Dominique Vinck et Pierre-Nicolas Oberhauser, « Approches sociologiques des bases de données »

La présentation est cette fois divisée en deux parties. Dans un premier temps, Dominique Vinck nous présente ce que dit la sociologie à propos des bases des données. Cette revue partielle de la littérature permet d'identifier plusieurs directions d'analyse faisant plus ou moins écho à nos propres réflexions. Dans un deuxième temps, Pierre-Nicolas Oberhauser nous fait part de ses réflexions quant aux enjeux liés à la collection de données numériques en sciences humaines.

Bases de données et sociologie : grand phénomène sociétal pour petit phénomène sociologique

Pour les sociologues, il semble admis que l'expansion des bases de données est un phénomène *sociétal* de grande ampleur. Nul besoin de trop creuser pour s'en rendre compte, la considération des expériences quotidiennes suffit : un assureur répond à son client paniqué : « quelle est votre numéro de dossier ? » ; un paquet de tomates est scannée au supermarché : « 2 francs 50, s'il vous plait » ; une lointaine amie est suggérée par Facebook : « Connaissez-vous Nathalie Janzer ? » ; je tape la lettre « m » dans la case « destinataire » de ma boîte mail et défilent aussitôt « Richard Marion », « Marco Prost », « Martin Grandjean », « m-recher-ladhul », etc., etc. À coup sûr, un minimum de sensibilité nous force à reconnaître les liens intimes que nous tissons chaque jour avec les bases de données numériques.

Sans trop creuser, le sens commun des sociologues (et des autres) poursuit aisément ses observations, le plus souvent sous le mode de la *sensation* :

« On *sent* bien que les bases de données débordent largement des questions informatiques ; on *sent* bien que les bases de données sont liées à des problématiques sociales puisqu'elles suggèrent de nouveaux agencements au sein du collectif ; on *voit* bien – sans trop creuser – que le support d'action de bon nombre d'administrations publiques sont des bases de données et que pour comprendre l'Etat d'aujourd'hui, il faut également comprendre la constitution des bases de données sur lesquelles il repose ; on *voit bien* que les sciences naturelles, expérimentales et sociales sont touchées de plein fouet par ce phénomène et qu'il suggère de nouvelles formes d'activité ; on *sent* bien que les bases de données sont chargées de vertus, qu'elles portent en

elles des espérances plus ou moins louables de rationalité, d'équité, de savoir ou encore de coordination ; on *sent bien* qu'elles ne sont pas neutres et que la mise en données suppose une certaine forme de violence ; on *voit bien* aussi que la mise en base de données peut être lourde de conséquence et qu'un individu peut voir des portes s'ouvrir ou se fermer en fonction de l'endroit où son nom est répertorié ; on le *voit*, on le *sent* ; il suffit de suivre le fil de nos expériences. »

Et pourtant – assez curieusement –, malgré l'ampleur du phénomène, les études sociologiques sur les bases de données sont rares, éparses et n'appartiennent à aucun courant unifié. Bien que les bases de données fassent à coup sûr émerger des enjeux nouveaux, elles ne retiennent que très peu l'attention des sociologues. Le paradoxe est ainsi saisissant entre l'ampleur admise du phénomène *sociétal* et la maigreur du phénomène *sociologique*.

Deux directions de recherche

Malgré la maigreur du champ d'étude, il existe tout de même un certain nombre de travaux qui laissent entrevoir deux directions de recherche. La première s'attache à traiter des bases de données en termes d' « effets » et tend ainsi à les considérer *en aval* de leur processus de conception. La deuxième direction de recherche – amorcée surtout par les études sociales sur les sciences – porte davantage sur l'*amont* des bases de données et les dynamiques collectives qui rentrent dans leur façonnage.

Regards sociologiques « en aval » des bases de données

Il y tout d'abord une **lecture politique des bases de données**, passablement inspirée des travaux de Michel Foucault sur la surveillance et l'assujettissement des individus. Cette lecture vise à dénoncer un scandale ; à lutter contre un certain pouvoir d'assujettissement et de contrôle des bases de données sur les individus (Lyon, 2003). En tentant de dénoncer l'asymétrie entre ceux qui conçoivent les bases de données (entreprises, pouvoirs publics) et ceux qui en sont les sujets-victimes, cette lecture cherche aussi à révéler les diverses façons dont les bases de données obligent certains individus à rentrer dans des catégories rigides aux attributs contraignants (Porter, 1995).

Pour autant, les sujets ne sont pas toujours considérés comme passifs et formatés par l'instrument de contrôle. Certains travaux (notamment Flichy, 2013) rendent compte des diverses façons dont les acteurs parviennent à jouer avec les bases de données et leur pouvoir contraignant.

Ce regard sociologique en aval des bases de données encapsule également une **lecture symbolique et culturelle** (Manovich, 1999 ; 2012). Ici, les bases de données sont considérées comme une nouvelle forme médiatique qui reflèterait la société d'aujourd'hui. L'interprétation globale du phénomène y est moins sombre que les précédentes et met davantage l'emphase sur les nouvelles possibilités d'action que les bases de données seraient à même de proposer.

Dans leur versant critique comme dans leur versant apologique, ces travaux en termes d'*effets sur* la société comportent tous un défaut non négligeable : ils ne nous disent rien sur ce qui produit ces effets. Que ceux-ci soient délétères ou appréciables ; qu'il faille les changer ou les préserver, ces études restent limitées dans leurs propositions pratiques de composition. Dit autrement, si elles participent certainement à une meilleure compréhension de ce qui se *passe*, elles ne peuvent faire des propositions sérieuses quant à ce qui serait appréciable (ou non) de *faire*.

Regards sociologiques « en amont » des bases de données

Il existe une série d'études historiques (Yates, 1993, 2005 ; Gardey, 2008) et sociologiques (Segrestin, 2004 ; Vinck et Penz, 2008) qui s'attachent à rendre compte des **dynamiques collectives de co-construction des bases de données**. Pour ce qui est du domaine industriel, Vinck et Penz montrent par exemple que si les bases de données tendent à être imaginées en fonction des nouvelles tendances managériales, ces mêmes tendances managériales se transforment dès qu'elles entrent en contact avec ces bases de données en-train-d'être-faites et les nouvelles opportunités qu'elles suggèrent. Difficile donc d'induire des relations entre des causes et des effets : les effets produits par l'apparition d'une base de données ne semblent pas réductibles à cette seule base de données ; il faut également prendre en compte toutes les transformations opérées lors de son façonnage collectif.

Une deuxième série d'études « en amont » concerne **l'exploration des décalages entre buts de conception et usages effectifs**. Ugettho (2013) montre par exemple que, du fait notamment des mécanismes intriqués de co-construction des bases de données, des écarts se creusent facilement entre les fonctions imaginées et les fonctions effectives. A l'image des tendances managériales qui se transforment en contact avec les possibilités suggérées par la base de données, l'architecture et les fonctionnalités des bases de données elles-mêmes se transforment au fur et à mesure des discussions et des séances au point parfois de ne plus servir l'impulsion qu'avait engendré leur mise en construction. Dans la même veine – mais cette fois pour le domaine du biomédical – Dagiral et Peerbaye (2013) montrent que les bases de données conçues au départ pour partager des informations se heurtent au problème de l'interprétation de ces mêmes informations, ce qui engendre un certain nombre de luttes afin de fixer une interprétation « standardisée ».

Une troisième série d'études s'intéressent aux **débats qui sous-tendent la conception des bases de données**. En prenant le cas d'une base de données administratives aux USA, Porter (1995) montre qu'il faut parfois des années pour se mettre d'accord sur les conventions sociales qui vont servir de bases aux catégories d'une base de données administrative. C'est toute une dynamique sociale, politique et économique qui s'invite dans le jeu de la conception des catégories : des entreprises, des administrations concurrentes, des agences fédérales, des chercheurs et des informaticiens essaient tous d'infléchir l'orientation de la base de données selon leurs critères spécifiques, le résultat étant un accord composite constitué d'intérêts de différentes natures.

Une quatrième série d'études s'intéressent **aux efforts qui sous-tendent la mise en données**. En prenant le cas de cartographes amateurs qui tentent de répertorier les aménagements cyclables au sein d'une base de données partagées, Denis et Pontille (2013) montrent par exemple que la mise en données repose sur des choix de catégorie contraignants, sur des classifications équivoques et sur des compétences visuelles et descriptives assez précises.

Une quatrième série d'études s'intéressent finalement aux **rapports entre bases de données numériques et activité scientifique**. Si l'on suit Latour (1987), l'activité scientifique consiste

en un processus de production de traces et d'inscriptions ; lorsqu'un phénomène est supposé, il faut obtenir une trace de ce phénomène qui soit à même d'être comparé, compilé et synthétisé en vue de produire de la connaissance. Mais qui dit traces, comparaisons, compilations et synthétisations dit également stockage, échanges et standardisations ; on le voit sans trop de peine : les bases de données sont coextensives à l'activité scientifique et ce bien avant l'avènement des technologies digitales (Desrosières, 2008). La majorité des travaux traitent du développement des bases de données dans le domaine des sciences naturelles et expérimentales (Hilgarter, 1995, 2012 ; MacKenzie, 2003 ; Beaulieu, 2004 ; Bowker, 2000, 2005 ; Hine, 2006). Sur le cas de la génomique, Burno Strasser (2011) montre par exemple que cette science qui se présentait comme *expérimentale* voit aujourd'hui ses pratiques de recherche ressembler de plus en plus à celles des sciences naturalistes (plus anciennes) précisément du fait de son emphase récente sur la collecte, le stockage et le partage de données. En effet, de par notamment l'avènement des technologies digitales et la mise en place d'énormes bases de données (*Big Data*), les praticiens de la génomique collectent, classent, standardisent et partagent des données davantage qu'ils mettent en place des expérimentations particulières, propres à leur laboratoire. C'est toute une nouvelle façon de concevoir la génomique qui se développe dans le sillon des nouvelles possibilités offertes par les bases de données numériques et évidemment, cela ne va pas sans résistances, appropriations, ruses et conflits. A noter que ces études concernent presque exclusivement les sciences « dures » et très peu a été fait sur les rapports entre sciences humaines et sociales et bases de données (Vinck, 2013). Les travaux existant proposent davantage des réflexions sur de nouveaux outils d'analyse en sciences humaines (Currie, 2012 ; Heftberger, 2012 ; Kirschenbaum, 2007) mais restent fortement entachés – à raison ! – d'un côté promotionnel. L'idée est surtout de présenter un outil d'analyse innovant et de convaincre les collègues de sa pertinence pour leur domaine d'étude particulier.

Quelques enjeux de la collection de données numériques en sciences humaines

Vient maintenant le tour de Nicolas Oberhauser qui nous présente quelques résultats provisoires de son mémoire de Master portant notamment sur les enjeux de la collection de données numériques en sciences humaines.

Problématique du partage des données numériques

Il est une chose qui change avec l'avènement du numérique : les données deviennent aisément *partageables*. Imaginons en effet un historien qui disposerait d'une collection d'extraits textuels rassemblant les écrits francophones du XVIIIe siècle traitant de la morale sensitive chez Rousseau. Il disposerait d'un index thématique très poussé mais ses données seraient confinées dans des classeurs et fonctionneraient à partir de fiches de carton, de post-it, d'empilements fragiles etc. Dans cette situation, notre historien devrait produire des efforts très concrets pour partager ses données avec un collègue et comme il aurait déjà passé certainement beaucoup de temps à produire ces données *selon cet ordonnancement-là*, il verrait certainement mal l'utilité de ce travail supplémentaire. Avec les données numériques, la numérisation elle-même va imposer une certaine organisation, une certaine mise en forme des données d'une façon telle qu'il sera très facile de les dupliquer. Les efforts concrets de partage, tout comme le risque de voir son classement modifié, disparaissent ; les données sont *en elles-mêmes* partageables et le coût de leur circulation tombe.

Mais ces possibilités de partage des données numériques n'interviennent pas seulement *en aval* de la production de données : elles interviennent aussi *en amont*. Cette propriété spécifique au

numérique suggère aux chercheurs de produire des données qui ne concernent plus seulement leur recherche en cours mais bien également une *autre*, plus large et plus collaborative. De cette tendance découlent deux ensembles d'enjeux, le premier relatif à la propriété des données : à qui appartiennent les données ? qui est en droit, ou non, de les travailler ? Un deuxième ensemble d'enjeux concerne cette fois la production effective de ces données : qui se charge *effectivement* de produire ces données ? Ne sont-ce pas souvent des doctorants ? Si oui, ont-ils encore le temps de s'investir dans une trajectoire intellectuelle qui leur serait propre ?

Problématique de la collectivisation du travail scientifique

La collectivisation du travail scientifique suggérée par les possibilités de partage des données numériques introduit de nouveaux enjeux d'organisation. Le premier est lié à ce que Bowker (1996 ; Bowker et al. 2010) nomme *bootstrapping problem* ou « problème de fonctionnement en circuit fermé ». Selon Bowker, pour que les données soient considérées comme adéquates, il faut les croiser avec des données produites par d'autres chercheurs à propos des mêmes objets. A partir de là, la collectivisation du travail scientifique va impliquer la constitution d'*accords* visant à définir les objets capables de se placer à l'intersection de préoccupations diverses. Reste à savoir sur quels *critères* sont établis les accords à propos de ces « objets-frontière » : quels sont les pouvoirs de ralliement de ces objets-frontières et qui est impliqué dans leurs définitions ?

Un deuxième enjeu lié à la collectivisation du travail scientifique concerne la standardisation des données et plus précisément les règles de formatage des données. Jusqu'à quel point ces règles sont-elles univoques ? Comment sont-elles produites ? Comment sont-elles mises en acte par les chercheurs concernés ?

Problématique des données produites

Une fois produites – c'est-à-dire mises en collection –, les données sont généralement soumises à des épreuves *épistémologiques* qui testent la robustesse du lien entre les données et les objets qu'elles ciblent. Mais de manière tout aussi importante, ces mêmes données sont également soumises à des épreuves *techniques* qui testent leur capacité à s'adapter aux besoins pratiques (et changeants) des chercheurs. Robustesse des données et souplesse de la collection : c'est bien l'articulation subtile entre ces deux attributs qui semble définir une *bonne* collection de données numériques en sciences humaines. A partir de là, une série de questions se pose : ces deux attributs parviennent-ils à être considérés symétriquement lors des processus de production de collection de données numériques ? Comment mettre en acte ce besoin d'articulation entre robustesse et souplesse au sein d'un univers collectif et passablement distribué ? Ces deux exigences induisent-elles un nouveau mode de collaboration avec le génie informatique ? Si oui, lequel ?

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

Bibliographie indicative

- Agar Jon (2003). *The gouvernment machine. A revolutionary history of the computer*. Cambridge, MIT Press.
- Atten Michel (2013). Ce que les bases de données font à la vie privée. L'émergence d'un problème public dans l'Amérique des années 1960. *Réseaux*, (178-179), 23-53.
- Beaulieu, Anne. 2004. From brainbank to database: the informational turn in the study of the brain, *Studies in History and Philosophy of Science* Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 35 (2), 367-390.
- Bowker Geoffrey (2000). Biodiversity Datadiversity, *Social Studies of Science*, 30 (5), 643-683.
- Bowker Geoffrey (2005). *Memory Practices in the Sciences*. Cambridge, MA : MIT Press.
- Bowker Geoffrey, Baker Karen, Millerand Florence, Ribes David (2010). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment, in J. Hunsinger et al. (eds.), *International Handbook of Internet Research*, 97-117
- Currie Morgan (2012). The Feminist Critique: Mapping Controversy in Wikipedia, in Berry David (Ed.), *Understanding Digital Humanities,* Houndmills: Palgrave Macmillan, 224-248.
- Dagiral Eric, Peerbaye Ashveen (2013). Voir pour savoir. Concevoir et partager des « vues » à travers une base de données biomédicales, *Réseaux*, (178-179), 162-196.
- Denis Jérôme, Pontille David (2013). Une infrastructure élusive. Aménagements cyclables et troubles de la description dans OpenStreetMap, *Réseaux*, (178-179), 91-125.
- Desrosières Alain (2008). *Pour une sociologie historique de la quantification*. Paris, La Découverte.
- Flichy Patrice (2013). Rendre visible l'information. Une analyse sociotechnique du traitement des données. *Réseaux*, (178-179), 55-89.
- Flichy Patrice, Parasie Sylvain (2013). Sociologie des bases de données : présentation, *Réseaux*, (178-179), 9-19.
- Gardey Delphine. (2008). *Écrire, calculer, classer. Comment une révolution de papier a transformé les sociétés contemporaines (1800-1940).* Paris, La Découverte.
- Heftberger Adelheid (2012) Do Computers Dream of Cinema? Film Data for Computer Analysis and Visualisation, in Berry David (Ed.), *Understanding Digital Humanities,* Houndmills: Palgrave Macmillan, 210-223.
- Hilgartner Stephen (1995). Biomolecular Databases: New Communication Regimes for Biology?, *Science Communication*, 17 (2), 240-263.
- Hilgartner, Stephen. 2012. Selective flows of knowledge in technoscientific interaction: information control in genome research, *The British Journal for the History of Science*, 45 (2), 267-280.
- Hine Christine (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social studies of science*, 36 (2), 269-298.
- Kirschenbaum Matthew G. (2007). *The remaking of reading: Data mining and the digital humanities*. The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD.
- Latour Bruno (1987). *Science in action*. Harvard : Harvard University Press.
- Lyon David (ed.) (2003). *Surveillance as social sorting: privacy, risk, and digital discrimination*, London, Routledge.
- MacKenzie Adrian(2003). Bringing sequences to life: how bioinformatics corporealizes

sequence date, *New Genetics and Society*, 22 (3), 315-332.
- Manovich Lev (1999). Database as symbolic form, *Convergence*, 5 (2), 80-99.
- Manovich Lev (2012). How to Compare One Million Images?, in Berry David (Ed.), *Understanding Digital Humanities,* Houndmills: Palgrave Macmillan, 249-278.
- Porter Theodore (1995). *Trust in numbers. The pursuit of objectivity in science and public life*. Princeton, Princeton University Press.
- Segrestin Denis (2004). *Les chantiers du manager*. Paris, Armand Colin.
- Strasser Bruno (2011). The Experimenter's Museum: GenBank and the moral economies of biomedecine, *ISIS*, 102 (1), 60-96.
- Ughetto Pascal (2013). Utiliser une base de données en organisation. La recherche de l'instrument, *Réseaux*, (178-179), 197-222.
- Vinck Dominique, Natale Enrico (2014). La transformation des sciences historiques. La part du numérique. In Leresche Jean Philippe, *Transformations des disciplines académiques : entre innovation et résistance*, Paris, Editions des Archives Contemporaines.
- Vinck Dominique, Penz Bernard (éds.) (2008). *L'équipement de l'organisation industrielle. Les ERP à l'usage*. Paris, Hermes.
- Vinck Dominique (2011), Taking intermediary objects and equipping work into account in the study of engineering practices, *Engineering Studies*, 3 (1), 25-44.
- Vinck Dominique (2013), Pour une réflexion sur les infrastructures de recherche en sciences sociales, *Revue d'anthropologie des connaissances*, 7 (4), 993-1001.
- Yates Jo-Anne (1993) *Control through Communication: The Rise of System in American Management*, The Johns Hopkins University Press.
- Yates Jo-Anne (2005). *Structuring the Information Age: Life Insurance and Technology in the Twentieth Century*, The Johns Hopkins University Press.

Florian Jaton

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL