowler, James, Gutmann, Myron, Jebara, Tony , King, Gary, Macy, Michael , Roy, Deb, Van Alstyne, Marshall. Computational social science, *Science*, 323 (5915), 721-3, 2009.

Manovich, L. Trending: the promises and the challenges of big social data, *Debates in the Digital Humanities*, ed. M. K. Gold, Press, Minneapolis, 2011.

Marres N. & Weltevrede E., Scrpaing the social ? *Journal of Cultural Economy*, 6(3), 313-335, 2013.

Marres N., The redistribution of methods: on intervention in digital social research, broadly conceived, *The sociological review*, 60, 139-165, 2012.

Rogers R., *Digital Methods*, MIT Press, 2013.

Romele A & Severo M, Une approche philosophique de la ville numérique : méthodes numériques et géolocalisation, *Devenirs Urbains*, Carmes M . & Noyer J-M ed., Presses de Mines, 2014.

Severo M & Zuolo E, Egyptian e-diaspora: Migrant websites without a network?, *Social Science Information* 51(521), 2012.

Severo M., Giraud T., Douay, N., The Wukan's protests: just-in-time identification of international media events, *Proceeding of Workshop Just-in-Time Sociology*, *SocInfo international conference*, 2012.

Severo M. & Venturini T. Intangible Cultural Heritage Webs: comparing national networks with digital methods. *New Media and Society*, à paraître.

Tukey, J. W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

Venturini, T. & Latour, B. The Social Fabric: Digital Traces and Quali-quantitative Methods, *Proceedings of Future En Seine*, 2010.

Venturini, T., Gemenne, F., & Severo, M. Des Migrants et des Mots. Une analyse numérique des débats médiatiques sur les migrations et l'environnement. *Cultures & Conflits*, 88(4), 2013.