



Introduction - Visualisation de données

David Reymond

► **To cite this version:**

David Reymond. Introduction - Visualisation de données. Les Cahiers du numérique, Lavoisier, 2016, La visualisation de données, 12 (4), pp.9-18. 10.3166/lcn.12.4.9-18 . sic_01808316

HAL Id: sic_01808316

https://archivesic.ccsd.cnrs.fr/sic_01808316

Submitted on 24 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INTRODUCTION

Visualisation de données

DAVID REYMOND

La question de la visualisation est au cœur de la pensée, de la philosophie occidentale et de la science. Elle est aussi au cœur des pratiques scientifiques (Goody, 1979 ; Hofstadter, Dennett, & Henry, 1987) et de leurs écritures (Latour, 1993). Elle est aussi centrale – et ce d’autant plus à l’ère de la numérisation du signe – au sein des pratiques socio-cognitives en général, des pratiques communicationnelles en particulier. La visualisation de données qui s’impose aujourd’hui comme instrument d’interface aux données massives en médiateur ou traducteur, et dont l’herméneutique modifierait nos pensées, est soutenue par une étendue applicative sans précédent (Bihanic, 2015). Réservée jusqu’à peu aux sciences expérimentales, elle s’ouvre aujourd’hui à l’ensemble des SHS « pour lesquelles les sciences de l’information et de la communication possèdent un ensemble de ressources pour développer de multiples pistes de recherche » (Plantin, 2014).

En effet, dans notre contexte, les données, source des humanités digitales (Berry, 2012), sont ces traces que nous laissons au cours de nos pratiques et comportements dans l’univers numérique, les traces que nous produisons au cours du vaste mouvement de déploiement de « l’internet of everything », qui expriment tout ou partie du vaste système relationnel dans lequel nous sommes inclus et que nous ne cessons d’alimenter. Ces traces d’édition, d’activité ou d’échange sont chacune, atomique, un témoin d’une réalité prise sous un certain angle de vue.

Ces empiries numériques sont à présent dominantes et, de leurs fabrications et traitements, la réflexion attend beaucoup. Du quatrième paradigme de Jim Gray (Hey, Tansley, & Tolle, 2009) en passant par les travaux séminaux de la scientométrie (Callon, Courtial, & Penan, 1993) et de l'infométrie, le traitement de ces traces s'appuie, on le sait, sur une algorithmique appropriée pour filtrer, augmenter, trier, agréger et rendre interopérables l'ensemble de ces données. Leur agrégation laissera peu de place à d'autres réalités alternatives au plan statistique et, à partir d'un seuil relatif, leur masse rendra ces signaux comme faits (Carmes et Noyer, 2014 ; Rouvroy et Berns, 2013) puis, en ce sens, comme preuves factuelles, des « obtenus » (Latour, 2011). Ainsi, la construction de ces obtenus et le travail – scientifique ou pas – d'extraction (Berry, 2008) définit une part essentielle de la recherche aujourd'hui (Abiteboul, 2013). Callon annonçait aux prémices de l'ère informationnelle que « dans cet environnement turbulent, dont l'évolution dépend des initiatives d'un grand nombre, la capacité de collecter des informations et de les traiter, avant de prendre les décisions, devient cruciale » (Callon *et al.*, 1991, 208). C'était alors sans compter sur l'explosion généralisée des données qui ouvre de nombreuses interrogations de types épistémologiques (Carmes & Noyer, 2014), politiques (Rouvroy & Berns, 2013), sociétales (Harcourt, 2007) mais aussi la nécessité de les observer dans leur totalité. La visualisation, qu'il convient de calibrer et d'adapter, se pose comme l'aboutissement de ces chaînes de traitement et d'exploration pour appréhender ces masses de données : apte à générer indicateurs, cartographies ou autres combinaisons synthétiques grâce auxquels l'interaction avec l'analyste sera informative et communicante. La visualisation des données en tant que processus et en tant que pratique devient alors décisive, mais la véracité informative résultante dépendra de la manière dont l'accès aux couplages données-technologies intellectuelles-problèmes, sera conçue et fabriquée.

Par la construction précédente, la visualisation devient un « Mieux voir » c'est-à-dire qu'elle permet d'étendre le spectre des conditions de visibilité, d'étendre le travail de l'intellect dans le contexte hypermédias et des écologies cognitives nouvelles. Sur le plan de ce que Pierre Lévy a nommé *l'idéographie dynamique* (Lévy, 2010) c'est une extension des nouvelles écritures et de la place des modes de perception dans les usages de la pensée : « le concept de l'idéographie dynamique ne consiste pas à faire appel à l'image pour illustrer le texte classique mais bien à inaugurer une écriture nouvelle (un instrument de pensée qui soit aussi image animée) ». Son exploitation constitue l'essentiel des activités intellectuelles pour produire ou extraire des savoirs divers et variés.

Rétrospectivement, la visualisation de données s'exprime aussi comme une généralisation de la notion de carte géographique : un instrument (Harley, Laxton, & Andrews, 2002) qui est « au géographe ce que le microscope est au

microbiologiste, par sa capacité à rétrécir la terre et à permettre des généralisations», s'avère être une «écriture dangereuse ou dotée d'une puissance inquiétante» (Farinelli, 1989, 36). La «puissance» de la visualisation de données (massives, multi-échelles, de provenance variée) est d'autant démultipliée tant par les procédés techniques mis en œuvre, que la variété scientifique des questions adressées aux obtenus «latouriens», ou encore que leur mise en pratique pluridisciplinaires et collective, ce qui (re)pose l'interrogation de la dangerosité de ses créations indirectes. La création de visualisations modernes (Farinelli *et al.*, 2009) est soumise à de nouveaux régimes de visibilité, à de nouvelles conditions d'exercice de la pensée et ce d'autant plus que les collectifs de pensée sont à présent de plus en plus différenciés et que les questions de la réflexivité prennent une ampleur sans précédent. De ce point de vue, la place des interfaces dans les processus collectifs de visualisation, d'une partie des hétérogénéités et des procédures, des modes d'orientation et de désorientation (Stiegler, 2006) est au cœur d'une nouvelle économie politico-cognitive de la perception de ses grammaires. Elle devient cette herméneutique (en trans-formation) des données complexes s'immiscant depuis les modes de pensée et les pratiques singulières, aux dispositifs de cognition distribuée, ou encore aux dispositifs d'intelligences collectives convoqués.

La visualisation est donc double processus à la fois comme mise en visibilité (l'aboutissement d'une instrumentation d'exploration de ces masses de données, devenant apte à générer indicateurs, cartographies, etc.) et une redéfinition du voir pour ouvrir vers de nouvelles connexions, de nouvelles analogies, de nouvelles conditions déductives, de nouvelles connaissances (Rasmus, 2013)... La visualisation des données comme problème revient donc à poser les fondements d'une nouvelle herméneutique singulière et/ou collective (Brunet, 1997), en évolution (Friendly, 2008) marquée aujourd'hui par les technologies.

Les articles de cet ouvrage laisseront le lecteur progressivement s'immerger vers cette nouvelle herméneutique. En commençant par des travaux exploitant l'utilisation de diverses formes de visualisations, sera abordée l'exploitation de nouvelles formes de visualisation de données, et conduite une réflexion sur l'instrumentation propre à la visualisation (boîte à outils) tout en considérant la visualisation sous ce double processus, comme médiation, entre approche critique, d'une part, et approche scientifique en extension de la cartographie, d'autre part. Nous résumons et commentons les travaux en ce sens ci après.

Après avoir rappelé les fondamentaux de l'intelligence économique sur lesquels elle construit ses indicateurs, le choix des sources d'information et leur traduction en problème de recherche d'information, N. F. Ndjock présente la construction d'un système décisionnel pour lequel la visualisation facilite

l'interprétation des données, par leur synthétisation. L'outil ODSE (observatoire dynamique du système éducatif), décline différents acteurs ayant accès à des fonctionnalités différentes construites en réponse aux besoins informationnels en agrégeant celles-ci depuis les sources réparties sur le territoire. Différentes visualisations préconstruites sur les données sont proposées en un tableau de bord pour les acteurs clefs afin de faciliter leur appréhension et la prise de décision. L'auteure, à l'aide d'exemples de terrain, montre que le croisement de visualisations basiques offre une lecture éclairante sur les données trop nombreuses pour être appréhendées sous leur forme initiale. Elle en souligne l'impact dans le système décisionnel en montrant l'instrumentation globale comme une combinaison d'outils facilitant l'intermédiation.

S. Gorla propose, quant à lui, d'explorer l'art et le jeu en particulier. Il rappelle le contexte des jeux sérieux et situe la pratique du *disengagement* consistant à détourner un jeu pour présenter une situation informationnelle particulière. L'auteur s'attache aux effets d'une photographie dans le cadre d'une partie de jeu et ce à un instant t pour mettre en valeur, mieux que d'autres graphiques, des informations pertinentes pour un problème. Il procède alors à la construction d'un exemple sur le marché automobile. Le choix du jeu sera pensé en vue de sa transformation en instrument cartographique et seront examinées et analysées les réactions multiples générées par le dispositif cartographique lui-même. L'auteur note que le jeu de plateau stratégique « Risk » est un bon support pour analyser la médiation d'information stratégique. Mais ce dispositif cartographique, pour être bien exploité, suppose la connaissance ou la référence au jeu d'origine, ce qui confirme que ce dispositif n'est opératoire que dans le cadre cartographique d'une culture spécifique plus large, celle du jeu de référence.

M. Peres aborde la visualisation et la représentation de connaissances complexes en étudiant un corpus d'infographies. Ce corpus contrôlé lui permet d'explicitier certains des effets majeurs du fonctionnement des outils informatiques descriptifs et statistiques dans la compréhension et l'analyse de ces visualisations. Pour satisfaire à une logique instrumentée, l'auteur déploie une chaîne instrumentale qui fait passer les infographies en représentations textuelles descriptives sur lesquelles une analyse factorielle des correspondances (AFC) des constituants des infographies mettra en exergue les variables séparant les différentes infographies. En outre, l'auteur montre l'intérêt d'une troisième dimension dans la représentation des AFC pour plus aisément expliquer les graphiques et interpréter les résultats. Interprétation qu'il convient de pondérer en regard du degré de représentativité de cet axe. Pour ce faire, le chercheur va établir une décomposition de chacune des infographies de son

corpus en objets, puis identifier leurs relations et leurs couleurs tout en élaborant une grille d'analyse. Chaque infographie est alors convertie par un outil en sa représentation textuelle qualifiant chacun des objets de la grille (titre, connecteur, logo, texte...) selon diverses modalités (présence, taille...). Les représentations obtenues sont à leur tour traitées par un outil de textométrie afin de mesurer les similitudes des profils descriptifs. L'auteur reconnaît, au niveau de cette phase pour laquelle le paramétrage impose l'élaboration d'une grille descriptive, que cela modifie la perception des infographies. Cette analyse permet toutefois de saisir des variations et ressemblances existant entre ces objets mêlant texte et image tout en éclairant le processus riche de construction de visualisations de données que constituent ces outils.

De notre point de vue, cette approche constitue les prémisses d'une méthode généralisable aux infographies, aux images, aux interfaces. La construction d'une grille de lecture est en effet comparable aux grilles d'analyse d'image utilisées en histoire, par exemple, que l'on pourrait rapprocher plus généralement, à l'élaboration d'une grammaire instanciée de la sémiotique visuelle. Les instruments informatiques de détection et classification des formes permettent d'envisager un traitement automatisé du processus de description par détection d'éléments de la grille et qualification automatique. In fine, l'originalité, côté lecture, consiste en l'utilisation de l'AFC textuelle sur ces transcriptions. L'on ajoute ainsi à une herméneutique liée à des conditions perceptives, cognitives variables et singulières pour partie des infographies, une herméneutique formalisée à l'aide de la grille d'analyse associée à une représentation visuelle pour en faciliter la lecture discriminante. La tension et le couplage entre les deux herméneutiques étant porteurs, du moins on l'espère, d'une herméneutique globale plus productive et créative.

Pour A. Kondratov, les dispositifs sociotechniques sont une clé pertinente pour comprendre le processus et les modalités de la médiation politique. Convoquant tour à tour Habermas, Dacheux, Miège, Cardon pour reconstruire l'espace public dans l'internet post-soviétique, l'auteur matérialise cet espace par la visualisation des échanges et des canaux de ces échanges. Il va opérer au classement systématique d'une sélection rigoureuse de blogs, en trois étapes, selon des critères de degré médiatique et un critère de « politisation » afin d'identifier cinq catégories de cibles des relations hypertextuelles qui en émanent. Ce faisant il évaluera qualitativement le pluralisme relationnel des différents blogs – en webométrie descriptive, cela est décrit comme la *dispersion* du site, qui détermine une typologie et la fréquence des liens sortants d'un site web en accord avec différents critères de distribution (Aguillo, 2002). Il intègre ainsi la dimension relationnelle des données analysées (l'hypertexte) en sus du problème de leur masse. Cette nouvelle dimension implique l'utilisation de

cartographies particulières issues de l'instrumentation des graphes. La représentation sous cette forme sert à montrer comment s'est formée la blogosphère et est devenue une partie de l'espace public. Les faits historiques et commentaires issus d'entretiens avec des personnalités majeures vont permettre à l'auteur de donner les clés de compréhension de sa dynamique et de sa fragmentation. L'auteur note toutefois la difficulté à saisir la dimension dynamique de ses projections, la nécessaire catégorisation pour faciliter la lecture et l'appréhension de quantités de données très importante.

M. Le Béhec développe de façon progressive le vocabulaire de la théorie des graphes, des algorithmes de positionnement en considérant le web comme un territoire à cartographier à l'aide de graphes, sans toutefois s'appuyer sur les méthodes de webométrie issues de l'infométrie et de la bibliométrie (Almind & Ingwersen, 1997, Aguillo, 2002, Björneborn & Ingwersen, 2004, Thelwall, 2009). Sur cette forme de visualisation de données, l'auteure souligne d'une part la limite de la projection des données relationnelles sur le plan à deux dimensions, d'autre part, que la cartographie des données web produit un instantané, une image arrêtée et statique, qui plus est, filtrée, excluant ces cartographies des systèmes de représentation de Hofstadter (Hofstadter *et al.*, 1987). L'auteure instancie la matérialisation d'une extraction du web en graphe et développe la réflexion sur la carte et les territoires. L'auteure présente le collecteur de donnée, la catégorisation manuelle des sites pour typer ces derniers (d'aucuns pourraient lui en vouloir de ne pas avoir proposé une méthode automatique de nous avions esquissée sur un autre terrain (Reymond & Pinède, 2010)) et convoque la théorie des agrégats. Elle souligne que cette dernière accompagne la spatialisation mais sort le lecteur de la cartographie de l'échelle à laquelle il est entré. L'auteure plonge ainsi dans les considérations herméneutiques de l'instrument cartographique de la relation qu'elle discute sur de nombreux points en montrant que cette visualisation oblige à comprendre non seulement les traitements et leur incidence sur les nœuds présentés, mais aussi à comprendre les biais inhérents qui s'en produisent lors des positionnements des nœuds (faisant abstraction du lien dans un premier temps). L'auteur montre enfin les limites de cette cartographie (et de l'instrumentation associée) pour explorer des contenus en relation à dynamique temporelle, ainsi que la nécessité de s'approprier la culture du graphe que la sémiotique de Bertin (2013) pourra compléter de façon performative.

N. Dugué, J.C. Lamirel et P. Cuxac abordent au cœur de la problématique de la dimension temporelle, manquante dans les approches précédentes (par construction initiale des données). Ils en montent la complexité et l'intérêt. Dans le champ de l'information scientifique et technique les auteurs situent leur objectif à identifier et décrire de façon textuelle et visuelle l'évolution de la

science. Ils nous offrent un panorama des méthodes de visualisation des données textuelles comparant leur apport en regard de technique d'identification thématique de l'état de l'art tout en présentant les visualisations associées. Les auteurs présentent l'algorithme de traitement par « maximisation des traits » au cœur de leur procédé permettant l'identification des thématiques tout en conservant l'importance des termes dans un sujet et de visualiser les interdépendances avec d'autres. Les auteurs indiquent différentes méthodes de visualisations puis l'analyse diachronique pour introduire l'outil Diachronic Explorer. Les différentes fonctionnalités sont présentées par un traitement d'un corpus de 9779 notices aboutissant à des visualisations les plus variées qui expriment simultanément de nombreux critères informationnels. Les auteurs concluent la possibilité d'application de cette instrumentation à de nombreux champs alliant une nécessaire synthétisation des données en amont des visualisations.

M. Boechat et T. Venturini quant à eux développent leur regard à partir de la sociologie et de la cartographie des controverses. Ils s'interrogent sur le statut de la visualisation au cours du processus de recherche en questionnant son rôle pendant le contexte de la découverte, lors de la présentation des résultats. Ils s'appuient sur le concept fondamental en philosophie des sciences de distinction du contexte de la découverte de celui de justification de l'activité scientifique.

Les auteurs dénotent l'existence d'une construction formelle amenant de l'analyse exploratoire à la présentation, pendant laquelle les processus de visualisation vont d'une part aider à soulever les idées qui prendront part de la découverte et, d'autre part, laisser entrevoir matière aux justifications. Les auteurs font référence au domaine de la sociologie numérique : collecte de données issues des médias sociaux et des sources de données variées qui fournissent aux chercheurs une représentation particulière de l'activité sociale en acte. Pour la cartographie des controverses, les auteurs parlent de traduction de ces données en une représentation dont l'objet n'est pas celui de statuer mais d'accompagner le débat. Le processus est alors un procédé récursif d'exploration et d'identification de clusters pour à nouveau explorer de nouvelles voies afin d'analyser les données jusqu'à une convergence satisfaisante. Les auteurs, sur la base d'exemples issus d'un projet international, vont discuter de diverses représentations visuelles des mêmes données qui satisferont les points de vue des différents chercheurs pour l'appréhension des résultats au cours des phases d'exploration. Ces diverses représentations ont pu être recombinaisonnées en réalisant une nouvelle visualisation qui n'est pas un procédé de simplification mais que les auteurs qualifient de procédé d'intégration des questions de recherche en tant qu'objet même de la

représentation (réification). Cette étape, complétée par un processus de communication pouvant faire partie de la phase exploratoire, convoque didactique et rhétorique pour présenter ou démontrer les idées (réenaction).

Dans l'ensemble les auteurs donnent un panorama d'applications des techniques de visualisations de données, applications réflexives sur leur mode d'utilisation, leur portée. Leur sémantique quelquefois très spécialisée nécessite un apprentissage. Ils montrent l'intérêt des dispositifs de « visualisation » non seulement comme support de rendu de synthétisation des données, mais aussi comme mettant en valeur de nombreuses fonctions complémentaires : présentation de grands volumes de données, exploration et fouille dans des données relationnelles complexes de dimensions multiples, ou dans les données temporelles.

Leur analyse de ces applications souligne que toutes ces opérations – fabrication des données (les « obtenues » de Latour), capture des données, visualisation des données et modèles d'intelligibilité – sont fondamentalement liées. La plupart des constructions proposées intègrent au moins une boucle de rétroaction, des procédés qualitatifs de réduction des données, ainsi que la définition d'un cadre d'interprétation (...). S'appuyer sur une épistémologie rigoureuse des pratiques cognitives et perceptives et sur une notion d'information pensée dans le cadre d'une pragmatique élargie serait du plus grand intérêt, les considérations des auteurs ayant sur ces points une certaine et souvent heureuse plasticité. Il s'agit en outre de permettre la « transportabilité », soit le déplacement des acquis dans d'autres domaines ce qui est d'un grand intérêt dans le contexte de l'émergence des *digital humanities* entendues ici comme transformations du procès de travail intellectuel au sein des SHS en particulier.

Références

- Abiteboul S. (2013). *Sciences des données : de la logique du premier ordre à la Toile*. (C. de France, Éd.) (Vol. 226). Paris, OpenEdition books.
- Aguillo I. (2002). Web Characterization for Cybermetric Purposes : Terminology and Definitions. In *Development of Web-Indicators* (p. 12–18). Projet Européen EICSTES.
- Almind T. C., & Ingwersen P. (1997). Informetric analyses on the World Wide Web : Methodological approaches to webometrics. *Journal of Documentation*, 4(53), 404–426.
- Berry D. M. (2012). *Understanding Digital Humanities*. Palgrave Macmillan.
- Berry G. (2008). Pourquoi et comment le monde devient numérique. Collège de France. sur <http://www.college-de-france.fr/site/gerard-berry/inaugural-lecture-2008-01-17-18h00.htm>

- Bertin J. (2013). *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. Paris, Éd. EHESS.
- Bihanic D. (2015). *New challenges for data design*. New York, Springer.
- Björneborn L., & Ingwersen P. (2004). Toward a basic framework for Webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216–1227.
- Brunet R. (1997). Champs & contrechamps, raisons de géographe (p. 319). Paris, Editions Belin.
- Callon M., Courtial J.-P., & Penan H. (1993). *La Scientométrie*. Paris, Presses Universitaires de France.
- Carmes M., & Noyer J.-M. (2014). L'irrésistible montée de l'algorithmique. *Les Cahiers du numérique*, 10(4), 63-109.
- Farinelli F. (1989). *Pour une théorie générale de la géographie*. Genève, Recherches géographiques.
- Farinelli F., Bienvenu K., Gruet B., & Fumey G. (2009). *De la raison cartographique*. Éd. du Comité de Travaux Historiques et Scientifiques.
- Friendly M. (2008). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Consulté sur <http://www.datavis.ca/milestones>
- Goody J., Bazin J., & Bensa A. (1979). *La Raison graphique : la domestication de la pensée sauvage*. Paris, Les Editions de Minuit.
- Harcourt B. E. (2007). *Against prediction profiling, policing, and punishing in an actuarial age*. Chicago, University of Chicago Press.
- Harley J. B., Laxton P., & Andrews J. H. (2002). *The New Nature of Maps: Essays in the History of Cartography*. Johns Hopkins University Press. Consulté sur <https://books.google.fr/books?id=dFWPuU2x0dkC>
- Hey A. J. G., Tansley S., & Tolle K. M. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research.
- Hofstadter D. R., Dennett D. C., & Henry J. (1987). *Vues de l'esprit : fantaisies et réflexions sur l'être et l'âme*. Paris, InterEditions.
- Latour B. (1993). « Les “vues” de l'esprit ». Une introduction à l'anthropologie des sciences et des techniques. In *Sciences de l'information et de la communication* (Larousse, p. 572-596). Paris, France, Daniel Bougnoux.
- Latour B. (2011). Sur la pratique des théoriciens. *Education et formation*, 131–146. <https://doi.org/10.3917/puf.barbi.2011.01.0131>
- Lévy P. (2010). *Idéographie dynamique: Vers une imagination artificielle ?* La Découverte.
- Plantin J.-C. (2014). L'avènement de la carte comme médiation. Généalogie des rencontres entre cartographie et théories de l'information. In *La ville, une œuvre ouverte ?* (Vol. 25, p. 309-326). OpendEdition.

Rasmus D. W. (2013). Visualizing knowledge. *KMWorld*, 22(9).

Reymond D., & Pinède N. (2010). Using a taxonomy based fingerprint: classification and recognition of the academic webspace. In *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting*. Mysore, India.

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

Stiegler B. (2006). *Mécréance et discrédit : Les sociétés incontrôlables d'individus désaffectés*. Galilée.

Thelwall M. (2009). *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*. Morgan & Claypool Publishers.