Les données numériques et *l'alchimie quali-quantitative*

Réflexions personnelles autour d'un thème récurrent en sciences humaines et sociales

Les dimensions d'un débat. Le débat sur les questions de méthodologie restent vifs en Sciences Humaines et Sociales, particulièrement depuis que les données numériques sont mobilisées comme *matériau de l'observation* de la société ou de ses acteurs. Pour un sociologue, un anthropologue ou un chercheur en sciences politiques, les immenses stocks de données potentiellement accessibles permettraient d'enrichir de façon inégalée dans l'histoire des sciences le "terrain" où se forge l'analyse et/ou la modélisation du "fait social", la vérification-reformulation des hypothèses. Dans un élan d'optimisme, on peut supposer que ces (*big-*)*data* vont enfin permettre aux sciences sociales d'embrasser aussi bien des phénomènes collectifs et massifs que de coller au plus près à la singularité sociale d'un "acteur", d'une "interaction" ou d'un "geste" (si l'on pense que certains se filment de façon continue avec un accès direct aux données en temps réel). On peut même rêver à l'avènement de *digital humanities* (autrement dit d'une forme de science sociale enfin réconciliée avec l'ingénierie numérique), surplombant du micro au macro l'ensemble des phénomènes et leurs imbrications dans une forme de zoom continu. En somme, le rêve d'une maîtrise complète du champ de l'observation.

Sans aller jusque-là, l'accès potentiel à ces masses de données (dont une grande partie sont des traces d'usage) représente aujourd'hui un enjeu stratégique pour les sciences humaines et sociales comme elles le sont déjà pour certaines sciences exactes. Si l'on peut noter çà et là un engouement certain pour l'invention de nouveaux dispositifs d'observation du "fait social", on ne peut contourner le fait que cet accès aux données - aux "traces" - produites par les différents dispositifs numériques dans un nombre très large de contextes d'usage contribuent d'abord et surtout à la résurgence d'un **débat contradictoire en sciences sociales** qui sépare historiquement, d'un côté, les tenants d'une approche qualitative et, de l'autre, ceux pour qui les "quantités" font apparaître des *patterns* que l'observation d'un seul élément (ou même de quelques-uns) ne saurait révéler. Ce débat n'est pas un accident ou le fruit du hasard: c'est tout le rôle des sciences humaines et sociales que d'imposer cette *suspension du temps* des projets de développement technologique pour interroger de façon critique aussi bien les modèles que les méthodes scientifiques mobilisées qui, à l'évidence, ne sont pas toujours explicités ou "objectivés". L'intensité actuelle de ce débat me semble même résonner à l'échelle de l'histoire où les disputes les plus vives, comme les bouleversement les plus profonds dans la connaissance du monde, ont eu lieu à des époques-charnières (Renaissance, Révolution industrielle à la fin du XVIIIe siècle). L'accélération des mutations sociétales et économique depuis l'avènement d'Internet paraît ainsi aller de paire avec la résurgence d'un questionnement global des méthodes scientifiques, hautement *prévisible*.

Sans entrer dans les débats érudits de la sociologie ou de l'anthropologie (dont je ne suis pas friand), les orientations *qualitatives* ou *quantitatives* me semblent représenter deux façons complémentaires de construire l'objet de science observable. A grands traits, on peut concevoir spontanément "l'orientation quantitative" comme l'univers des mesures et des variables, de la recherche de patterns

sociaux 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 passés 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* (sociales), les interaction 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.

L'irruption récente de la culture des *data* dans le champ scientifique (soit parce qu'elles représentent un moyen d'investigation dans les *e-sciences*, soit parce qu'elles constituent elles-mêmes un objet à connaître sous la forme du *data scientist*) est ainsi venue faire resurgir, et sûrement enrichir, quelques-uns des multiples aspects de ce débat fécond entre "approches qualitatives" et "quantitatives". L'on aurait tord de croire que cette double orientation du travail de construction scientifique de l'objet ou du phénomène à investiguer se cantonner au seul secteur limité de la sociologie, voire peut-être à une problématique spécifiquement liée à l'épistémologie des SHS. A y regarder de près, le débat résonne à travers toutes les sciences dès qu'il s'agit d'interroger les méthodes d'analyse, les instruments d'observation et le travail de recueil des données expérimentales. Et les *digital data* en soulignent l'urgence tant les dispositifs numériques démultiplient les possibilités de production des données expérimentales, sans oublier les masses d'informations qui circulent sur les réseaux, parfois même accessibles sous forme de flux, et que l'on peut mobiliser comme données de l'observation du fait social.

Ecueils. Le débat peut être passionnant, encore faut-il en démêler les dimensions épistémologiques, méthodologiques et techniques. Aussi, je n'ai pas l'intention, ni l'érudition nécessaire, pour aborder la question de la quête de la "nature" du fait social, des problématiques riches et anciennes de l'articulation de l'individuel et du collectif, des usages et de l'habitus, du faire société et de la révolution historique de ses formes. Et les sciences des réseaux m'ont éloigné depuis longtemps de l'horizon des SHS. Cependant, il me semble possible, en restant sur les questions de méthodes de traitement, de reformuler les termes du débat ancien entre "approches qualitative et quantitatives", en particulier dans le contexte de l'ingénierie des *data* qui me semblent en renouveler les contours. En d'autres mots, de le concentrer sur ce qui en fait *l'origine*: comment s'empare-t-on des données (expérimentales) pour les organiser en "objet" interprétable?

Dans les débats actuels, l'écueil principal semble résider dans le fait que l'on assimile souvent (et trop rapidement) l'approche quantitative aux démarches volontairement "ascendantes" ou exploratoires, qui ne procèdent pas d'une démarche théorique ou qui la mettent en "suspend" le temps d'isoler les patterns contenus dans les données (*data driven methodology*). Personnellement, j'adhère à ce genre de posture de type *bottom-up*, faite d'inventivité technique et de découvertes inédites et récentes, où les concepts des disciplines "classiques" (sociales ou non) se révèlent trop limités pour expliquer les propriétés des réseaux ouverts d'informations. Mais, comme on le verra, "l'approche *data*" se constitue par un travail aussi bien qualitatif que quantitatif sur les informations par lesquelles nous saisissons les "faits" ou les "phénomènes" (notamment à travers le travail de conception de modèles de données). En tout état de cause, assimiler "approche quantitative" et démarche exploratoire en *data sciences* (ou en *network sciences)* me paraît réducteur. J'ai tendance à penser qu'une forme de méconnaissance de la richesse des mécanismes en jeu dans un "simple" algorithme ou, plus largement, d'une « chaîne de traitement de l'information » conduit à ce genre de confusion. C'est là que la reconduction régulière du débat historique entre approche "quali" et "quanti" fait pour moi *symptôme*: faut de s'être suffisamment emparé du champ des données numériques et des réseaux (du "fait informatique"), la sociologie comme les SHS en général ont aujourd'hui bien des difficultés à concevoir la richesse de l'intelligence algorithmique et du *data-processing* qui reposent sur l'exploitation de phases tant qualitatives que quantitatives de traitement des données (explicitement conçues *comme telles* par les ingénieurs notamment).

Il en va de même quand on réduit l'orientation qualitative à un travail sur les "concepts", les "hypothèses" ou les modèles théoriques. Certes, une "perspective qualitative" suppose un travail fin et attentifs sur les données et leurs modèles de traitement, éventuellement en amont du travail de recueil de façon abstraite, mais le débat gagnerait en clarté (au moins dans un premier temps) si l'on dissociait les réflexions de type épistémologique (l'articulation dans le travail scientifique entre "hypothèses" et données "expérimentales", avec ses processus à priori antagonistes de type hypothyético-déductif d'un côté et démarche ascendante de l'autre) de l'observation attentive des mécanismes *logiques* et *techniques* par lesquels nous travaillons *qualitativement* et *quantitativement* sur les informations qui donnent corps aux "phénomènes" ou aux "objets" de science. En somme, en revenir aux deux grands types d'opérations logiques que nous mobilisons pour recueillir et analyser les données de l'observation, tant logiquement pour les traiter que techniquement pour les produire. Et, en sciences humaines et sociales comme ailleurs, c'est à ce niveau-là seulement que me semble fondé le débat entre "approches qualitatives" et "quantitatives". En d'autres mots, il m'apparaît manifeste que *qualité* et *quantité* ne sont guère opposables, ni même isolables en tant que telles et quelle que soit l'échelle envisagée.

La culture des *data* et de l'hybridation. Si, donc, on se donne la peine de plonger dans cette ingénierie contemporaine des *data*, on s'apercevra rapidement à quel point **on ne peut pas figer la distinction** entre "qualitatif" et "quantitatif", quelles que soient les postures. Avec des données numériques entre les mains, ce qui relève d'une approche très qualitative (par exemple, l'observation d'un acteur engagé dans une tâche qu'une une foule de capteurs peuvent enregistrer) peut rapidement déboucher sur un travail très quantitatif. Dans l'univers des données numériques en réseau, il est toujours possible d'enrichir les informations les plus locales par des ajouts ou des corrélations continus (avec d'autres sets de données par exemple pour isoler des "types" de comportements), de multiplier à l'infini le support de travail (en faisant varier les méthodes de

traitement d'un jeu restreint de données), de vérifier ailleurs sur le réseau la "singularité" d'une pratique ou d'un phénomène considéré comme "social"... Et cette *plasticité* du support numérique et des données joue dans les deux sens, incitant à fusionner les deux approches. Ainsi dans une approche *big data* la réduction des masses repose surtout (en première approche) sur l'identification d'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. 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 (éventuellement valorisées sous forme de services). C'est sur cette **intrication native** des deux approches qu'est basée une partie 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 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).

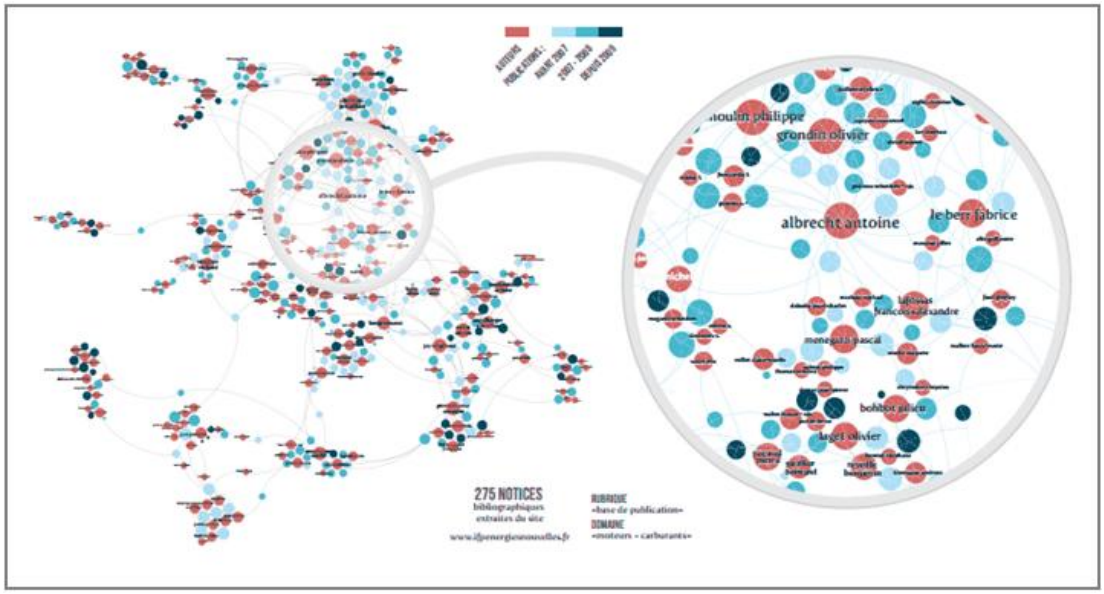
Ainsi, ce que l'on appelle *l'intelligence des données* n'est pas une métaphore et il faut être attentif aux différentes façons dont sont conçus et testés les **modèles** de données, depuisleslaboratoires où ils sont conçus jusqu'aux applications dont nous servons en ligne. Ils constituent un enjeu central en phase d'exploration des corpus "bruts" parce qu'ils agrègent autant des **règles qualitatives** (repérage, sélection, tri) de réduction des masses tout comme ils mobilisent des **règles quantitatives** de validation (seuils statistiques, recherche d'échelles pertinentes). Encore faut-il ne pas confondre, à chaque étape de traitement de l'information (car les données circulent et sont reprises en différents points d'accès du réseau, subissant chaque fois de nouvelles transformations), les "données brutes" et les modèles d'intelligibilité qu'on leur applique. Le renouveau du débat en SHS sur les questions de "méthodes" et d'épistémologie des dispositifs numériques me semble souvent symptomatique d'un corps de disciplines dépossédées de leur prérogative expérimentale. Hormis dans quelques cas (comme le *médialab* de Sciences-Po-Paris où l'on essaie de contribuer habilement aux *digital humanities* en termes de méthodologie d'extraction et d'analyse des données numériques), les univers de *data* n'ont pas été construits dans les champs d'origine des SHS comme un ensemble contrôlé de données d'observation. Faute de s'être dotés d'une forme originale d'ingénierie et d'instruments dédiés d'observation, de nombreux chercheurs en SHS ont ainsi naturellement tendance à réduire les "machines", les algorithmes et les *big data* à un l'univers (supposé) des processus automatiques et des "méthodes statistiques" et quantitatives. La distinction entre "données" et "modèle" (par exemple sous forme d'un **algorithme** permettant de contrôler tant qualitativement que quantitativement un *process*) devrait au contraire conduire à reconnaître la richesse et la complexité des mécanismes en jeu dans un processus de "traitement de l'information".

La cartographie: une machine logique artisanale. La pratique de la cartographie d'information me semble ici exemplaire si l'on s'attache au travail élémentaire sur les données. Le plus souvent, 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	48.8566I4	2.3522219	
A2	Acteur	P37	47.2I8371	-1.553621	

Combien *quantitativement* de lignes (d'objets à compter), combien *qualitativement* de colonnes (de descripteurs, d'angles sous lequel éclairer le set)? Je réduis à peine l'horizon: en SHS, comme ailleurs, "l'objet de science" est souvent construit logiquement en articulant deux curseurs: 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. L'exploration des données sous forme de graphes relationnels représente à cet égard un exercice type de construction: en croisant les colonnes entre elles pour faire émerger, à travers le nombre de lignes des principes de redondance ou de récurrence.

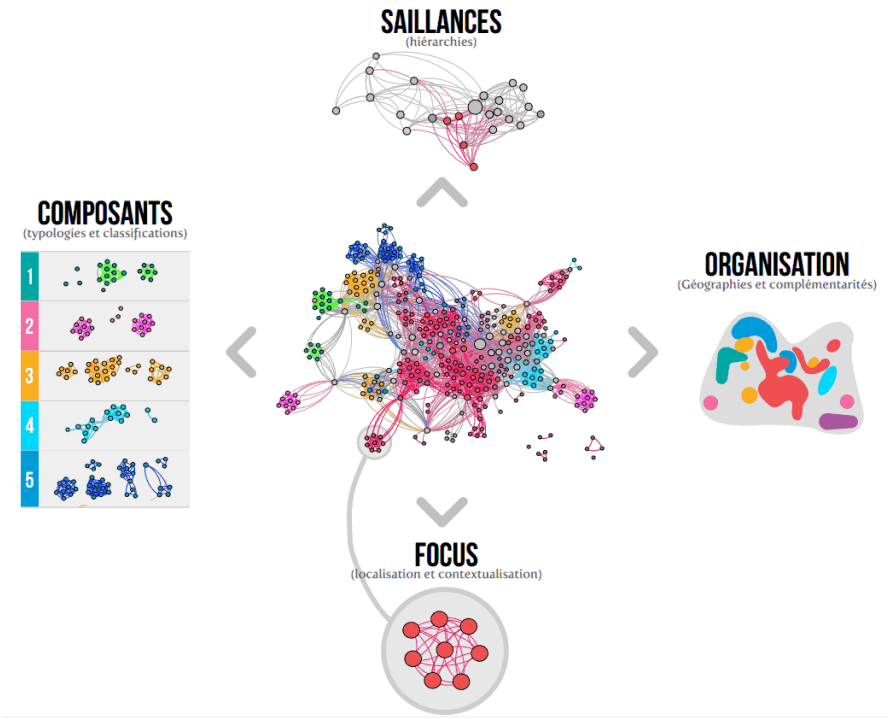
Si le réglage des fenêtres de saisi de l'objet sont trop massives (du point de vue cognitif comme du point de vue des instruments), des outils de (pré)traitement automatique et de réduction sont nécessaires pour identifier des clusters, analyser des composantes multiples, agréger des dimensions, visualiser des croisements ou des résultats de seuillage. Dans son atelier, le cartographe polit ses productions avec ces deux types d'outils qui font les deux facettes d'un métier unique; le **contrôle qualitatif** des **frontières de son corpus** à partir de critères de pertinence et d'exhaustivité (que connaissent bien les documentalistes et les bibliothécaires) et le **contrôle quantitatif** de **l'étendue** en veillant à pouvoir parcourir le corpus à tout instant (que connaissent les bien les ingénieurs avec leur art du requêtage sur les bases de données pour balayer les plus rapidement possible tous les objets d'une collection).



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.

Peu importe, d'ailleurs, la nature des phénomènes scientifiques visés dans l'exploitation de ce qu'il faut bien appeler une **machine logique** qui, de la simple feuille de papier avec ses listes et ses tableau à la table *Excel* jusqu'aux systèmes "temps réel" sur le web, nous inscrit (manuellement ou automatiquement) dans un travail de croisement **systématique** du qualitatif et du quantitatif. Nos deux curseurs dessinent une fenêtre de saisi de phénomènes décrits avec des informations et s'appliquent aussi bien à des gènes, des publications scientifiques, des organisations sociales, des molécules, des abonnés, des concepts qu'à des "acteurs" ou des "interactions". Fondamentalement, la dichotomie "qualitatif/quantitatif" renvoie pour moi aux **deux dimensions logiques** par lesquelles se construisent les objets de connaissance, ces deux curseurs qu'associe le cartographe, souvent avec difficulté (ni *trop*, ni trop *peu* de dimensions dans les données pour pouvoir manuellement jouer sur leur corrélation; ni *trop*, ni trop *peu* d'éléments ou de "lignes" pour pouvoir régler manuellement là aussi les échelles appréhendables localement). Ce sont là deux conditions pour assurer à la cartographie de l'information sa pertinence à titre de "cliché photographique" sur un jeu de données: a-t-on suffisamment réduit les masses pour faire apparaître un pattern identifiable, a-t-on identifié suffisamment d'identités remarquables pour les voir apparaître dans l'ensemble des masses? La problématique du "quantitatif" et du "qualitatif" s'inscrit d'abord 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.

Classer et composer. Evidemment, une dimension peut imprimer à la seconde ses contraintes. En pratique, certains vont se concentrer plutôt sur les opérations d'identification, de sélection, de *ranking* et recomposer petit à petit des quantités (représentatives). Par exemple, on peut construire 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 automatiques)? 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* (et oui, je l'utilise encore!), j'archive à la fois un corpus de pages web mais aussi celles que j'ai exclues.



Mais l'exercice nécessaire du contrôle qualitatif sur les quantités ne s'arrête pas à la construction des "frontières" (si l'on a à le faire) mais se poursuit aussi dans la recherche de **composants internes** des masses et leur distribution dans une **géographie générale**. 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. Il s'agit, en quelques sortes, d'explorer les masses en les segmentant selon différents critères (par exemple en recherche de clusters), de "faire tourner" les *data* et les appréhender dans leur ensemble selon le nombre de dimensions (ou composantes) présentes dans la table de données. Sous certains angles, les quantités envisagées se segmentent et l'ensemble apparaît parfois (ce qui fait aussi la valeur de l'angle choisi) comme une sorte de "puzzle" où chacun des éléments semble trouver "naturellement" sa place. C'est ce principe que j'avais appelé il y a quelques temps à propos d'un travail de cartographie sur les brevets, les "composants" et "l'organisation". C'est l'une des figures à mon avis a essentielle du travail de cartographie: faire varier la palettes des métriques qualitatives (par exemple le type d'attributs associés aux éléments d'un système) pour multiplier les facettes sous lesquelles envisager le corpus, et faire ainsi émerger différents modes de découpage des quantités de données. La recherche de "clusters" participe de cette recherche des formules de complémentarité qui président, notamment, à leur assemblage.

De façon complémentaire, les quantités permettent de valider le degré de généralité d'un phénomène - voire son "universalité". Et c'est toute la question des approches "data", donc aussi de la démarche de cartographie de l'information: ce n'est qu'à partir d'un **certain seuil quantitatif** qu'apparaissent ces *patterns* que l'on cherche à faire apparaître et que, parfois, personne n'avait jamais vraiment supposés. 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 défini ou

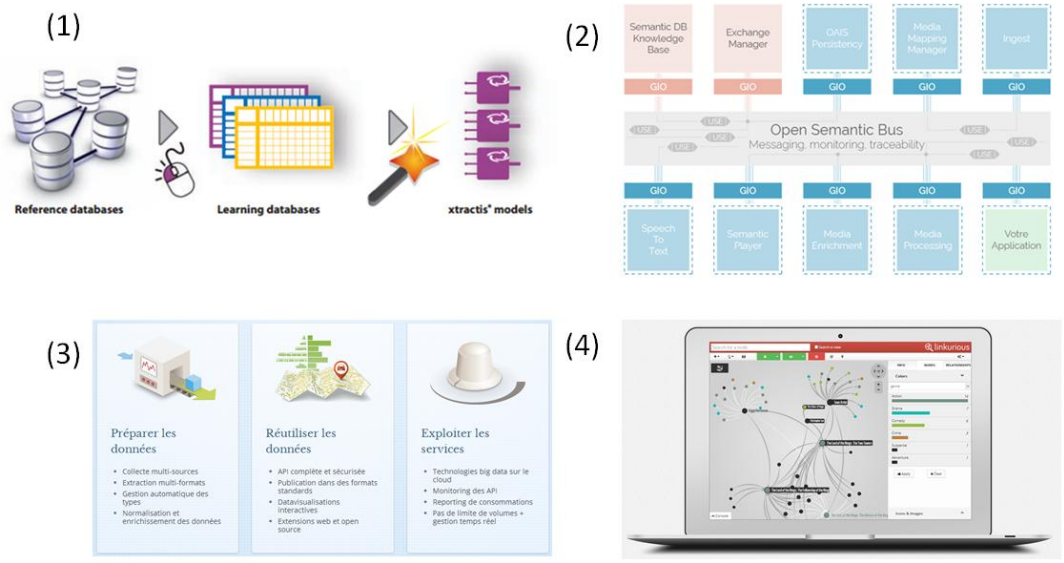
sélectionné qualitativement. On réduit souvent le travail sur les quantités à l'usage des outils statistiques et mathématiques, conduit par des ingénieurs focalisés sur les grandeurs. C'est vite oublier, comme en *network sciences*, que les résultats de calculs statistiques (par exemple, *le small world 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 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 (et donc qu'un dispositif pourrait traiter sous forme visuelle de "zoom" en assumant qualitativement l'angle adopté ou le "point de vue" sur les données).

L'art de la formule. Cette maîtrise des deux curseurs logiques est essentielle dans le travail de conception (et de communication) d'une cartographie, qui est une forme de **modèle visuel et statistique 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, 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 débat en 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 souvent cité dans ce blog 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.

$$\mathrm{auth}(p) = \sum_{i=1}^{n} \mathrm{hub}(i)$$

L'objectif, évidemment, est de disposer en phase exploratoire des corpus d'une batterie suffisamment étendue d'algorithmes (ou de formules opérationnelles), quitte à les associer à leur tour dans des ensembles plus vastes, et qui fonctionnent rapidement. Si nos instruments informatiques ont contribué à élargir de façon massive nos fenêtres qualitatives et quantitatives de saisi de l'objet scientifique, ils nous permettent surtout d'automatiser leur croisement ou leur *triangulation*, jusqu'à une grande complexité. 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 réaliser avec de vastes dispositifs de stockage et de traitement des données. **L'art de la formule**: voici ce qui caractérise selon moi le développement des algorithmes à l'heure du *big data* et de l'accès à des corpus distants, eux-mêmes déjà analysés et enrichis.

Certaines formules, d'ailleurs, "valent de l'or" si j'ose dire car il s'agit d'une question centrale dans le domaine des technologies numériques de l'information. La conception de formules ou de modèles de traitement de données concentre les 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. Il s'agit d'une activité centrale en recherche et développement mais aussi au plan scientifique où on peut les considérer comme des **modèles d'intelligibilité des données**. 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. Comme dans d'autres domaines industriels, les formules peuvent être protégées juridiquement (brevets) et valoriséee à hauteur de leur 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.



Différentes technologies actuelles développées par des sociétés françaises dans le domaine du traitement et du management de l'information 1) *xtractis* développé par *Intellitech* dans le domaine de l'aide à la décision 2) la solution de gestion des connaissances développée par *perfect memory* 3) la plateforme de gestion et

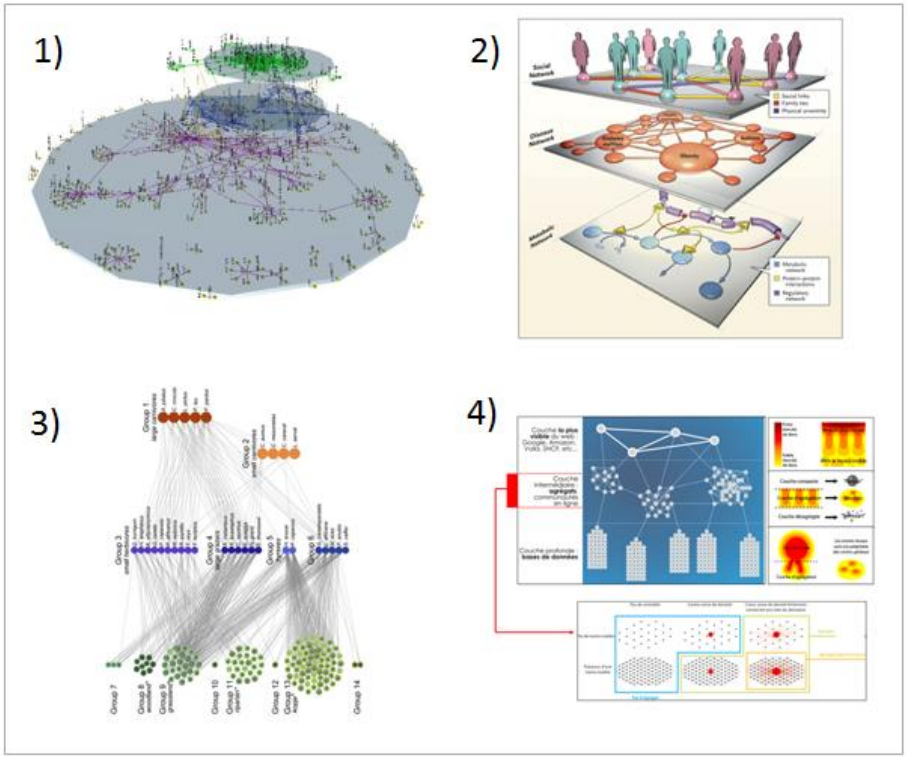
d'accès *opendatasoft* sur le cloud 4) Les web services développés par *linkurious* dans le domaine de l'exploration des grandes bases de données sous forme de graphes dynamiques ([1]).

L'*alchimie quali-quantitative* et les changements d'échelle. L'engouement actuel en *data sciences* pour la construction de nouveaux modèles de données (qui donneront naissance pour certains à de nouveaux services) conduit nécessairement à embrasser des volumes de données de plus en plus grands, hétérogènes et dynamiques. Vue de loin, cette effervescence peut faire penser à une course technologique vers un monitoring global et "temps réel" des informations produites ou transitant via nos instruments numériques en réseau. Cependant, derrière les questions de volumétrie des données et de puissance de calcul, se profile un autre aspect important de l'alchimie quanti-qualitative (ou quali-quantitative) qui préside à la conception de nouveaux modèles de données: les **changements d'échelle**. L'accumulation quantitatives des données peut en effet conduire, dans certains cas, à la *modification qualitative* de l'objet de science. En sciences des réseaux, cela est certain: l'apparition des outils informatiques et des masses de données analysées a propulsé l'univers de la théorie mathématique des graphes vers les *network sciences* actuelles, en particulier depuis les développement des technologies web et les données indexées dans les moteurs de recherche. A partir du milieu des années 90, les travaux de J. Kleinberg, D. Watts, M. Neuman ou A.-L. Barabasi (et bien d'autres) ont commencé à révéler quelques propriétés fascinantes des grands systèmes complexes (comme les pages web reliées entre elles par liens hypertextes, les réseaux de distribution électrique, les acteurs de cinéma participants à des films communs, la circulation des billets de banque à l'échelle d'un pays...). L'étude de deux propriétés statistiques majeurs de ces réseaux (le *small worl phenomenon* et le rôle majeur des "hub" dans la distribution de la connectivité entre chacun des éléments du système) ne sont perceptibles qu'à un *certain niveau quantitatif* de données mais elles ouvrent aussi de fait l'exploration scientifique à un ensemble de problématiques nouvelles, comme celle de l'évolution temporelle de ces systèmes, celles des mutations globales, de l'analyse des flux, de leurs niveaux de décomposition en clusters ou, *in fine*, celle de leur contrôle. Bien au delà des rapports qui lient historiquement un objet à sa discipline (les gènes et la biologie, les acteurs et la sociologie, les pages web et les computer sciences, les articles scientifiques et la scientométrie...), les *network sciences* interrogent ainsi aujourd'hui les propriétés transversales (et en partie communes) de tous ces "objets" considérés comme systèmes d'interaction. Et l'un des apports majeurs de l'étude des réseaux à grande échelle est ainsi constitué par l'avènement de nouvelles métriques dont certaines peuvent être classées comme de mesures qualitatives reposant sur la maîtrise des quantités. Il en est ainsi des procédés de *ranking* (classement) des éléments d'un systèmes en fonction de leurs propriétés topologiques (scores de différentes formes de centralité, de *PageRank*, d'*Authorities* avec HITS...).

C'est en cela que le terme de *topologie* m'a toujours paru justifié pour désigner la distribution des interactions à grande échelle entre tous les éléments d'un système mais aussi la façon dont il se décompose en ***layers* superposés** (en "couches" si l'on veut). La réalité des univers numériques, notamment celle du web, ressemble étrangement à celle des nouveaux matériaux, composites

[1] (1) http://www.xtractis.fr/pdfDoc/Brochure%20xtractis_UK.pdf
(2) http://www.perfect-memory.com/nos-solutions/les-offres/
(3) www.opendatasoft.com/fr/
(4) http://linkurio.us/tour/

(principe de corrélation entre les "liens" et les "mots" ou contenu-structure) et multi-échelles (principe de couche et des agrégats). Je ne sais ce que peut valoir la comparaison entre formalisation quali-quantitative de l'information et la question des nouveaux matériaux, un domaine central de la révolution industrielle actuelle. Il me semble cependant qu'une même conception de la complexité est à l'oeuvre dans les deux cas: celle de l'étude des interaction entre différents niveaux d'échelle, une propriété située à une nano-échelle pouvant être exploitée à plusieurs niveaux macroscopiques, tout comme la distribution des liens hypertextes sur le web peuvent s'analyser en différentes "couches". Les questions de changement d'échelle dans les systèmes complexes (voire même de réseaux de réseaux – NoN, *networks of networks*) apparaissent aujourd'hui comme centrales dans bien des domaines de recherche, depuis l'étude des biotopes naturels jusqu'aux écosystèmes numériques en passant par l'étude du vivant depuis l'ADN jusqu'au phénotype.



Le principe des changements d'échelle dans les données dans différents domaines scientifiques: 1) et 2) les réseaux biochimiques d'interaction - métaboliques-protéines/protéines-régulations génétiques 3) réseaux trophiques et interactions entre espèces du Serengeti 4) modèle "en couches" en web-mining.

D'un bout à l'autre des échelles de complexité, des éléments font système à chaque niveau à travers une série de ruptures. Un blog, en soi, peut constituer un champ d'étude avec ses *posts* et ses commentaires (autrement dit un espace que l'on cherche à étudier tant qualitativement en faisant varier les points de vue de l'analyse que quantitativement en les validant sous forme d'ensembles d'informations alors cohérentes) mais il faut dépasser le seuil des centaines de blogs pour apercevoir les premiers contours de "communautés actives sur le web", puis les continents formés sur *Facebook* ou, dans une autre direction, la distribution de la connectivité des liens hypertextes à grande échelle pour construire "Le" graphe du web (ou sa "carte"). Dans le domaine des masses de données numériques et de leur mobilisation potentielle en sciences et humaines et sociales, cela devrait conduire à supposer qu'il n'y a pas plus de continuité dans le "social" qu'ailleurs et qu'une série de

"sauts" qualitatifs (ou au contraire de décomposition dans le sens inverse) règle l'univers des changements d'échelle, et donc des objets d'étude. Du micro au macro, de l'individu à la société ou à l'histoire (ou quel que repère que ce soit), on ne peut augmenter à l'infini les données d'observation sans apercevoir à un moment donné émerger un pattern (ou un ensemble de patterns par exemple des "groupes sociaux" ou des affiliations enchâssées) qui constitue en soi un nouvel objet d'étude à son propre niveaux. Pour moi, l'examen des masses de données numériques attachées à l'observation d'un phénomène (social ou non) ne débouche pas sur cette forme fantasmée de *continuum*, de zoom optique continu depuis les strates les plus fines jusqu'aux effets massifs des patterns statistiques et collectifs. Un épistémologue m'avertirait qu'il s'agit là, peut être, d'un phénomène produit grâce à l'artifice des instruments ou des méthodes déployés. Au contraire, j'y vois la manifestation d'un principe d'organisation réellement inscrit dans les données et ce n'est pas le moindre des apports de "l'approche réseau" que d'obliger à revenir au travail de fonds sur les données et leurs formalisations tant qualitatives que quantitatives. Le principe des ruptures des échelles logiques d'observation (puisqu'à chaque niveau on change d'objet) englobe d'ailleurs aussi l'exploration des phénomènes temporels puisqu'il s'agit là d'un champs d'expérimentation majeurs actuellement. L'examen des phénomènes temporels associés aux données numériques commence à laisser entrevoir des ruptures par transitions de phase souvent brusques. D.Watts, S. Strogatz ou V.-L. Barabasi ont entamé depuis une quinzaine d'années l'étude de ces phénomènes temporels dont les données d'observation portent la trace (et c'est peut-être là l'une des grandes avancées de la manipulation de données numériques en masses: l'observation des variations dynamiques dans le temps). En termes d'approche réseau, la topologie des systèmes observés (sociaux ou non) imprime aux données une organisation aussi bien en *layers* (logiques) qu'en périodes (dénombrement quantitatif) ponctuées "d'événements" (qualitativement repérables). Dans les deux cas, il me semble que la conception de nouveaux modèles de données repose sur tous ces croisements potentiels encore imaginer entre dimensions qualitatives et quantitatives des informations. Une alchimie à laquelle le numérique et le *big data* apportent indéniablement de nouvelles perspectives...

Franck Ghitalla.