Le territoire comme un graphe

Mariannig Le Béchec

LE TERRITOIRE COMME UN GRAPHE

Pratiques, formes, éthique

Mariannig Le Béchec

Pour citer cet article :

Le Béchec, M. (2016). Le territoire comme un graphe: Pratiques, formes, éthique. *Les Cahiers du numérique*, vol. 12(4), 131-156. DOI:10.3166/LCN.12.4.131-156

résumé

Le graphe peut être mobilisé par les chercheurs comme une forme de visualisation de données en lien avec un territoire. Cet article interroge les pratiques qui aboutissent à cette forme de visualisation, les biais inhérents à l'utilisation de logiciels et porte un regard critique sur les manières de faire. Dans le graphe, les connexions entre les nœuds, les « lignes » deviennent source de sens et leur forte densité, des clusters. En reprenant l'analyse de réseaux et les théories des réseaux sociaux, mais également l'interprétation par cette ligne, cet article propose de mettre en perspective ces mises en forme et les modèles de pensée qu'elles mobilisent.

Territory as a graph. Practices, forms, ethic

Researchers can use the graph as a form of data visualization in connection with territory. This article examines the practices and the uses of software that lead to this form of visualization and will take a critical look on way of doing. In the graph, connections between nodes – the « lines » – become source of meanings and their high-density shape clusters. This article focuses on the meaning of this line when reading a graph. Based on network analysis and social network theory, this article exposes biases that may alter the interpretation of data. It underlines the ethic when designing the graph.

1. Introduction

La visualisation de données complexes peut être considérée comme une discipline, à l'image de la photographie et du cinéma (Lima, 2013). Les expressions telles que *big data* ou *open data* sous-entendent, selon le sens commun, des données massives qu'il convient de représenter pour mieux les maîtriser ou anticiper des comportements futurs. Ces bases de données majoritairement sous leur forme tableau, une « raison graphique » (Goody, 1982), apportent peu de réponses sans une interrogation de leur contenu via des dispositifs de requête, de traitement et de calcul mobilisés par les chercheurs et les professionnels. hercheur comme le professionnel utilise également des dispositifs de traduction envers le plus grand nombre de ses bases de données à l'aide de logiciels qui permettent leur visualisation. L'objet de cet article porte sur une forme de visualisation qu'est un graphe.

Un graphe est un ensemble de sommets nommés nœuds et d'arcs que sont les liens entre ces nœuds. Les indices pour les repérer sont donc les points pour les nœuds et les flèches ou lignes pour les arcs, dans le cas de graphes dirigés. Ces graphes sont obtenus à partir de bases composées de données, parfois nommées traces, extraites du web à l'aide de logiciels qui passent de sites en sites web ou de comptes en comptes utilisateurs de plateformes numériques. Les données une fois extraites peuvent alors être visualisées à l'aide de logiciels. Parfois ce qui est nommé « cartographie du web » est légendé par les chercheurs avec la mention d'un nom, d'une région ou d'une ville, renvoyant ainsi à une forme de territorialisation du web et

à une cartographie d'un territoire sous forme de graphes. Nous souhaitons donc interroger dans cet article cette visualisation d'un territoire sous forme de graphe, c'est-à-dire cette forme constituée d'un ensemble de points reliés par des lignes qui est construite à partir de données extraites du web.

Cette forme de visualisation de données n'est pas un domaine proprement lié au design informationnel. Elle devient quasiment inhérente au travail empirique du chercheur avec les outils numériques, surtout dans une approche quali-quantitative. La recherche que nous avons conduite sur le livre papier nous a fait passer d'entretiens semi-directifs à des noms de domaines de sites web, de blogs que les lecteurs consultaient ou éditaient. Au final, le graphe présenté dans cet article contient 3 286 sites web, appelés nœuds.

Si le web peut être visualisé sous forme d'un graphe, en quoi le territoire et les pratiques informationnelles qui y sont associées peuvent-ils eux aussi être visualisés sous forme de graphes ? Le territoire comme un graphe, pratiques, formes et éthique fait référence au premier article qui nous a amené vers cette forme de visualisation. Cet article s'intitule « *The web as a graph : measurements, models and methods* », publié par Kleinberg *et al.* (1999). Nous montrerons que les travaux menés par ces chercheurs influent les interprétations faites avec le logiciel d'analyse de réseaux et de spatialisation, Gephi (Heymann, 2014). La visualisation qui en résulte interroge à la fois nos pratiques au sens d'activités de recherche, d'interprétation des résultats et également notre éthique.

Nous reviendrons tout d'abord sur les travaux de recherche dans le champ des sciences de l'information et de la communication (SIC) où les graphes servent à alimenter les « explorations » d'un espace documentaire qu'est le web. Cette visualisation pose deux questions : la première est sur la dimension panoptique par la vue d'ensemble et la projection en deux dimensions qu'elle propose (Latour, 1993) ; la seconde, *a contrario,* est le zoom qui fait passer d'une vue d'ensemble de sites web, à un agrégat, un sous-ensemble pour finir à ce nœud, un site web qui semble isolé même s'il est relié à d'autres (Cardon, 2015). Enfin, dans cette visualisation des données, deux autres questions se posent face à la possible performativité de ces productions scientifiques : quelle éthique (Rogers, 2015) et quelles « méthodes d'interface » (Marres, Gerlitz 2015) sont-elles à l'œuvre ?

2. La visualisation de la topologie du web territorial

Cette première partie interroge les pratiques des chercheurs, ce que Rieder nomme la socionumérisation, à partir du concept de sociodigitization développé par Latham et Sassen.

> *« La socionumérisation a pour effet d'ancrer certains aspects d'une pratique dans le monde des machines où chaque entité, relation ou processus est représenté en forme de donnée ou d'algorithme. »* Rieder (2010, 92)

La forme de visualisation interrogée ici est le graphe qui repose sur des données et des traitements algorithmiques et statistiques. Ce graphe peut-il permettre de visualiser un territoire ?

2.1. Quelques récits d'explorations du web

En 1736, Leonhard Euler, mathématicien suisse vivant à Saint-Pétersbourg, crée le Königsberg graphe. Königsberg est une ville comportant une rivière à deux branches et sept ponts qui l'enjambent (Barabási, 2003 ; Watts, 2004). Le problème de l'époque est de savoir s'il est possible de franchir les sept ponts sans passer deux

fois par le même. Euler infirme cette proposition, en formalisant graphiquement des nœuds (des points de la ville) et des liens (les ponts), donnant ainsi naissance à la théorie des graphes. Il est à noter que ce réseau invisible, puisqu'imaginé, prend pour terrain d'observation et de démonstration le territoire physique d'une ville. Barabási et Watts mentionnent ces travaux pour l'appliquer à d'autres disciplines et d'autres terrains de recherche que celui de la scientométrie ou de l'infométrie. Dans la pensée d'Euler et de son modèle c'est la relation, le lien entre deux points, qui est privilégiée. Le modèle n'est pas celui de l'arbre mais du graphe. Et la pensée hypertextuelle dans les travaux de Bush (1945) par exemple renvoyait à cette notion de réseau, de lien entre documents soit un graphe.

Comme nouvel espace, le web n'offre plus les mêmes repères, les mêmes échelles, ni le même système d'orientation que dans le territoire. Le graphe d'Euler se compose de points et de lignes, il ne nécessite pas d'avoir en toile de fond les tracés des rivières et des sept ponts de la ville. Le chercheur sans cette toile de fond sur le web devient alors un « explorateur » selon ses propres termes (Plantin, 2012). Le graphe lui permet de surplomber des « territoires » inconnus, de se repérer dans ses hypothèses et de créer une carte. Cette référence au terme exploration est étonnante et peut renvoyer à la manière dont les chercheurs extraient des métadonnées liées aux sites web. L'outil et la manière de faire sont l'usage de *crawler*, littéralement un collecteur. Ce logiciel par les liens hypertextes passe de page web en page web, d'image en image. Lors de cette exploration automatique, les attributs des éléments de la page web (URL, balise, etc.) sont extraits. La maîtrise et la vision panoptique reposent en partie sur l'utilisation d'un autre logiciel d'analyse et de spatialisation de réseaux, comme Gephi (www.gephi.org) lancé en 2008 (Heymann, 2014). Mais, il faut convenir que les repères visuels du territoire au web sont modifiés. Ghitalla (2008) ajoute :

> *Au niveau où nous sommes ici, celui des documents numériques distribués en réseau, il n'y a aucune chance de trouver l'équivalent d'un territoire que l'on pourrait arpenter, mesurer, étalonner, baliser pour le cartographier.*

Un type sémio-graphique se met donc en place pour la maîtrise des territoires numériques. Une « carte du web » ne possède pas d'échelle ni de système d'orientation, mais la « géodésie hypertexte du web » avec un « effet de pouvoir » reposant en partie sur cette représentation sous forme de graphes et les systèmes d'orientation qu'ils mettent en place (Ghitalla, 2008).

Quand les chercheurs travaillent à partir de cette visualisation sous forme de graphe, lorsqu'ils étudient des controverses en ligne, ils ne mobilisent pas obligatoirement dans leur système d'orientation le fond de carte du territoire mais le nom d'une ville, comme par exemple « débat en ligne sur la radiation après Fukushima » (Plantin, 2013). Le graphe accompagne dans ces exemples l'exploration d'un web en lien avec un territoire de référence précisé par un nom de lieu, de ville ou de pays. Le territoire devient alors un graphe. Ce graphe repose que une « base de données », c'est-à-dire un ensemble de données qui ne sont pas simplement juxtaposées les unes aux autres, mais qui sont liées entre elles. Ainsi la métrique au sein du cyberespace peur se mesurer par des distances caractérisées par l'accès ou le non-accès (Mitchell, 2005) ou, la présence ou la l'absence de lien hypertextes, nommé également hyperlien. Le web est considéré comme un territoire que peut cartographier à l'aide de graphes, forme qui rend compte graphiquement de l'hyperlien.

2.2. L'extraction et la visualisation sous forme de graphes statiques

Dans les graphes du web, les attributs extraits par un *crawler* ne sont pas obligatoirement conservés. Ils peuvent être retirés par les chercheurs, supposant une erreur de la part du logiciel collecteur ou une erreur de stratégie de la part de l'éditeur du site web. En effet, la métrique dans les graphes est celle de la connexion, combien de chemins peut-on emprunter du nœud A au nœud B ? Imaginons un profil sur les plateformes *Facebook* ou *Twitter* sans lien, c'est quelque peu incongru par rapport aux objectifs des responsables de ces entreprises qui parlent de graphe social. Nous allons donc montrer que cette forme de visualisation par le graphe est sélective et statique. Nous nous intéresserons donc ici à la socionumérisation à l'aide de graphe.

2.2.1. Extraction et pratiques automatisées de constitution d'un corpus

Le graphe suivant est issu d'une recherche menée en 2010-2011, au medialab de Sciences Po, dans le cadre du projet SOLEN (FUI, fonds unique interministériel). L'intérêt porte sur la circulation d'un bien, le livre, et sur la circulation de la conversation-livre sur le web. Lors des 141 entretiens avec les lecteurs, ils nous ont mentionné environ 343 noms de domaines. Ce nombre comprend également nos propres observations de sites web grand public, de réseaux d'échanges, de ventes de livres, d'institutions, d'éditeurs, de libraires, d'événements littéraires, de syndicats, de bibliothèques principalement.

Ce premier corpus a été ensuite associé avec 440 blogs, plus ou moins littéraires du corpus « Linkscape » de la société Linkfluence. Cette société a réalisé un crawl, i.e. une extraction automatique en décembre 2010 avec récupération des URL des sites web à n+1, c'est-à-dire à 1 de distance d'hyperlien du corpus initial et possédant au moins deux hyperliens avec celui-ci. Ainsi dans les réglages, tout site web ne possédant pas au moins deux hyperliens vers un des 783 sites web du corpus initial est éliminé. Il faut comprendre ici que les métriques dans le graphe avec un nœud déconnecté vont être nulles. Ces nœuds déconnectés vont abaisser la densité, c'est-à-dire qu'à densité égale à 1, tous les liens possibles entre les nœuds sont présents dans un graphe. L'objectif du chercheur est de s'approcher le plus possible de 1, comme gage de pertinence de son corpus. Ainsi, le graphe comme système d'orientation oblige à la connexion à l'intérieur du corpus défini par le chercheur et suppose une élimination d'ensembles plus ou moins variables d'URL extraits par le logiciel collecteur. Ce critère éliminatoire renvoie également à la loi d'attachement préférentiel définie par Barabàsi (2003) comme le fait qu'un nouveau site web développe toujours deux connexions dont l'une vers le site web le plus connecté du réseau.

Après ce premier crawl, le volumineux fichier obtenu se compose de 7 187 sites web et de 120 217 hyperliens. Il a été « nettoyé » selon quatre critères que sont les sites web hors de la thématique, pas en langue française, inactifs en fonction de la dernière date de mise à jour, et faiblement connectés. Ce ne sont donc pas uniquement des critères liés à la thématique de la recherche qui définissent le corpus final mais également des propriétés propres à la théorie des graphes, où la faible connexion devient un critère éliminatoire. L'exploration des pratiques informationnelles et sa restitution sous forme de graphe ne représentent donc pas un quelconque réel.

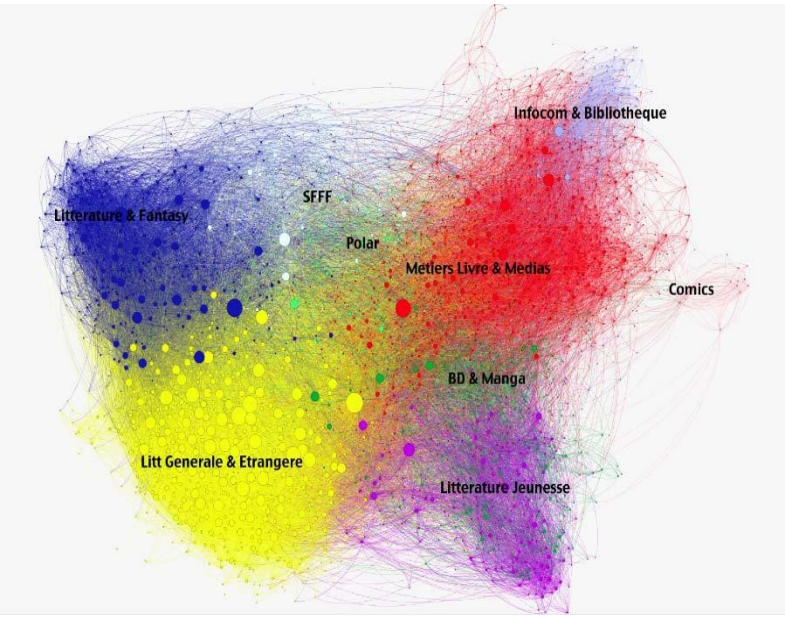

Figure 1. Graphe du « Web du livre en France », 12/2010-03/2011,
http://www.medialab.sciences-po.fr/publications/webdulivre/

Le graphe « web du livre en France » présenté ci-dessus se compose de 3 289 sites web reliés entre eux par 52 884 hyperliens. Dans la base de données disponible dans l'onglet « laboratoire de données du logiciel Gephi, ces URL sont des lignes possédant un identifiant (ID). Lors de la vérification de la cohérence des sites web par rapport à la thématique, une opération manuelle est réalisée. Elle consiste à catégoriser manuellement avec des mots-clés le type d'acteurs (blogeur, libraire, éditeur, e-book, etc.) et le genre (BD, Fantasy, étranger, etc.) en étudiant chaque site web. Ces mots-clés sont nommés descripteurs dans la base de données. Suite à ce travail manuel, une visualisation avec le logiciel est effectuée. Obtenir le graphe de la figure 1 oblige à utiliser les algorithmes de spatialisation et de modularité soit d'agrégation auquel s'ajoute un coefficient de clusterisation. Dans le premier, l'algorithme positionne les sites web dans un espace vectoriel. Dans le second, l'algorithme calcule la densité des hyperliens externes qui relient des sites web entre eux et différencie des zones de connexions plus denses entre sites web du graphe, nommées ici clusters. L'interprétation de ce graphe a fait l'objet d'une précédente publication que nous reprenons partiellement (Le Béchec et *al.*, 2014). Notre intérêt porte sur la visualisation des données extraites à partir de l'onglet « Vue d'ensemble » du logiciel.

Ce corpus visualisé sous forme de graphe s'appuie sur des entretiens avec des lecteurs. Le relevé des sites web à n+1 étant automatisé, le chercheur ici n'explore pas à proprement parler le web. La construction de ce corpus suppose plus un prélèvement à partir de variables que le chercheur choisit, valide à partir de l'extraction et de ses hypothèses.

2.2.2. L'agrégat ou quelle échelle d'analyse dans les graphes ?

Regarder un graphe avec 3 289 nœuds et 52 884 liens peut laisser perplexe en termes d'analyse visuelle. À partir des années 1990, la rencontre entre les réseaux et la visualisation sert à

rendre compte de la complexité avec des formes simples : points, lignes et courbes. Dans la figure1, un score est calculé pour chaque nœud. L'information va être traduite par « plus ce score est important plus la taille du nœud est importante ». La variation visuelle de la taille du nœud traduit donc une grandeur. Pour rendre compte de cette grandeur, le chercheur travaille sur les connexions entre les nœuds, c'est-à-dire la topologie du web.

La topologie du web oblige à tester un certain nombre d'indicateurs, et à s'intéresser à la théorie des agrégats de Kleinberg et son algorithme HITS (2006). Depuis 1950, des méthodes quantitatives se référant à un langage mathématique sont appliquées en sociologie et en anthropologie pour des données ethnographiques (Degenne et Forsé, 2004). Ainsi les termes tirés de la théorie des graphes comme « centralité de l'acteur », « distance », « cliques », « éléments connectés » viennent décrire l'influence, la cohésion, les rôles sociaux, les « identités » dans les réseaux sociaux. Mais un réseau social n'est pas une allégorie de l'acteur collectif. Il est construit méthodologiquement pour un ensemble fini d'acteurs (Lazega, 2007).

La théorie des agrégats se retrouve dans la visualisation des graphes. Le site web qui est un nœud a des propriétés que sont le nombre d'autres sites web qui le citent (nombre de liens entrants) et le nombre d'autres sites web qu'il cite (nombres de liens sortants). Les hyperliens résultent de, ou pour reprendre le terme de Kleinberg et al. (1999), ils encodent, un « jugement humain latent » ou caché. C'est sur ce jugement qu'ils fondent la notion d'autorité. L'attribution d'un lien hypertexte d'un acteur A à un acteur B, est une mesure de l'autorité de B, même s'il existe un certain nombre de négligences par omission dans cette proposition. À travers l'algorithme HITS, Kleinberg montre que le « poids » d'un nœud dans un graphe varie suivant le nombre de liens hypertextes sortants et le nombre de liens hypertextes entrants, distinguant en ce sens les nœuds pivots, appelés hubs (liens sortants), des nœuds d'autorité, authority, (liens entrants). Cette variable s'appuie sur la présence de lien entre A et B, donc une dyade mais elle ne traite pas de la pragmatique du lien. La motivation d'un lien a été étudiée ultérieurement par Mohammadi et al. (2015) à partir des signets sur la plateforme Mendeley. Sur la plateforme Twitter, Alloing en analysant les pratiques de recommandation et les attributs des comptes, définit l'autorité réputationnelle de ces derniers (le Béchec, Alloing, 2016).

Cette variable ne nous permet ici que d'observer des sous- ensembles cohérents, que nous présentons comme des clusters en fonction de la densité des liens. Dans la figure 1, on peut observer des clusters distants et représentés à l'échelle du graphe complet, par exemple, *comics* et littérature générale et étrangère. Dans le graphe consultable sur le web, il est possible de sélectionner un site web et de voir un graphe égocentré. Mais que veut dire un graphe égocentré et des dynamiques propres à un site par rapport à un agrégat. Il est possible de produire des typologies à partir de matrices reposant sur le nombre de liens entrants et sortants et des attributs (Le Béchec, 2010) ou selon la recherche de notoriété des éditeurs de sites web (Cardon *et al.,* 2014). Le principe est donc d'individualiser le rôle du site web au sein d'un graphe et de l'associer à une stratégie de présence sur le web de la part de l'éditeur par le nombre d'hyperliens sans interroger obligatoirement les contenus mis en scène.

2.2.3. Dépasser la dyade

Afin de dépasser le nœud ou sa dyade, sa relation avec un autre site web, ce sont les densités des liens dans le graphe qui vont être recherchées. L'agrégat qualifie les nœuds connectés et traitant d'une même thématique, comme le montrent les clusters de la figure 1, quelques nœuds en bordure des clusters exceptés. Dans le logiciel Gephi, il existe une possibilité de calcul appelée modularité, définie comme « un algorithme de détection des communautés ». Cette notion de communauté soulève

alors d'autres enjeux si nous nous référons à sa définition comme communauté de sans ou de voisinage par Tönnies en 1887, comme communauté en réseau (Proulx, Latzko-Toth, 2000), comme communauté imaginée (Anderson, 2002) ou comme communauté d'utilisateurs d'un nouveau nom de domaine de 1er niveau pour l'*Internet Corporation for Assigned Names and Numbers*. Le web offre donc la possibilité d'un voisinage d'un genre nouveau.

Dans la figure 1, les clusters sont ceux de genres littéraires ou de la typologie d'acteurs construite dans le cadre de l'analyse et à partir de cette fonction de modularité. La position des clusters les uns par rapport aux autres, comme l'opposition entre le cluster des lecteurs dans la blogosphère littérature et Fantasy avec les éditeurs présents dans le cluster « métiers du livre » est aisée à expliquer. L'autorité d'un site web, ici *amazon.fr*, par le nombre de liens entrants explique cette saillance au centre du graphe. Mais cette position s'explique également selon la structure du graphe et les biais de construction du graphe. En l'état sa stratégie numérique ne peut être interprétée sans une connaissance des pratiques des libraires d'occasion dans le *market place* de ce site web. De même, la catégorisation réalisée par le chercheur d'un site web ne présume pas des regroupements faits par l'algorithme de modularité qui s'appuie sur la densité des liens. Les sites web des deux catégories en BD/mangas et littérature jeunesse s'entremêlent, des sites web catégorisés dans un autre genre apparaissent alors dans le cluster d'un autre genre.

Cette fonction de modularité regroupe des nœuds dans des clusters et se confronte aux catégories définies par le chercheur ou les conforte. L'algorithme de Louvain utilisé pour détecter des regroupements au niveau de l'ensemble du graphe n'est pas adapté pour définir la position d'un site web, qui peut se situer entre deux clusters (Guillaume, 2015). Les chercheurs travaillent donc sur les triades, présentes en sociologie par exemple avec les travaux sur le tertius gaudens de Simmel (1999), la théorie des liens faibles de Granovetter (1973) ou les réseaux structuraux de type Exponential Random Graph Model (Lazega, 2007). Ces dernières recherches portant sur les triades et l'homophilie permettent d'affiner cette position des nœuds dans un cluster et dans un graphe (Guillaume, 2015).

Mais montrer les relations à un niveau de granularité ascendant et descendant se confronte à la représentation statique à un temps donné de ce type de représentation. L'agrégat dont dépend cette visualisation par le logiciel Gephi n'est pas adapté à présenter des dynamiques dans le temps ni à présenter les stratégies propres à l'éditeur du site web ou à des poids économiques qui dans ce secteur du livre sont très importants avec le numérique.

2.3. Ce graphe est-il un « système de représentation » ?

Un système de représentation est : « *un regroupement actif de structures se mettant automatiquement à jour et organisées de façon à « refléter » le monde au fur et à mesure qu'il évolue »* (Hofstadter, 1987).

Ce graphe n'est pas automatiquement mis à jour. Et la formalisation sous cette forme ne correspond pas aux temps d'extraction lors du *crawl* automatique, ni de traitement du corpus (élimination, catégorisation). Cette visualisation sous forme de graphe statique n'est donc pas un « système de représentation » selon la définition d'Hofstadter. Les dernières mises à jour du logiciel Gephi en 2015 tendent effectivement à rendre compte des dynamiques de connexions dans le graphe en présentant leur évolution dans le temps.

Comme il a pu être constaté lors de précédents travaux sur un web territorial représenté à l'aide de graphes (Le Béchec, 2010) ou d'autres recherches sur la propagation de conversation sur le web (Highfield, 2011), la difficulté est de rendre

compte des dynamiques par des graphes statiques. En effet, un graphe représente un instant t du corpus, c'est-à-dire un ensemble de sites web et la présence d'hyperliens entre eux. Cependant, ces graphes ne rendent pas compte de la localisation, des flux, ni de la morphologie de ce corpus à un instant *t+1* ou *t+2*. En effet, un site web peut être mis hors-ligne. Pouvoir suivre la propagation, la circulation de contenus est un enjeu dans l'analyse du web et des pratiques ordinaires de publication qui s'y déroulent. Un graphe ne rend pas compte non plus des refus de mise en relation, des limites, des hiérarchies entre les éditeurs de sites web. Anderson *et al.* (2015) cherchent à comprendre sur une plateforme de réseau socionumérique professionnelle telle que LikedIn, les connexions dans le temps à partir de cascades d'homophilie. L'intérêt est de comprendre comment dans ces réseaux invariants d'échelle (Barabàsi, 2003), ces plateformes attirent toujours plus de connexions, c'est-à-dire leur propension à développer l'attachement préférentiel, qui fait qu'un nouveau nœud vient principalement se connecter au nœud le plus connecté du réseau.

Comment cet attachement préférentiel peut-il être traduit dans un graphe statique ? La « nouvelle » science lie la description des données, de la plus petite relation à l'échelle d'un graphe et la proposition de modèles sur les réseaux sociaux. Newman et al. (2006) relèvent les travaux d'Erdos et Rényi (1960) sur l'analyse des réseaux aléatoires appliqués à l'information, suite aux travaux de Solomonoff et Rapoport (1951) qui ont montré l'influence et les différents modes de propagation de l'information et des maladies avec la méthode des graphes. Dans le même temps, Sola Pool (chercheur en SHS) et Kochen (mathématicien) rédigent un article en 1958 qui questionne les réseaux sociaux. Cet article inspirera les travaux sur les petits mondes (Milgram, 1967). Mais les structures de ces réseaux (la relation) ne sont pas statiques, elles sont évolutives. Si l'intérêt porte sur le degré de connexion, la connectivité, ces réseaux ont pour principale caractéristique d'être dynamiques. La propagation d'un virus informatique varie selon les connexions des machines dans le temps.

Avec le logiciel Gephi, la taille des nœuds module selon les deux statistiques appliquées sur le corpus que sont principalement le degré de liens entrants et le degré de liens sortants pour chaque nœud. Le degré de liens entrants calcule un score d'autorité tandis que le degré de liens sortants calcule un score de hub. Cette distinction est reprise à Kleinberg (2006) et son algorithme HITS décrit précédemment. Ces scores une fois calculés influent sur la position des nœuds dans l'espace vectoriel du graphe. Nous pourrions prendre l'exemple d'une carte dans laquelle la visualisation des distances entre deux villes varie en fonction du temps de parcours de cette distance en train. Une ville avec le plus de liaisons ferrées aurait alors un poids supérieur à une ville avec peu de connexions. Cette ville, quelle que soit sa population, pourrait donc être placée côté de la ville densément connectée par l'algorithme de visualisation sans tenir compte de la distance en kilomètre.Il existe donc une forme de distorsion dans ce type de graphe que le chercheur peut comprendre mais qu'il ne peut modifier lui-même en recourant à ce type de logiciel. Nous noterons que les différentes mises à jour de l'algorithme Force Atlas de Gephi sont menées avec une réflexion pour l'équipe de développement sur ce point. Toutefois nous admettons nos limites à en traduire tous les enjeux ici et renvoyons à un article d'un des créateurs du logiciel (Heymann, 2014).La visualisation sous formes de graphes oblige tout de même le chercheur à comprendre :

-l'origine des calculs statistiques et algorithmiques qui définissent le « rôle » joué par le nœud dans un graphe ;

- et, les biais inhérents à la détection et à la définition de ce rôle et de son positionnement dans le graphe.

Le chercheur doit également attendre les mises à jour pour pouvoir mettre en évidence des phénomènes de propagation, de circulation des images par exemple. Car ce web territorial, « Web du livre en France » prend en compte un territoire de référence « »physique » mais rend compte également des propriétés hypertextuelles de ce média qu'est le web (Davallon, 2012). Un web territorial est un territoire qui dépend de l'adhésion des internautes à créer des contenus sur cette thématique et dans une dimension temporelle. Le territoire se construit sur des temps longs. Le temps des hommes est différent du temps de la société, du rythme de vie des sciences et des techniques (Braudel, 1984). Or si le web n'est pas défini temporellement par le politique, ou par un artefact tel que le journal, c'est le bien le temps des acteurs qui fait exister un « web » territorial. Ainsi, le web est à la fois un espace et un temps mobilisés, construits, composés entre territoire « physique » et web. Mais le logiciel de visualisation de graphes n'étant pas un système de représentation, il ne peut pas rendre compte de cette dimension temporelle des territoires. Il permet uniquement d'aider à visualiser des topologies du web. Il faut donc chercher d'autres attributs que la connexion pour visualiser un web territorial.

3. La ligne est-elle performative ?

Il existe une limite dans la visualisation d'un territoire comme un graphe qui va amener à rechercher des conventions de lecture. Nous passons du niveau du site web et de ses hyperliens à des ensembles de plus en plus vastes, plus ou moins cohésifs, vus successivement et de manière statique. La focalisation sur la relation entre les nœuds peut-elle induire une performativité de ces graphes ? Nous nous éloignerons de la linguistique (Austin, 1962) pour comprendre la capacité d'action de cette forme de visualisation qu'est le graphe.

3.1. La ligne comme lien, la ligne comme contour

Nous avons précédemment montré que les graphes mettent en valeur la relation entre des éléments. Les hyperliens dans un graphe sont représentés par des lignes entre deux points. Nous souhaitons interroger dans cette partie deux rôles de la ligne, celui du lien, de la relation et celui du contour, de l'isoligne. Car la ligne et sa démultiplication permettent de constituer des clusters aux contours identifiables dans le graphe du « Web du livre en France ».

3.1.1. La ligne comme lien

Appliqué à la notion de réseaux sociaux présents dans les territoires et étudiés par la sociologie, cet objet graphique soulève d'autres enjeux. La représentation graphique qu'est le graphe repose sur trois temps, qui consistent à décrire la fragmentation, puis à valoriser les liens, les intermédiaires, et enfin, le lien devient la relation qui permet de lire le social (Musso, 2003). Le réseau obtenu par le logiciel devient alors un mésoconcept pour identifier le passage, la relation. La ligne comme lien rationalise la relation, définit le groupe, son identité et son fonctionnement, une « liaison graphique » accompagnée par des calculs et une représentation graphique du réseau social. Ces lignes connotant des relations entre des sites web conduisent à opérer parfois une analogie avec le lien social sur le web. La ligne vient alors révéler des concepts abstraits. Cette surinterprétation de la relation est le passage des attributs ou statuts aux relations, à leur nombre, à leur fréquence, à leur direction qui caractérisent les sociétés contemporaines depuis l'article de White, Boorman et Breiger en 1976 (Mercklé, 2004).

3.1.2. La ligne comme contour

Les représentations du territoire passent ou passeraient d'une topographie, une vue d'ensemble d'une surface à une topologie, une vue d'ensemble d'un ensemble de points et de lignes dans les graphes. La ligne joue un rôle important dans la manière de représenter les territoires et les lieux avec les techniques de l'imprimé. Ces techniques ont permis de dessiner les contours, les frontières d'un pays. La carte dans ses vertus heuristiques permet d'identifier le pays et son territoire. « Le territoire devient une forme aux contours identifiables » facilement déclinable sur différents supports comme les timbres, les cartes planes, les affiches politiques, etc. (Anderson, 2002).

Selon Grevsmühl (2014), la ligne de contour intègre à la fin du XVIII$_{e}$ siècle la cartographie marine la topographie terrestre. Puis dans la géographie du XX$_{e}$ siècle, les frontières terrestres issues du relevé topographique sont dépassées et parallèlement une vision holiste de la terre se met en place, le « visiotype de la Terre vue de l'Espace ». Par simplification, la ligne comme contour ou isoligne est associée avec les missions spatiales aux visualisations du « trou d'ozone ». Ces visualisations sont réalisées en associant l'art de la cartographie et notamment des couleurs et ce que nous nommerons « l'art de la computation », c'est-à-dire l'aide d'algorithme informatique. Pour visualiser les « trous d'ozone », la ligne de contour relie les points de même valeur et crée une forme.

La ligne doit également s'associer à une « sémiologie graphique » (Bertin, 2005) pour faire sens et créer des conventions. Le relevé de données, les algorithmes de traitement des données et de visualisation permettent de créer cette convention, le « trou d'ozone ». Il en est de même avec un graphe, la compléxité de leur construction est difficile à transmettre et le logiciel produit un effet de continuité entre les éléments, voire des regroupements à partir de patterns visuels formés de points et d'une densité forte de lignes, devenues relations. La ligne de contour comme l'isoligne a un pouvoir visuel, celle de l'illusion des mesures continues tout en cachant son processus de construction. Le graphe et sa dimension panoptique peuvent également simplifier la complexité d'un objet technique et des usages qui en sont faits. Les sites web et les hyperliens forment des regroupements aux contours flous mais facilement identifiables visuellement, qualifiés de clusters. La signification dépasse alors le contexte de production.

3.1.3. Rechercher des conventions de lecture

Une des difficultés est de trouver des conventions de lecture qui lient ces résultats visualisés avec un graphe à quelque chose qui fasse sens pour tous les acteurs (auteurs ou lecteurs). L'une de ces conventions est souvent la carte du territoire, et plus précisément les contours de ce territoire que sont ses frontières externes. D'une part, parce que des travaux qui retracent l'histoire du web montrent que nous n'avons pas un web mais des webs (Rogers, 2013). D'autre part, parce que l'imprimerie, la gravure et les mathématiques ont produit un « cadre de référence commun » à partir du XV$_{e}$ siècle (Olson, 1998). Autrement dit, notre habitude de colorier les fonds de carte à l'école selon les frontières des territoires, nous amène peut-être, nous chercheurs, à considérer que c'est le seul terrain d'entente avec notre lectorat, surtout s'il est élu. Les données pour qu'elles sortent de leur pauvreté sémiotique des statistiques, des algorithmes et des graphes doivent rentrer dans le désirable par le biais de la narration. Ces données sont localisées, territorialisées, re-territorialisées peut-être pour appuyer une institutionnalisation du web tout autant qu'un marketing territorial (Carmes, Noyer, 2015). La localisation aurait donc valeur de validité, de vérité, de désirabilité pour l'élu ou le chercheur lui-même.

La métadonnée de l'inscription spatiale des acteurs dans le fond de carte suppose également l'inscription spatiale de leurs liens, ces lignes qui associent des points sur la carte. La méthode SWARMS, *Spatial Web Automatic Reasoning and Mapping System* permet de visualiser et d'analyser dans l'espace et le temps les informations, les concepts publiés sur les sites web. Une de ces étapes est la visualisation par localisation sur fond de carte (Tsou *et al.,* 2012). Cette localisation est une « économie épistémique » de visualisation des données en lien avec un territoire.

« Chaque procédé de visualisation produit sa propre économie de visibilités et d'invisibilités, économie que nous avons nommée ailleurs une « économie épistémique » pour désigner justement ce qui relève d'un savoir méthodologique et instrumental propre à chaque procédé de visualisation. » (Grevsmühl, 2014, p. 167).

La visualisation sous forme de graphe et son procédé amènent à une « économie épistémique » en négligeant le savoir méthodologique et instrumental indispensable à sa conception et sa lecture. Le nord, le sud, l'ouest et l'est sur un graphe ne font plus sens commun, et le fond de carte pourrait ainsi devenir une convention de lecture d'un graphe territorialisé. De plus, les graphes semblent à même de rendre compte de la complexité du web car ils fournissent au même titre que la carte une vue d'ensemble qui dépasse alors l'échelle unique du territoire-pays (Lévy, 2013). Toutefois, pour être compris par les lecteurs, les graphes doivent être accompagnés d'une mise en récit des données qui le composent.

3.2. Visualiser un territoire à l'aide de graphes ?

Ce procédé de visualisation semble s'éloigner de la notion de territoire qui s'est construite dans le temps en référence à un territoire « physique », à un espace « concret », avec une métrique, c'est-à-dire un calcul de la distance, une mesure d'un écart entre deux points selon le plan de projection. Trivialement, sur le papier, la métrique euclidienne peut se mesurer avec un double décimètre. Mais sur le web, il semble qu'une métrique différente vienne qualifier les distances.

> *Il existe deux grands types de métriques : topographiques (territoires) et topologiques (réseaux). L'opposition entre topographie et topologie offre un principe de différenciation économique puisque simple et efficace. Soit la métrique est continue et ce qui compte est la mesure, comme pour un objet dénombrable, de la distance entre deux points ; soit elle est discontinue, et l'important est le nombre d'unités de distance discrétisée qui séparent deux points.* (Lévy, 1994, p. 50)

Analyser un web dans sa dimension territoriale comme dans le graphe du « Web du livre en France » repose encore sur des méthodes qui cherchent des conventions. Ces méthodes sont cependant indispensables car le terrain d'observation qu'est le web repose sur un système de liens hypertextes. Visualiser un territoire à l'aide de graphes s'inspire des travaux de la « nouvelle » science définie par Newman *et al.* (2006). Cette science propose une autre approche de terrains empiriques et des modèles mathématiques. Ses fondements reposent sur les structures et les dynamiques des réseaux sur le web passant d'une analyse topologique des réseaux à une analyse de systèmes dynamiques distribués en train de se construire. Cette science n'est pas une pure théorie des graphes pour des modèles mathématiques, ni une approche « design » ou « ingénieure », mais une analyse de réseaux non planifiables et décentralisés comme les réseaux biologiques, les réseaux d'information (la citation dans les articles scientifiques ou le *World Wide Web*, par exemple) qui sont nommés des réseaux invariants d'échelle (Barabàsi, 2003). Par exemple, les travaux de Zook et *al.* (2011) qui s'inscrivent dans les Websciences s'intéressent à la variation de la distance entre des villes mondiales à partir de la distance « physique » et sur le web. Ils

ont conçu des modèles statistiques à partir des repères générés et indexés par les utilisateurs sur *Google maps.* Ces auteurs travaillent alors sur ce qu'ils nomment des « patterns » invisibles de flux d'informations. Si les données permettent de dépasser les frontières et d'intégrer un espace global, le territoire n'est pas obligatoirement déterritorialisé. Les graphes permettent ainsi d'identifier des saillances visuelles ou patterns, comme dans le cadre des clusters de la figure 1.
Les hyperliens entre sites web extraits à partir de logiciels collecteurs, pour ensuite être visualisés à l'aide d'un logiciel de visualisation utilisant des algorithmes ont des inconvénients et des limites. Il existe des erreurs dues à l'absence de relevés de certaines pages web par l'application web, à l'absence de visualisation de nœuds et de liens par le logiciel de visualisation par exemple. Toutefois, face aux volumes des corpus et à leurs limites indéterminées, ces outils demeurent indispensables afin d'établir une base de données et d'en proposer une représentation. Cependant, les graphes présentés ne résultent pas d'une matrice d'adjacence, ni d'une application de la sociométrie. Ainsi il conviendrait de parler non pas de graphes mais de dessins automatiques de graphes, d'un objet graphique (Cristofoli, Vilaça, 2002). Ce dessin automatique des graphes pose des problèmes de lecture quand il est dense, car la position des sommets (des nœuds) se fait selon des forces (attractives et répulsives) définies par l'algorithme de visualisation. Il existe donc une sélection par la lisibilité des données ou par ce que le graphe doit montrer selon les hypothèses du chercheur. Les auteurs empruntent alors à Christian Grataloup la notion de « mode d'écriture ».

> *« Le dessin d'un graphe est, de la même manière qu'une carte, une technique, un mode d'écriture des données. »* (Cristofoli, Vilaça, 2002)

Toutes les recherches présentées travaillent de nouvelles métriques pour analyser un territoire et les graphes permettent ici de signifier une certaine discontinuité dans la mesure des distances. Les trois tendances méthodologiques (Rogers, 2015) d'analyse du web sont : l'analytique culturelle (métrique des images, des *selfies* sous forme de grille) ; la culturnomique (métrique linguistique visualisée sous forme de graphiques) et la cybermétrie (métrique des citations entre sites web visualisées sous forme de graphes). Les analyses culturelles (*cultural analytics*) ou culturomiques développées par Manovich (2001, 2014) s'intéressent aux formes, niveaux de gris, luminosité etc. des images. Tifentale (2014) lors de son analyse quantitative et sémiotique des images exportées à partir de la plateforme *Instagram* sur une durée de 144 heures et localisées à Kiev, autour de la place Maïdan plus précisément, souligne la présence et la mise en scène du drapeau national par les internautes. Selon nous, une méthode d'analyse d'un web territorial défini comme tissé à partir de liens créés entre des attributs, des formes élémentaires qui font tenir des collectifs à géométrie variable, oblige à associer différentes métriques pour analyser ces contenus en réseaux.

3.3. La performativité du graphe

Cette visualisation sous forme de graphes par les agrégats, ces patterns visuels qu'elle engendre, peut-elle être performative ? Austin (1962) définit la notion de performativité au sein d'une discipline qu'est la linguistique et afin de décrire le langage ordinaire. À travers le concept de *performative agency*, Butler (2010) démontre que ce n'est pas l'émetteur, une institution mais le discours lui-même qui a cette capacité d'action. La notion d'*agency* (Hoskins, 2006) souligne cette capacité des discours, des objets à se détacher de leur contexte de production pour circuler. Notre proposition est d'interroger comment la visualisation et la spatialisation sous forme de graphes ont un caractère performatif à travers la réflexivité, le changement de comportement qu'il peut susciter chez les acteurs. Nous ne pourrons pas montrer

que le graphe est un acte perlocutoire ayant des effets par exemple. De ce fait, nous interrogerons la manière dont ce graphe est construit par le chercheur.
Dans la cartographie conventionnelle, la carte est « un instrument de « mise en visibilité » de son territoire par un État qui maîtrisait toutes les étapes de contrôle de cette conception cartographique. » (Noucher, 2015). Il construit les bases de données via le recensement et des services comme l'Institut national de la statistique et des études économiques qui fournissent ce qui est communément nommé des *hard data*. Puis, il projette sa maîtrise du territoire sur un fond de carte produit par exemple par l'Institut géographique national, *i.e.* un espace défini et délimité par des frontières. L'État projette des données qu'il a lui-même choisi de récolter sur un fond de carte. Son but est étatique. C'est l'intentionnalité cartographique, celle devant objectiver le réel qui doit être interrogée pour ce chercheur qui s'intéresse aux modes d'écriture des cartes.
Les graphes opèrent-ils également cette même mise en visibilité ? Après le « c'est quoi ? » qui est la première réaction des personnes qui découvrent un graphe, la seconde est « où suis-je ? » Quelle est alors la réflexivité ? Le « monde sur papier » (Olson, 1998) rentre ici dans le graphe : « Vous êtes ici ». Or ce qui devrait être indiqué est plutôt : « voilà l'état de votre site web ou de votre profil sur telle plateforme et de vos hyperliens à la date à laquelle ces métadonnées ont été extraites par un logiciel dans notre recherche ». Le graphe est une mise en forme qui ne repose pas sur des conventions de lecture, quel que soit l'âge des interlocuteurs. Le même étonnement se rencontre chez les étudiants face à une image de graphe du web. La réponse est souvent de décrire cette image comme une constellation. Or dans la sémiologie graphique de Bertin (2005), l'image doit être compréhensible instantanément et être globale. Le graphe qui s'appuie sur les relations, *i.e.* les liens entre les nœuds, attache à un ensemble ce qui n'est qu'un point. Cet ensemble fait rarement sens visuellement. Ce sont donc les statistiques appliquées sur ce graphe qui donnent un ensemble de repères de lecture principalement aux chercheurs. Les textes, appelés labels des nœuds dans Gephi, la légende ou les couleurs attribuées aux nœuds selon les descripteurs choisis par le chercheur ou selon des statistiques du graphe contribuent à l'identification de patterns visuels et à leur interprétation. La taille des nœuds peut varier selon le degré de liens entrants par exemple, et le chercheur se doit d'expliquer la manière dont il les a catégorisés et dont il en rend compte par le graphe, une honnêteté méthodologique soulignée par De Maeyer (2011). Nous pouvons ajouter : qu'en est-il de la publication publique de la base de données à partir de laquelle ce graphe est généré ?
Dans le cadre supposé que ce graphe puisse avoir des effets sur les décisions des acteurs concernés, il convient d'expliciter les critères de retrait des sites web. Le chercheur intervient alors directement sur le degré des nœuds du graphe en retirant ainsi des hyperliens. Or, ces graphes peuvent être lus comme une représentation précise de ce qu'est le web car il est aisé de croire que tout devient traçable *a posteriori* dans nos pratiques quotidiennes de ce média. Mais cet instrument de « mise en visibilité » de la stratégie d'un nœud, d'un cluster dépend tout autant des hypothèses (nous osons l'espérer) que des modes d'extraction, de collecte, d'élimination, de catégorisation, de visualisation, de traitement statistiques, de filtrage des données à partir de seuils définis par le chercheur. De plus le chercheur peut sélectionner de ne restituer que les nœuds ayant un degré entrant précis ou définir des requêtes plus complexes, notamment par les options de filtrage proposées dans Gephi. Le rendu est donc une mise en forme des données effectuée par le chercheur.

4. Quelle éthique pour le chercheur ?

Si ces graphes peuvent être performatifs, quelles responsabilités prend en charge le chercheur dans le cadre de la visualisation et de l'extraction, notamment avec les modes

de collecte par automatisation ou via des API, *Application Programming Interface,* littéralement interface de programmation ? L'analyse du web de plus en plus réalisée via ces plateformes interroge ce que Marres et Gerlitz nomment les « méthodes d'interface » (2015), c'est-à-dire des hypothèses de recherche guidées par les plateformes elles-mêmes.

4.1. La donnée : entre démesure et indéfinition

Le web n'est pas lié uniquement à l'exploration d'un espace documentaire, mais à l'exploration de contenus produits principalement par des internautes ou des robots, comme les *crawlers* qui peuvent laisser des « traces » de leur passage.
Nous avons souligné que ce sont tout autant les théories applicables à ces réseaux que les attributs du chercheur qui vont définir le rôle d'un nœud dans un graphe. L'un des calculs appliqués dans un graphe est celui de la connectivité. Or, la première difficulté en SIC est de dépasser l'approche media comme dispositif et d'oublier ses propriétés. La notion d'hypertexte soulève des enjeux épistémologiques (Davallon, Jeanneret, 2004), mais elle est constitutive de la pensée en graphe du web. La seconde est de prélever un objet et de l'analyser comme tel en présumant qu'il est l'indice d'un lieu ou d'une pratique. L'élaboration des indices dans ces dispositifs de visualisation demeure complexe. Il faut donc sortir des cas choisis par Peirce où « l'indice se donne immédiatement parce que la causalité et l'identification existent (comme pour la girouette) » (Davallon, 2004). Les évidences doivent être discutées, le web n'est pas un territoire, le site web n'est pas un lieu, les agrégats de nœuds dans un graphe ne sont pas des communautés, le profil n'est pas un individu. Il convient de sortir de l'évidence qui lie ces objets que l'on découpe selon une certaine connaissance d'un alignement entre « objet concret, objet scientifique, objet de recherche » (Davallon, 2004). Cette première mise en garde pourrait être couplée à celle d'une recherche uniquement guidée par l'accessibilité des données sur les plateformes, les « capta » décrites par Drucker (2011). Ces notions d'indices ou de capta montrent qu'il est difficile de qualifier ce qui est extrait et visualisé. Dans notre propos, nous avons utilisé le terme métadonnée. Mais le terme le plus couramment mobilisé est celui de données.
La définition la plus simple que l'on peut attribuer au terme donnée est un élément brut qui ne supposerait aucune mesure, ce qui demeure très complexe à repérer. Gitelman dans un ouvrage collectif qualifie ces données brutes d'oxymore (2013). La donnée est le plus généralement issue d'une mesure effectuée comme la température de l'air. En informatique, la donnée est structurée selon son langage de traitement. Mais nos mobilisations en SHS du terme « donnée » ne sont jamais au singulier. Ce sont des données, voire des bases ou jeux de données.
Le rapport du Comets, Comité d'éthique du Centre national de la recherche scientifique en France (CNRS), créé en 1994, le souligne dès 2009. Son guide édité en 2014 s'appuie sur quatre points : 1) « collecte, traitement et archivage de données personnelles » en renvoyant à la Commission nationale de l'informatique et des libertés (CNIL) ; 2) « les données massives (Big Data) » en renvoyant à une charte éthique et big data, publiée en 2013 ; 3) « l'archivage des données » et 4) « les conduites inappropriées associées au traitement et à l'archivage des données ». Ainsi il est conseillé au chercheur de protéger les données et dans le même temps, d'assurer la possibilité d'exploitation des données par des tiers, en raison de la reproductibilité de la recherche et d'assurer la conservation pour des générations futures, de données « brutes ». Toute la question est de savoir si des données « anonymes » sont des

données brutes alors même qu'elles ont été partitionnées par un logiciel. Ce sont donc les pratiques, les activités informationnelles qui sont la source de ces données. À plusieurs reprises le terme de traçabilité est également utilisé.
« *Mettre la main sur la trace, ce serait à la fois tenir un modèle performant de la communication et pouvoir manipuler le réel. »* (Jeanneret, 2011).
Les graphes deviennent pour cet auteur « une réalité sémiotique homogène » qui cartographie des traces qui pourraient être nommées des reflets, et qui fait passer de documents à des collectifs de personnes. Or, ces traces sont des indices dans un contexte d'éditorialisation particulier. La trace est un signe, dont l'éditorialisation sous forme de graphes confère une maîtrise, voire un certain pouvoir notamment par l'interprétation. Les traces seraient donc facilement enregistrées, conservées et transmises (Venturini, 2012). Toutefois, nous pouvons constater que qualifier ces traces laissées par les utilisateurs du web en dehors des indicateurs fournis par les grands opérateurs, comme G*oogle Analytics* est complexe. La mise en bases de données de ces pratiques demande un investissement important de la part des entreprises qui possèdent pourtant le back-office d'un site web. Dans une étude en cours sur le financement participatif, une entreprise a accepté de créer une vue de cette base de données selon les hypothèses des chercheurs. L'analyse des commentaires sur des pages web intéresse le chercheur en SIC qui définit ainsi ce qu'il souhaite observer. Une fois la base de données récupérée, tout un travail de nettoyage, de traitement manuel et automatisé des données sera nécessaire pour obtenir un corpus. Il sera également nécessaire de la compléter par des extractions connexes sur d'autres plateformes afin de vérifier des hypothèses.
La visualisation des données demande tout un travail de l'extraction à la présentation des résultats. Latour (1993) les nomme « obtenues » et non « données ». Nous pourrions même ajouter que ces « obtenues » ne sont en rien des informations que nous pouvons transformer en connaissance sans poser sur ces « obtenues » un ensemble d'hypothèses dont découlent les catégorisations utiles à la visualisation sous forme de graphes. Ces « obtenues » deviennent vite pour le chercheur des « tortueuses » en raison du traitement manuel et fastidieux qu'elles supposent. Ces données ne parlent donc pas d'elles-mêmes, et elles ne sont pas brutes lors de leur visualisation sous forme de graphes.

4.2. L'obligation de la donnée et le croisement des données

Rogers (2015) introduit le risque d'observer plus le fonctionnement des réseaux sociaux numériques qu'une quelconque tendance sociétale en raison de cet accès foisonnant aux données. Les graphes reposent sur un processus méthodologique simple : extraire, accéder, traiter, analyser, produire et présenter les données. Boullier (2015) définit même en ce sens les 3e générations de quantification en SHS qui reposent sur le traitement de données, de traces. Mais l'automatisation de la collecte néglige dans de trop nombreuses études l'éthique sur l'utilisation de ces données à caractère personnel, tandis que le traitement par agrégation ou par localisation des données pour présenter des résultats peut conduire à une « cécité computationnelle », celle de leurs conditions de production au nom de la perspicacité pour Rogers.
Qui du chercheur ou de la plateforme, telle que Facebook, influence le choix de la méthode, s'interrogent Marres et Gerlitz (2015). Tout autant que l'éthique lors de l'extraction, la catégorisation ou la visualisation, ce sont les méthodes qui sont à interroger. Face à ce « troublant mystère méthodologique », les auteurs développent la notion de méthodes d'interface. Ces méthodes émergentes demeurent proches des intérêts et des démarches scientifiques et reposent dans le même temps, et

indubitablement sur les outils créés pour le grand public par les plateformes web. Ces instruments des chercheurs et des industriels entrent en résonance. La volonté d'obtenir des données issues de plusieurs plateformes amène également à travailler sur la mise en forme, l'interopérabilité des données selon les hypothèses du chercheur afin d'automatiser et de réduire le temps de traitement. Dans ce type de corpus qui commencent à se constituer, d'autres interrogations sur l'extraction, le traitement, le croisement des données et la visualisation viendront s'ajouter.
Ainsi, l'analyse qui en résulte doit également admettre que les catégories définies pour décrire le social, le territoire, malgré toutes les précautions d'interprétation se délitent. Pourtant, cette remise en question des catégories pour décrire l'homme moyen créé au XIXe siècle dans le traitement des données issues du web et la mise en double statistique du sujet sont critiquées (Rouvroy, Berns, 2013). En se focalisant sur la relation entre deux nœuds, ce qui pour ces auteurs finit par valoir pour le sujet, le chercheur évacue toutes les contraintes sociétales. Nous avons souligné qu'il peut introduire des biais ou retirer des données, il prétend donc rarement à expliquer le social via cette visualisation sous forme de graphes du territoire. Effectivement, ces auteurs soulignent à juste titre que ces traces sont laissées et ne sont pas des données transmises. Pour nous, se focaliser sur l'individu qui aurait laissé ces traces ne doit pas être l'objet de toutes les recherches mobilisant la visualisation de données et des graphes. Analyser la circulation des images, de signes à partir de métrique plus que la position des individus produit certes moins de modèles mais interroge tout autant ce qu'est l'espace public numérique, dans sa dimension collective. L'éthique repose ici tout autant sur des procédures de déclarations de traitement et de stockage des données que des méthodes d'extraction, de traitement, de visualisation, d'analyse et de transmission des résultats et des données elles-mêmes.

5. Conclusion

Dans cet article, deux niveaux d'interrogation s'entremêlent : l'extraction des données et les pratiques de visualisation. Les données en quantité de plus en plus importante qui participent à la recherche en SHS nécessitent des nouvelles formes de visualisation. Les graphes ne sont pas nés avec le web, mais les logiciels deviennent aisément accessibles aux chercheurs en SHS depuis quelques années. La lecture de ces formes de visualisation qui sortent des conventions demeure complexe. Barabási (2003, p. 238) en convient dans un propos quelque peu déterministe.
« Une fois que nous aurons découvert par hasard la bonne vision de la complexité, il ne sera pas très difficile de la développer. Mais la question quand ça arrivera est l'un des mystères sur lequel beaucoup d'entre nous travaillons. »
Dans la volonté du chercheur de donner une forme, un sens commun accessible au plus grand nombre, la performativité de cette mise en discours des données ne doit pas être évacuée. Nous analysons des données nativement numériques qui deviennent nos données en tant que chercheur, pour analyser des phénomènes dits « sociaux » (Rogers, 2013). La volonté de la relation est formalisée par la ligne dans les graphes et est indispensable à l'analyse dans un logiciel comme Gephi. Les statistiques appliquées au corpus issu de la théorie des agrégats participent à cette mise en relation et c'est là l'intérêt de recourir à la visualisation sous forme de graphes. Mais dans cette visualisation où les repères, les conventions ne sont pas connus de tous, la volonté d'inscrire les données issues du web dans un fond de carte montre que la ligne est également contour. Toutes ces visualisations doivent questionner l'art computationnel et la « cécité des modèles computationnels » (Rogers, 2015) mobilisés pour rendre continues des données qui sont extraites automatiquement de manière parcellaire, et que le chercheur associe dans la construction de son corpus. Les degrés d'agencement de données hétérogènes,

leur sélection ou leur abandon dans le cadre de l'analyse importent tout autant que les algorithmes de spatialisation ou de modularité utilisés avec les logiciels.
La visualisation de données interroge tout autant la démarche méthodologique majoritairement empirique mais également l'éthique, qui doit s'inscrire dans les hypothèses mêmes du chercheur. L'abondance des données, leur accessibilité, relativement aisée par l'extraction automatique ne doivent pas évacuer les enjeux épistémologiques (Carmes, Noyer, 2015), politiques (Rouvroy, Berns, 2013) et sociétaux exposés dans cet article. Mais si les catégories amènent à être repensées c'est qu'elles ne se dupliquent pas à l'identique sur le web. L'intérêt selon nous est que ces visualisations et les méthodes qu'elles induisent devraient permettre également de sortir du sujet pour analyser des matériaux plus fins que sont les contenus produits, leur matérialité propre et leur circulation comme la capacité de propagation, de circulation d'images sur le web. La visualisation sous forme de graphes nous permet d'en rendre compte plus aisément. De notre point de vue, ce sont les objets étudiés et les catégories dupliquées de l'homme moyen au web qui devraient être interrogés.

Bibliographie

Anderson A., Huttenlocher D., Kleinberg J., Leskovec J., and Tiwari M. (2015). Global diffusion via cascading invitations : Structure, growth, and homophily. *Proceedings of the 24th International Conference on World Wide Web*, p. 66-76.

Anderson B. (2002). *L'imaginaire nationall :réflexions sur l'origine et l'essor du nationalisme*, La Découverte, Paris.

Austin J. L. (1962). *How to Do Things with Words*, Clarendon, Londres.

Barabási, A. -L. (2003). *Linked*, Plume, New-York.

Bertin J. (2005). *Sémiologie graphique*, Éditions de l'EHESS, Paris.

Boullier D. (2015). L'écume des territoires numériques. *Traces numériques et territoires,* Presses des Mines, Paris, p. 113-134.

Braudel F. (1984). *Écrits sur l'histoire,* Champs-Flammarion, Paris.

Butler J. (2010). Performative agency, *Journal of Cultural Economy*, vol. 3, n°2, p. 147-161.

Bush V. (1945). As we may think, *The Atlantic.*

Cardon D. (2015). *A quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Coédition Seuil-La République des idées, Paris.

Cardon D., Roth C., Fouetillou G. (2014). Topographie de la renommée en ligne : un modèle structurel des communautés thématiques du web français et allemand. *Réseaux*, vol. 6, n°188, p. 85-119.

Carmes M., Noyer J-M. (2015). Désirs de data. *Traces numériques et territoires,* Presses des Mines, Paris, p. 179-212.

Cristofoli P., Vilaça O. (2002). Le Monde comme réseau. *Jeu de Cartes, nouvelle donne : cartographier aujourd'hui les espaces d'aujourd'hui.* Projet CartogrAm, Rapport Datar, Paris, p. 165-188.

Davallon J. (2004). Objet concret, objet scientifique, objet de recherche. *Hermès*, n° 38, p. 30-37.

Davallon J. (2012). *L'économie des écritures sur le web*, Lavoisier, Paris.

Davallon J., Jeanneret Y. (2004). La fausse évidence du lien hypertexte. *Communication et langages,* n°140, p. 43-54.

De Maeyer J. (2010). Methods for mapping hyperlink networks: Examining the environment of Belgian news websites. *International Symposium of Online Journalism,* Austin, TX, April 2010.

Degenne A., Forsé M. (2004). *Les réseaux sociaux*, Armand Colin, Paris.

Drucker J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quartely*, n° 1, vol. 5, http://www.digitalhumanities.org/dhq/vol/5/1/000091/ 000091.html.

Ghitalla F. (2008). La « Toile Européenne » Parcours autour d'une cartographie thématique de documents web consacrés au thème de l'Europe et à ses acteurs sur le web francophone. *Communication & langages,* n°158, p. 61-75.

Gitelman L. (2013). « *Raw data » is an oxymoron*, Mit Press, Cambridge.

Goody J. (1982). *La raison Graphique*, Librairie des Méridiens, Paris.

Granovetter M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, n° 78, p. 1360-1380.

Grevsmühl S. V. (2014). *La terre vue d'en haut*, Seuil, Paris.

Guillaume J-L. (2015). Conférence, École thématique *Étudier les réseaux sociaux : espaces, mobilités,* Oléron, 21-25/09/2015.

Heymann S. (2014). Gephi. *Encyclopedia of Social Network Analysis and Mining*, Springer, New York, p. 612-625.

Highfield T., Kirchhoff L., Nicolai T. (2011). Challenges of Tracking Topical Discussion Networks Online. *Social Science Computer Review*, vol. 29, n° 3, 340-353.

Hofstadter D. (1987). Prélude …. *Vues de l'esprit.* Interéditions, Paris, p. 155-203.

Hoskins J. A. (2006). Agency, Biography and Objects. *Handbook of Material Culture*, Sage /Routledge, London, UK, New York, NY, p. 74-85.

Jeanneret Y. (2011). Complexité de la notion de trace. De la traque au tracé. *L'Homme trace. Perspectives anthropologiques des traces contemporaines*. CNRS Éditions, Paris, p. 59-86.

Kleinberg J. M. (2006). Authoritative Sources in a hyperlinked environment. *The structure and dynamics of networks*, p. 515-542.

Kleinberg J. M., Kumar S. R., Raghavan P., Rajagopalan S., Tomkins A. (1999). The web as a graph: Measurements, models and methods. Invited survey at the *International Conference on Combinatorics and Computing,* http://www.cs.cornell.edu/home/kleinber/.

Latour B. (1993). Le topofil de Boa Vista ou la référence scientifique–montage photophilosophique. *Raison Pratique,* n°4, 187-216.

Lazega E. (2007). *Réseaux sociaux et structures relationnelles*, PUF, Paris.

Le Béchec M. (2010). *Territoire et communication politique sur le « web régional breton ».* Thèse en SIC, Université Rennes 2.

Le Béchec M., Alloing C. (2016). Les humanités numériques pour repenser les catégories d'analyse. *Revue française des sciences de l'information et de la communication*, n°8. http://rfsic.revues.org/1804

Le Béchec M., Boullier D. (2014). Communautés imaginées et signes transposables sur un « web territorial ». *Études de communication,* n° 42, p. 113-125.

Le Béchec M., Crépel M., Boullier D. (2014). Modes de circulation du livre sur les réseaux numériques. *Études de communication*, n° 43, p. 129-144.

Lévy J. (1994). *L'espace légitime*, Presses de la Fondation nationale des sciences politiques, Paris.

Lévy J. (2013). *Réinventer la France*, Fayard, Paris.

Lima M. (2013). *Cartographie des Réseaux*, Eyrolles, Paris.

Manovich L. (2001). *The Language of New Media*, MIT Press, Cambridge.

Manovich L., Tifentale A., Yazdani M., Chow J., 2014. *The exceptionnal and the everyday : 144 hours in Kiev,* http://www.the-everyday.net.

Marres N., Gerlitz C. (2015). Les méthodes d'interface. *Traces numériques et territoires,* Presses des Mines, Paris, p. 33-62.

Mercklé P. (2004). *Sociologie des réseaux sociaux*, La Découverte, Paris.

Mitchell W. J. (2005). *City of bits*, MIT Press, Cambridge.

Mohammadi E., Thelwall M., Kayvan Kousha K. (2015). Can Mendeley bookmarks reflect readership ? A survey of user motivations, *Journal of the Association for Information Science and Technology*, vol. 67, n° 5, p. 1198-1209.

Musso P. (2003). *Critique des réseaux*, PUF, Paris.

Newman M., Barabási A.-L., Watts D. J. (2006). *The structure and dynamics of networks,* Princeton University Press, Princeton.

Noucher M. (2015). De la trace à la carte et de la carte à la trace. *Traces numériques et territoires,* Presses des Mines, Paris, p. 215-226.

Olson D. (1998). *L'univers de l'écrit*, Éditions Retz, Paris.

Plantin J-C. (2012). D'une carte à l'autre : le potentiel heuristique de la comparaison entre graphe du web et carte géographique. *Analyser le web en Sciences Humaines et Sociales*, Armand Colin, Paris, p. 228-242.

Plantin J-C. (2013). Qu'y a-t-il à côté d'un graphe de sites web ? *Communication et organisation*, n° 25, p. 59-70

Proulx S., Latzko-Toth G. (2000). La virtualité comme catégorie pour penser le social : l'usage de la notion de communauté virtuelle. *Sociologie et sociétés*, vol. 32, n° 2, p. 99-122.

Rieder B. (2010). Pratiques informationnelles et analyse des traces numériques : de la représentation à l'intervention. *Études de communication*, n° 35, p. 91-104. http://edc.revues.org/2249

Rogers R. (2013). *Digital Methods*, MIT Press, Cambridge.

Rogers R. (2015). Au-delà de la critique big data. *Traces numériques et territoires,* Presses des Mines, Paris, p. 13-32.

Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation : le disparate comme condition d'individuation par la relation ? *Réseaux*, 31(177), p. 163-196.

Simmel G. (1999). *Sociologie*, PUF, Paris.

Tönnies F. (1977). *Communauté et société*, Retz, Paris.

Tsou M-H., Lusher D., Yang J-A., Gupt D., Gawron J. M., Spitzberg B. H., An L., Wandersee S. (2012). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing) : A case study in 2012 U.S. presidential election. *AutoCarto International Symposium on Automated Cartography Proceedings* (Columbus, OH): Mt. Pleasant, South Carolina, Cartography and Geographic Information Society.

Venturini T. (2012). Great expectations : méthodes quali-quantitative et analyse des réseaux sociaux. *L'Ère Post-Media. Humanités digitales et Cultures numériques.* Hermann, Paris, p. 39-51.

Watts D. J. (2004). *Six degrees*, W.W. Norton & Company, New-York-London.

Zook M., Devriendt L., Dodge M. (2011). Cyberspatial Proximity Metrics : Reconceptualizing Distance in the Global Urban System. *Journal of Urban Technology,* vol. 18, n° 1, p. 93-114.