De l'Open Data à l'information, quelques points d'appui depuis les SHS

Françoise Paquienséguy

▶ **To cite this version:**

Françoise Paquienséguy. De l'Open Data à l'information, quelques points d'appui depuis les SHS. Données Urbaines et smart city, 2020. hal-03163034

HAL Id: hal-03163034
https://hal.science/hal-03163034

Submitted on 12 Mar 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.

Conclusion
De l'Open Data à l'information, quelques points d'appui depuis les SHS

Françoise Paquienséguy
Sciences Po Lyon / Elico EA4147

Résumé :
Cette conclusion met en tension deux termes forts de notre approche comme de l'actualité des données ouvertes : Open Data et information, qui pourraient aussi se formuler différemment données brutes et valeur ajoutée. Nous le savons, la réutilisation des données ouvertes se niche, ou se développe, à l'intersection de ces deux termes qui nomment des processus longs et de faible transparence de transformation et d'éditorialisation.

Mots-clés : data – information – éditorialisation – data science – SHS

Le thème de l'Open Data est fécond et sensible de bien des façons car il entraîne réflexions et travaux sur le partage, les licences, sur la *privacy* et la protection, sur les modèles économiques et les acteurs qui les portent, sur l'idéologie et les concrétisations de la ville intelligente, ou encore sur le web sémantique. L'ouvrage qui s'achève ici a tenté d'en explorer certaines à la fois à partir du terrain qui a été le nôtre et de notre discipline car ils restituent les apports et les problématiques de chercheuses en Sciences de l'Information et de la Communication, à l'œuvre aux côtés d'informaticiens, de codeurs et de développeurs. C'est pourquoi cette conclusion met en tension deux termes forts de notre approche comme de l'actualité des données ouvertes : Open Data et information, qui pourraient aussi se formuler différemment données brutes et valeur ajoutée. Nous le savons, la réutilisation des données ouvertes se niche, ou se développe, à l'intersection de ces deux termes qui nomment des processus longs et de faible transparence de transformation et d'éditorialisation.

A ce jour, comme l'a montré le chapitre 2, les principaux réutilisateurs des données ouvertes sont des développeurs, des chercheurs des data journalistes ou *data scientists*. Autrement dit, mis à part les premiers qui généralement les exploitent pour en produire d'autres, les autres ont pour mission de les traiter de façon à produire de l'information, ce qui signifie créer de la valeur ajoutée en transformant les données, ici ouvertes, en information.

Cette transformation est au cœur du partage et de la consommation des données. La pratique est plus que fréquente et massive puisque bien des applications transports/mobilité utilisées quotidiennement s'appuient sur des données ouvertes par les métropoles ou par ces applications elles-mêmes (Uber, Air BnB, Vinci par exemple). L'individu hyperconnecté régule et ajuste ses actions quotidiennes sur la base d'informations qui sont de plus en plus calculées, qu'il s'agisse d'horaires en temps réel, de conseils d'achats ou de fils d'actualité. Ces informations résultent d'un traitement (généralement automatisé) des données (souvent issues de capteurs) et s'organisent fréquemment de façon prédictive afin de faciliter notre quotidien. Les données ouvertes tout particulièrement dans le contexte des métropoles, sont emblématiques de ce type d'informations pour la mobilité, l'énergie ou les services aux usagers. C'est la logique d'accès à l'information qui prévaut et les efforts des métropoles la valorisent en soutenant partage et réutilisations afin d'engendrer des applications et services aux résidants.

Ainsi les citadins, les travailleurs pendulaires et autres urbains mobiles, sont-ils accompagnés, ou cernés, au quotidien, par des données plus ou moins traitées, dans une logique d'accès dominante et permanente qui soulève trois questions fondamentales pour les chercheurs de notre discipline, les sciences de l'information et de la communication :

1. Comment l'évolution des pratiques des professionnels de la donnée, que nous avons étudiée dans l'ANR *OpenSensingCity* et vécue dans l'équipe pluridisciplinaire qui l'a portée, comme dans nos relations actives avec les fonctionnaires territoriaux de certaines métropoles, pèse-t-elle sur la transformation de ces données en information ?
2. Comment, en tant que chercheurs en sciences humaines et sociales pouvons-nous prendre en compte la nature computationnelle de l'information, au-delà de ses sources ?
3. Comment associer ou dissocier les data, leur transformation en information, puis leur diffusion *via* leur éditorialisation ?

C'est grâce aux résultats de l'ANR *OpenSensingCity* achevée en septembre 2018 et aux travaux récemment conduits sur les Humanités Numériques[1] que la réflexion a mûri sur la base de :

- deux enquêtes de terrain auprès des professionnels de la donnée (producteurs, développeurs, *data scientists* et data journalistes) ;
- deux expérimentations proches de la démarche de création / conception des hackathons, les données recueillies et nos analyses sont allées bien au-delà de la problématique de l'ANR, tout particulièrement liée à la mobilité urbaine ;
- d'un travail de fond, épistémologique sur les notions fondamentales des sciences de l'information et de la communication qui nous aide à penser l'usage des data et à penser leur place dans les pratiques professionnelles ;
- d'un chantier transversal sur l'hétérogénéité des données conduit dans le cadre de la Chaire Industrielle du Labex IMU[2] et de l'UDL[3] intitulée « DataServices pour une ville durable ».

1. Comment l'évolution des pratiques des professionnels de la donnée pèse-t-elle sur la transformation de ces données en information ?

Toujours convaincue que pour comprendre le présent il faut en maîtriser les conditions de réalisation, nous souhaitons insister sur quelques points.

Toutes les données ont un passé !
Le paradoxe ici est très lourd à porter chez les professionnels de la data et il faut le prendre en compte car ce sont leurs données qui nourrissent nos applications et soutiennent de plus en plus nos décisions. Y compris dans nos propres pratiques professionnelles directement à des fins de recherche lorsqu'elles constituent nos corpus, ou pour nous aider dans notre travail d'analyse. Ce paradoxe est simple à saisir à partir du postulat que la production de la donnée brute n'existe pas. Les données portent toutes la trace de leur processus de génération, de ce pour quoi elles ont été récoltées et de la façon dont elles l'ont été. Leur hétérogénéité complexifie la donne, nous le savons bien. Le paradoxe se formule donc ainsi à propos du passé des données : pour traiter des données, leur contexte de production et d'ouverture doit être connu, mais le professionnel de la data souhaite parfois s'en débarrasser pour revenir à des données brutes qui seraient alors autrement traitées.
C'est une double injonction qui découle de ce paradoxe puisque d'une part l'Open Data se nourrit des univers du libre, du libre accès et de l'ouverture (Ibekwé, Paquienséguy : 2015) ce qui signifie la diffusion de données déjà produites et donc partagées, en l'état ! Donc il faudrait

[1] Paquienséguy F., Pélissier N., (dir) *Questionner les Humanités Numériques : positions et propositions des Sciences de l'Information et de la Communication*, SFSIC/CPDirSic, 2020
[2] Intelligences des mondes urbains
[3] Université de Lyon

de facto renoncer à l'idée de données brutes. Mais, d'autre part, pour les données mise en partage circulent et que l'utilisation suive la diffusion, il faudrait produire des données « brutifiées ». Autrement dit, selon Denis et Goëta mettre en œuvre un processus de « brutification » des données déjà produites (Denis, Goëta : 2016) avant leur ouverture.

We want raw data !

Pour mieux comprendre cette injonction c'est ici aux SHS qu'il faut revenir, et plus particulièrement à la différence que fait Lévi-Strauss entre le cru et le cuit, titre d'un de ses ouvrages, paru en 1964. Il cherche à « montrer comment des catégories empiriques telles que celles de cru et de cuit, de frais et de pourri, de mouillé et de brûlé, etc. (...) peuvent néanmoins servir d'outils conceptuels pour dégager des notions abstraites et les enchaîner en propositions » (Lévi-Strauss : 1964, 8). L'analyse de Lévi-Strauss « se situe au niveau du signe qui transcende l'opposition du sensible et de l'intelligible » (*idem*, 22) et vise à mettre en évidence un emboîtement de codes. Or un code ne sert-il pas justement à traduire ? En fait, le code lui-même n'est pas utilisé dans son usage commun qui est de permettre le passage une fois pour toutes des termes d'un signifiant à ceux d'un signifié auquel on s'arrête (Isambert : 1965, 392-394). Autrement dit, les formats des données, les métadonnées, les fréquences de collecte, de mise à jour ou de production, la longitudinalité des données ou encore les jeux qui les associent et les thématiques qui les structurent doivent se percevoir comme des catégories empiriques car il n'y a à ce jour aucune formalisation officielle ou homogène au-delà des thématiques listées par la Directive Inspire de 2007. Quant au format premier tel que pensé par Chignard ou Goëta pour désigner les données nées ouvertes, il n'existe pas encore. Ces catégories empiriques, dont le sens pèse lourd, dégagent des notions abstraites comme par exemple la qualité de vie, le bien-être urbain, mais aussi de communs comme l'a développé le chapitre 4. De même la catégorie tout aussi impalpable des « données brutes » à partir desquelles, effectivement des propositions s'enchaînent sous forme algorithmique, calculée, prédictive.

A l'évidence, aucune *raw data* ne peut être asociale, c'est dire sa force et son ancrage. Elle dépend toujours d'un environnement sociotechnique qui la génère, la crée ou l'instaure en montrant tout le côté illusoire des données brutes, puisque les données ouvertes se catégorisent dès leur production, catégories qui pèsent sur l'analyse à venir, quelle qu'elle soit.

Selon J. Denis et S. Goëta (2016, 26), il faudrait alors « appréhender le travail d'ouverture des données comme un travail de 'brutification' ». Sinon l'ouverture est l'opération invisible produite par le code qui ouvre le passage du signifiant au signifié et renforce l'évidence : celles du poids des stratégies d'acteurs qui font les choix de l'ouverture (pourquoi ouvrir certains jeux et pas d'autres ?) et les choix des modalités de l'ouverture (sous quelle forme les ouvrir ? sous forme de jeu ou de visualisation ?). Pour marquer, signifier l'ouverture, résultant du travail du code, et la rendre visible, il faut intégrer le « déjà-là » des données.

Des données « déjà-là » : le substrat de l'Open Data

Les données ouvertes nous sont données. C'est ce « donné » que nous devons interroger car il reste difficile à identifier. Le donné s'impose à nous car il précède toute initiative, de compréhension, d'analyse, de traitement. Le donné, comme les données ouvertes, a ses spécificités et son propre sens, « puisque cela a été créé, produit, construit en une opération mentale [de traduction, de transformation] à laquelle on a accès » (Dewitte : 2001, 399).

Plusieurs types de chercheurs, spécialistes de la data ou philosophe en l'occurrence, proposent de matérialiser ce déjà-là à la fois porteur d'opérations automatisées et mentales. « Mais la question est alors de savoir ce que nous faisons de ce donné, ou selon quelles modalités existentielles, pratiques, théoriques nous nous rapportons à lui. Et à ce niveau très originaire, il peut s'agir tout aussi bien de choses naturelles ou humaines (sociales, culturelles) » (*Idem*, 395).

Bien visible, le lien est fort : les spécialistes de l'OD parlent de la nécessité de dégager un

format premier et le philosophe pense en termes de niveau « très originaire ». A l'origine de l'ouverture étaient les données ? Oui bien sûr ! et c'est justement ce déjà-là et la nature des données (pour tenter le parallèle rationaliste avec la nature et ses ressources, le monde, qui nous sont donnés). Déjà-là, les données sont porteuses d'une vision du monde et de sa restitution, déjà pensées et structurées pour en rendre compte d'une certaine façon, orientée et située dans une action d'ouverture qui nous échappe et qui relève d'une part de l'action publique et de l'autre des objectives politiques de métropoles. Ce déjà-là devrait nous conduire à la plus grande prudence et vigilance et nous pourrions, peut-être, mieux considérer ces manipulations et ces traitements pour reconsidérer les catégories empiriques proposées par les professionnels de la data. Autrement dit envisager l'origine de la production mais aussi de l'ouverture des données, sans « avaler tout rond » comme on pourrait le dire pour rester dans l'univers du cru et du cuit. Un parallèle pourrait se dessiner pour situer cette posture critique et la vigilance qui l'accompagne dans le petit univers des sciences de l'information et de la communication, de façon à mieux percevoir la construction du déjà-là, même relevant d'une déconstruction puisque la brutification est un retour à la source *a posteriori*. Les choix antérieurs dont les données ouvertes résultent (formats, jeux, modalités d'accès) sont en fait l'équivalent des médias pour la communication, c'est par eux que nous accédons à l'information, une information qu'ils produisent à partir des données dont ils disposent et des choix éditoriaux qu'ils effectuent lesquels relèvent finalement d'une opération de tri. Le parallèle fonctionne entre format/jeux/data et médias/sources/information.

Pour conclure ce premier point, nous relevons donc que les sources, les formats, les modes de collecte et de restitution des données soulèvent trois questions dont la communauté pourrait se saisir : comment gérer le contexte associé à la donnée (en présence ou en absence) ? Comment prendre en compte le processus d'ouverture : brutification, modification, premier traitement de la donnée pour la rendre ouverte ? Comment prendre la mesure des choix techniques et politiques antérieurs qui formatent l'ouverture et le partage des données, car le lien est fort entre métropole et data ?
En fait, les SHS sont déjà grandement nécessaires et présentes avant même l'ouverture et elles seules peuvent identifier et analyser le contexte qui définira ensuite le déjà-là. Encore une fois, elles sont outillées pour penser ces cycles de transformation qui feraient passer du cru au cuit ! De la donnée brute à la valeur informationnelle, toutes deux substrats de la connaissance. Cette conclusion n'étant pas le lieu idéal pour une revue de littérature, elle s'en tiendra à un texte fondateur et fondamental, produit en 1988 par Madeleine Akrich, Michel Callon et Bruno Latour : *A quoi tient le succès des innovations ?* Ces trois auteurs y présentent le modèle tourbillonnaire, qui introduit deux concepts pour penser l'innovation : l'itération (autrement dit des choix successifs qui en visant le consensus formatent peu à peu l'objet final d'une part et son contexte de l'autre, lequel est modifié par ces choix) et l'intéressement (ne prennent finalement part au processus d'innovation que les acteurs qui y trouvent un intérêt, en notant que ces intérêts peuvent être de nature et de rythme différents). Ainsi en va-t-il des données ouvertes lissées au fil de l'eau entre le cadre législatif européen et national, les politiques publiques locales, les moyens et contraintes techniques, les ressources humaines disponibles et les partenariats économiques issus des métropoles. L'itération est manifeste puisque la plupart des métropoles proposent régulièrement de nouvelles versions de leur portail Open Data, celle de Lyon ouvrait en septembre 2019 sa troisième version, en mode bêta, Montpellier propose aujourd'hui sa quatrième version pendant que Rennes transforme complètement son offre avec la deuxième version, toujours en préparation, et que Dijon brûle les étapes avec une première version qui cherche à coiffer toutes les autres métropoles au poteau ! Quant à l'intéressement, le problème est transverse et préoccupe toutes les métropoles qui ont la compétence et donc la mission d'agréger les données de différentes catégories d'acteurs (régie des transports,

communes de la métropole, établissements publics, voirie, parkings, chantiers, etc.) dont les intérêts et buts ne convergent pas, sans parler de la catégorie des réutilisateurs dont l'intéressement reste complètement à motiver pour les intégrer dans les boucles itératives.

2. Comment, en tant que chercheurs en sciences humaines et sociales, pouvons-nous prendre en compte la nature computationnelle de l'information ?

La notion de computation paraît fondamentale pour comprendre la transformation de la data en information qui pour certaines données se joue justement à l'ouverture, dans le processus d'ouverture. Bien avant les technologies que nous côtoyons, Edgar Morin se penchait déjà sur le lien entre la computation et l'information :

> « Ainsi, en vertu des principes/règles qui la gouvernent, en fonction des modes association/séparation qu'elle combine, la computation effectue ce qu'indique bien l'origine latine computare : supputer ensemble, com-parer, con-fronter, com-prendre. La computation ne peut donc se limiter au calcul numérique. De même, la computation ne peut se réduire à l'information. L'information ne devient une information que par rapport à une computation, et n'est, sinon, qu'une marque ou une trace. » (Morin : 1986, 38)

Autrement dit, la masse d'informations collectées et calculées constitue une base statistique, quantitative qui sert de ressource à leur computation (*Idem*). C'est celle-ci, centralement, qui permet de lire des tendances ou récurrences à l'œuvre sur la plateforme en procédant *« à des opérations d'association (conjonction, inclusion, identification) et de séparation (disjonction, opposition, exclusion) »* (*Ibid*., 38) qui classent, caractérisent et prescrivent en fonction des profils des usagers sur la plateforme. Ce recueil systématique d'information est couplé à une algorithmie omniprésente depuis la captation jusqu'à la sélection des données et en passant par leur traitement (Paquienséguy : 2017). Pour Anne-Marie Dujarier, il naît du « travail cognitif taylorisé » (Dujarier : 2016) effectué par les membres de la plateforme et engendre des « échanges non-marchands à but lucratif » (*Idem*). Le lien se révèle donc fort et indéfectible entre algorithmie, computation et monétarisation ; les données ouvertes n'en sont pas exclues. Plusieurs atouts peuvent nous aider à saisir cette transformation et ses acteurs dont nous ne présentons ci-après que les plus forts.

Quelques atouts

En effet, nous devons dans doute rappeler ici l'importance et le positionnement de la science de l'information, discipline qui nous échappe quelque peu en France en tant que telle et que le raccourci d'information-documentation ne saurait remplacer. La science de l'information se définit comme « un champ disciplinaire ayant pour objet scientifique l'information ; lequel est principalement concerné par l'analyse, la collecte, la classification, la manipulation, le stockage, la récupération, la circulation, la diffusion et la protection de l'information. » (Stock & Stock : 2013, 26). Trois paradigmes la structure : la production, le traitement et l'usage de l'information. La computation se situe donc dans le processus de production, voire dans certains cas avancés des machine learning ou de l'intelligence artificielle, dans celui du traitement ce qui rend les SHS plus que légitimes à son étude et positionne les sciences de l'Information et de la Communication en première ligne. C'est ici sans doute aussi au web sémantique, ses métadonnées et ses ontologies qu'il faut évoquer (et qui était présent dans la problématique de l'ANR Opensensingcity) car il est pensé dès les années 2000 comme le web.3 qui permettra à la machine comme à l'humain de trouver et partager des données. Son format, le RDF a donc

pour objectif de transformer toutes les documents HTML en données exploitables ; ici c'est à la fois la machine et l'humain qui en sont destinataires et sources de computation pour passer des données au sens[4].

Il convient ensuite de convoquer la statistique, dans toute la puissance du quantitatif et du traitement du chiffre qui en font la force, elle sait agréger, trier et traiter des données pour fournir différents types d'informations. Nous la retrouvons d'ailleurs, bien qu'elle ne soit plus guère enseignée dans nos cursus, au cœur de plusieurs logiciels que nous utilisons pourtant, parfois sans bien comprendre comment ils calculent et quelle computation ils opèrent pour restituer de l'information, comme le logiciel industriel *N'Vivo* qui promet « des méthodes de recherches qualitatives et combinées. Il est conçu pour […] organiser, analyser et trouver du contenu perspicace parmi des données non structurées ou qualitatives telles que des interviews, des réponses libres obtenues dans le cadre d'un sondage, des articles, des médias sociaux et des pages Web ». De même, Iramuteq ou les logiciels R, issus de la communauté scientifique qui permettent des analyses multidimensionnelles de textes ou de questionnaires dont les ressorts statistiques doivent être maîtrisés. Le lien doit être pensé entre statistique et Open Data, sans oublier le domaine très particulier de la statistique publique.
Une des modalités familières de transformation de la data en information repose sur la puissance de la data visualisation qui présente les résultats du traitement des données, le plus souvent uniquement statistique, sous une forme schématique ; dont les graphiques d'Excel seraient l'archétype en passant automatiquement d'un tableau de chiffres à un abaque par exemple. Simplificatrice, parfois fascinante, la data visualisation est considérée aujourd'hui comme un outil de communication (*story telling* des données) en s'appuyant sur le principe qu'une image, qu'une visualisation sera plus simple à interpréter par le cerveau humain qu'un tableau Excel, format assez fréquent dans les jeux de données ouvertes. La data visualisation opère donc pour partie la transformation en mettant la statistique en image, en visuel, si possible coloré. Comme le dit un professionnel[5] expert en visualisation de données *corporate* : « la data visualisation, c'est l'art de raconter des chiffres de manière créative et ludique, là où les tableaux Excel échouent. C'est en quelque sorte mettre en musique l'information chiffrée ». Justement, les dangers de la data visualisation sont liés au quantitatif qui nous dépasse et nous oblige à recourir à la machine pour traiter nos données et nous en fournir une synthèse, visuelle. Dans leur ouvrage *Information Visualization - Human-Centered Issues,* Daniel Keim *et alii* (2008) soulignent les trois principaux : produire à partir des données des visualisations résultant de traitement inapproprié (à l'objectif de la recherche comme au contexte d'analyse), sans rapport signifiant avec la tâche en cours (autrement dit sans opération de tri pensée en fonction de l'objectif et du contexte) et donc finalement proposer des types de visualisation inappropriés (aux résultats comme aux données qui les engendrent). La situation sans aucun doute dans les outils de communication du chiffre et de la statistique, les dangers, les limites de la data visualisation se dessinent clairement et nous font préférer l'analyse visuelle qui cherche justement à porter trace du traitement des données.
L'analyse visuelle, ou *visual analytics*, consiste à transformer les données quantitatives en représentation visuelle, comme le fait la data visualisation, mais avec comme premier objectif de rendre visible le traitement subi par les données pour produire de l'information au-delà de la visualisation des résultats qu'est la data visualisation, quand elle ne produit pas, en plus, de distorsion de la perception des résultats nous l'avons dit. Dans « *Illuminating the Path* » (2005),

[4] Bachimont Bruno, Gandon Fabien, Poupeau Gautier *et al.*, « Enjeux et technologies : des données au sens », *Documentaliste-Sciences de l'Information*, 2011/4 (Vol. 48), p. 24-41.
[5] Interview de Charles Miglietti promoteur du « *data story telling* »
https://toucantoco.com/blog/en/author/charles-miglietti/ consulté le 2 décembre 2019

James J. Thomas et Kristen A. Cook définissent l'analyse visuelle comme la science du raisonnement analytique facilitée par des interfaces visuelles interactives.

Fig.2 Intégration étroite des méthodes d'analyse des données visuelles et automatiques avec les bases de données pour un support interactif évolutif – Daniel Keim *et alii*

Comme le schéma de Keim (2008, 156) le montre, le processus est itératif, comme nous l'avons déjà vu avec les boucles du modèle tourbillonnaire d'Akrich et le retour se fait entre le modèle qui génère la data visualisation et les données à partir desquelles il est produit afin d'être le plus adapté et spécifique possible, sur la base de différents niveaux d'itération afin d'en améliorer la pertinence, de sorte que la visualisation produise des connaissances. L'analyse visuelle cherche donc à dépasser l'isomorphisme des visualisations et des datas qui les soutiennent.

Science des données et *data scientists*

La richesse et la complexité de cette transformation de la data en information dont nous venons d'évoquer quelques éléments constitutifs correspondent aujourd'hui à des compétences professionnelles qui s'expriment dans l'expression « *data scientist* » qui nous est sans doute plus familière que son inversion, instituant une science des données portée à l'analyse des données et à la production d'information *via* des processus de computation. En France, la science des données s'inscrit dans les efforts d'accompagnement du numérique, en lien avec la mission Etalab et la création des postes de *Chief Data Officer* (administrateur général des données) dans la fonction publique territoriale et le recrutement de *data scientists* pour « accélérer la possibilité de politiques publiques « augmentées » par les données et leur analyse. C'est à ce point précis que se croisent donc : les données ouvertes, la statistique publique et l'idéologie première de l'ouverture des données traitée dans le premier chapitre de cet ouvrage : la transparence de la gouvernance[6] sous couvert de science des données. Le site d'Etalab[7] rappelle en effet que

> « Faisant office de *Chief Data Officer* de l'État, Etalab bâtit et met à disposition de tous une infrastructure des données publiques, socle des services numériques publics ou privés, et porte une véritable politique et gouvernance de la donnée, reposant sur plusieurs volets :
> - l'ouverture : rendre accessibles à tous les données publiques qui peuvent l'être (« Open Data »),
> - le partage et circulation des données entre administrations, socle notamment du "Dites-le nous une fois" et de la simplification des démarches administratives notamment,
> - l'exploitation des données par les datasciences et l'intelligence artificielle afin de mieux concevoir, piloter et évaluer les politiques publiques, d'améliorer le service public, mieux cibler des contrôles, etc.»

Il instaure ainsi comme centrale au fonctionnement de l'état, la science des données qui s'appuie sur des outils mathématiques, de statistiques, d'informatique et de visualisation des données pour produire de l'information computationnelle comme aide à la décision publique. Une transformation de l'action publique sans doute à réfléchir en complément de celle des métropoles sous le poids de l'ouverture et des données territoriales (Courmont : 2019).

---

[6] Et le lien ne peut pas ne pas être alors fait avec une gouvernance algorithmique. Rouvroy Antoinette, Berns Thomas, « Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ? », *Réseaux*, 2013/1 (n° 177), p. 163-196.

[7] https://www.etalab.gouv.fr/politique-de-la-donnee consulté le 2 décembre 2019

Nous sommes maintenant en mesure de formuler le deuxième paradoxe : l'exploitation, le traitement, l'analyse des données conduisent à plusieurs états depuis la donnée soi-disant brute à l'information éditorialisée, nous y reviendrons. Cet état, et la forme qui en résulte, dépendent du contexte dans lequel les données sont produites, traités, formalisées et éditorialisées pour produire de l'information, et celui de l'ouverture est particulier comme le rappelait Tim Berners Lee avec sa formule « data is the new link[8]: with the ever-expanding world of data-driven products, and the explosion of graphs […], the benefits of Open Data would only be realized by a positive attitude to sharing the data in an accessible way, without expecting too much in return[9] ».

De plus, dans ce processus fluctuant et difficilement perceptible, l'information peut être produite automatiquement par un traitement confié à une machine, comme elle peut résulter de l'exploitation de données par un humain aux compétences de *data scientist*. L'abondance des informations ainsi disponibles valorise et favorise ce dernier car aujourd'hui la valeur n'est plus dans la donnée, qui n'est plus rare, la rareté s'est déplacée, elle « réside à présent dans l'expertise et le savoir-faire de la mise à disposition de l'information » comme souligne le rapport sur la fiscalité de l'économie numérique (Collin et Colin : 2013). Elle s'appuie sur la computation prise comme un méta-calcul, comme la capacité à relier et traiter des data pour produire des informations, elles-mêmes reliées entre elles et éditorialisées. Ainsi la computation[10] serait au final le résultat d'un calcul selon des règles formelles (code et algorithmes) qui donne lieu à la fois à une sémantique et à de symboles (Varela *et alii* : 2006, 73) et produit une pensée calculante (Morin :1986). *« Une computation est une opération effectuée ou accomplie sur des symboles, c'est-à-dire sur des éléments qui représentent ce dont ils tiennent lieu. »* (*Idem*, 80). Du code et des algorithmes à l'information sociale, c'est là que le grand œuvre opère, dans ce procès de transformation transformant data/information computationnelle (prédictive, stratégique) en information médiatique (éditorialisée, appropriée), et si nos concepts et théories savent se saisir des deux termes le passage reste complexe à appréhender. Le lecteur sera alors questionné par le dernier paradoxe qui affecte les chercheurs en sciences sociales : doivent-ils développer leurs propres outils d'analyse de la transformation (dans la lignée des logiciels R avec par exemple Iramuteq ou Mediaswell) ? En ce sens, nous remarquons que de plus en plus de thèses sont menées sur la base d'une double compétence, et que les projets pluridisciplinaires de recherche allient fréquemment sciences sociales et informatique, comme dans le cas de la recherche ANR à l'origine de ce projet éditorial ; ou bien doivent-ils développer les compétences d'un *data scientist* pour « ouvrir la boîte noire » et comprendre les processus de production, de transformation et d'exploitation de la donnée ?

Fig 3 : les compétences de la data science

---

[8] https://techcrunch.com/2008/02/28/data-is-the-new-links-tim-berners-lee-says-sites-that-dont-give-users-their-data-back-are-boring/ consulté le 2 décembre 2019

[9] « *La data est le nouveau lien : avec un monde en constante expansion des produits axés sur les données, et l'explosion des graphiques […] les avantages des données ouvertes ne seraient réalisés que par une attitude positive à partager les données de manière accessible, sans trop attendre en retour…* » Traduction F. Paquienséguy

[10] Sans doute faut-il se souvenir ici des origines de la computation : « l'une des grandes contributions d'Alan M. Turing (1937) fut que cette propriété de décidabilité et calculabilité est équivalente à des procédures effectives de manipulation de symboles réalisées par un type particulier de machine physique (appelée plus tard « machine de Turing »). On dit alors que cette machine « computait » cette formule. » Meunier Jean-Guy, « Humanités numériques et modélisation scientifique », *Questions de communication*, 2017/1 (n° 31), p. 19-48.

Finalement, la computation (Paquienséguy : 2017) dans sa généralisation marquée par le terme, central, de *data scientist*, ouvre la porte à la manipulation des données comme source d'information mais aussi de valeur. Même si l'ensemble des compétences relève de la pluridisciplinarité, l'analyse et la conceptualisation des modèles économiques des industries qui s'y rapportent ont toujours fait partie l'approche des Sic et doivent continuer à y être étudiées, de même les processus de médiatisation de l'information, dont la première étape reste l'éditorialisation, phase conclusive de la transformation.

3. Comment considérer et positionner l'éditorialisation des données ?

Pour devenir une information telle que la comprennent les sciences de l'information et de la communication, pour quitter la data et le monde de la machine et de ses algorithmes et intégrer celui des médias et de du social, la data doit, entre autres, être éditorialisée. Autrement dit l'éditorialisation des données les transforment en information, quantitative ou qualitative.

Les travaux de Marcello Vitali-Rosati, collègue québécois, littéraire proche de la philosophie, en charge de la chaire sur les écritures numériques à l'Université de Montréal constituent la référence la plus proche de nos questionnements disciplinaires pour comprendre la nature de l'éditorialisation. Il en propose la définition restreinte, mais élargie aux contenus disponibles en connexion :

> « l'éditorialisation désigne l'ensemble des appareils techniques (le réseau, les serveurs, les plateformes, les CMS, les algorithmes des moteurs de recherche), des structures (l'hypertexte, le multimédia, les métadonnées) et des pratiques (l'annotation, les commentaires, les recommandations via les réseaux sociaux) permettant de produire et d'organiser un contenu sur le web. En d'autres termes, l'éditorialisation est une instance de mise en forme et de structuration d'un contenu dans un environnement numérique » (Vitali-Rosati : 2018).

Dans son développement et sa « propagation », elle permet aujourd'hui de rendre compte des recours aux outils et capacités du numérique pour élaborer et façonner des représentations ou des « visions du monde » (Vitali-Rosati : 2016). L'éditorialisation, comme concept et comme démarche, renvoie donc à un processus qui réussit à faire dialoguer potentialités du numérique, architectures conceptuelles et praxis documentaire, dans un but de co-construction de la connaissance.

C'est dans les quatre natures de l'éditorialisation, que Vitali-Rosati définit, que nous trouvons notre rôle à jouer en tant que chercheurs en sciences humaines et sociales spécialistes de l'information et de sa communication.

« Processuelle », l'éditorialisation se veut ouverte et continue dans le temps et dans l'espace.

À nous de prendre en compte la longitudinalité indispensable à l'analyse et de retrouver le temps long malgré celui qui s'accélère (Rosa : 2013) pour suivre les processus à l'œuvre. La structuration des portails, des administrations qui les proposent et de leur écosystème en montrent tout à fait l'absolue nécessité.

Impliquant plusieurs acteurs, elle est aussi « collective » et complique ainsi les tentatives d'identifier des actes d'éditorialisation perpétrés par un seul individu, car ces actes sont tous liés les uns aux autres.

A nous à travailler les stratégies d'acteurs en lien avec la notion de dispositif (Larroche : 2018) qui permet d'agréger plusieurs types d'action et de finalité ou de revenir aux fondamentaux de la théorie de l'acteur réseau sur la base de deux de ses concepts au moins : l'intéressement et l'itération.

Vitali-Rosati fait ensuite état de la « performativité » inhérente à l'idée d'éditorialisation. Puisque l'environnement numérique dépasse aujourd'hui le cadre du web, il affirme que l'éditorialisation « tend à agir sur le réel plutôt qu'elle ne le représente ».
A nous à questionner les représentations et les formes discursives (Collet : 2018) qui se dégage de la rhétorique numérique (Saemmer : 2015) et de comprendre les processus de conditionnalité du sens (Merzeau : 2016) ou comment le sens se crée, entre autres à partir de catégories et de symboles et d'opérations.
Finalement — du fait que l'environnement numérique implique désormais un web des objets (ou un web 3.0 sémantique) où « il n'est plus approprié de séparer le discours sur le réel du réel lui-même » — l'éditorialisation a aussi une nature « ontologique ».
À nous bien sûr de considérer la multiplicité des origines, des représentations, à chaque fois reconstruite et réinterprétée car Vitali-Rosati pense la nature ontologique de l'éditorialisation comme méta-ontologique, qui accepte cette multiplicité originaire, le caractère multi-essentiel du réel.
Et c'est bien toutes les problématiques à propos de l'identité numérique (d'un humain) qu'il faut investir également à propos des contenus, des informations et des datas, au-delà de leur source, afin d'en comprendre l'alchimie au creuset de la computation et des formes d'éditorialisation en lien avec les modèles économiques présents.

Les SHS ont connu la barrière du « déclaratif » qui laissait toujours planer le doute, mais elles ne peuvent pas plus faire confiance aux données. Par contre à l'évidence, le problème est différent ! À elles de s'adapter [11]?
OUI ! produire de la connaissance sur ces processus qui croisent des questions que nous connaissons bien en Sic : notre versant sciences de l'information se penche sur les données à travers les architectures structurées, les infrastructures numériques et/ou des standards ouverts, et notre versant communication sur les processus, les acteurs et les stratégies à l'œuvre.
Autrement dit, sans en faire un plaidoyer ni une revendication, les sciences de l'information et de la communication, qui incluent celles de la documentation, et le terme aujourd'hui retrouvé de bibliothéconomie, sont tout à fait outillées pour traiter la question de l'ouverture, du partage l'éditorialisation et de la valorisation des données dans toute sa complexité.
A nous donc de produire de la connaissance sur la production de la connaissance comme cet ouvrage a tenté de le faire.

[11] Rappelons ici la très intéressante hypothèse d'Olivier Le Deuff traitée dans son habilitation à diriger des recherches : les humanités digitales seraient la réponse des chercheurs SHS à la datafication du monde.

BIBLIOGRAPHIE

Akrich, M., Callon, M. et Latour, B. (1988). A quoi tient le succès des innovations ? *Gérer et Comprendre*. p. 4-29.

Boukacem-Zeghmouri, C. et Paquienséguy, F. (à paraître 2020). La genèse des humanités numériques en Sic. Dans F. Paquienséguy et N. Pélissier (dir). *Questionner les Humanités Numériques : Positions et propositions des SIC.* Paris, Sfsic/CPDirSic.

Collin, P. et Colin, N. (2013). Mission d'expertise sur la fiscalité de l'économie numérique. Repéré sur : http://www.economie.gouv.fr/files/rapport fiscalitedunumerique_2013.pdf.

Dewitte, J. (2001). Le déni du déjà-là. *Revue du MAUSS*, 1(17), p. 393-409.

Francony, J.-M. (2017). *Dispositifs info-communicationnels : des traces numériques d'usage aux données d'analyse* (Mémoire d'HDR). Université Grenoble Alpes.

Isambert F.-A. (1965). Lévi-Strauss Claude. Mythologiques. Le cru et le cuit, *Revue française de sociologie*, 6-3. p. 392-394.

Keim, D., Andrienko, G., Fekete J.-D., Görg, C., Kohlhammer J., *et al*. (2008). Visual Analytics: Definition, Process and Challenges. Dans A. Kerren, J. T. Stasko, J.-D. Fekete et C. North. *Information Visualization - Human-Centered Issues and Perspectives*. Springer, p.154-175, LNCS . Repéré sur : https://hal-lirmm.ccsd.cnrs.fr/lirmm-00272779

Larroche V. (2018). *Le dispositif, un concept pour les sciences de l'information et de la communication*. Londres : Iste éditions.

Le Deuff, O. (2017). Documentation et Humanités Digitales (Mémoire d'HDR). Université Bordeaux Montaigne.

Meunier, J.-G. (2017). Humanités numériques et modélisation scientifique. *Questions de communication* 31. doi : 10.4000/ questionsdecommunication.11040

Morin, E. (1986). *La Méthode, La Connaissance de la connaissance*. 3. Paris : Seuil.

Paquienséguy, F. (2017). L'usage prescrit renouvelé, ou l'injonction socio-algorithmique. *Études de Communication*, 49, 2017, p 13-32. Repéré sur : https://journals.openedition.org/edc/6989

Paquienséguy, F. et Pélissier N. (dir.) (à paraître 2020). *Questionner les humanités numériques : positions et propositions des Sciences de l'information et de la communication*. Paris : Sfsic/CPDirSic.

Rosa, H. (2013). *Accélération. Une critique sociale du temps, Paris : La découverte*

Saemmer, A. (2015). *Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques*. Lyon : Presses de l'enssib.

Stock, W. et Stock, M. (2013). *Handbook of Information Science*. Berlin, Boston : De Gruyter Saur.

Thomas, J.J. et Cook, K.A. (2005). *Illuminating the Path*. Los Alamitos : IEEE Computer Society Press.

Varela, F., Thompson, E. et Rosch, E. *(*2006). *L'inscription corporelle de l'esprit. Sciences cognitives et expérience humaine.* Paris : Seuil.

Vitali-Rosati, M. (2016). Qu'est-ce que l'éditorialisation ? *Sens Public*. Repéré sur : http://www.sens-public.org/article1184.html