La cartographie d'information

ou *l'alchimie quali-quantitative*

> "Puisque vous avez désobéi au commandement que Dieu vous fit, allez parcourir une vie de peines et de fatigues" Adam, que Dieu avait choisi pour notre père à tous, qu'il avait si noblement doué, reconnut sa faute et quitta l'idée de sciences pour en revenir au travail des mains qui fait vivre. Il prit la bêche, et Eve commença à filer. Plusieurs arts nés du besoin suivirent, tous différents l'un de l'autre. Celui-ci entraînant plus de science que celui-là, ils ne pouvaient tous être égaux; car **la science** est le plus noble. Après elle en vient qui lui doit son origine et la suite de près, il vient de la science et se forme par l'opération des mains. C'est un art que l'on désigne par le mot de ***peindre***: il demande la fantaisie et l'habileté des mains; il veut trouver des choses nouvelles cachées sous les formes connues de la nature, et les exprimer avec la main **de manière à faire croire que ce qui n'est pas, soit**.", Cennino Cennini, *Il Libro dell'arte*, Florence, 1437.

La cartographie d'informations semble tout avoir de l'art de l'illusion. On la tient parfois en suspicion dans le milieu scientifique comme un objet déporté, relatif ou contingent parce qu'elle relève du champ de l'image ou du design graphique et non du «concept» articulé, autrement dit des «mots». Ce qui fait la force visuelle et esthétique de la cartographie dans la mise en scène des *data* semble aussi faire sa faiblesse comme instrument cognitif où, souvent, on ne la tolère que sous la forme de la «visualisation» censée «traduire les données» (la *dataviz* et ses mirages) ou «accompagner» le discours, la pensée ou la décision. En somme, un jeu d'illusions visuelles comme on aime à définir les images au pays de Descartes. Telles que je me remémore nos expéditions cartographiques au cours de toutes ces années, les points de vue que nous avons produits sur les différents *data sets* pourraient effectivement sembler relever massivement des «trucs et astuces» dignes de l'art de la mise en scène, *par magie*. Comme dans un théâtre vitruvien où les événements vont apparaître dans un cadre réglé par l'illusion perspective, nos variables graphiques comme les liens courbes ou la distribution des couleurs pour créer des zones (les fameux *clusters* que nous aimions qualifier de *patatoïdes* visuels) et nos filtres de spatialisation avec les choix imposés par l'utilisation systématique des algorithmes de type *force-directed* (dont le *ForceAtlas2* évidemment) ont constitué les rouages nécessaires à la manifestation de ce que nous croyons être des *patterns* dans les data. Et, au-delà, certaines propriétés que nous sommes

parvenus à identifier à propos du web en dépendaient aussi.

Pourtant, l'illusion cartographique qui aboutit *ici* sur un poster, *là* sur une interface numérique, repose sur une machinerie logique qui fait résonner tout l'appareillage conceptuel et technique de la construction de l'objet de science. La cartographie d'information, qu'on la pratique soi-même ou qu'on l'observe, apparaît aussi comme une *machine logique productrice de topologies interprétables*. Y compris en dehors de la question des réseaux, elle recèle des propriétés particulières, *logiques* si l'on veut, qui la rangent aux côtés de tous les autres instruments d'investigation et d'analyse, donc de *cognition*. A y regarder de près, l'appareil cartographique qui me permet de prendre des vues sur les *data* se compose d'une série de filtres, dont une partie est destinée à préparer l'espace graphique pour le déploiement d'une interprétation. Il ne s'agit donc pas que d'une mise en scène des data au sens d'une forme de *rhétorique visuelle* mais aussi d'un traitement logique où s'équilibrent des processus de traitement de l'information tant *qualitatifs* que *quantitatifs*. En position de pivot entre les données et les hypothèses (ou l'interprétation), la cartographie constitue donc un *lieu* d'observation, ou un *moment*, très éclairant de ce deux mécanismes logiques par lesquels nous saisissons les objets de science, alors observables. C'est en ce sens que j'aime à définir la cartographie d'information comme une *machine logique*.

La problématique du "quantitatif" et du "qualitatif" s'inscrit dans cet espace de la pratique scientifique où nous construisons les phénomènes étudiés, à mi-chemin entre l'hypothèse qui le surplombe et les "data", brutes, sur lesquelles ils émergent de façon logique et organisée. En ce sens, elle n'est pas spécifiquement liée à la sociologie, à l'anthropologie, ni même aux sciences humaines et sociales: elle renvoie à ce jeu où s'associent tant des *qualités* que des *quantités* dans un modèle d'intelligibilité des données, et un seulement parmi d'autres possibles quand on adopte les procédures d'un champ de recherche comme je le fais avec l'approche réseau. Ce qui se joue localement en ingénierie des données et des connaissances sous forme de curseurs de saisie d'un phénomène relève d'une problématique plus générale en science, et au-delà. On peut réduire à grands traits les principes des orientations qualitatives ou quantitatives qui me semblent représenter deux façons complémentaires de construire l'objet de science, tel que l'on peut l'observer ou l'analyser et dont la cartographie est une forme technique. Spontanément, on peut concevoir *l'orientation quantitative* comme l'univers des mesures et des variables, de la recherche de patterns (sociaux par exemple) indirectement observables, des effets de masses et de seuils statistiques. Le travail scientifique sur des hypothèses quantitativement argumentées peut-être d'une grande complexité dans la manipulation des opérations mathématiques, comme la recherche des "axes" ou des "composants" dans des univers de données multidimensionnelles. Elle me semble fondamentalement liée à la quête de régularités, voire de lois accessibles une fois seulement dépassés certains seuils dans les masses de données. Ce qui ne semble pas poser de problème méthodologique ou critique en physique ou en biologie, se révèle particulièrement discuté en sciences humaines et sociales où l'usage des données chiffrées et en nombre serait sujet à une forme inhérente de "relativité" (du chercheur, des instruments, des cadres théoriques sous-jacents) et à des formes de "positivisme" dont les fondements théoriques ou idéologiques n'ont pas été suffisamment interrogés. On peut ainsi opposer à l'orientation quantitative, une *démarche qualitative* orientée vers les singularités, les interactions locales ou le suivi longitudinal d'un nombre restreint, mais richement décrit, d'acteurs ou de petits groupes. Contrairement aux supposées réductions quantitatives, le curseur qualitatif me semble gouverné par la recherche foisonnante des moyens de dilater le "phénomène" ou "l'objet social", par

exemple en y participant soi-même comme en ethnométhodologie, en invoquant la complexité du fonctionnement humain avec ses dimensions culturelles ou psychanalytiques, ou alors (comme je l'ai observé en ergonomie cognitive) en consacrant beaucoup de temps et de moyens d'observation ou d'interprétation à quelques minutes d'activité d'un acteur devant une console d'ordinateur.

Oui, je sais: j'entre là dans les débats érudits de l'épistémologie, de l'anthropologie, des sciences de la cognition. Le débat entre démarche *qualitative* et démarche *quantitative*, à lui seul, occupe une place majeure en sciences humaines et sociales, surtout en sociologie et en sciences politiques. Il s'y joue des postures, des hypothèses et des méthodes à priori irréductibles, comme si la dichotomie *qualitatif-quantitatif* ouvrait sur deux champs distincts de science. Et les sciences exactes n'y échappent pas non plus, bien que le débat y soit moins sensible. Ce débat ancien trouve aujourd'hui un écho particulier, et très polémique, quand on envisage, de surcroît, les technologies numériques, les masses d'information et les puissances de calcul. Cette résurgence du débat a été alimenté récemment par certaines perspectives induites par le *big data*, en particulier dans la découverte et l'exploitation de tendances statistiques massives dans les données numériques aujourd'hui accessibles. Des provocations de Chris Anderson dans *Wired* ([1]) qui annonce la fin du modèle scientifique expérimental avec l'apparition des *data sciences* ("*But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete*") aux réactions conservatrices des tenants des modèles abstraits, théoriques et érudits en sciences humaines face à cette foule d'ingénieurs et leurs infrastructure de traitement à grande échelle, les postures des uns et des autres s'arc-boutent sur des distinctions que le cartographe que je suis a du mal à discerner dans la pratique. Ma perplexité face à ce débat redouble quand je m'aperçois des autres problématiques que la distinction qualitatif-quantitatif emporte avec elle dans son sillage comme celle du *manuel* et de *l'automatique* (dès qu'il y a des machines et des ingénieurs, certains s'imaginent qu'il ne s'agit que de traitement quantitatif de l'information) ou celle des rapports de l'hypothèse ou du *modèle théorique* avec les *données expérimentales* (l'effort de théorisation étant présenté comme un exercice qualitatif).

Quand bien même voudrais-je échapper à ce débat généralisé, il m'est imposé de toutes les façons quand je parle de cartographie de l'information. En particulier, c'est systématiquement le cas lors des conférences où je présente mes activités: je rattache bien évidemment la cartographie de l'information (et la science des réseaux) à la culture actuelle des data, en termes de méthodes comme de "posture" (sous l'angle de l'open-data, du *big data analytics*, du *data scientist*, du *data intelligence* voire du *grid-computing* auxquelles ont fait référence depuis quelques temps des revues comme *Nature*, *Science* ou *O'Reilly Radar*). La cartographie, après tout, est une façon de traiter des masses de données et les *network sciences* de les expliquer. Mais souvent, à peine ai-je eu le temps d'afficher un poster sur un mur, que tout un ensemble de questions récurrentes pleuvent et m'obligent à développer des débats sans fin sur le rôle des "machines" et des "statistiques" à l'heure des réseaux distribués d'information. Les procédures d'analyse automatisées de vastes ensembles de données modifie-t-elle le rôle des experts, voire leur place dans la construction des connaissances? L'ingénierie quantitative en *data sciences*, avec leur démarche inductive, peut-elle se passer de modèles qualitativement construits? La cartographie, avec ses techniques et ses méthodes, est assurément un instrument de réduction des masses mais sa *relativité* (ce n'est qu'une méthode parmi d'autres) n'impose-t-elle pas une *réduction* des

1 Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired, 2008.

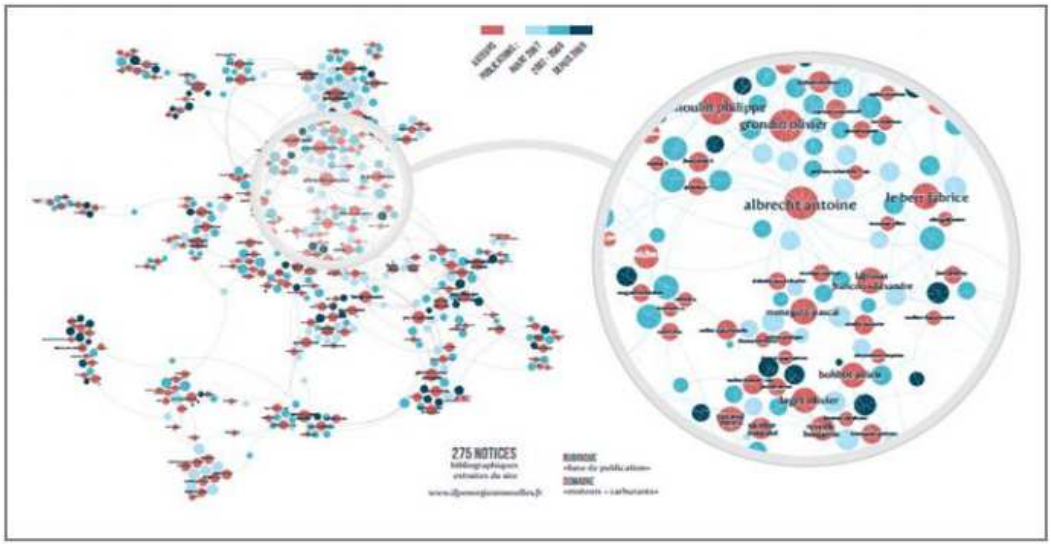
pistes de réflexion, comme une forme *d'appauvrissement*? C'est dans ces moments que j'aimerais pouvoir transporter mes interlocuteurs dans mon atelier de cartographie et leur montrer, à travers les gestes et les procédures, que ce qu'ils considèrent comme des limites ne sont que des contraintes de départ pour accéder à des possibles dont, souvent, ils ne soupçonnent pas la rationalité et la dynamique.

*****

Les débats gagneraient en pertinence si l'on s'attachait quelques instants à la description des traits saillants de cette ingénierie des données, du moins tels qu'ils peuvent m'apparaître, et où la cartographie d'information tient sa place. Cela permettrait peut-être d'éviter les caricatures que l'on voudrait parfois dresser à propos de l'ingénierie actuelle des data supposée *quantitative* (le culte des machines, l'obsession des statistiques, l'absence de recul critique sur les pratiques et les méthodes). Pour apporter un peu de clarté au débat, il me semble que l'exercice de la cartographie d'information constitue un excellent terrain d'interrogation de ces problématiques dont la dichotomie *qualitatif-quantitatif* semble être le nœud central. Quand on regarde de près une chaîne de traitement de l'information pour produire des cartographies, les opérations manuelles et artisanales menées sur les données (depuis l'extraction jusqu'aux patterns visuels cartographiés) mettent à jour de façon explicite *l'imbrication* des procédures tant qualitatives que quantitatives que l'on fait porter sur les données. Dans l'univers de l'artisanat de haute-technologie dans lequel je baigne, la mise au point de prototypes oblige décomposer des dimensions et des procédures qu'un algorithme industrialisé masquerait sous la forme d'une «boîte noire», éventuellement propriétaire et cadenassée par un brevet. L'observation du travail minutieux du cartographe dans son atelier éclaire le processus de croisement systématique du qualitatif et du quantitatif dont dépendra, au final, la nature du phénomène scientifique visé. Ce sont les deux dimensions constitutives de l'information dans le milieu dans lequel je travaille. De la simple feuille de papier avec ses listes et ses tableaux à la table *Excel* et ses fonctionnalités de calcul jusqu'aux systèmes "temps réel" sur le web, ces deux formalismes logiques imposent leurs dimensions à nos instruments techniques de traitement au rang desquels figure la cartographie d'informations.

Si on la considère sous l'angle d'une *machine logique*, apparaissent alors les deux curseurs quantitatif et qualitatif qui dessinent une fenêtre de saisie de phénomènes décrits avec des informations. Comme avec un métier à tisser, les motifs générés dépendent de formes avancées d'hybridation de qualités (couleurs, textures...) et de quantités (sur combien de lignes?). En ce sens, la cartographie d'information comme les autres instruments scientifiques imposent leur grille de saisie aux «objets» visés qui s'applique aussi bien à des gènes, des publications scientifiques, des organisations sociales, des molécules, des abonnés, des concepts qu'à des "acteurs" ou des "interactions". Dans la pratique, le curseur qualitatif et le curseur quantitatif doivent être réglés avec soin. C'est un moment important pour un cartographe qui doit au final associer les propriétés de l'espace des données prises en compte avec les propriétés du plan graphique et ses variables. La pertinence de la cartographie d'information comme instrument d'investigation n'est effective que dans certaines limites liées quantitativement aux masses d'informations et qualitativement au nombre de dimensions. Quantitativement, il faut ni *trop*, ni trop *peu* d'éléments ou de

"lignes" pour pouvoir régler manuellement là aussi les différentes échelles de l'information. Dans le cadre des graphes, un nombre trop faible de nœuds ne fait pas un réseau, qui plus est si le nombre de liens est également réduit. Les patterns organisationnels recherchés (la topologie d'un système) n'apparaissent qu'une fois passé un certain seuil de données. Qualitativement, il faut ni *trop*, ni trop *peu* de dimensions dans les données pour pouvoir manuellement jouer sur leur corrélation (jeu manuel dans le cadre de la cartographie d'information relationnelle telle que je la pratique mais on peut en faire un jeu très abstrait et/ou automatisé dans ce que l'on désigne par méthodes d'analyse multdimensionnelle). Ces réglages simultanés des deux curseurs de saisie de l'instrument conditionnent la pertinence du cliché cartographique: a-t-on suffisamment réduit les masses pour faire apparaître un pattern identifiable? A-t-on identifié suffisamment d'identités remarquables (ou de dimensions associées) pour les voir apparaître dans l'ensemble des masses?



Cartographie pour un laboratoire de recherche. Ici ont été croisés les données "auteurs" et les données "publications" parmi les nombreuses corrélations possibles dans une table de données. Cette carte intègre aussi un croisement des "publications" avec les "dates de parution" perceptible à travers les différentes colorations de bleu attribuées aux publications. Une cartographie de ce type implique un double travail de manipulation d'une table de données qui peut s'avérer d'une grande complexité.

Les mécanismes sur lesquels repose le cliché cartographique supposent l'exploitation d'une structure de données dont la table est l'archétype. Au sens large, elle peut être réduite à une «simple» liste (CSV) ou élargie à une structure de base de données (éventuellement distante et interrogeable via une A.P.I.). Dans tous les cas, la production de cartographie dépend d'une combinatoire de saisie (interrogation, requêtage, extraction, transformation...) des informations contenues dans les tables de données. Encore une fois, il convient de souligner que cette combinatoire ne relève pas seulement du choix ou de la posture théorique du chercheur (souvent déterminante en sciences humaines et sociales) mais aussi de la façon dont techniquement nous construisons les objets étudiés à travers des modèles de données. Ainsi, tout commence par ce *prêt à découper-composer* qu'est le fichier *XLS*, disons une structure de tables de données.

Noeuds	Types	Liens	Coordonnées GPS		Date
P1	Projet				2006
P2	Projet				2007
MC1	Mot-clé	P1, P2			
A1	Acteur	P1	48.856614	2.3522219	
A2	Acteur	P37	47.218371	-1.553621	

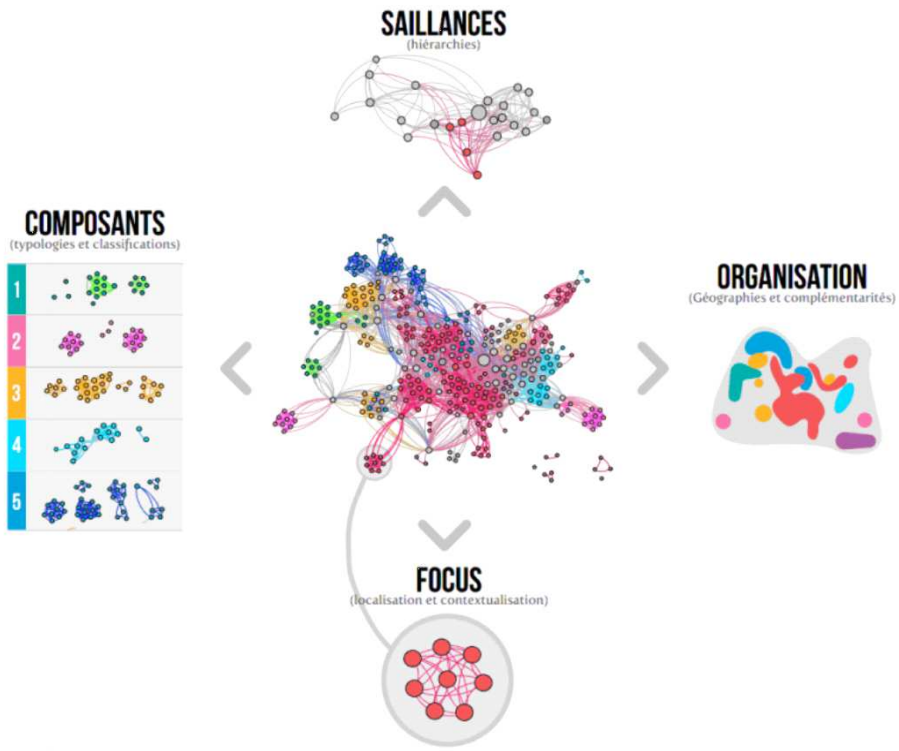
Combien quantitativement de lignes (d'objets à compter), combien qualitativement de colonnes (de descripteurs, d'angles sous lequel les éclairer?). La structure de données livre donc deux curseurs à manipuler pour construire l'objet d'observation: la fenêtre des lignes dont plus le nombre est grand plus le potentiel de détection de patterns est important selon les échelles; la fenêtre des colonnes qui fixe un potentiel de corrélation des dimensions dans les données. La production de structures de graphes constitue une façon particulière de «prendre les données» en croisant deux colonnes parmi les autres qui partagent des objets ou des propriétés communes (principe de redondance de l'information ou de récurrence). C'est là toute la dynamique de la cartographie à base de graphe. Dans le cas des bases de notices bibliographiques, les publications peuvent être reliées entre elles par des liens de citation (directement accessibles depuis l'une des colonnes du tableau) ou bien par des liens calculés dans la table (présence de descripteurs communs, mêmes auteurs, mêmes laboratoires). Si le réglage des fenêtres de saisi de l'objet sont trop massives (du point de vue cognitif comme du point de vue des instruments), des outils de (pré)traitement automatique et de réduction sont nécessaires pour identifier des clusters, analyser des composantes multiples, agréger des dimensions, visualiser des croisements ou des résultats de seuillage. Les données imposent leurs modèles que l'on peut, en retour, façonner.

*****

La double saisie des informations mobilisées n'est pourtant qu'un point de départ, le premier mouvement de la *machine logique.* Les lignes et les colonnes ouvrent potentiellement sur un jeu infini de croisements, de comparaison, de corrélation. En termes de compétences humaines, il s'agit bien d'intelligence des données (*data intelligence*) et, à regarder travailler les ingénieurs en systèmes d'informations, on se rend compte du degré *d'intelligence embarquée* que peut contenir un algorithme, lui-même intégré dans un processus ou une bibliothèque plus vaste. Là où les Sciences Humaines et Sociales semblent avoir en partie figé la distinction entre les deux processus de production comme d'analyse de l'information, l'ingénierie des data en exploite le potentiel, si possible *tout* le potentiel en tentant d'automatiser les corrélations possibles comme dans le *big data* avec l'idée de chercher à identifier dans de grandes masses de données un ou plusieurs "traits" communs, révélant un pattern statistique qui n'est qu'une façon parmi d'autres d'isoler des identités partielles et partagées. Cette méthode *pattern based* peut être qualifiée de quanti-qualitative. Mais la méthode inverse, quali-quantitative, n'est pas incompatible; elle est même souhaitable: une propriété repérée localement (par exemple une information particulière associée à un nœud dans un graphe) est-elle repérable aussi à grande échelle? C'est l'agilité avec laquelle on manie les deux opérations qui détermine souvent le nombre et la richesse des prises que l'on se donne sur les corpus de données numériques. Cette intrication des deux approches est

pour ainsi dire native des recherches opérationnelles de type *data analytics*: dans les domaines de l'analyse des flux (par exemple les connexions quotidiennes sur réseau de téléphonie mobile ou les informations liées aux "parcours patients" entre des unités médicales dans un hôpital), les objectifs consistent souvent soit à modéliser les flux principaux, soit à "détecter des événements" qualitativement identifiables-calculables (curieux, remarquables, inédits...). Cette recherche repose sur la production d'un ou plusieurs modèles qualitatifs du phénomène (combinaison de traits spécifiques, distribués dans une configuration probable) appliqués à différentes échelle des masses de données (les quantités de données réunies pour valider le modèle pouvant donc se trouver à leur tour mobilisées comme traits qualitatifs à un niveau supérieur d'intégration).

Pour le moment, la cartographie d'information reste encore en marge de l'univers des réseaux dynamiques de données et rares sont les projets où l'on essaie de la produire automatiquement. C'est peut-être sa nature-même: figer le processus de traitement, au moment propice d'un cliché pertinent. L'image me paraît pertinente si je calcule le nombre d'essais nécessaires pour caler la prise de vue en alignant précisément tout une série de filtres! Dans son atelier, le cartographe que je suis façonne encore artisanalement sa chaîne de traitement de l'information, qui n'est jamais exactement la même d'un chantier à l'autre. Mais l'artisanat de haute technologie comme la cartographie d'informations a des vertus, disons, «pédagogiques» car on peut y décrire soigneusement, *au ralenti*, ce que l'automatisation cache définitivement: le cartographe qui règle ses curseurs et aligne différentes procédures de construction qualitative et quantitative de l'information. Et c'est maintenant dans la dynamique d'hybridation des deux procédures que prend forme la carte.



La carte, presque achevée, porte un scénario exploratoire calqué sur les grandes directions d'analyse de données. L'exploration de la carte, donc de certaines propriétés des données du corpus, est réglée par la façon dont qualités et quantités *se contrôlent mutuellement*. La carte permet en premier lieu de faire apparaître certaines propriétés *qualitatives*, qui peuvent être vérifiées globalement ou localement en fonction de la façon dont elles *résonnent* dans les masses quantitatives. C'est le principe qui permet de classer les éléments d'un système, d'un point de vue ou d'un autre. Dans l'exemple choisi, il s'agit d'une cartographie de brevets basée sur les liens de citation au sein du corpus. Je peux faire apparaître des propriétés globales, des *saillances* indiquant l'ossature générale du jeu de données: les brevets les plus importants en grosse taille – le plus souvent en couleur grise – ainsi que les principaux flux de citation. Je peux aussi «zoomer» sur une zone particulière du graphe et y vérifier des propriétés locales, valables seulement en contexte. Ce qui fait donc la variation qualitative de généralité ou de particularité des propriétés de corpus dépend de la façon dont elle est quantitativement vérifiée ou contrôlée dans les masses.

On pourrait en illustrer le principe de nombreuses façons. Par exemple, certains vont se concentrer plutôt sur les opérations d'identification, de sélection, de *ranking* et recomposer petit à petit des quantités (représentatives). C'est le principe de la recherche des types regroupant différentes identités (la notion de "profil" en *social data mining*) et les classer de différentes façons selon des critères qualitatifs (âge croissant des internautes, distribution de leurs liens affinitaires, métiers...). Mais c'est dans sa confrontation avec les quantités (ou les "masses") que ce travail sur les critères qualitatifs se trouve validé - en réalité *contrôlé*: le *type* construit est-il suffisamment récurrent? Mon ou mes critère(s) de sélection suffisent-ils à classer tous les objets de mon système (comme avec les algorithmes de classification automatique)? Sur le web, parmi les blogueurs, les "communautés voisines" de celle que j'étudie font-elles partie ou non de mon "corpus"? Sélectionner (des traits), choisir (des types ou des familles), découper ou segmenter (le continu des masses), autant d'opérations qui guident la maîtrise des quantités. On pourrait illustrer le principe de mille et une façons mais il paraît d'autant plus prégnant (et observable) quand il s'agit d'extraire des données d'un espace comme le web qui ne livre pas de lui-même des principes clairs de découpage. Certaines opérations en *web mining* illustrent bien toute cette mécanique de la sélection, avec ses essais, ses erreurs ou ses incertitudes, dans le réglage par exemple du "focus sémantique" d'un crawler web (est-ce que je choisis les "bons" termes pour sélectionner les pages pertinentes?) ou bien encore quand, avec le *navicrawler* ([2]) (eh oui, je l'utilise encore!), j'archive à la fois un corpus de pages web mais aussi celles que j'ai exclues. Les "masses de données" font rêver pour leur grandeur estimée mais il me semble qu'elles ne deviennent "quantités" que lorsqu'on leur applique des seuils, que l'on y repère des saillances statistiques, des effets de redondance ou qu'on les considère comme "représentatives". En somme, qu'elles constituent un moyen de contrôle ou de validation de ce qui est défini ou sélectionné qualitativement. On réduit souvent le travail sur les quantités à l'usage des outils statistiques et mathématiques, conduit par des ingénieurs focalisés sur les grandeurs. C'est vite oublier, comme en *network sciences*, que les résultats de calculs statistiques (par exemple, le *small world phenomenon* et une distribution des connexions en loi de puissance) ne sont que des signatures, des indices qui ouvrent sur la recherche d'un principe d'ordre qui gouvernerait des réseaux ouverts, de taille gigantesque et dynamiques dans le temps. Encore faut-il mobiliser ou maîtriser un nombre suffisamment important de "quantités" pour faire

2 http://webatlas.fr/wp/navicrawler/

émerger des propriétés ou, à l'autre extrême, que l'on dispose bien des algorithmes et de la puissance de calcul nécessaires à leur exploitation. Entre les deux, tous les niveaux de "zoom" sont possibles, du plus petit élément singulier aux effets de structure, à condition de spécifier que les "masses" ont été contrôlées comme des "quantités" calculables, des ensembles composés d'unités dénombrables.

Mais la carte permet, parallèlement, de faire apparaître certaines propriétés *quantitatives* qui peuvent être vérifiées quand on décompose un système en *composants* ou en clusters (a-t-on distribué tous les éléments dans des groupes homogènes?) et, au-delà, quand on cherche à recomposer le corpus selon une logique de complémentarité, comme pour en dessiner le plan global ou l'*organisation*. Notre capacité à décomposer/recomposer selon des frontières dans les masses dépend des critères adoptés. Dans l'exemple choisi, la distinction des *composants* ou des clusters est effectuée est basée sur les variations de distribution des liens de citation, auxquelles sont très sensibles certains algorithmes de détection de communautés sociales comme *Modularity*. De même, ces derniers s'assemblent en une organisation spécifique selon un ou plusieurs critères de complémentarité. Ainsi, dans les cartographies à base de graphes relationnels, la restitution d'un modèle d'intelligibilité du corpus passe par l'identification de sous-ensembles constitutifs, ses "régions" si l'on veut en termes de cartographie, et par la compréhension de la ou des logiques qui préside(nt) à leur distribution. Mais le découpage des masses peut s'effectuer selon d'*autres* critères présents dans la table de données (ou que j'ai moi-même produits au cours de l'analyse) de façon un trouver un critère optimal de segmentation des masses. Par exemple, je peux produire avec le même jeu de notices de brevets une cartographie des liens entrants et/ou sortants, isolant les brevets les plus anciens et/ou les plus stratégiques (liens entrants qui s'accumulent au cours du temps) ou les brevets les plus récents (l'actuel «front de recherche»). Je peux aussi faire appel à d'autres dimensions du jeu de données comme les mots-clés communs entre les brevets, ou même les inventeurs. De là, il apparaît que notre capacité à découper quantitativement les masses d'information dépend de la façon dont nous faisons *tourner* le corpus sous différents angles en l'analysant sous différents dimensions qualitatives, comme avec un prisme. Ce sont ces opérations successives qui font une partie du travail de data scientist. Et le travail peut être parfois long pour un cartographe, même accompagné par un dispositif de calcul: d'ailleurs, a-t-on jamais abordé un corpus de données à toutes les échelles et selon toutes les dimensions, d'un coup? Peut-on même le faire?

*****

Cette maîtrise des deux curseurs logiques est essentielle dans le travail de conception d'une cartographie, qui est une forme de modèle statistique et visuel des données. Evidemment, il s'agit d'une pratique artisanale et très contextuelle mais qui s'apparente selon moi au domaine plus général du *data analysis*, de la science des réseaux et, au-delà, des *computer sciences*. Certaines formules d'un atelier de cartographie, maintes fois éprouvées manuellement, peuvent être en partie automatisées, voire intégralement. L'automatisation ne change rien fondamentalement à la grammaire ou aux formules quali-quantitatives ou quanti-qualitatives sur lesquelles repose la machine logique, seulement l'ampleur et la

vitesse de traitement de l'information. L'automatisation permet le déploiement de procédures plus complexes, *embedded*, de mêler les types de traitement ou les algorithmes mobilisés, de croiser toujours plus vite les processus tant qualitatifs que quantitatifs de production de "l'information". Ce qui fait donc débat du côté des sciences humaines et sociales, fonctionne donc "à plein régime", si j'ose dire, du côté des métiers du *data analysis* et de l'innovation technologique.

Dans ces domaines, les avancées expérimentales reposent sur l'exploitation continue de la dynamique du croisement des dimensions quantitatives et qualitatives, pour embrasser de plus vastes ensembles, pour peu que des formules pertinentes de traitement aient été figées dans un algorithme. Le terme de *formule* me paraît ici adéquat pour résumer le travail de conception d'une combinatoire complexe qui vise à isoler un type précis de propriété dans des données en faisant varier ou alterner processus qualitatifs et quantitatifs. Ainsi, de nombreux algorithmes mobilisés en *web mining* reposent sur des "recettes" ou des "formules" largement éprouvées, avec un "préréglage" des deux curseurs (seuils minimum et maximum de données web, dimensions qualitatives prédéfinies de l'analyse): c'est le cas de HITS conçu par J. Kleinberg qui opère à partir de différents degrés de corrélation entre liens hypertexte et détection de contenu permettant de découper des "masses d'informations" en ensembles cohérents. Certaines formules, d'ailleurs, "valent de l'or". Elles sont au coeur des technologies numériques de l'information à l'image de ce qu'elles sont dans le domaine des cosmétiques et de la chimie fine. Souvent protégées par des brevets, certaines formules fonctionnent sous forme de «boîtes noires», propriétaires, volontairement inaccessibles. On prête même à certaines un pouvoir prédictif en identifiant les variables qui, parmi tant d'autres, semblent jouer un rôle majeur dans l'évolution d'un phénomène ou d'un système. D'autres, sont conçues comme des systèmes évolutifs, *auto-apprenants* et capables de modifier eux-mêmes l'alignement des filtres qu'ils intègrent. Au sens propre, les formules concentrent des formes artificielles d'intelligence.

Toutes ces formules algorithmiques contribuent à élargir de façon massive nos fenêtres qualitatives et quantitatives de saisi d'un phénomène. Beaucoup d'entre elles nourrissent aujourd'hui les applications les plus courantes et les univers quotidiens de manipulation de l'information. Sans être spécialiste de l'ingénierie data, chacun reconnaîtra l'étendue des formules de croisements complexes d'une simple application locale comme *Excel*, tout-à-fait étonnante, et l'on imagine ce que l'on peut réaliser avec de vastes dispositifs de stockage et de traitement des données dans l'univers du *big data* actuel et de l'accès à des corpus distants, eux-mêmes déjà analysés et enrichis. La conception de formules ou de modèles de traitement de données concentre ainsi de nombreux efforts d'innovation, dans tous les secteurs et les métiers de l'information. En termes de services innovants, la formule peut être incarnée dans un dispositif né de l'agrégation originale d'une série de filtres analytiques qui portent sur ou plusieurs dimensions des données et qui peut être concentrée dans une interface. En science, l'alignement d'une série de filtres, testé sur de nombreux data sets, peut être considéré comme un objectif en ingénierie de l'information et l'on ne compte plus les publications qui en détaillent la grammaire technique. En un mot, la formule préfigure le prototype (scientifique ou industriel) et fonctionne parfois en *RetD* comme une «boussole» pour l'orientation de l'innovation technologique et la mise en place de nouveaux services.

Dans son atelier, le cartographe d'information dispose désormais aujourd'hui d'une batterie étendue d'algorithmes ou de formules opérationnelles, qu'il combine en solutions hybrides, en les associant parfois à leur tour dans des ensembles plus vastes. Il m'arrive par exemple de faire appel aux compétences ou à la puissance de calcul d'une entreprise pour préparer un grand jeu de données. Une fois travaillé dans mon atelier quotidien, les données enrichies retournent parfois chez un autre partenaire, une autre entreprise qui va à son tour les passer au prisme de sa technologie...pour finalement être visualisées depuis ma machine. Ce que l'on désigne «intelligence des données» me semble relever massivement en phase exploratoire d'un processus de plus en plus étendu, géographiquement ou temporellement, incluant aussi bien des compétences humaines diverses que des applications distantes et coopérantes. L'art de la formule me semble constituer le cœur de la démarche cartographie que j'aime à considérer comme une des formes de l'artisanat de haute technologie dédié aux data. Quand on le regarde précisément travailler dans son atelier, avec ses méthodes et ses instruments, le cartographe cherche seulement à comprendre la matrice des croisements ou en éprouver le potentiel. De là, émergeront parfois quelques formules prometteuses d'où l'on pourra extraire un modèle de données plus vaste et plus robuste. Mais ailleurs ou plus tard car, pour le moment, le cartographe reste plongé dans ses tentatives d'hybridation de filtres, faisant naître parfois un phénomène au hasard de ses manipulations combinatoires. Comme un alchimiste...