Olivier Ricou

DONNÉES TRANSPARENCE & DÉMOCRATIE

Exploitons les données publiques ouvertes

Culture numérique

Site du livre : https://opendata.ricou.eu.org/

Version : 1.0.5
ISBN : 978-2-9581873-0-9
Dépôt légal : juin 2022

Avant-propos

Vous avez entendu parler des *Big data* [1], ce nouvel or noir du numérique qui permet des gains économiques fantastiques, que ce soit en améliorant les processus ou en découvrant de nouveaux marchés. L'idée est que des informations précieuses sont cachées au sein de la masse de données produite en continu par nos ordinateurs et nos capteurs. Ainsi les compteurs Linky fournissent des mesures instantanées et très précises au gestionnaire du réseau électrique qui peut ainsi anticiper les besoins et optimiser les flux sur son réseau. Netflix analyse l'activité de ses 150 millions d'abonnés pour proposer à chacun un produit personnalisé avec des suggestions pertinentes (probablement la clef de son succès). De tels exemples sont nombreux, aussi, on peut légitimement se demander s'il n'y a pas quelque chose à faire avec toutes les données qui nous entourent pour améliorer le fonctionnement de nos sociétés.

Parmi les données, celles qui nous intéressent en particulier sont les données publiques, à savoir celles que possède l'État, les collectivités territoriales, les différentes administrations, mais aussi les établissements à mission de service public. Ce sont justement ces données que la France a choisi d'ouvrir de plus en plus dans un effort de transparence, ainsi que pour des raisons économiques.

L'effort de transparence est une bonne chose et l'ouverture actuelle va dans le bon sens, mais restons vigilants. La transparence est inscrite dans le principe démocratique depuis la révolution de 1789 sans pour autant être pleinement appliquée aujourd'hui. Si des difficultés pratiques ont pu freiner la diffusion de données,

1. Données de taille colossale décrites par la règle des 3 V : volume, variété, vélocité. La vélocité soulignant la vitesse prodigieuse de flux de données qui augmentent quotidiennement le volume.

cette excuse n'a plus de sens avec Internet. « On nous cache des choses » semble une raison plus logique, qui renforce la défiance entre les citoyens et leurs dirigeants. Aussi le choix de la transparence démocratique est une voie intéressante pour réparer une relation abimée. Reste à savoir si l'ouverture des données publiques ira assez loin pour être efficace.

Aujourd'hui le portail `data.gouv.fr` propose des dizaines de milliers de jeux de données publiques de tout type sous licence libre (données ouvertes). Des entreprises, des associations, des universitaires et des particuliers ouvrent aussi leurs données au grand public. L'Union Européenne participe au mouvement comme de nombreux autres États. Ces nouvelles données renforcent la position des citoyens qui peuvent contrôler l'action des élus, mais aussi en tirer profit dans leur vie personnelle et professionnelle. Il ne s'agit pas seulement de lutter contre la corruption, ou autres manquements à la probité, mais aussi de simplifier les démarches administratives, de diffuser l'information, de lutter contre la désinformation, de permettre la création de produits commerciaux, voire d'entreprises. La suite logique serait notre participation active à la gouvernance du pays, les liens entre les données, la transparence et la démocratie étant étroits.

Ce livre montre en quoi l'ouverture des données peut aider à vivre dans un monde meilleur, comment des jeux de données permettent des usages et inversement, comment des usages génèrent des données. Si les exemples peuvent sembler impressionnants et réservés à une élite, il est important de savoir qu'il n'en est rien. Wikipédia a commencé avec un article. Tout le monde peut participer. Savoir que tout cela existe est déjà un premier pas vers une démocratie plus participative.

Le premier chapitre tente un panorama des avantages moraux, pédagogiques et économiques de la transparence, puis en indique les limites intrinsèques et celles imposées par la société dans le respect des autres droits. Le second commence par les aspects légaux de l'ouverture des données avec un point sur les licences de réutilisation. Il se poursuit par un tour d'horizon des données déjà disponibles en fonction de leurs sources. Enfin, il dessine ce que pourrait être une exploitation commune de ces données pour une

gouvernance plus directe, pour finir sur les communs. Le dernier chapitre traite des aspects pratiques de la création, de l'analyse et de la diffusion des données. Il se termine en indiquant comment lancer une alerte, façon extrême de pratiquer la transparence lorsque l'intérêt public est en jeu.

Mes objectifs sont tant de promouvoir la transparence que d'inciter chacun et chacune à découvrir ce monde des données, à en tirer parti, voire à prendre part au mouvement en produisant des données, des analyses et des services.

Les associations déjà au contact de données trouveront dans ce livre des éléments pour s'intégrer dans le mouvement de l'*open data*[g], pour interagir entre elles, avec l'État, les administrations et avec le grand public. Pour les autres, l'ouverture des données publiques est l'occasion de prendre du recul et de s'interroger sur les bénéfices possibles qu'elles peuvent en tirer. Cet ouvrage s'adresse aussi aux enseignants, transmetteurs de la connaissance, qui peuvent utiliser les données factuelles disponibles en ligne pour montrer aux élèves comment retrouver un résultat par soi-même. Enfin, les journalistes, les chercheurs, les professionnels des données et les citoyens désirant participer y trouveront des conseils pratiques ainsi que des raisons pour accompagner ce mouvement de la transparence numérique.

Notes :

- les références au glossaire sont indiquées par un [g],
- les liens cités sont disponibles sur `opendata.ricou.eu.org`.

TABLE DES MATIÈRES

Chapitre 1

La transparence

La transparence consiste à diffuser publiquement les informations en rapport avec la politique au sens large, c'est-à-dire tout ce qui concerne le fonctionnement de la société. À l'échelle d'un État, ces publications sont aussi multiples que diverses : rapports, données administratifs, vidéos de conseils municipaux, déclarations de patrimoine, appels d'offre, des mesures diverses comme celles des différents types de pollution, de la propagation d'une maladie, de l'occupation des routes, etc.

L'idée première est que les journalistes, les associations, mais aussi les citoyens, puissent vérifier que personne ne triche. Cela participe à la lutte contre la corruption, les malversations et autres actions illégales.

La transparence apporte aussi la lumière là où le doute existe, que ce soit avec des faits ou avec la publication des mécanismes de prise de décision [1]. Elle permet aussi de lutter contre la désinformation et d'améliorer l'enseignement en lui donnant des données brutes.

Un troisième axe est l'efficacité. En ouvrant leurs données, les administrations évitent la redondance et en améliorent la qualité grâce aux retours des utilisateurs. Elles réduisent la bureaucratie qui les pénalise aussi, en tant qu'actrices et en tant que consommatrices de données d'autres ministères. Elles changent le regard qu'on porte sur leur travail et favorisent les collaborations avec la société civile. Elles ouvrent aussi des opportunités économiques.

Enfin, il ne faut pas oublier l'imprévu. L'un des effets de la

1. Comme l'algorithme qui répartit les bacheliers dans les établissements d'enseignement supérieur. Ainsi chacun peut comprendre son affectation.

transparence est de permettre à d'autres d'analyser l'information. En publiant des données, on ouvre la porte à des usages que même leurs créateurs n'imaginent pas toujours. C'est d'ailleurs un des principes du mouvement de l'*open data*[g] : « Ouvrez vos données, si vous ne savez pas pourquoi, d'autres trouveront ! »

1.1
Au nom de la justice

La transparence n'est pas un élément naturel de la gouvernance, l'histoire montre que le pouvoir préfère le secret [2]. L'ouverture vers la transparence vient de la Réforme protestante et des philosophes des Lumières. En France, cela s'est concrétisé dans la constitution française de 1789 en son article 15 : « la société a le droit de demander compte à tout agent public de son administration ». Dans les pays scandinaves le mouvement va plus loin : une loi suédoise de 1766 introduit le principe de transparence en même temps que la liberté de la presse. Tout citoyen peut avoir accès aux documents administratifs y compris financiers comme les salaires et les notes de frais des élus [3]. Le but était de contrôler les abus.

Corruption & Co.

En 2017, le Fond Monétaire International (FMI)[g] a estimé le coût des détournements de fonds publics à 2600 milliards de dollars, soit plus de 5% du PIB mondial. En Europe le coût de la corruption était estimé entre 179 et 990 milliards d'euros pour 2016 [4]. À titre de comparaison, le budget de l'État français en 2020 est de 250 milliards d'euros en recettes et de 343 en dépenses. Il s'agit

2. Cela se retrouve dans le mot secrétaire qui vient de secret. Le secrétaire d'État est celui qui détient les secrets de l'État.

3. https://www.la-croix.com/France/Politique/En-Suede-lultra-transparence-prix-confiance-2019-07-29-1201038149

4. https://www.lefigaro.fr/conjoncture/2017/12/09/20002-20171209ARTFIG00019-trois-chiffres-edifiants-demontrent-que-la-corruption-gangrene-le-monde.php

donc d'un problème qui coûte cher, très cher, même en Europe où la corruption est relativement faible.

Le phénomène touche aussi la France. L'organisation Transparency International[g] la classe en 23e position mondiale en 2019 suivant son indice de perception de la corruption, avec un score de 69/100 soit 3 points au-dessus de la moyenne européenne et au même niveau que les États-Unis [5]. En 2019, les parquets ont traité 813 affaires de manquements à la probité [6] impliquant 1263 personnes dont 242 personnes morales. Il en a résulté 274 condamnations dont 126 pour corruption et 68 pour détournement de bien public [7]. Un sondage de 2015 auprès d'acheteurs français a révélé qu'un quart de ces acheteurs publics a été confronté à une tentative de corruption [8]. Pour autant, seulement 4,4 % des communes avaient mis en œuvre un plan ou des mesures anticorruption en 2018 d'après l'Agence française anticorruption (AFA)[g] [9]. Ce pourcentage montait à 40 % pour les départements et 85 % pour les régions. La taille d'une collectivité locale semble donc contribuer à mobiliser les ressources humaines nécessaires, mais rend les contrôles plus difficiles. Une enquête européenne publiée en 2020 par le Groupe d'États contre la corruption (Greco)[g] [10] indique que le risque existe jusqu'au sommet de l'État. Ainsi, dans un pays où le président est la personne la plus puissante, il serait légitime que les membres de son cabinet soient soumis aux mêmes contrôles que

5. Transparency International publie un indice de perception de la corruption des pays. Cet indice varie de 0, totalement corrompu, à 100, pas de corruption. En 2019 les 2/3 des pays sont en dessous de 50, la moyenne mondiale est à 43, la moyenne européenne est à 66. Le pays le mieux classé est le Danemark avec 87/100, le moins bien classé est la Somalie avec 9/100. Cf https://www.transparency.org/fr/cpi/2019

6. Les manquements à la probité sont la corruption, le trafic d'influence, la concussion, la prise illégale d'intérêt, le détournement de fonds et le favoritisme.

7. Cf page 14 du rapport d'activité 2020 de l'Agence française anticorruption https://www.agence-francaise-anticorruption.gouv.fr/files/files/RA_AFA_2020_V2_WEB.pdf

8. Communication personnelle de l'association Anticor[g]

9. https://www.agence-francaise-anticorruption.gouv.fr/files/files/Rapport_danalyse_-_enquete_service_public_local.pdf

10. https://rm.coe.int/cinquieme-cycle-d-evaluation-prevention-de-la-corruption-et-promotion-/16809969fd

les membres des cabinets ministériels. Ce n'est pas le cas. De plus, si le président, comme les ministres, doit soumettre une déclaration de patrimoine, la sienne n'est que déclarative et ne sera pas contrôlée par la Haute autorité pour la transparence de la vie publique (HATVP)[g] contrairement à celles des ministres. Autre faille, les ministres sont jugés par la Cour de justice de la République, composée en majorité de personnalités politiques.

Le problème n'est pas que financier : en cassant les échelles de valeur, la corruption mine la société. La réussite n'est plus liée au mérite, le travail rapporte moins que les magouilles. Dès lors, le découragement et le sentiment d'injustice se développent.

> *« La corruption (l'abus d'une charge publique à des fins personnelles) nuit à l'activité de l'État et compromet les chances de parvenir à une croissance économique durable et inclusive. La corruption permet à certaines personnes de se soustraire aux impôts, tandis que d'autres finissent souvent par en payer plus. Les pertes de recettes peuvent également entraver la capacité de l'État à réaliser des dépenses sociales. En outre, la qualité des services publics et des infrastructures est réduite lorsque les décisions des pouvoirs publics sont mues par les pots-de-vin ou le népotisme. De surcroît, la corruption érode la confiance envers l'État et peut provoquer une instabilité sociale et politique. »*
>
> Moniteur des finances publiques, FMI[g], avril 2019

Heureusement la lutte contre les manquements à la probité, dont la corruption est l'élément le plus visible, progresse. Les créations récentes de la HATVP[g] et de l'AFA[g] ainsi que l'ouverture des données au sein de l'administration en sont des témoignages. Transparency International[g] a d'ailleurs hissé la France en 2020 dans la catégorie « Application modérée » pour sa lutte contre la corruption (2e catégorie sur 4) [11].

11. https://transparency-france.org/actu/plus-des-trois-quarts-des-exportations-mondiales-sont-touchees-par-la-corruption-exporting-corruption-un-rapport-qui-revele-les-insuffisances-de-la-lutte-contre-la-corruption-internationale/

L'évasion fiscale est un autre type d'injustice. En France, son coût financier est estimé à plus de 100 milliards par an [12], plus que le déficit budgétaire français, plus que le budget de l'Éducation nationale ou plus que ce que rapporte l'impôt sur le revenu. Là encore, ce manque à gagner pour l'État a des répercussions sur l'ensemble de la société : moins de service public et une plus grande pression fiscale sur l'ensemble du pays.

Alors que peut faire la transparence ?

En premier, projeter la lumière là où l'obscurité favorise les actes illégaux. Pour certains, la publicité peut être pire qu'une sanction fiscale. Or depuis 1920 le « verrou de Bercy » permettait au ministère des Finances de décider seul et sans justification la poursuite éventuelle d'un fraudeur coupable d'évasion fiscale. En 2018 la loi a été modifiée : les dossiers de fraude caractérisée atteignant au moins 100 000 € et assortis de pénalités lourdes (40 % ou plus) sont systématiquement transmis au parquet. Cette avancée serait louable si elle ne demeurait insuffisante et incomplète : la nouvelle loi a créé en effet deux procédures spéciales ayant pour but d'éviter le procès public, la « Comparution sur reconnaissance préalable de culpabilité » et la « Convention judiciaire d'intérêt public ». Le risque de mauvaise publicité est de fait considérablement réduit [13].

L'opacité fiscale ne s'arrête pas à Bercy. L'Europe n'est pas parvenue à un accord sur une transparence fiscale qui imposerait aux multinationales de publier leurs données fiscales pays par pays (chiffre d'affaires, bénéfices, impôts payés), afin de garantir l'imposition locale. Plus largement, la lutte contre les paradis fiscaux ne semble guère active comme le montrent les fuites régulières [14].

Et le domaine fiscal n'est pas le seul concerné. Des scandales ont aussi eu lieu, entre autres, dans les secteurs industriels, sanitaires et politiques.

---

12. https://www.socialistsanddemocrats.eu/sites/default/files/2019-01/the_european_tax_gap_en_190123.pdf

13. cf l'article d'E. Vergès, *L'outil procédural au service de l'efficacité (à propos de la lutte contre la fraude)* publié dans la Revue de science criminelle en 2019.

14. L'International Consortium of Investigative Journalists (ICIJ)[g] a rassemblé dans une base les contenus des *Panama papers*, *Offshore Leak*, *Bahamas Leaks* et des *Paradise Papers*, cf https://offshoreleaks.icij.org/

Les scandales récurrents dans l'alimentation illustrent les difficultés de ce secteur avec la transparence. Les analyses fournies aux laboratoires pourraient avantageusement être rendues publiques au lieu d'être remontées aux seules autorités sanitaires. Ainsi l'affaire du lait contaminé illustre le défaut dramatique de transparence qui a permis à Lactalis de taire la présence des salmonelles pourtant repérées dans ses bâtiments.

Au niveau sanitaire, l'affaire du Mediator a mis au jour non seulement la tromperie des laboratoires Servier mais aussi les connivences entre les laboratoires pharmaceutiques et l'Agence nationale de sécurité du médicament [15] ainsi que la corruption de médecins. Aussi l'une des conséquences de cette affaire est-elle l'obligation de publier l'ensemble des dons faits au corps médical [16].

Tous ces scandales ternissent l'image de professions entières. Leur impact dépasse les victimes directes et touche tous les membres de ces professions, y compris ceux qui sont honnêtes.

La transparence a donc une triple action contre la corruption et autres dysfonctionnements : dénoncer et mesurer le phénomène, contrôler l'efficacité de la lutte, mettre la lumière sur les coupables.

Sentiment d'injustice

À l'injustice s'ajoute le sentiment d'injustice. Porté à son paroxysme, il peut se transformer en révolte comme celle des Gilets jaunes. À cette occasion le gouvernement a fait quelques concessions [17] mais la confiance n'est pas pour autant revenue. Les dirigeants et les hauts fonctionnaires restent accusés de se vautrer dans les ors de la République quand le peuple souffre. L'apaisement social semble loin. Ce sentiment d'injustice peut être fondé

15. Le jugement du 29 mars 2021 a reconnu les laboratoires Servier coupables de « tromperie aggravée » et d'« homicides et blessures involontaires ». L'Agence nationale de sécurité du médicament a été condamnée pour avoir tardé à suspendre la commercialisation du Mediator. Les laboratoires Servier ont fait appel.

16. 98 millions d'euros de dons pour les médecins hospitaliers en 2018, cf le travail du collectif de journalistes Data + Local sur `https://collectif-datalocal.github.io/`.

17. 17 milliards d'euros.

ou non. S'il n'est pas fondé, il provient d'un manque d'information, d'un manque de confiance et d'un manque de transparence.

Dans le domaine fiscal, la libre consultation par tous des revenus et des impôts payés par chacun, telle qu'elle est pratiquée dans les pays scandinaves, permettrait de vérifier que chacun paie sa part et de comparer les revenus [18]. Elle permettrait ainsi d'apprécier l'effort fiscal des plus riches et de souligner des disparités de revenu peu justifiables.

Hélas en France la situation est telle que même le Sénat ne peut pas avoir accès aux données fiscales des Français, ce qui ne lui permet pas d'évaluer l'impact de propositions de lois fiscales et donc réduit son contrôle de l'action du gouvernement [19].

La transparence devrait aussi permettre de savoir où va l'argent public. Malheureusement, là encore, la communication est faible. Les graphiques du site *ad hoc* [20] laissent le lecteur sur sa faim. Il n'est pas possible d'avoir des informations détaillées, comme voir comment l'argent est réparti dans un secteur [21]. Il serait pourtant utile d'obtenir des détails, de savoir quel est le budget de l'école primaire de son enfant, d'où viennent les recettes, où vont les dépenses. La déception est la même pour les commissariats, les musées, les administrations, etc. Pour l'instant, les données financières sont rares et souvent dispersées ce qui rend très difficile la compréhension globale des flux financiers de l'État.

C'est pourtant en comprenant où va l'argent que l'on pourrait réduire ce sentiment d'injustice qui veut que l'herbe soit toujours plus verte ailleurs.

---

18. La Norvège permet à ses citoyens de lire en ligne les revenus globaux et le total des impôts payés de tous les citoyens. La Suède et la Finlande permettent d'avoir l'information par téléphone. En France, il faut se déplacer dans les locaux de la direction locale des Finances publiques pour obtenir une information orale restreinte aux personnes du district fiscal du requérant. En Italie le ministre des Finances italien a décidé en 2008 de procéder comme en Norvège, mais les pressions l'ont contraint à fermer le site web.

19. https://www.publicsenat.fr/article/parlementaire/le-senat-demande-au-gouvernement-l-acces-aux-donnees-fiscales-des-francais

20. https://www.aquoiserventmesimpots.gouv.fr/

21. https://www.economie.gouv.fr/aqsmi/comment-largent-public-est-il-utilise

L'accès aux services publics est un autre point sensible qui peut procurer un sentiment d'injustice. On se doute que les raisons de l'État pour fermer une école ou une administration dépendent de données démographiques et des ressources financières. Mais quels sont les critères exacts ? Existe-t-il une norme pour l'implantation de services publics tous les x kilomètres ou pour n habitants ? La réponse est bien sûr plus compliquée : tout le monde n'a pas les mêmes besoins. Avoir un bureau de Pôle-Emploi pour n habitants sans prendre en compte le taux du chômage local serait une erreur. Mais alors comment les décisions sont-elles prises ? Là encore, la transparence est importante non seulement pour éviter des injustices, l'influence d'un élu local par exemple, mais aussi pour permettre des retours qui amélioreront les processus.

Dans ce cas la transparence s'appelle *ouverture des algorithmes*. Il s'agit d'indiquer les étapes qui mènent au résultat (l'emplacement optimal des bureaux de Pôle-Emploi dans notre exemple) [22].

Que ce soit pour les flux financiers ou pour la prise de décision, l'explication remédie au sentiment d'injustice, qui disparait lorsqu'on sait... mais peut aussi se transformer en colère légitime.

1.2
Éclairer notre monde

La transparence donne accès à des informations, c'est donc aussi un bon outil pédagogique. Dans un monde qui doute, cette aide est appréciable.

Un outil pédagogique

En offrant la possibilité de vérifier par soi-même, la transparence est un outil citoyen ainsi qu'un outil pédagogique pour les disciplines qui touchent au fonctionnement de notre société. Malheureusement, elle est encore trop peu utilisée à l'école, la tradi-

22. L'ouverture des algorithmes a fait parler d'elle avec le système Admission post-bac devenu Parcours Sup, les candidats voulant comprendre sur quels critères ils sont sélectionnés.

tion voulant qu'on s'appuie sur les manuels et les enseignants, et non pas qu'on remonte aux sources pour retrouver les résultats. Cet usage est dénoncé par le projet GapMinder[23], qui montre à quel point nous avons une vision passéiste, voire totalement fausse de notre monde, que ce soit sur l'évolution de la démographie des pays, sur la pauvreté, sur l'accès à la médecine, etc. Les tests qu'il a pu effectuer sur des Européens ou des Nord-américains montrent que nous n'avons pas vu que le reste du monde a fortement changé. Les raisons sont principalement :

— un biais personnel lié à notre expérience de la vie,
— des manuels scolaires et des enseignants qui n'enseignent que le passé,
— des médias qui faussent notre jugement en mettant en avant l'exceptionnel (sans parler du fait que les journalistes sont aussi ignorants que nous, d'après les tests de ce projet).

Heureusement Internet permet de vérifier nos intuitions et de progresser, si l'on sait éviter le piège de la désinformation. Wikipédia, qui est un outil de transparence, permet à chacun de disposer d'une encyclopédie, ce dont peu de gens disposaient auparavant. Plus encore, elle offre des données fondamentales qui permettent de comparer des éléments semblables. Par exemple, pour les entreprises ces données sont le chiffre d'affaires, les bénéfices ainsi que les filiales, la maison mère, les dirigeants, les actionnaires principaux, etc. Les villes, les régions, les monuments, les personnes célèbres ont aussi leurs données fondamentales. Toutes ces données sont stockées dans une base de données ouverte ce qui permet de faire des recherches globales et d'obtenir la liste des églises construites au 13e siècle ou de classer les entreprises suivant leurs bénéfices. Ce passage d'encyclopédies propriétaires en papier à une encyclopédie numérique ouverte est plus qu'une simple démocratisation, c'est une évolution majeure qui permet désormais de travailler sur ses données pour en extraire des informations difficilement accessibles avant.

Wikipédia n'est que la partie émergée de l'iceberg, l'ouverture de la connaissance est nettement plus large. Dans le milieu péda-

23. https://www.gapminder.org/ignorance. À tester!

gogique, le phénomène des Massive Open Online Course (MOOC)[g] a fait beaucoup de bruit, certains y voyant la fin de l'enseignement tel qu'il est pratiqué aujourd'hui. Si les choses n'ont pas tellement changé en France, il faut admettre que le libre accès aux cours des plus grandes universités du monde est, là encore, une avancée majeure. Toute personne sur Terre peut à présent suivre librement des milliers de cours et construire son cursus à la carte pour obtenir une formation supérieure de qualité.

Le niveau suivant consiste à extraire de la connaissance des données en ligne. Malheureusement, si l'école enseigne de citer ses sources, elle n'indique pas souvent l'importance de donner des liens web vers ces sources et vers les données brutes (lorsqu'elles sont en ligne). Les manuels scolaires affirment, mais ne proposent pas à l'élève de vérifier par lui-même en récupérant les données sur un site institutionnel, puis en les travaillant. La recherche reproductible, au sens où un article scientifique donne toutes les informations (et données) pour reproduire ses résultats, n'est pas encore arrivée dans le monde de l'enseignement. Trop souvent, les sources utilisées ne citent pas leurs propres sources brutes ou n'en donnent pas l'accès. Aussi, il est très encourageant de voir le réseau pédagogique Canopé proposer depuis peu d'étudier deux jeux de données brutes sur la première guerre mondiale [24]. Il est à espérer que cette expérience incitera plus d'enseignants à transmettre ce réflexe de l'analyse de la donnée brute [25].

Le complotisme

On nous cache tout, on nous dit rien. Cette vieille chanson de Jacques Dutronc écrite en 1967 semble être toujours d'actualité si on se réfère à ses commentaires sur YouTube. Pourtant, depuis cette époque, Internet a démultiplié l'accès à l'information et s'il reste des secrets (dans le domaine financier comme on a vu), une partie notable de l'information est en ligne.

24. `data.gouv.fr/fr/organizations/reseau-canope`

25. Voir aussi le combat du créateur du Web pour l'accès aux données brute : `https://www.wired.co.uk/article/raw-data`

Les démocraties rendent l'information publique plus en plus accessible. L'Union Européenne met en ligne les textes en cours d'examen par le parlement, les directives et de multitudes de rapports souvent de grand intérêt. En France, le site Legifrance permet depuis 1999 d'accéder librement à l'ensemble des textes de droit applicables. Les sites de l'Assemblée nationale et du Sénat présentent les projets de loi. Les ministères, les administrations proposent de plus en plus de contenu lié à leur activité [26]. Localement, nombre de mairies diffusent les vidéos des conseils municipaux et permettent l'accès aux documents structurants de la ville (budget, plan d'occupation des sols, demandes de subvention associative, etc).

Plus encore, grâce aux sites web, blogs, podcasts et vidéos en ligne, on accède à une information quasi exhaustive, à tel point que désormais cette surabondance d'informations, plus ou moins exactes, peut poser un problème. En permettant à chacun de s'exprimer publiquement, Internet offre le pire et le meilleur. Dans le pire, on trouve les théories du complot qui peuvent aller jusqu'à déstabiliser un pays.

Cela nous mène aux complotistes. Il y a les extrémistes, ceux pour qui « rien n'arrive par accident ; tout ce qui arrive est le résultat d'intentions ou de volontés cachées ; rien n'est tel qu'il paraît être ; tout est lié, mais de façon occulte » [27]. L'accès à l'information n'est plus le remède pour eux, leur problème dépasse notre sujet. Les autres sont surtout des personnes qui doutent. Pour elles la transparence et la logique peuvent être utiles.

Prenons l'exemple de la crise sanitaire du Covid-19. Cette pandémie a deux aspects : le virus et la gestion de la crise. Le premier point a fait de chacun de nous un expert en chloroquine. Il faut dire que l'on a tout pour s'informer si l'on en a vraiment la vo-

26. Seule la Documentation française en charge de la publication des rapports et travaux produits par l'administration semble être restée bloquée au 20e siècle et demande toujours de payer pour des documents que l'on peut trouver gratuitement sur le site web des auteurs. Ainsi le rapport au format PDF de la Cour de cassation est librement téléchargeable sur le site de cet organisme, mais coûte 12 € sur le site de la Documentation française.

27. Pierre-André Taguief, L'imaginaire du complot mondial, Ed. Mille et une nuits, 2006

lonté. Autrefois, l'accès aux revues se payait au prix fort, à moins de se déplacer dans une bibliothèque universitaire. Désormais, la majorité des articles scientifiques sont accessibles en ligne gratuitement [28], chacun a largement de quoi se faire un avis. Si on a besoin de rafraîchir ses connaissances en virologie, on peut suivre gratuitement des cours en ligne. On peut aussi écouter les scientifiques qui font l'effort de vulgariser leur domaine à travers des blogs et des vidéos. On a tout ce qu'il faut pour être un expert... en y passant quelques années.

Le second point, celui de la gestion de la crise, est plus intéressant, car plus abordable. En fait, lorsque les défenseurs de la chloroquine soupçonnent les grands groupes pharmaceutiques de manipuler l'État pour qu'il achète des médicaments plus chers, on est dans le domaine de la gestion de la crise. La question est donc de savoir si l'État fait bien son travail. En quoi la transparence aurait-elle aidé à le déterminer ?

Prenons l'exemple la gestion des masques chirurgicaux. Le gouvernement a déclaré qu'ils n'étaient pas utiles pour les personnes en bonne santé avant de changer d'avis, puis il a déclaré en avoir toujours eu assez en stock alors que les élus se battaient pour en acheter en Chine. Il a fait tout ce qu'il faut pour faire naitre des rumeurs. Aucune transparence des données, aucune transparence dans la communication. Au même moment, à Taïwan, le nombre de masques en stock dans les pharmacies était accessible en ligne et mis à jour quasiment en temps réel. Durant le pic, lorsque les masques étaient rationnés, il était simple d'estimer le nombre de masques achetés dans une pharmacie en fonction du nombre de clients et de vérifier, sur son ordiphone, qu'il correspondait avec la baisse des stocks de cette pharmacie. Ce calcul local permet de valider les données fournies par les dirigeants, ce qui rassure la population, tue dans l'œuf les rumeurs et responsabilise chacun.

Pour avoir une telle efficacité, il faut avoir les données et les ouvrir. La comptabilité des morts en France avec l' « oubli » des

28. Google scholar, DBLP référencent, HAL, ArXiv, MedRxiv, BioRxiv hébergent des articles ouverts et pour les autres, il existe Sci-Hub, une source illégale déployée par des scientifiques qui militent pour l'ouverture de la science sans barrière financière.

EHPAD a montré que la France n'avait même pas un système de comptabilité de ses morts en temps réel. Sans données, il ne reste que la confiance, or l'État français a trop souvent menti à ses citoyens pour être digne de confiance. Rappelons-nous du nuage radioactif de Tchernobyl officiellement stoppé à la frontière française. Si, au lieu de communiqués rassurants, l'État avait choisi de publier les mesures brutes de radioactivité sur le territoire, la présence du nuage aurait été évidente et les éventuelles mesures à prendre auraient pu l'être. Dans le milieu médical, l'affaire du sang contaminé ou celle du Médiator donnent deux raisons de plus de douter de l'État et de croire aux théories du complot. Ajoutons à cela que de nombreux élus et dirigeants mentent régulièrement ce qui n'ajoute rien à la confiance dans les discours officiels [29].

Que faire pour aider ceux qui doutent ?

Le plus souvent, il est possible de désamorcer une rumeur en cherchant sur Internet. Les sites dédiés au traitement des intox [30], ainsi que Wikipédia, offrent un bon commencement. Citons :

— Hoax Buster `https://www.hoaxbuster.com`
— le Décodex du Monde
`https://www.lemonde.fr/verification/`
— CheckNews de Libération
`https://www.liberation.fr/checknews`
— France Info Vrai-Faux
`https://www.francetvinfo.fr/vrai-ou-fake/`
— AFP Factuel `https://factuel.afp.com/`
— l'IA Véra [31] par téléphone au 09 74 99 12 95

Si ces sites associatifs et de médias n'ont pas la réponse, il est également possible de détecter une fausse rumeur en croisant différentes sources. C'est une technique, certes longue, mais très efficace si on l'associe aux techniques usuelles d'autodéfense intellec-

29. *« Il est incontestable que ce sont plutôt des menteurs éhontés et récidivistes qui parviennent à être élus aux plus hautes responsabilités. »* Thomas Guénolé, auteur du Petit guide du mensonge en politique, Hachette, 2017.

30. appelées aussi *Fake news* ou infox.

31. Véra est développée par l'ONG LaReponse.Tech et intègre de nombreux sites de vérification et des journaux, cf `https://www.askvera.org/`.

tuelle [32]. Enfin Internet permet d'interagir avec des spécialistes sur leurs sites, en répondant à leurs vidéos ou via Twitter. La discussion est un élément important pour découvrir ses erreurs et profiter de l'expérience des autres. Des forums comme Reddit[g] sont aussi intéressants, tout en sachant qu'ils peuvent être le repaire de trolls[g] et de complotistes.

L'un dans l'autre, on dispose d'outils pour vérifier qu'une rumeur est fausse ou qu'un élu ment. Il va cependant de soi que plus le cas est difficile, plus il est nécessaire de prendre le temps de s'informer et de sortir de sa bulle de confort pour discuter avec les autres, ceux qui pensent différemment. Il ne faut pas oublier que parfois des complotistes ont raison [33] et donc les rejeter a priori n'aide pas à trouver la vérité.

Même pour les cas les plus compliqués comme les cas médicaux, il est souvent possible de vérifier une intuition. Par exemple la comparaison des morts d'une année sur l'autre ne demande aucune connaissance médicale, mais offre à tous un bon indicateur pour évaluer l'impact de la Covid depuis 2020.

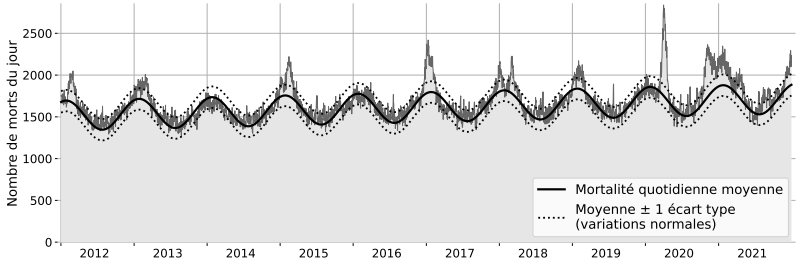

FIGURE 1.1 – Mortalité quotidienne en France (données : INSEE)
Note : les pics hivernaux sont dus à la grippe. Pas de grippe durant les hivers 19-20 & 20-21.

Ainsi, plus on possède d'information, de données brutes, plus il est possible de séparer le vrai du faux. Toutefois, il faut faire attention à la qualité des données. Elles peuvent être fausses involontairement (erreur humaine, appareil de mesure déficient, chan-

32. Cf *Petit cours d'autodéfense intellectuelle* par Normand Baillargeon chez Lux.

33. Combien de lanceurs d'alertes ont été jugés complotistes au début ?

gement de définition au fil du temps...) ou volontairement (pour justifier un point de vue ou semer le doute). Aussi, il est nécessaire de prendre la peine d'analyser une information, et ce, d'autant plus que les fausses rumeurs prennent souvent le dessus : on préfère l'incroyable faux au véridique ennuyeux. Une étude a montré que les tweets véhiculant une information fausse sont nettement plus relayés que ceux dont l'information est juste [34]. Là encore l'école a un rôle à jouer et elle en est consciente [35], mais la lutte semble impossible. Pourtant, sa réussite est nécessaire pour apaiser notre société et améliorer notre démocratie.

Pour la science

Les scientifiques aussi désirent plus de données ouvertes.

La science des données a pris son envol au début du siècle. Elle a pour but d'extraire de nouvelles connaissances des données. Pour cela, il faut beaucoup de données afin d'avoir des informations statistiquement valables. Les domaines les plus concernés sont la physique, les sciences du vivant, l'environnement et les sciences humaines. Ainsi l'étude du réchauffement de la planète nécessite des mesures de données physiques sur toute la planète et sur de longues périodes de temps. Suivre l'évolution des villes, de leurs transports, demande d'avoir des cartes dynamiques et des mesures des flux urbains. La santé en général est un autre exemple de domaine grand consommateur de données.

Pour que le plus grand nombre de scientifiques puisse exploiter des données, elles doivent être massives, correctes et ouvertes. Sans cette ouverture les progrès fulgurants de l'intelligence artificielle de ces dernières années n'auraient pas eu lieu [36]. Mais aujourd'hui ce sont les GAFAM[g] (et d'autres entreprises équivalentes)

34. The spread of true and false news online, S. Vosoughi, D.Roy et S.Aral., Science 9/03/2018, https://science.sciencemag.org/content/359/6380/1146

35. cf https://www.reseau-canope.fr/developper-lesprit-critique.html et https://eduscol.education.fr/cid107295/former-l-esprit-critique-des-eleves.html

36. L'Intelligence Artificielle (IA) a besoin de grandes quantités de données pour apprendre.

qui disposent de la plus grosse masse de données, ce qui les place en avance sur le milieu académique dans le domaine de l'intelligence artificielle.

En France, l'État dispose de données considérables qui pourraient faire de lui un acteur scientifique important. Malheureusement, la coopération entre les chercheurs français et les services de l'État n'est pas toujours simple, et fort au-dessous des standards internationaux d'après le rapport Bothorel[g] qui cite l'exemple d'une chercheuse française : « Souhaitant évaluer le travail détaché en France, [elle] n'a ainsi à ce jour pas reçu les données de l'administration malgré une demande effectuée il y a plus de deux ans, et l'accord du comité du secret statistique. Elle n'a en revanche eu aucune difficulté à obtenir ces mêmes données de la part de la Belgique, du Luxembourg et du Portugal ». Outre le fait que l'ouverture des données publiques résoudrait ces problèmes administratifs, elle permettrait aussi de constituer des jeux de données nettement plus conséquents à condition que tous les pays fassent de même. La quantité de données accessibles améliore la qualité des résultats. D'autre part, le simple fait de devoir présenter une demande administrative suffit à décourager des chercheurs, qui utiliseront d'autres données déjà ouvertes (américaines souvent) avec pour conséquence des résultats pas adaptés à la situation française. Enfin l'ouverture de données permet à des citoyens et à des associations d'alerter les scientifiques. Par exemple les données sur la pollution de l'air publiées par Air Parif[37] et ses consœurs peuvent permettre à chacun de noter un phénomène inhabituel, que les scientifiques expliqueront ou étudieront.

L'exemple le plus actuel de l'intérêt de l'ouverture des données pour la science est celui du Covid. Il s'agit d'un problème mondial et les enjeux liés à la découverte des vaccins sont tels qu'on aurait pu craindre que chacun garde son information pour soi. Cela n'a pas été le cas et de nombreuses bases de données sur cette épidémie ont été mises en ligne pour permettre à chacun de l'étudier. Cependant, force est de constater que la qualité des données mises en ligne est des plus inégales et l'on verra que c'est l'un des pro-

37. http://www.airparif.asso.fr/

blèmes récurrents de l'analyse des données massives.

Les exemples ci-dessous témoignent de la diversité des types de données et des émetteurs :

— Le site dédié de l'Organisation Mondiale de la Santé (OMS)[g] présente une carte du monde pour visualiser les différents paramètres et permet aussi de télécharger les données brutes de base de l'épidémie [38],

— Santé publique France fournit de nombreuses données nationales sur l'épidémie [39],

— Un dépôt français collaboratif agrège des données de différentes sources sur l'épidémie [40],

— Un site web propose liste de pré-publications [41] sur le Coronavirus déposées sur MedRxiv et BioRxiv [42],

— Un dépôt canadien collaboratif a des images aux rayons X de poumons de malades pour entrainer des IA [43],

— Des chercheurs hospitaliers iraniens offrent un jeu de 63849 images médicales pour 377 patients et décrivent leur méthode de classification pour établir un diagnostic [44],

— Le projet international GISAID propose de suivre la phylodynamique de la pandémie et offre des outils pour analyser les génomes des Coronavirus [45].

Tout n'a pas été aussi transparent. L'institut Pasteur, l'AP-HP, l'Inserm et d'autres institutions spécialisées ont développé des modèles de propagation de l'épidémie, utilisés par les autorités, mais non diffusés au grand public, ni entre eux. Inversement l'Impe-

38. https://covid19.who.int/

39. https://www.data.gouv.fr/fr/organizations/sante-publique-france/

40. https://github.com/opencovid19-fr/data

41. Pré-publication : article soumis à une revue ou une conférence mais pas encore évalué par le comité de lecture, donc à valider soi-même. Les pré-publications permettent d'avoir l'information plus rapidement.

42. https://connect.medrxiv.org/relate/content/181

43. https://github.com/ieee8023/covid-chestxray-dataset

44. https://github.com/mr7495/COVID-CTset

45. https://www.gisaid.org/

rial College a publié son modèle dès avril 2020 sur le dépôt public GitHub[g].

Pour conclure sur les données étatiques à caractère scientifique, regardons les données médicales dont dispose la France. L'Assurance Maladie possède une base médicale assez unique au monde qui comprend l'ensemble des ordonnances délivrées à tous les assurés ainsi que l'ensemble des actes remboursés à des générations de patients. Les hôpitaux, les laboratoires d'analyse et les centres médicaux disposent, eux aussi, d'une masse de données médicales. En combinant toutes ces sources, on peut tracer à grande échelle le dessin de la vie médicale des Français et analyser l'efficacité des traitements et des médicaments. Ces données sont donc une source progrès potentiellement importants. Bien sûr, il n'est pas possible de les ouvrir pour des raisons d'éthique et de protection de la vie privée. Cependant, on peut extraire des informations totalement anonymes à partir des données agrégées sur des cohortes ou pour l'ensemble des Français, comme par exemple la liste les médicaments prescrits [46]. Enfin, pour les travaux qui ont besoin de données spécifiques, le gouvernement a construit le Health Data Hub [47] (en français dans le texte) qui a pour mission de collecter et de mettre à disposition ces données aux projets de recherche sélectionnés.

1.3
Un gain d'efficacité

La transparence a aussi un impact en termes de morale et d'efficacité.

La mise en œuvre de l'ouverture des données publiques durant la décennie 2010 a généré des perturbations importantes au sein de nombreuses administrations. Leur premier réflexe a pu être un rejet de « ce truc » mal compris [48], qui génère des risques

46. `http://open-data-assurance-maladie.ameli.fr/medicaments/`

47. `https://www.health-data-hub.fr/`

48. « L'utilité et le potentiel des données et des codes sont mal compris, et restent identifiés dans l'imaginaire collectif comme des sujets purement *tech-*

légaux, qui peut entacher la réputation sans parler des enjeux de pouvoir. Alors parfois elles ont triché, elles ont trouvé des façons de respecter les ordres de transparence tout en brouillant assez les pistes pour tuer le processus. On a vu une administration, forcée de transmettre des chiffres, les imprimer puis les faxer pour bloquer tout traitement informatique. Une variante consiste à fournir l'information oralement au demandeur [49]. Parfois une administration transmet des informations manuscrites. Cela a été le cas de la première version de déclaration de patrimoine des élus, publiée telle quelle, non pour des raisons techniques, mais pour qu'en soit freinée l'exploitation. Un ancien vice-président du Conseil d'État [50] faisait remarquer qu'il devient fréquent qu'un document ayant vocation à être ouvert « soit expurgé de la partie la plus sensible des constatations faites ou des jugements portés, cette information sensible étant réservée à des commentaires oraux de l'auteur du rapport à l'autorité destinataire du rapport. De même, les procès-verbaux de réunions peuvent être établis selon des termes qui masquent ou altèrent la réalité ou sous une forme à ce point synthétique que leur valeur informative s'en trouve affectée ». Et ces blocages ne concernent pas seulement le grand public. On les constate à l'intérieur même de l'État, à l'intérieur même de chaque administration [51]. En 2020, le Rapport de la mission Bothorel[g] soulignait que « le partage de données entre administrations de l'État est scandaleusement faible, au point que certaines directions ressaisissent des données disponibles dans une direction du même ministère, ou que l'*open data*[g] est parfois le seul moyen pour une administration de connaître l'existence puis d'accéder aux don-

niques, qui ne sont pas au service de l'élaboration et du pilotage des politiques publiques. », Rapport de la mission Bothorel[g]

49. L'article de Christian Quest donne d'autres exemples, `https://medium.com/@cq94/mission-donn%C3%A9es-et-codes-sources-dd684c4b8410`

50. Renaud Denoix de Saint Marc, colloque *Transparence et secret*, 2003.

51. « Sur les aspects organisationnels, l'AGD[g] signale que les administrations refusent souvent de partager les données entre, elles, car, d'une part, le partage des données ne fait pas partie des missions, des objectifs et du budget des services et, d'autre part, la protection des secrets relatifs à la vie privée ou à la sureté de l'État ferait l'objet de « précautions excessives. » », thèse de doctorat de Samuel Goëta, 2016, `https://pastel.archives-ouvertes.fr/tel-01458098`

nées d'une autre administration ».

Pourtant, élus et fonctionnaires ont beaucoup à gagner de l'ouverture des données, que ce soit vers le public ou vers les différentes administrations. Moralement, il est satisfaisant de servir les citoyens (ce qui n'est pas la même chose que de servir l'État) et de lutter contre les manquements à la probité. Il est aussi agréable de pouvoir faire son travail efficacement sans devoir se battre contre d'autres administrations pour accéder aux données dont on a besoin. Mais que pèsent ces avantages face à une hiérarchie qui freine, par inertie, par peur de ternir l'image de son service, que ce soit en laissant apparaître un dysfonctionnement ou en publiant des données de mauvaise qualité ? Que faire lorsque les ressources techniques ne sont pas suffisantes ? Il y a donc un travail d'accompagnement à mener pour que cette transparence soit bien acceptée et mise en œuvre sans trop de souffrance. Pour l'acceptation, le travail a été simplifié par la décision de l'État d'inscrire l'ouverture des données publiques dans la loi pour une République numérique[g] de 2016. L'accompagnement est proposé par Etalab[g], l'administration chargée de coordonner cette ouverture et d'aider à sa mise en œuvre au sein des administrations. En 2021, la feuille de route *Transformer l'action et la fonction publiques par la donnée* [52] du ministère de la transformation et de la fonction publiques continue dans la même direction en soulignant l'accompagnement nécessaire des administrations et la mise à disposition d'outils, Etalab étant toujours la cheville ouvrière de cette aide. Cette feuille de route rappelle aussi l'importance politique de l'ouverture de données.

> *« La mise à disposition de données publiques en open data ouvre des opportunités inédites de rendre compte de l'action de l'État, de faire que cette transparence puisse être utilisée par la société civile - qui peut ainsi croiser ces données et conduire ses propres analyses - et d'engager par là même une démarche de dialogue avec ces réutilisateurs. »*

52. https://www.numerique.gouv.fr/uploads/feuillederoute_MTFP.pdf

L'avantage moral de la transparence se traduit aussi en efficacité : dès lors qu'il y a transparence, et donc obligation de rendre des comptes, les actes illégaux ou immoraux tendent à diminuer [53]. Or ces actes vont à l'encontre de l'intérêt général et génèrent une atmosphère de travail néfaste. La lutte n'est pour autant pas simple. Celui qui dévoile les irrégularités au sein d'un service peut vivre un enfer : perçu comme un traître qui salit la réputation du service, il risque le placard ou pire le tribunal. Devenir un lanceur d'alerte est un engagement fort [54]. Lorsque la tendance générale va vers plus de transparence, que les textes officiels confirment ce mouvement, alors la situation est plus simple, même si elle demande toujours un effort.

D'autres avantages de la transparence, dans notre cas liés à l'ouverture des données publiques, sont d'ordre pratique et financier. Lorsqu'un jeu de données est ouvert, il n'est plus nécessaire de faire une demande administrative pour y accéder. L'ouverture permet aussi de découvrir des jeux de données en doublon ce qui permet de n'en garder qu'un, réduisant du même coup les frais de gestion. De plus l'utilisation de ces données par plusieurs administrations et par le grand public permet de repérer les éventuelles erreurs plus rapidement. L'unicité permet aussi une plus grande cohérence entre les différentes administrations. Enfin, les données publiques ouvertes peuvent servir de support à des projets d'utilité publique.

Un exemple illustre cette utilité publique de l'ouverture de données : la création de la Base d'Adresses Nationale [55] qui permet d'associer une adresse postale à sa position géographique (latitude, longitude). Pour le service public, il s'agit d'un outil de gestion du

53. Cette constatation est géographique,les pays les plus transparents sont les moins corrompus. Cf *Lutter efficacement contre la corruption : la Suède comme exemple* par S.Paquin et J-P.Frady dans la revue Éthique publique, 2018.

54. cf l'histoire de N.M. Meyer `http://mindbreak.fr/unpourtous/single.php?id=56`. Depuis l'évolution de la loi, les lanceurs d'alertes sont mieux protégés, cf section 3.4.

55. `https://www.data.gouv.fr/fr/datasets/base-adresse-nationale/`

territoire qui concerne de nombreux ministères. Pour les entreprises, cette base permet le développement de logiciels affichant sur une carte des informations dont les données sont exprimées sous forme d'adresse postale (magasins, théâtres, fontaines, etc.). Elle est aussi nécessaire pour la livraison par drone. Il s'agit donc d'un projet qui intéresse de nombreux acteurs. Il a regroupé la Poste, Etalab[g], l'Institut national de l'information géographique et forestière (IGN)[g], des acteurs locaux, des collectivités, des communes et des citoyens liés au projet OpenStreetMap[g]. Cette collaboration entre des acteurs détenant chacun une partie de l'information permet de construire à moindre coût une référence de qualité. Son format ouvert règle le problème du contrôle de l'accès et permet toute innovation technique et commerciale sans frein administratif ou financier.

La baisse des coûts, les opportunités économiques et les gains de productivité sont donc des avantages de l'ouverture des données que la France a mise du temps à intégrer. Son modèle historique était la vente des données publiques, contrairement aux États-Unis qui considèrent depuis toujours que le travail de l'administration fait partie du domaine public [56]. Certes la vente de données génère des revenus directs [57], mais elle bride l'innovation et le service rendu au public. Le Danemark a fait le calcul. Dans le cadre de son plan 2011-2015 d'eGovernement, il a décidé d'ouvrir les données fondamentales comme le cadastre, les cartes topographiques, les adresses et certaines données sur les individus. Cet accès simplifié à l'information, ainsi que l'impact structurant de cette ouverture, permettent de générer des économies d'environ 100 M€ par an pour les secteurs privé et public (70 et 30 M€ respectivement). Le coût de l'ouverture de ces données a été de 135 M€ [58]. Il s'agit donc d'un investissement rentabilisé en 5

56. Légalement depuis 1895 avec la *Printing law*, mais auparavant la question n'avait pas été soulevée.

57. 10 M€ pour l'Institut national de la statistique et des études économiques (INSEE)[g] et autant pour l'IGN[g] en 2013 ce qui représentait 70 % du total des redevances portant sur l'utilisation des données publique. Le problème de la gratuité ne concernait donc que l'INSEE et l'IGN, cf thèse de Patricia Rahmé *Open data locale : acteurs, pratiques et dispositifs*, `https://www.theses.fr/2016LYSE3045`

58. `https://eurogeographics.org/wp-content/uploads/2019/12/LEIF-T`

ans pour l'État avec, en prime, un gain financier conséquent pour le secteur privé. En 2011, la Commission Européenne évaluait le potentiel économique des données publiques à 40 milliards d'euros par an [59]. La Fondation internet Nouvelle Génération (Fing)[g] considérait en 2020 que « les bénéfices économiques de l'ouverture des données en France sont évalués *a minima* autour de 3 à 5 milliards d'euros » par an [60].

Stationnement illégal

Enfin l'ouverture des données permet des retours constructifs de citoyens.

En 2014 la mairie de New-York a décidé de publier la liste des procès-verbaux pour stationnement illégal mentionnant le montant, la date, l'heure et le lieu de l'infraction. Un internaute a eu l'idée d'utiliser ces données pour générer une carte des emplacements interdits les plus rentables pour la ville. Il a compilé les données et, à sa grande surprise, il a découvert que certains emplacements, répartis de façon uniforme dans la ville, généraient un montant d'amendes considérablement plus élevé que les emplacements voisins. Ne trouvant pas d'explication à cette anomalie, il s'est rendu sur le terrain et a découvert que ces emplacements étaient situés face à des bornes d'incendie placées sur le trottoir. Or ces bornes justifiant l'interdiction de stationner étaient peu visibles et rien n'indiquait, en dehors de leur présence, que le stationnement était interdit. Il a alors contacté la mairie, qui fit la même constatation et décida de marquer de zébrures ces emplacements interdits [61].

---

he-Danish-Basic-Data-Initiative_10_15VI20_eng.pdf

59. cf thèse Patricia Rahmé, voir note 57.

60. *Dix ans d'ouverture des données publiques, Un bilan critique*, Fing 2020, https://fing.org/wp-content/uploads/2020/02/Open-Data-impact-bilan-2010-2018-de-l-open-data.pdf

61. source : https://www.ted.com/talks/ben_wellington_how_we_found_the_worst_place_to_park_in_new_york_city_using_big_data

1.4
Limites et mise en œuvre

On voit que l'impact de la transparence va bien au-delà de la lutte contre la corruption. La transparence change notre société en profondeur. Cependant, elle peut aussi présenter des dangers.

Les limites de la transparence

Pour certains la transparence viole la vie privée, le secret des affaires, le secret médical et bien d'autres secrets, volontairement ou pas.

Un jeu de données peut avoir un effet collatéral non souhaité. Lorsque l'État publie la liste des ventes immobilières incluant le prix de vente de chaque parcelle, il permet certes de stabiliser le marché et de mieux comprendre l'évolution des prix, mais il permet aussi de savoir le prix payé par les voisins pour leur maison. Certains peuvent y voir une violation de leur vie privée.

Dans le domaine politique, la transparence oblige les élus et les responsables politiques à faire une déclaration d'intérêts comprenant des informations personnelles ainsi que la profession et l'employeur de leur conjoint. Il est aussi demandé des informations financières dont le patrimoine, ce que certains n'apprécient pas. Le dernier en date, et non des moindres, est le ministre de la Justice, M. Dupond-Moretti qui, face à cette obligation, a déclaré :

> *« Je n'aime pas la transparence, je n'aime pas la transperçance (sic). Je pense que l'on vit dans une époque où l'on exige tout et où le secret et en particulier le secret professionnel devient suspect. Je n'aime pas l'ère de la suspicion.*
>
> *...*
>
> *Mes enfants ne savent pas ce que j'ai gagné, pour une raison simple, c'est que j'avais envie de leur donner le goût de l'effort et le goût du travail. Je ne voulais pas qu'ils se disent : « On est des fils de bourgeois et ça va ». Et voilà que tout cela va être dévoilé à la France entière*

et une partie de ceux qui vont aller là vont y aller avec gourmandise. J'estime que c'est du ressort de ma vie privée. »

Éric Dupond-Moretti, août 2020

Donc oui, la transparence peut entrer en conflit avec la vie privée. Oui, maintenant un ministre doit déclarer ses biens lors de son entrée en fonction pour que l'on puisse vérifier qu'il n'y a pas eu d'enrichissement personnel à la fin de son mandat. Cela peut effectivement être assimilé à de la suspicion, mais à qui la faute ? La création de la loi exigeant la déclaration de patrimoine et d'intérêts fait suite à l'affaire Cahuzac en 2013, qui vit l'intéressé condamné pour fraude fiscale et blanchiment (nombre de ses collègues l'avaient précédé dans la faute sans être toujours pris la main dans le sac).

Lorsque les affaires se multiplient, il devient légitime de faire glisser le curseur vers plus de transparence et moins de vie privée pour le groupe concerné. D'autres solutions pourraient être imaginées, mais quand le groupe concerné est *ceux qui nous gouvernent*, tout autre solution est d'une façon ou d'une autre sous leur contrôle.

Plus généralement, lorsque différents droits entrent en conflit, alors la nécessité de trouver un bon équilibre apparaît.

« La protection de la vie privée d'un individu est une raison importante pour faire exception au droit à l'information. Cependant, cela ne signifie pas que l'accès à un document doit être systématiquement refusé s'il contient des données personnelles. Transparence et vie privée sont deux droits fondamentaux d'importance égale, aucun des deux ne prévaut sur l'autre. Un examen attentif de ces deux principes est la clé d'une solution appropriée. »

Peter Hustinx, Contrôleur européen de la protection des données, 2005

Les juges sont aussi régulièrement interrogés sur le positionnement du curseur. Lorsque la justice européenne a créé le droit

à l'oubli, qui permet à chacun de demander que les moteurs de recherche ne référencent plus des pages le concernant, elle a aussi posé des limites à ce droit pour les personnes publiques, la transparence prenant le dessus en fonction de « l'intérêt du public à disposer de cette information ».

Le secret de la vie privée n'est pas le seul frein à la transparence. Il existe d'autres secrets, le secret des affaires, le secret médical, la protection des sources des journalistes, le secret professionnel des avocats ou le secret défense qui peuvent entrer en conflit avec la transparence. Là aussi la ligne de séparation ne peut être gravée dans le marbre et doit pouvoir s'adapter aux différents cas. Le but n'est pas d'avoir une transparence absolue, les conséquences en seraient terribles pour l'équilibre personnel de chacun. Il s'agit de trouver le bon équilibre pour notre société. Beaucoup de Français considèrent que l'argent est du ressort de la vie privée, voire de l'intime, alors ce n'est pas le cas dans d'autres pays (les pays anglo-saxons et nordiques en particulier). Doit-on mettre en ligne les impôts payés par chacun pour améliorer notre société ?

La transparence n'a pas non plus pour but de détruire l'économie. Les entreprises ont des secrets à juste titre. Le secret des affaires est important. Mais jusqu'où peut aller ce secret ? Fin 2020, le gouvernement a fait passer un amendement (rattaché à la loi ASAP) qui place le secret des affaires au dessus du code de l'environnement qui permet aux citoyens d'avoir accès aux éléments qui sont de nature à avoir un impact environnemental ou sanitaire [62]. Il choisit donc, en positionnant le curseur, de faire passer l'économie avant l'environnement et la santé publique. C'est un choix de société. Dans le domaine financier, les entreprises cotées en bourse doivent appliquer des règles de transparence et sont contrôlées par l'Autorité des marchés financiers (AMF)[g]. En pratique, elles évitent de plus en plus ces règles en se tournant vers d'autres sources de financement, le *shadow banking* [63]. Quant aux GAFAM[g] qui influencent nos vies et en savent plus sur nous que

62. https://www.marianne.net/economie/en-catimini-le-gouvernement-etend-le-champ-du-secret-des-affaires

63. Ensemble des entités ou des activités qui participent au financement de l'économie en dehors du système bancaire traditionnel et de la bourse.

l'administration, elles sont loin de supporter les mêmes contraintes en matière de transparence.

La transparence a donc ses limites pour préserver un équilibre indispensable au bien-être de chacun ou pour des raisons « supérieures ».

Enfin, elle ne peut pas se poser en remède à tous nos maux. Ainsi, en politique, elle est utile, mais changer de mode d'élection et/ou de gouvernance serait peut-être plus efficace pour redonner confiance aux citoyens.

Qui s'en charge ?

Le gain apporté par la transparence à nos sociétés, dans le respect des limites évoquées ci-dessus, est incontestable. Comment la mettre en œuvre, quelles modalités de développement préconiser ?

Contrairement à ce qu'on pense trop souvent, la transparence n'est pas seulement l'affaire de l'État et des collectivités. Leurs contributions sont certes capitales, mais d'autres acteurs peuvent et doivent intervenir : les associations, les entreprises, de multiples organismes. Chaque citoyen aussi peut apporter sa pierre à l'édifice. Essayons-nous à une liste :

— Les élus peuvent demander l'ouverture de données de l'administration à travers la loi [64]. Lorsqu'ils sont au gouvernement, ils peuvent inciter leur ministère à être plus transparent. Ils montrent l'exemple (en principe).

— L'administration publie ses données de façon brute (pas toujours de façon utilisable malheureusement). Elle gagnerait expliquer comment les utiliser pour lutter contre les fausses rumeurs, pour aider à l'éducation de tous, pour rapprocher les citoyens de l'État.

— Les organismes de contrôle, de surveillance ou de garantie des droits, non seulement diffusent des données, mais publient aussi leurs travaux d'analyse. Grâce à leurs rapports

64. La loi pour une République numérique[g] qui impose l'*open data*[g] par principe est la référence en la matière.

qui soulignent les problèmes, ils forment une force de proposition pour permettre à notre société de s'améliorer.

— Les entreprises peuvent avoir intérêt à ouvrir leurs données, comme outil de communication (données économiques à destination des actionnaires par exemple), pour redorer leur image, pour promouvoir l'usage de leurs produits (composants, manuels, spécifications techniques...) voire pour inciter d'autres entreprises à innover en s'appuyant sur leur activité.

— Les journalistes, les chercheurs, les associations travaillent sur des données ouvertes ou ouvrables [65] pour y retrouver des problèmes souvent constatés par ailleurs. Un tel travail, fondé sur des données publiques, a une portée nettement supérieure à un simple ressenti.
La publication en libre accès de leurs travaux et leur reproductibilité [66] participent aussi à la transparence.

— Le citoyen peut apaiser les débats, démonter une fausse rumeur, en utilisant pleinement les informations sur Internet. C'est l'étape qui suit le réflexe Wikipédia pour trancher lorsqu'on n'est pas d'accord sur un point simple. Elle demande certes de creuser plus en profondeur, mais c'est toujours possible avec une connexion à Internet et un peu de temps.
Le citoyen peut aussi ajouter de l'information, l'exemple le plus marquant étant la mise en ligne d'un témoignage vidéo dont on a vu l'importance dans la dénonciation des violences policières.

La palette est large, chacun peut participer.

65. Parfois il faut forcer la porte car avoir la loi pour soi ne suffit pas toujours.
66. C.-à-d. la mise à disposition de tous les éléments qui mènent aux résultats.

Chapitre 2

État des lieux

Faire un état des lieux demande de regarder dans différentes directions : regarder l'existant, ce qui est déjà en ligne, sous quel format et avec quels droits, ainsi que les indices qui permettent d'imaginer le futur, comme les communs (section 2.3) et la gouvernance ouverte (section 2.4).

Les aspects légaux concernant les données et les codes sources sont importants et complexes, ils seront abordés dans la section 2.1. La liste de ce qui existe déjà permet d'apprécier la diversité des données ouvertes et leur potentiel, elle sera présentée dans la section 2.2. Mais avant tout, commençons par ce qu'est une donnée.

C'est quoi une donnée ?

Une donnée est un point de départ connu pour mener un raisonnement. Les données qui nous intéressent sont des descriptions élémentaires d'une réalité. On les appelle brutes lorsqu'elles n'ont pas subi de traitement mathématique, humain ou autre. Une donnée a un type (nombre, texte, son, image…). Elle a aussi un format qui décrit comment la stocker. Pour simplifier, on ne parlera pas du format d'une donnée, mais seulement du type d'une donnée et du format qui groupe un ensemble de données dans un document (PDF, Excel, csv…).

Il existe différents assemblements de données. Il y a des données structurées : celles d'un tableur par exemple, des données non structurées : un roman, une vidéo, ainsi qu'un continuum de données semi-structurées entre les deux. Les premières sont facilement exploitables par un ordinateur, les secondes moins. Il existe

aussi des données naturellement numériques et des données numérisées. Ce qui est écrit avec un clavier est naturellement numérique et n'est pas une approximation comme l'est une photo ou une bande sonore stockée sur un ordinateur [1]. Là encore l'ordinateur est à l'aise avec les premières, mais souffre pour comprendre les secondes (les progrès de l'intelligence artificielle devraient changer cette situation rapidement). Un troisième critère est celui de la taille. Un site web, un rapport en PDF, un tableur pas trop grand sont exploitables par un humain. À l'inverse une base de données [2] avec des milliards de chiffres demande d'utiliser un outil pour pouvoir extraire de l'information. Dans le premier cas, le grand public peut exploiter les données ; dans le second cas, il faut avoir des compétences spécifiques. En regardant ces critères, on voit que l'analyse d'un ensemble de textes administratifs à partir de millions de photocopies revient à traiter des données non structurées, numérisées et de grande taille, ce qui est très difficile pour un ordinateur et impossible pour un humain.

Si la taille augmente la difficulté d'analyse par un humain, elle offre une meilleure vision de la situation et peut permettre de trouver plus d'information que prévu. On peut découvrir des informations cachées, voire inconnues. Une analogie pourrait être une photo de la mer. En très basse résolution, c'est un gros pixel bleu. C'est une information tout comme le PIB est une information sur un pays, c'est grossier. Avec davantage de données, donc davantage de pixels, on commence à voir des vagues et des reflets. En augmentent la résolution, un avion militaire peut apparaître dans les reflets, ce que ne désiraient pas montrer les diffuseurs de l'image. Enfin, à très haute résolution, on va peut-être découvrir une personne qui se noie, ce que personne ne savait.

Le format d'enregistrement de ces données a aussi son importance. Un texte peut être stocké dans un fichier simple, sous Word ou en PDF. Le premier format est naturellement lisible par un logiciel, le second demande de connaître le format de Word quant au troisième, il ressemble plus à une image donc à un document nu-

1. La représentation d'une photo par un tableau de pixels ne montre pas ce qu'il y a entre deux pixels, c'est une approximation de la réalité représentée.

2. Une présentation des bases de données est proposée page 88, section SQL.

mérisé ce qui rend son traitement difficile (c'est surprenant, mais c'est le cas). Des données numériques [3] structurées peuvent être enregistrées dans un tableur ou dans une base de données. Si la taille du jeu de données est grande, le tableur sera difficilement exploitable, car il n'est pas fait pour cela.

On comprend alors que si une administration est forcée de publier des données, elle peut s'arranger pour que leur exploitation soit très difficile. C'est pour cela que la loi qui encadre l'ouverture des données publiques souligne que le format choisi doit faciliter leur réutilisation.

La technique a donc son importance même si elle n'est pas le facteur déterminant pour aller vers une société plus transparente. La technique est la logistique de l'armée, elle ne combat pas, mais elle ne doit pas être négligée pour autant.

2.1
Le droit

En matière de diffusion des données, le droit comporte deux volets : l'un porte sur l'obligation de diffuser, l'autre sur les modalités de cette diffusion. Des textes règlementent les obligations faites aux administrations de publier les données dont elles disposent et en décrit les usages autorisés avec les licences d'utilisation.

Les lois

Obligation de publication

Le code des relations entre le public et l'administration (CRPA)[g] est la référence concernant la publication de données publiques. Créé en 2015, il reprend les éléments de la loi n° 78-753 du 17 juillet

3. Normalement une *donnée numérique* couvre tout ce qui est stocké sur un ordinateur puisqu'il s'agit de 0 et de 1, mais ici, on appelle données numériques celles qui sont composées de nombres ou de champs prédéfinis que ce soient des bilans financiers, des résultats de sondages, des relevés de consommation, etc.

1978, dite loi CADA[g] et des autres lois concernées. Il a été largement modifié en 2016 par la loi pour une République numérique[g].

Dans son livre III, il stipule que les administrations peuvent diffuser leurs données au grand public, avec les réserves qu'on verra, et doivent les transmettre à qui en fait la demande, sauf dans le cas où ces données sont déjà diffusées publiquement.

Les administrations concernées par ce code sont

— l'État,
— les collectivités territoriales,
— les entités chargées d'une mission de service public.

Les documents concernés sont les dossiers, rapports, comptes-rendus, études, procès-verbaux, statistiques, instructions, avis, circulaires, notes et réponses ministérielles, correspondances, prévisions, codes sources et décisions que l'administration publie ou reçoit. Les bases de données régulièrement mises à jour et les codes sources et algorithmes qui prennent des décisions sont également concernés. Cependant, seuls sont concernés les documents en lien avec la mission de l'administration. Sont exclus les documents relatifs au seul fonctionnement interne.

De plus doivent être occultées les parties de documents administratifs dont la consultation ou la communication porterait atteinte :

— au secret des délibérations du Gouvernement et des autorités responsables relevant du pouvoir exécutif,
— au secret de la défense nationale,
— à la conduite de la politique extérieure de la France,
— à la sûreté de l'État, à la sécurité publique, à la sécurité des personnes ou à la sécurité des systèmes d'information des administrations,
— à la monnaie et au crédit public,
— au déroulement des procédures engagées devant les juridictions ou d'opérations préliminaires à de telles procédures, sauf autorisation donnée par l'autorité compétente,
— à la recherche et à la prévention, par les services compétents, d'infractions de toute nature ou sous réserve de l'ar-

ticle L 124-4 du code de l'environnement, aux autres secrets protégés par la loi,

mais aussi

— à la vie privée,
— au secret médical,
— au secret des affaires.

Enfin les parties portant un jugement sur une personne identifiable ou faisant apparaitre un comportement dont la divulgation pourrait lui porter préjudice doivent aussi être occultées.

Le résultat doit être publié de façon à être facilement réutilisable, ce qui implique d'utiliser des formats ouverts.

Les administrations de plus de 50 personnes, à l'exclusion des collectivités de moins de 3 500 habitants, doivent publier en ligne les documents demandés.

La procédure de diffusion est double : soit l'administration publie d'elle-même des données, soit elle le fait en réponse à une demande d'un citoyen ou d'une autre administration. Une fois que les données sont publiées, l'administration n'a plus à les envoyer aux personnes qui en font la demande (un lien vers le lieu où sont les données suffit). Si les données évoluent avec le temps, il est bien sûr légitime d'en demander la nouvelle version.

La liste des données déjà diffusées est bien trop longue [4] pour être citée ici, mais en voici cependant quelques éléments en vrac :

— le cadastre, les parcelles des AOC, les PLU
— les transports,
— la base de données nationale des vitesses maximales autorisées sur le domaine public routier,
— la qualité de l'environnement : air, eau, sol, paysage, sites naturels, zones côtières et marines...
— les subventions publiques,
— les marchés publics,
— les budgets,

4. Plus de 40 000 jeux de données sur `https://data.gouv.fr` en 2022 et d'autres ailleurs.

— les délibérations des conseils municipaux, des départements, des régions,
— les arrêtés municipaux, préfectoraux,
— les comptes des associations ayant obligation de les publier,
— les statistiques nationales et locales [5],
— les statistiques de Parcoursup, du bac et d'autres concours et examens école par école,
— les résultats d'élections,
— la production d'énergie, par type, par lieux,
— l'annuaire de l'administration [6],
— l'emplacement des défibrillateurs public.

Si cette liste – surtout celle sur `data.gouv.fr` – semble impressionnante, elle ne représente qu'une petite partie des données publiques. Beaucoup reste à faire. En juin 2021, seuls 11,12 % des organismes et collectivités concernés par la loi ont ouvert leurs données [7]. De plus certains jeux de données ne sont pas mis à jour, ou disparaissent, ou sont publiés sous un format difficilement exploitable. Si globalement la situation est en net progrès sur les dix dernières années, il reste du chemin à parcourir tant en quantité que dans la qualité et la pérennité des données publiées.

C'est dans ce cadre que la circulaire n°6264/SG du 27 avril 2021 relative à la politique publique de la donnée, des algorithmes et des codes sources prend toute son importance. Non seulement elle rappelle aux services de l'État le devoir d'ouverture, mais elle impose la création d'un AMDAC (administrateur général des données, algorithmes et codes sources) pour chaque ministère et l'équivalent pour les préfectures. Ces responsables doivent faire appliquer le principe de l'ouverture et favoriser les échanges entre les administrations. L'ensemble des AMDAC sont chapeautés par

5. Des statistiques européennes sont disponibles sur `https://ec.europa.eu/eurostat/`

6. La base de l'annuaire est sur `data.gouv.fr`, une interface est disponible sur `https://lannuaire.service-public.fr/`

7. Chiffre de septembre 2020 de l'observatoire open data des territoires, `http://www.observatoire-opendata.fr/resultats/`

l'AGDAC (administrateur général...), qui dirige aussi la direction interministérielle du numérique (DINUM)[g].

Si cette nouvelle structure devrait permettre d'améliorer l'efficacité de nos administrations, il sera important de veiller à ce que les données publiques et les algorithmes qui les traitent ne se retrouvent pas enfermés dans les ministères, sous prétexte de protection de la vie privée ou d'autres fausses raisons qui reviendront à réduire la transparence.

> *« La politique de la donnée doit constituer une priorité stratégique de l'État dans ses relations avec tous ses partenaires, notamment les collectivités territoriales et les acteurs privés »*
>
> Circulaire n°6264/SG du 27 avril 2021

Création de nouvelles données publiques

Publier l'existant est certes nécessaire, mais il arrive souvent que les données recherchées n'existent pas. Il faut donc les créer, ce qui peut nécessiter une nouvelle loi.

Ainsi la base de données Transparence-Santé [8] a-t-elle été créée avec mission d'enregistrer tous les cadeaux offerts au personnel soignant à l'occasion du scandale du Mediator. Malheureusement les données en sont incomplètes et de mauvaise qualité, ce qui illustre encore le fait qu'il ne suffit pas que des données soient publiées pour permettre la transparence. Reste qu'une base mal faite vaut mieux que rien, car on pourra toujours la nettoyer et l'améliorer [9].

Une autre illustration est offerte par la création des données déclaratives d'intérêt des personnalités politiques. Elle a fait suite à l'affaire Cahuzac et permis la promulgation de la loi n°2013-907 du 11 octobre 2013 relative à la transparence de la vie publique. Elle fait obligation à certains responsables publics (candidat à l'élection présidentielle, membres du Gouvernement, parlementaires, membres des cabinets et quelques autres) de procéder auprès de

8. https://www.data.gouv.fr/fr/datasets/transparence-sante-1/
9. cf. le travail de Euros For Doc, https://eurosfordocs.fr/.

la HATVP[g] à une déclaration d'intérêts et / ou à une déclaration de patrimoine. On a ainsi créé de nouvelles données avec ces déclarations qui permettent, en fin de mandat, de vérifier que la personne n'a pas abusé de sa position pour servir son intérêt personnel. La loi ajoute que c'est aussi le moment de vérifier que la nouvelle situation de responsable public n'entre pas en conflit avec d'autres activités.

Notons que les premières déclarations ont été l'enjeu de combats entre certains députés et les défenseurs de cette loi. Les députés voulant limiter le plus possible leur publication ont obtenu que ces déclarations soient manuscrites et que seules des images soient mises en ligne. En réaction, le collectif Regards Citoyens[g] a mis en place un site web afin de créer une base de données à partir des images. Pour cela, le site proposait des images des déclarations avec à côté un formulaire dans lequel l'internaute recopiait ce qu'il arrivait à lire. En donnant plusieurs fois les mêmes images à différents internautes, la base a été rapidement remplie en minimisant le risque d'erreur [10]. Maintenant les déclarations sont remplies en ligne et les données diffusées par la HATVP[g] sont dans un format exploitable par des ordinateurs. On peut ainsi connaitre facilement les revenus moyens des députés, avoir la liste des personnes qui cumulent le plus, etc.

Les licences

Une licence fixe les conditions d'utilisation d'une production : logiciel, document, données... Elle est définie par le propriétaire de la production dans le respect de la loi (tout ne peut pas être diffusé, toutes les clauses ne sont pas légales). Elle peut être dite ouverte ou fermée. Une licence fermée n'autorise pas la diffusion des données brutes dans le domaine public. À l'inverse une licence ouverte (ou libre) autorise la rediffusion et permet de travailler librement les données afin d'en faire ressortir un problème, ou simplement pour

10. Cette technique qui fait travailler en parallèle un grand nombre de personnes pour remplir un but s'appelle le *crowdsourcing* ou la production participative. Il s'agit d'une technique très efficace pour lutter contre ce type de blocage.

aider à leur compréhension [11].

Il est important de noter que toutes les données librement accessibles sur Internet ne sont pas ouvertes, c'est-à-dire ne sont pas sous une licence ouverte. Un grand nombre sont proposées avec des conditions d'utilisation extrêmement fermées. Regardons différents cas à travers l'exemple d'une personne qui désire acheter une maison.

Outils d'aide à l'achat d'une maison

— Les sites de vente en ligne comme SeLoger (annonces de professionnels) ou leboncoin (pro et particuliers) permettent de voir des biens en vente et de se faire une première idée.

— Géoportail avec ses différentes cartes permet de localiser la maison recherchée. Pour y arriver, il faut imaginer à quoi peut ressembler le toit de la maison, les éléments qui l'entourent (route, arbre...) et chercher avec l'image satellite. Le calque carte IGN est aussi très utile, car plus lisible lorsqu'on veut prendre en compte le relief, une rivière ou un élément caractéristique comme un terrain de tennis.

— Avec Google Street View on regarde depuis la rue si c'est bien la maison trouvée est la bonne. On peut en profiter pour se "promener" dans le quartier.

— Vient ensuite le cadastre. Il permet de vérifier que les bâtiments ont une existence légale. L'outil de mesure de surfaces de Géoportail aide à trouver l'ensemble des parcelles du bien lorsqu'on connaît sa surface globale. Le cadastre est téléchargeable sur `data.gouv.fr` [12] et consultable sur GéoPortail.

— Enfin la base de Demandes de valeurs foncières (DVF) [13] donne le prix des maisons proches vendues récemment, ce

11. Une liste des stations de carburant avec les prix dans un fichier texte est difficilement exploitable par contre une application qui affiche les prix par catégorie sur une carte est nettement plus utile. La licence associée au fichier d'origine peut en autoriser, ou en interdire, son usage par une application.

12. `https://www.data.gouv.fr/fr/datasets/cadastre/`

13. Dont voici une interface : `https://app.dvf.etalab.gouv.fr/`

qui sera utile pour la négociation.

Lorsque qu'on a trouvé une maison intéressante, on peut appeler la mairie pour avoir le nom du propriétaire (c'est une information ouverte donc on doit vous la donner). Ainsi, il est possible de contacter directement le vendeur sans avoir à payer des frais d'agence.

Regardons les droits d'utilisation des différents sites cités. Ils sont indiqués dans les conditions générales d'utilisation (CGU) [14].

— SeLoger interdit la copie même partielle de son site, mais indique quand même que l'on peut regarder le site (on a dû leur expliquer qu'accéder à une page web oblige à avoir une copie sur son ordinateur). Par contre, il n'est pas permis d'envoyer une description ou une photo d'une maison à un proche.
— Leboncoin est encore plus absurde puisqu'il interdit en plus les liens hypertextes vers son site sans son accord écrit, donc pas de mail à sa mère avec un lien vers une annonce intéressante. De toute façon le site interdit de faire une recherche pour un autre.
— Google Street View est nettement plus souple puisque, par défaut, on peut faire ce qu'on veut avec les images qu'il montre tant que ce n'est pas à but commercial ni de reconstitution 3D et qu'on cite la source. De plus, il permet les liens et en génère de simples à transmettre sur demande.
— GéoPortail est un site public édité par l'IGN[g] qui diffuse des données dans le cadre de sa mission de service public. Cela étant, il ne permet pas « le stockage local des données au-delà de la session ni la réutilisation des contenus ». Il est néanmoins autorisé de faire des copies d'écran pour un blog ou un site web, pour impression ou pour un usage documentaire (avec des limites sur la taille des images).
— Les données du cadastre sur GéoPortail ont les mêmes contraintes donc on ne peut pas les stocker. Ces mêmes données sont en format ouvert sur le site de data.gouv.fr ce qui permet

14. CGU lues en novembre 2020.

d'en faire ce qu'on veut, y compris un usage commercial, tant qu'on cite la source.

— La base de données de transactions immobilières [15] est aussi diffusée sous licence ouverte. On notera qu'il est néanmoins indiqué que la réutilisation « ne doit pas permettre la réidentification des personnes concernées, de manière indirecte ». Ce point est délicat puisque les informations sur les transactions sont liées aux parcelles cadastrales et que la loi permet de savoir à qui appartient une parcelle. Une interprétation est qu'il ne faut pas simplifier plus l'identification. On retrouve la limite entre la transparence et la vie privée.
L'application citée pour visualiser les parcelles vendues est développée par Etalab[g]. Son code source[g] est référencé sur site des codes publics : `https://code.gouv.fr/`. Sa licence libre permet d'en faire ce qu'on veut dès lors que les parties non modifiées restent sous la même licence.

On voit qu'il existe une grande variété de licences liées aux données (et aux logiciels) et qu'avoir accès avec son navigateur à une donnée ne signifie pas qu'on peut l'utiliser librement. Seule une donnée ouverte ou libre peut être réutilisée pour un usage personnel ou professionnel sans trop de contrainte. Il en est de même pour un logiciel dont on dispose du code source.

Code source

Sur Internet les premières licences ouvertes ont été écrites pour les logiciels. Il s'agissait des licences GPL[g] écrite en 1989 par la Free Sotfware Foundation (FSF)[g] et BSD[g] écrite en 1990 par l'université de Berkeley. Le but de la première est de garantir qu'un code ouvert le reste ainsi que les codes dérivés (aspect viral appelé copyleft ou gauche d'auteur) alors que la seconde donne la liberté de réutiliser comme on le souhaite le code sous cette licence. Depuis, ces licences ont évolué et ont eu de nombreuses petites sœurs [16]. L'État

15. `https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/`

16. `https://fr.wikipedia.org/wiki/Liste_de_licences_libres`, liste non exhaustive

français en reconnait 10 pour les logiciels qu'il diffuse dont 3 GPL et 2 BSD (il existe des variantes et des versions différentes)[17].

Les licences libres garantissent au minimum quatre libertés, que l'on retrouve pour les documents et les bases de données :

— liberté d'utilisation,
— liberté de modification,
— liberté de redistribution,
— liberté de publication.

La redistribution est celle du code d'origine quand la publication est la diffusion du code modifié.

On utilise aussi le terme d'*ouvert* pour logiciels et les données. La différence avec le terme *libre* est nulle dans la majorité des cas. Les puristes pourront étudier les différences sur le site de l'Open Source Initiative (OSI)[g].

Lorsqu'on parle de transparence, il faut aussi penser au code source[g]. Son rôle est décisif pour comprendre le fonctionnement des décisions automatisées, d'où l'importance de le publier. Le code source permet également de créer des logiciels dérivés. Ainsi dans l'exemple de la recherche de maison ci-dessus, le travail d'Etalab peut être recopié et modifié pour présenter d'autres données qui s'appuieraient sur le cadastre.

Élargissons

Si les logiciels libres existaient avant qu'Internet soit ouvert au grand public, les licences libres pour les autres types de production sont arrivées plus tard. La première licence libre marquante pour les documents et plus largement pour les œuvres culturelles est la Creative Common (CC) écrite en 2002. Elle reprend les principes de base des licences libres pour les logiciels auxquels elle ajoute l'obligation de citer l'auteur (BY) et offre trois options qui permettent de créer sa licence à la carte en retirant des droits :

NC pas d'utilisation commerciale. (il est interdit de revendre ou revendre une œuvre dérivée d'une œuvre ainsi protégée),

17. https://www.data.gouv.fr/fr/licences

SA partage dans les mêmes conditions aussi appelé copyleft (la licence utilisée pour retransmettre l'œuvre ou une œuvre dérivée doit être la même. Attention, même sans cette option l'œuvre ou une œuvre dérivée ne peut pas être partagée avec une licence moins restrictive que l'originale),

ND pas d'œuvre dérivée (donc pas de modification, mais on peut l'inclure dans une autre œuvre)

La licence incluant toutes les options (`CC-BY-NC-SA-ND`) est très proche du copyright, il est donc préférable de choisir ce dernier.

La CC0 (Creative Common Zero), ajoutée en 2009, est une licence « zéro droit », qui permet de mettre sa production dans le domaine public dans la limite de ce qu'autorise la loi [18]. Cette licence est conseillée par la FSF[g] et l'Open Knowledge Foundation (OKFN)[g] pour abandonner tout droit.

Pour les jeux de données que l'on veut libérer sans pour autant les mettre dans le domaine public, il existe des licences adaptées établies par l'OKFN :

— Open Data Commons Attribution License [19] : elle demande de transmettre les informations permettant de retrouver la base d'origine et sa licence. En Europe la droit d'auteur ne s'applique pas aux base de données ce qui fait que le BY n'a pas le même sens que pour la Creative Commons.

— Open Data Commons Open Database License (ODbL) [20] : en plus du point précédant, il est demandé de partager avec la même licence, presque comme la `CC-BY-SA` sauf que la contrainte de licence identique ne se propage qu'aux bases de données, ce qui permet de faire une œuvre dérivée qui n'est pas une base de données (mais qui utilise la base de données) avec une licence totalement différente.

18. En France où le droit de paternité existe et où le droit moral est inaliénable, la CC0 n'a pas la même saveur que dans les pays anglo-saxons. Cela peut être un problème lorsque le document en CC0 circule sur Internet.

19. `https://opendatacommons.org/licenses/by/`

20. `https://opendatacommons.org/licenses/odbl/`

— Open Data Commons Public Domain Dedication and License (PDDL) [21] : domaine public.

Données publiques

L'État a fait évoluer les règles de réutilisation des données publiques ces dernières décénies pour finalement les ouvrir [22].

> *« Toute mise à disposition effectuée sous forme électronique en application du présent livre se fait dans un standard ouvert, aisément réutilisable et exploitable par un système de traitement automatisé. »*
>
> Article L300-4 du CRPA[g]

La liste des licences possibles pour les administrations est disponible sur `https://www.data.gouv.fr/fr/licences`. La licence de base est celle conçue par Etalab[g] à savoir la « Licence Ouverte / Open License » [23].

> *« La « Licence Ouverte / Open License » présente les caractéristiques suivantes :*
>
> 1. *Une grande liberté de réutilisation des informations :*
>
> — *Une licence ouverte, libre et gratuite, qui apporte la sécurité juridique nécessaire aux producteurs et aux réutilisateurs des données publiques ;*

21. `https://opendatacommons.org/licenses/pddl/`

22. En 1978 la loi Commission d'accès aux documents administratifs (CADA)[g] définit le droit d'accès aux documents administratifs. L'évolution commence avec la directive 2003/98/CE du 17 novembre 2003, dite directive PSI pour *Public Sector Information*. Elle a été transposée en 2005 par décret puis la directive évoluant, elle a donné la loi du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, dite loi Valter. Enfin la loi pour une République numérique du 7 octobre 2016 ouvre les données publiques.

23. `https://www.etalab.gouv.fr/licence-ouverte-open-licence`.
Note : l'anglais US écrit *License*, avec un *s* alors que le français et l'anglais britanique l'orthographient avec un *c*. L'usage de plus courant en anglais est avec un *s*.

— *Une licence qui promeut la réutilisation la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ;*

— *Une licence qui s'inscrit dans un contexte international en étant compatible avec les standards des licences* open data[g] *développées à l'étranger et notamment celles du gouvernement britannique (Open Government Licence) ainsi que les autres standards internationaux (ODC-BY, CC-BY 2.0).*

2. *Une exigence forte de transparence de la donnée et de qualité des sources en rendant obligatoire la mention de la paternité.*
3. *Une opportunité de mutualisation pour les autres données publiques en mettant en place un standard réutilisable par les collectivités territoriales qui souhaiteraient se lancer dans l'ouverture des données publiques.*

»

Licence Ouverte / Open License version 2.0

On peut également utiliser la licence ODC Open Database License (ODbL) version 1.0. Ces deux licences ont en commun de permettre une réutilisation très large, y compris commerciale, la contrainte principale étant la citation des sources. Notons que la licence d'Etalab souligne sa compatibilité avec d'autres licences ouvertes à travers le monde, ce qui facilite le travail de ceux qui veulent travailler sur les données d'autres pays ayant d'autres licences ouvertes. D'autre part les options NC, SA et ND de la licence CC, ne sont pas citées, car incompatibles avec l'esprit de l'ouverture des données, à savoir permettre la libre réutilisation et la création de valeur économique.

Pour conclure, voici la liste des licences les plus utilisées par les collectivités territoriales en septembre 2019 pour leurs données

ouvertes [24] :

Licence Ouverte / Open License v 2.0	42,7 %
Open Data Commons Open Database License	24,0 %
Licence Ouverte / Open License	18,5 %
Aucune licence	13,6 %
Autres	1,2 %

Il est regrettable qu'autant de données figurent sans indication de licence. Si cela ne change pas leur statut de données ouvertes, qu'en serait-il si la loi venait à changer ?

2.2
Des données ouvertes

Pour mieux comprendre l'importance des données ouvertes, quittons la technique et la loi pour entrer dans le concret avec des exemples. Regardons la diversité des données déjà en ligne, leurs formes et leurs auteurs.

Les données publiques

Le site `data.gouv.fr`

Le site web `data.gouv.fr`, géré par Etalab[g], est le portail des données publiques ouvertes en France. Il regroupe plus de 40 000 jeux de données en 2022. Environ 3 700 organisations y publient leurs données dont presque 600 déclarées comme étant de service public. Parmi elles, les trois plus importants sont :

1. La région Île-de-France avec 934 jeux de données [25],
2. Le ministère de l'Intérieur avec 665,
3. Toulouse métropole avec 614.

24. cf `http://www.observatoire-opendata.fr/resultats/`

25. Les directions départementales des territoires proposent parfois plus de jeux de données. mais, étrangement, elles ne sont pas référencées comme organisation de service public.

On trouve des services de l'État comme l'INSEE[g] ou le Système d'information sur l'eau (SIE)[g] et des entreprises d'État comme la SNCF. À noter que l'INSEE qui ne propose que 44 jeux de données est l'organisation qui a le plus d'abonnés sur le site.

1. L'INSEE a 618 abonnés,
2. Le ministère de l'Économie et des finances en a 401,
3. Le ministère de l'Intérieur et des outre-mer en a 360.

Un autre critère est le nombre d'utilisations déclarées d'un jeu de données. Avec ce critère, on obtient le classement suivant :

1. Données hospitalières relatives à l'épidémie de COVID-19, 144 réutilisations,
2. Première Guerre mondiale - Les Poilus morts pour la France, 62 réutilisations,
3. Demandes de valeurs foncières, 56 réutilisations.

56 jeux de données sont souvent réutilisés, 79 assez réutilisés, 1 300 une ou deux fois et les 96 % restant n'ayant pas été déclarés réutilisés. Notons que l'analyse des données brutes des jeux de données stockés stockées sur `data.gouv.fr` ne donne pas exactement les mêmes valeurs de réutilisation que celles du classement pris sur le site web.

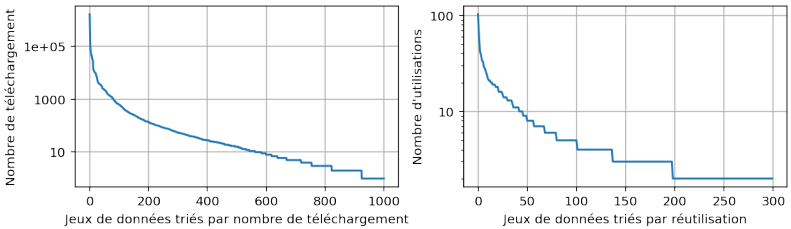

FIGURE 2.1 – Téléchargements et utilisations déclarées des jeux de données de `data.gouv.fr`. Décroissances plus qu'exponentielles !

Même en prenant en compte l'erreur de mesure, on peut être déçu par ces chiffres qui ne montrent pas un engouement massif pour les données publiques ouvertes. Cette impression est en

partie juste, d'où ce livre. Cela étant, il faut noter que tous les utilisateurs ne se déclarent pas et qu'un jeu de données peut être intégré dans des applications pouvant avoir des millions d'utilisateurs. Ainsi l'application *Essence & Co*, qui permet de connaître le prix de l'essence dans les stations les plus proches, a plus de 1,5 million d'utilisateurs. Elle s'appuie sur la base de données *Prix des carburants* [26] et ne compte que pour une seule réutilisation.

Parfois une application a un nombre plus modeste d'utilisateurs, mais ces derniers sont des intermédiaires. C'est le cas de CommoPrices [27], un service qui donne le cours des matières premières en s'appuyant, entre autres, sur le jeu de données *Statistiques nationales du commerce extérieur*. Cela lui permet d'avoir des prix de matières premières absentes des cours de bourse. Ainsi, ses entreprises clientes peuvent mieux négocier leurs achats de produits destinés au grand public. Là encore, avec une seule réutilisation impacte un grand nombre d'utilisateurs.

Concernant les organisations qui déposent des données, la très grande majorité est « de service public » ou « liée » au service public. Ainsi les DTT [28] sont d'importants fournisseurs de données. Des associations comme Regards Citoyens[g] et l'OKFN[g] France publient aussi leurs données. L'entreprise Navitia [29] dont le métier est de proposer l'accès à des données liées au transport, propose des jeux de données qui agrègent les informations de différents réseaux de transport. De nombreuses autres organisations connues, Pôle Emploi, la RATP, l'IGN[g], le CNC [30], l'ADEME [31], Météo France et d'autres, sont aussi présentes sur `data.gouv.fr` [32].

Malheureusement toutes les données publiques publiées ne figurent pas sur ce site, même lorsqu'elles appartiennent à des organisations inscrites. Il faut alors aller chercher sur les sites des

26. `https://www.data.gouv.fr/datasets/prix-des-carburants-en-france/`
27. `https://commoprices.com/`
28. Direction départementale des Territoires
29. `https://navitia.io/`
30. Centre national du cinéma et de l'image animée
31. Agence de la transition écologique
32. Le moteur de recherche du site permet de trouver les organisations inscrites.

organisations concernées ce qu'elles n'ont pas déposé sur le portail. Il arrive aussi que les données soient à jour sur le site web de l'organisme et obsolètes sur `data.gouv.fr` [33]. Plus généralement la mise à jour des données est faible, moins de la moitié ont été mises à jour durant l'année, pour les administrations centrale ce mauvais score monte à 60 %. Enfin la majorité des acteurs publics ne publient pas encore leurs données [34]. La tâche est plus ardue que prévu pour avoir des données propres, complètes, cohérentes et maintenues à jour. Néanmoins, le mouvement continue avec, peut-être, une plus grande attention pour les données sites de référence, quitte à délaisser d'autres données moins fondamentales.

Données publiques en dehors de `data.gouv.fr`

D'autres organismes qui disposent de jeux de données ouverts :

— L'Union européenne
`https://data.europa.eu/euodp/en/data`
— Le portail des données ouvertes des pays européens
`https://www.europeandataportal.eu/`
— La banque mondiale avec une grande variété de données
`https://data.worldbank.org/`
— Data World, une entreprise qui agrège des données de tout types du monde entier
`https://data.world/`
— Météo France offre certaines données et en vend d'autres
`https://donneespubliques.meteofrance.fr/`.
— Des données financières par la Banque de France
`http://webstat.banque-france.fr/`

Les régions ont leur portail :

— Auvergne-Rhône-Alpes : `https://www.datara.gouv.fr/`
— Bourgogne-Franche Comté :
`https://ideo.ternum-bfc.fr/`

33. Par exemple l'annuaire de l'éducation est à jour sur `https://data.education.gouv.fr/` mais périmé sur `data.gouv.fr` (vérifié le 03/01/22)

34. Point développé dans la section 2.2 du Rapport de la mission Bothorel[g].

— Bretagne : https://data.bretagne.bzh/
— Centre-Val de Loire : http://data.centrevaldeloire.fr/
— Corse : http://www.data.corsica/
— Grand Est : http://datagrandest.fr/
— Haut de France : http://opendata.hautsdefrance.fr/
— Ile-de-France : http://data.iledefrance.fr/
— Normandie : http://opendata.normandie.fr/
— Nouvelle-Aquitaine : https://portail.pigma.org/
— Occitanie : http://data.laregion.fr/
— Pays de la Loire : http://data.paysdelaloire.fr/
— Provence-Alpes-Côte d'Azur : https://www.datasud.fr/

De nombreux pays aussi dont nos voisins :

— Le Royaume-Uni : https://data.gov.uk/
— La Belgique : https://data.gov.be/
— L'Allemagne : https://www.govdata.de/
— La Suisse : https://opendata.swiss/fr/
— L'Italie : https://www.dati.gov.it/
— L'Espagne : https://datos.gob.es/
— Les États-Unis : https://data.gov/

Il peut être intéressant de regarder leurs données publiques à titre de comparaison, voire pour demander les données manquantes en France.

Enfin citons quelques villes :

— Paris : https://opendata.paris.fr/ et sa circulation,
— Strasbourg https://data.strasbourg.eu/ et sa cartothèque,
— Nantes Métropole https://data.nantesmetropole.fr/ pour sa volonté de faire participer les citoyens,
— Boston https://data.boston.gov/ et ses finances dont les impôts locaux avec valorisation et description précise de chaque maison.

Les API

Les différents jeux de données présentés jusqu'à présent sont stockés dans des fichiers statiques. Pour exploiter des données dynamiques [35], il est préférable d'avoir un service qui réponde en temps réel et de façon compréhensible par une machine à travers une interface de programmation d'applications (API)[g]. Il peut être aussi préférable d'utiliser une API pour récupérer une donnée précise lorsque le jeu de données statique sous-jacent est gros, ce qui induit un temps de chargement long.

Voici quelques exemples d'API :

Appels d'offre (BOAMP) `https://api.gouv.fr/les-api/boamp`	Pour interroger les appels d'offre en cours et connaître les résultats.
Base adresse nationale `https://geo.api.gouv.fr/adresse`	Complète une adresse postale. Convertit une adresse vers sa position GPS et l'inverse.
API service public `https://api.gouv.fr/rechercher-api`	Les deux API précédentes et plein d'autres.
INSEE `https://api.insee.fr/`	Données de l'INSEE et l'accès à la base des entreprises Sirene.
Bibliothèque nationale de France `http://api.bnf.fr/`	Interroger et télécharger les catalogues et les collections numérisées de la BnF.
SNCF `https://www.digital.sncf.com/startup/api`	Itinéraire, horaires, prochain train, recherche de gare.

35. Les données dynamiques sont générées à la volée, le plus souvent en temps réel. Par exemple un capteur donne la température actuelle à tel endroit. Les anciens se souviendront de l'horloge parlante.

NASA `https://api.nasa.gov/`	Données astronomiques, les météorites proches de la terre, la météo sur Mars...
Aviation Stack `https://aviationstack.com/`	Information sur tous les vols en cours. API commerciale, mais gratuite pour moins de 500 interrogations par mois.

L'usage de ces API se base le plus souvent sur la technologie REST qui s'appuie sur HTTP (langage réseau du Web). Cela veut dire qu'il suffit d'une URL pour poser sa question au service derrière l'API. Ainsi l'URL qui permet d'obtenir l'adresse de la place Bauvau [36] à Paris et d'avoir ses coordonnées géographiques est [37] :

```
https://api-adresse.data.gouv.fr/search/?q=place+bauvau+paris
```

La réponse est formatée pour une machine, mais reste lisible :


```
{"type": "FeatureCollection",
 "version": "draft",
 "features": [{
       "type": "Feature",
       "geometry": {"type": "Point",
                    "coordinates": [2.316647, 48.87119]},
       "properties": {"label": "Place Beauvau 75008 Paris",
                      "score": 0.7348970163329845,
                      "id": "75108_0799",
                      "type": "street",
                      "x": 649874.54, "y": 6863679.52,
                      "importance": 0.5049198112417776,
                      "name": "Place Beauvau",
                      "postcode": "75008",
                      "citycode": "75108",
                      "city": "Paris",
                      "district": "Paris 8e Arrondissement",
                      "context": "75, Paris, Île-de-France"}}],
 "attribution": "BAN",
 "licence": "ETALAB-2.0",
 "query": "place bauvau paris",
 "limit": 5}
```


36. Avec la faute d'orthographe.
37. À tester dans son navigateur en recopiant l'URL.

En utilisant ce mécanisme de façon répétée, on peut générer ses propres bases de données [38]. C'est encore plus intéressant lorsque l'API transmet des données en temps réel, cf. page 80.

Les œuvres collectives

La création de base de données n'est pas l'apanage des administrations ou des entreprises. Les individus peuvent aussi participer à des projets collaboratifs qui construisent de telles bases.

Wikipédia est le premier exemple qui vient à l'esprit, même si pour beaucoup il s'agit d'une encyclopédie et non d'une base de données. Outre le fait que le sens du mot «données» est large, on a vu page 17 que Wikipédia est aussi une base de données qu'il est possible de télécharger pour la consulter localement. Cela permet aussi des usages innovants. Par exemple des chercheurs en intelligence artificielle l'injecte à des réseaux neuronaux qui apprennent à écrire des articles.

OpenStreetMap[g] est un autre exemple de projet collaboratif. Il s'agit de créer une carte du monde la plus précise possible. Pour cela, chacun devient cartographe et édite la carte en s'appuyant sur sa connaissance de son quartier, sur son GPS et sur des données en ligne comme le cadastre et les images satellites. Dans certains cas des programmes peuvent aspirer automatiquement des données extérieures et modifier la carte. On est donc dans un projet qui mélange la participation à la main et l'automatisation grâce à des outils développés en interne. Comme Wikipédia, OpenStreetMap est utilisé pour de nombreux projets. Certains langages de programmation peuvent même utiliser sa carte pour y superposer des résultats.

Dans le domaine de la nourriture, le projet collaboratif Open Food Facts [39] indexe les produits alimentaires du commerce avec leur composition et des informations nutritionnelles.

Le milieu académique est un autre producteur d'œuvres col-

38. Ce procédé n'est pas toujours légal, cf. les conditions d'utilisation de l'API.
39. `https://fr.openfoodfacts.org/`

lectives ouvertes. Certaines sont scientifiques [40], d'autres sont plus en relation directe avec la transparence démocratique, comme la base mondiale des inégalités [41] mise en avant par Thomas Piketty.

Tous ces exemples produisent des données ouvertes pour des raisons pratiques et idéologiques, mais ce n'est pas toujours le cas. Parfois une entreprise s'appuie sur une large communauté pour développer une base de données privée consultable plus ou moins librement.

FlightRadar24 [42], par exemple, est un service qui permet de localiser en temps réel les avions en vol, avions de passagers, avions-cargos ainsi que quelques avions privés. L'histoire a commencé avec deux passionnés d'aviation qui ont décidé de construire un réseau de récepteurs ADS-B. Ce protocole permet aux avions d'indiquer en permanence aux aéroports leur identité, leur position en même temps qu'un certain nombre d'autres informations. Comme l'achat de récepteurs ADS-B est libre, chacun peut écouter les informations qu'envoient les avions à portée. Aussi FlightRadar24 a mis en place une procédure qui permet aux volontaires de déposer les données de leur récepteur. Ainsi aujourd'hui FlightRadar24 est connecté à plus de 20 000 récepteurs ADS-B qui écoutent le ciel et permettent de générer en temps réel une carte mondiale du trafic aérien. La vision de la carte est gratuite, mais certaines options, comme l'historique des vols, sont payantes, comme l'historique de tous les vols sur les 3 dernières années. Ce service nécessiterait probablement une multitude de démarches fastidieuses.

Faire travailler des bénévoles pour développer son entreprise est évidemment un modèle économique tentant, mais pas toujours simple à mettre en œuvre. L'entreprise doit offrir quelque chose en retour, un quelque chose qui, finalement, peut contribuer à plus de transparence.

Avec cet exemple qui s'appuie sur des récepteurs, on entre dans le monde de l'Internet des objets, qui peut contribuer à la transparence en apportant des mesures en temps réel.

---

40. Fair sharing et re3data en regroupent un très grand nombre, cf. `https://fairsharing.org/databases/` et `https://www.re3data.org/`.

41. `https://wid.world/`

42. `https://www.flightradar24.com`

L'Internet des objets

On prévoyait que l'Internet des objets (IoT en anglais pour *Internet of Things*) révolutionnerait notre monde comme l'a fait l'ouverture d'Internet au grand public, mais pour l'instant, son impact demeure modeste. Cela ne signifie pas qu'il n'a ou n'aura aucun d'impact, c'est simplement que la route suivie n'a pas été celle qu'on avait prévue.

Dans le monde de la transparence, les objets qui nous intéressent le plus sont les capteurs, tout ce qui mesure et observe notre monde comme les récepteurs ADS-B. Ils mesurent la qualité de l'air, la qualité de l'eau, le bruit, la densité de circulation, la ponctualité des transports publics, etc. Ces capteurs sont des témoins neutres de notre environnement et du fonctionnement de notre société. L'exemple de la qualité de l'air est intéressant en termes de transparence, car il met en opposition les activités humaines à la santé publique. Pour certaines industries, la lutte contre la pollution est coûteuse, ce qui peut les pousser à sous-estimer ou même dissimuler les taux de pollution. Les restrictions de circulation lors des pics de pollution génèrent d'autres mécontentements. À l'inverse pour les habitants, la pollution est un danger sanitaire dont il faut se protéger, ce qui nécessite un accès à des mesures fiables. L'ouverture des mesures dans ce domaine a trois conséquences : interdire la publication de fausses mesures, permettre à chacun de constater et mettre les autorités devant leurs responsabilités. Tricher devient impossible avec la publication de mesures locales en temps réel, car les habitants peuvent vérifier que ces données correspondent à leur ressenti ou aux mesures de leurs appareils personnels. La confiance s'installe chez les particuliers lorsque les mesures publiques correspondent à leurs mesures. Elle et se généralise ensuite aux mesures de niveau national ce qui permet aux citoyens, ainsi qu'aux élus, d'avoir une vision globale. Ce partage de l'information oblige les élus à agir en conséquence.

Aujourd'hui la publication des mesures de la qualité de l'air provient des autorités à travers des organismes dédiés ainsi que d'associations qui regroupent les mesures faites par des particuliers. Côté service public, la France a agréé les associations de la

fédération Atmo France, dont AirParif est le représentant en Île de France. Côté associatif, citons Sensor Community [43], un projet allemand qui dépasse largement les frontières. Il existe aussi des expériences locales, comme celle de la ville de Paris qui a déployé ses propres stations de mesures dans le 20e arrondissement [44]. Dans tous ces cas les données sont ouvertes ce qui permet de les regrouper afin d'avoir une vision mondiale comme le propose le site World's Air Pollution [45]. Par contre, certaines entreprises vendent des capteurs dont les mesures ne sont pas ouvertes. Lorsque ces capteurs mesurent la pollution à l'intérieur des habitations c'est compréhensible, cela relève de la vie privée. En revanche, d'autres appareils comme Flow [46] de PlumeLabs mesurent la pollution extérieure. Or PlumeLabs (qui publie pourtant une carte mondiale de la pollution en s'appuyant en particulier sur les données d'Atmo en France), fait de sorte qu'il soit impossible de récupérer librement les mesures des appareils de ses clients. C'est de la captation de données.

Cette captation par les entreprises est encore plus patente dans le monde de la médecine du sport avec les capteurs d'activités qu'utilisent les sportifs. Ces bracelets connectés mesurent le rythme cardiaque, la température, le mouvement et demain, d'autres paramètres. On pourrait penser que les données issues de ces capteurs sont stockées sur l'ordinateur personnel de l'utilisateur. Il n'en est rien, elles sont enregistrées sur les serveurs de l'entreprise commercialisant l'appareil. La raison en est évidente : l'exploitation et la revente de ces données sont sources de profits, aussi les entreprises les gardent précieusement. Pourtant l'intérêt de telles données en termes de santé publique justifierait pleinement leur ouverture (après anonymisation) ainsi que leur transfert à l'assurance maladie.

L'application Strava est une autre illustration de captation des données à des fins commerciales. Elle permet d'enregistrer et de

43. https://sensor.community/fr/

44. https://opendata.paris.fr/explore/dataset/respirons-mieux-dans-le-20eme-donnees-mini-stations/information/

45. https://waqi.info/

46. https://plumelabs.com/en/flow/

partager ses trajets et temps de parcours à pied ou à vélo. Comme l'indique son site web : « Grâce à des millions d'athlètes à travers le monde, le réseau de routes et de chemins sur Strava est sans égal », ce qui est probablement exact sachant que ses 50 millions d'utilisateurs enregistraient 19 millions de parcours par semaine en 2020. Ces données ont une valeur commerciale évidente, si bien que Strava les vend indirectement à ses abonnés en proposant des parcours populaires de proximité. Mais elles ont aussi une valeur publique : elles permettent aux collectivités territoriales d'avoir des informations sur les usages de la route par les cyclistes et piétons et donc d'adapter en conséquence les chemins et les pistes cyclables, de mettre en place des signalisations et d'en tirer un avantage touristique. Strava y a pensé et commercialise Strava Metro [47] qui permet d'accéder aux parcours des utilisateurs de Strava dans une zone géographique pour 0.80 € par utilisateur et par an [48].

Pour souligner le potentiel de ces données, Strava propose une carte d'activité, `https://strava.com/heatmap`, qui montre les chemins les plus fréquentés selon la discipline [49]. Elle est cependant moins précise que les données brutes, elle n'indique ni les temps d'arrêt aux intersections ni le tracé de chaque parcours. La revente des données aux particuliers, la revente aux collectivités, sont des illustrations des avantages économiques de la captation de données.

En réaction, la loi oblige les entreprises à restituer les données personnelles depuis 2016. Il s'agit du droit à la portabilité des données, inscrit dans le Règlement général sur la protection des données (RGPD)[g], qui permet à chacun de récupérer ses données collectées par une application ou un site web. De plus il est possible de demander à ce que les données soient transmises à un organisme de son choix. Dans le cas des capteurs sportifs ou de Strava, les utilisateurs pourraient donc récupérer leurs parcours, leur his-

47. `https://metro.strava.com/`

48. `https://www.lemonde.fr/economie/article/2016/08/06/strava-veut-rentabiliser-les-donnees-fournies-par-les-cyclistes_4979235_3234.html`

49. Ce qui a permis de connaitre l'existence de bases militaires secrètes, les soldats lançant l'application pour leur jogging.

torique cardiaque et leurs autres données personnelles pour les déposer sur un site public ou associatif. Mais il s'agit là d'une hypothèse toute théorique d'autant que la confiance des citoyens est aussi limitée envers l'État qu'envers les entreprises.

C'est ainsi qu'on assiste à un glissement de données d'intérêt général [50] vers le privé. Pourtant, les États avaient une avance importante avec de nombreux capteurs (radars, capteurs météo, caméras de surveillance...) en activité avant même qu'on parle de l'Internet des objets. Mais la démocratisation des objets connectés a rebattu les cartes. Si les associations n'en ont que peu profité, si les collectivités locales de taille importante ont pu mettre en place des capteurs, ce sont surtout les entreprises qui en ont tiré le plus grand avantage. Le pourcentage de données publiques dans l'Internet des objets ne cesse de baisser au profit des données captées par les entreprises.

La ville intelligente

L'utilisation de l'Internet des objets afin d'améliorer le fonctionnement du territoire a été baptisé « la ville intelligente ». Il s'agit d'utiliser les différents capteurs à disposition pour optimiser les flux (transport, électricité, eau, déchets), les usages (occupation des infrastructures sportives, bibliothèques, bureaux) ou l'exploitation du matériel (meilleur rendement, prévention des pannes).

Le domaine de l'électricité permet d'avoir un aperçu des possibilités offertes par l'Internet des objets. Le recours croissant aux énergies douces, solaire et éolien, rend la production moins prévisible. Tout le défi pour les exploitants est de trouver l'équilibre entre production et consommation. C'est là que peut intervenir l'Internet des objets. Un compteur intelligent peut indiquer aux appareils de la maison qu'il y a un surplus de production et donc que c'est le bon moment pour lancer la machine à laver. Pour in-

50. La loi pour une République numérique[g] indique que les données privées d'intérêt général doivent être ouvertes. Toute la question est dans la définition de l'intérêt général. Sont explicitement nommées les données de transport et les données de consommation d'énergie.

citer les consommateurs à jouer le jeu, il suffit que ce surplus de production se traduise en une baisse du prix et que le particulier puisse indiquer à sa machine à laver de démarrer quand le prix sera inférieur à tel niveau. Les voitures électriques pourront revendre leur électricité lorsque les prix seront élevés et permettre ainsi d'amortir les pics. On retrouve le principe des tarifs jour/nuit avec plus de souplesse. En diffusant librement les mesures du réseau en temps réel, on élargit le champ des usages. Des études médicales sur l'électrosensibilité pourront vérifier si le ressenti des malades est corrélé aux flux électriques. Les villes pourront adapter leur activité à celle des habitants, etc.

Autre exemple, l'ouverture des données des capteurs mesurant le trafic routier. Ces mesures [51] permettent à la ville de mieux gérer le trafic, de façon statique en adaptant le réseau routier et de façon dynamique en couplant l'ensemble des mesures aux feux tricolores, afin de favoriser les voies congestionnées [52]. L'ouverture de ces mesures permet à n'importe qui de proposer des améliorations, voire de développer des applications à destination des citadins [53], éventuellement pour développer une activité commerciale.

En fait chaque mesure peut être utilisée directement ou indirectement. Au siècle dernier, on mesurait l'audience des chaînes de télévision grâce aux mesures du réseau d'eau. À la fin du film, les spectateurs allaient aux toilettes ce qui génère un pic de consommation d'eau à cet instant dont on déduit le nombre de spectateurs de la chaîne. Ces effets indirects confortent le plaidoyer de l'*open data*[g] pour l'ouverture généralisée des données, même si l'on ne voit pas à quoi elles vont servir.

---

51. Pour Paris : `https://parisdata.opendatasoft.com/explore/dataset/comptages-routiers-permanents/dataviz/`

52. Testé à Perpignan `https://www.francebleu.fr/infos/societe/perpignan-des-feux-tricolores-auto-regules-en-fonction-du-trafic-1423237990`

53. Les piétons aussi peuvent avoir intérêt à savoir où sont les bouchons et donc les pics locaux de pollution. Les habitants, les commerçants peuvent tirer profit de statistiques établies à partir de ces données pour défendre leurs intérêts.

2.3
Les communs

Les données diffusées gratuitement ont un point faible, elles peuvent disparaitre ou ne plus être mises à jour. Seul leur propriétaire décide. S'il s'agit de données publiques, l'État peut revenir en arrière ou le délégataire de service public qui les génère peut se les approprier. Si les données sont sous licences ouvertes, elles ne peuvent pas être retirées, mais elles peuvent ne plus être mises à jour et devenir rapidement obsolètes. Aussi une solution pour garantir un libre accès à des données publiques à jour, passe par les communs.

Les communs sont des ressources gérées par une communauté pour une communauté. Il s'agit d'un droit entre la propriété privée et le bien public. Historiquement des pâturages, des zones de pêche, l'eau et d'autres ressources communes étaient gérées par ce droit. Par exemple, les éleveurs d'un village se mettaient d'accord sur l'usage des champs communaux, la fréquence d'utilisation de chacun, l'ordre, les devoirs de maintenance, etc. Cet usage du bien était défini par l'ensemble des acteurs concernés, ce qui impliquait une méthode de gouvernance ouverte. Cette gouvernance est fondamentale dans la définition des communs. Sans elle, on parle de biens communs (l'air est un bien commun, mais pas un commun car aucun groupe ne contrôle l'accès à cette ressource).

Si ce système a fonctionné jusqu'au Moyen Âge en occident et fonctionne encore dans d'autres régions du monde, l'appropriation des communs par des grands propriétaires l'a fait disparaitre au fil des siècles. En 1968 un article dans le journal Science intitulé « Tragédie des communs » a détruit le concept en montrant que les communs sont inefficaces économiquement, un acteur pouvant abuser du système pour son avantage personnel. Le néolibéralisme des années 80 a ajouté la touche finale en privatisant des pans entiers des services publics. Pourtant, durant la même période deux initiatives indépendantes ont émergé et permis de remettre les communs sur le devant de la scène :

— Les travaux de l'économiste Elinor Ostrom qui ont montré

que contrairement à ce qu'affirme la « Tragédie des communs », les communs peuvent produire de meilleurs résultats que les biens privés (intérêt personnel) ou le bien public (trop lourd).

— Les logiciels libres qui modernisent le principe des communs et soulignent leur importance pour les biens immatériels.

Aujourd'hui, dans le monde physique, les excès du capitalisme, les problèmes écologiques et sociaux poussent au retour des communs pour s'organiser localement et reprendre le contrôle de ressources locales.

Dans le monde numérique, les communs ayant largement construit Internet, sont un principe pleinement reconnu. De plus, ils n'ont pas souffert de la « Tragédie des communs » puisque les biens numériques sont inépuisables. Il s'agit d'une différence majeure avec les communs historiques à laquelle s'ajoute une différence de communautés : les projets libres sont le plus souvent constitués d'une petite communauté qui gère le projet et d'une grande communauté qui l'utilise [54]. Les données numériques et les logiciels peuvent être utilisés et recopiés sans limites, sans les abîmer. On peut développer son propre projet à partir d'une copie, ce qu'on appelle créer sa branche en informatique [55]. Chacun peut créer sa propre version de Wikipédia et la faire évoluer comme il le désire (dans le respect des licences d'origine). Ainsi un schisme est sans conséquences pour les communs en question. Cela étant, il existe quand même une ressource rare dans un projet libre : les acteurs. Aussi la gouvernance des projets libres est un point particulièrement important qui est d'autant moins négligé que tout schisme réduit le nombre de participants et donc la capacité à continuer le projet. En conséquence, la communauté du libre a développé au fil du temps des formes de gouvernances ouvertes, fondées le plus souvent sur le consensus, qui ont permis à des projets de perdurer

54. Des projets comme Linux, Wikipédia ou OpenStreetMap sont des références de communs numériques. Ils ont tous les trois leur grande communauté d'utilisateurs, leur petite communauté de gouvernance et un système ouvert de gouvernance.

55. La plateforme GitHub[g] a un bouton *fork* sur chaque page de projet pour créer son projet dérivé.

sur des décennies [56].

Les communs numériques à travers les logiciels libres et les données libres sont une réalité fonctionnant pour le mieux. Aussi la consultation publique préparatoire à la loi pour une République numérique[g] a proposé de définir dans la loi les communs numériques, pour les ressources communes appartenant au domaine public. Le but était d'éviter les appropriations.

> *« Article 8 - Définition positive du domaine commun informationnel*
>
> *I. Relèvent du domaine commun informationnel :*
>
> 1. *Les informations, faits, idées, principes, méthodes, découvertes, dès lors qu'ils ont fait l'objet d'une divulgation publique licite notamment dans le respect du secret industriel et commercial et du droit à la protection de la vie privée, et qu'ils ne sont pas protégés par un droit spécifique, tel qu'un droit de propriété ou une obligation contractuelle ou extracontractuelle ;*
> 2. *Les œuvres, dessins, modèles, inventions, bases de données, protégés par le code de la propriété intellectuelle, dont la durée de protection légale, à l'exception du droit moral des auteurs, a expiré ;*
> 3. *Les informations issues des documents administratifs diffusés publiquement par les personnes mentionnées à l'article 1 de la loi n° 78-753 du 17 juillet 1978 et dans les conditions précisées à l'article 7 de la même loi, sans préjudice des dispositions des articles 9, 10, 14 et 15 de ladite loi.*
>
> *Les choses qui composent le domaine commun informationnel sont des choses communes au sens de l'article 714 du Code civil. Elles ne peuvent, en tant que tels, faire l'objet d'une exclusivité, ni d'une restriction de l'usage commun à tous, autre que l'exercice du droit moral.*

56. Des décennies, c'est un peu l'éternité à l'échelle d'Internet qui n'a que 50 ans.

Les associations agréées ayant pour objet la diffusion des savoirs ou la défense des choses communes ont qualité pour agir aux fins de faire cesser toute atteinte au domaine commun informationnel. Cet agrément est attribué dans des conditions définies par un décret en Conseil d'État. Il est valable pour une durée limitée, et peut être abrogé lorsque l'association ne satisfait plus aux conditions qui ont conduit à le délivrer. »

Malheureusement cet article n'a pas été retenu dans le projet de loi pour une République numérique[g]. Si la loi reconnait les logiciels et les données libres et si l'État pousse pour leurs usages dans le cadre des données publiques, le gouvernement semble ne pas vouloir des communs qui pourraient restreindre l'activité des entreprises du numérique.

Pour autant l'idée continue à faire son chemin. L'ADEME [57] a lancé un appel à projets qui s'appuie sur le communs :

« Résilience des Territoires 2021

Nous considérons qu'à l'ère du numérique, il est nécessaire de relier les acteurs de la résilience afin de construire ensemble des ressources communes : plateformes technologiques, données ouvertes, logiciels libres, connaissances, retours d'expérience, protocoles, territoires d'expérimentation. Le rôle de l'appel à communs (AAC) est de rassembler tous les acteurs volontaires pour produire des ressources ouvertes – les communs - et ainsi faire évoluer la résilience des territoires dans une logique d'adaptation aux changements climatiques par la création et le partage de ces communs. »

`https://agirpourlatransition.ademe.fr/entreprises/dispositif-aide/20210319/resilience2021-57`

Elle finance les projets retenus et demande aux candidats de publier la description de leur projet sur un wiki afin d'« Indexer et

57. Établissement public en charge de la protection de l'environnement et de la maîtrise de l'énergie. `https://www.ademe.fr`

organiser la production de communs »[58]. Il s'agit donc de rendre l'appel à projet transparent, de permettre à chacun de découvrir les autres projets et de créer des synergies.

Cette tendance se retrouve dans d'autres services de l'État. Certains hauts fonctionnaires sont même d'ardents défenseurs des communs. Parmi eux se trouve sans surprise les plus concernés par le bon fonctionnement d'Internet, comme Sébastien Soriano, ancien directeur de l'Autorité de régulation des communications électroniques, des postes et de la distribution de la presse (ARCEP)[g] et depuis janvier 2021, directeur de l'IGN[g]. À ce nouveau poste, il promeut la création de géo-communs :

> *« Aujourd'hui, en France comme à l'étranger, les pouvoirs publics sont de plus en plus nombreux à se saisir de la dynamique des « communs » issus de l'initiative citoyenne, comme tierce dimension au public et au privé.*
>
> *...*
>
> *Dans ce contexte, l'IGN s'inscrit dans le mouvement des communs pour accroître son impact en nouant des alliances avec un large écosystème d'acteurs et à cette fin a mis en place le dispositif « Fabrique des géocommuns » selon deux approches principales :*
>
> — *un incubateur de services publics numériques s'appuyant sur la mécanique des communs*
> — *un programme de valorisation et de soutien des géocommuns*
>
> »
>
> https://www.ign.fr/la-demarche-geocommuns

En transférant ainsi des données de l'État dans les communs, non seulement la transparence progresse, mais aussi la participation citoyenne. Il ne manque que le cadre légal afin de pérenniser le processus pour rentre difficile tout retour en arrière et bloquer toute privatisation. Si on arrive à franchir cette étape, le mouvement actuel laisse à penser qu'on peut y arriver, alors il s'agit d'une (r)évolution très prometteuse pour notre société.

58. https://wiki.resilience-territoire.ademe.fr/

2.4
Le gouvernement ouvert

Dès lors qu'on dispose des données étatiques, il devient possible de donner son avis en connaissance de cause. La question de la gouvernance en commun arrive naturellement. Comment dépasser le cycle des élections pour avoir une interaction continue entre les citoyens et les dirigeants ?

En 2009 le président Obama a lancé aux États-Unis la *Open Government Initiative* dont le but était de rendre plus transparent le fonctionnement de l'État et de permettre la participation du public. Deux ans après le « Partenariat pour un gouvernement ouvert » [59] a été créé afin de regrouper les pays désirant aller dans le même sens. La France a rejoint ce groupe en 2014. Outre le but d'impliquer davantage les citoyens dans la gouvernance, il s'agit aussi (surtout) de regagner leur confiance.

En pratique

La gouvernance ouverte nécessite d'avoir accès aux données pour s'informer. Il faut un processus pour établir les lois avec une participation la plus large possible. Enfin, il faut décider. Bien sûr, ces étapes sont liées. On comprend que la seconde étape n'est possible que si chacun a accès aux données nécessaires à l'élaboration des textes, donc l'étape deux dépend de la première. Quant à la dernière étape, le vote, elle s'appuie non seulement sur les textes à voter, mais aussi sur les traces des débats qui ont mené aux textes ainsi que sur les données initiales. On a donc :

La première étape a déjà été largement abordée, aussi regardons les deux suivantes.

59. https://www.opengovpartnership.org/

La collaboration ouverte

Le travail collectif, pas toujours simple en petit groupe, devient vraiment difficile à grande échelle. Les discussions sur Twitter ou autres forums montrent à quel point nous sommes plus doués pour nous insulter que pour chercher à comprendre le point de vue de l'autre. La bonne nouvelle est qu'en se concentrant sur un sujet précis avec les informations adéquates, le risque de conflit est largement réduit, sans toutefois disparaître.

Aussi, il faut une méthode pour avoir une discussion constructive qui passe à l'échelle. Elle n'existe pas encore pour un pays comme la France, mais pourrait déjà être envisagée au niveau d'une petite ville. Wikipédia est un exemple intéressant. Son niveau de satisfaction auprès du grand public est probablement bien supérieur à celui qu'on a de nos élus. Elle est ouverte au sens où chacun peut contribuer, elle dispose de règles et d'un gouvernement garantissant un bon fonctionnement [60]. Pour éviter les contributions négatives, des administrateurs [61] appliquent des règles précises et transparentes. Bien sûr, écrire ensemble une page Wikipédia n'est pas la même chose que de définir une politique. Autant pour un article la vérité existe, ou semble accessible, autant il n'y a pas de vérité lorsqu'on fait le choix d'investir dans un secteur plutôt qu'un autre. Aussi pour profiter pleinement de l'expérience de Wikipédia, il est plus intéressant d'observer la gestion des cas difficiles, à savoir les pages polémiques. La discussion constructive y atteint vite ses limites, les administrateurs doivent intervenir, effacer des interventions, bloquer l'édition de la page, voire bannir certaines personnes avec toute la subjectivité que cela implique. Ce sont les règles, elles génèrent des insatisfactions, mais Wikipédia est toujours en vie et la qualité de ses articles reste bonne.

Une autre expérience plus proche de la gouvernance ouverte a

60. Ce gouvernement a un pouvoir exécutif, les décisions se prennent par consensus et rarement par vote.

61. Les administrateurs sont des contributeurs très actifs élus par les contributeurs assez actifs. La page `https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Administrateur` décrit leur rôle. Il existe aussi les bureaucrates et les stewards qui ont plus de droits techniques.

été la consultation publique mise en place par la secrétaire d'État Axelle Lemaire fin 2015 pour préparer la loi pour une République numérique[g]. Il s'agissait alors de co-créer une proposition de textes de loi avec les internautes, proposition qui serait ensuite soumise au parlement. Pendant 3 semaines, chacun pouvait enrichir le document initial, y apporter des modifications, des commentaires et voter pour les différentes contributions. Avec 21 000 contributeurs auteurs de 8 500 contributions et ayant voté 150 000 fois, cette première expérience a été un succès de participation. La méthode mise en place pour permettre ce travail collaboratif était fondée sur une plateforme numérique [62] qui permettait non seulement d'ajouter des contributions, mais aussi de donner des arguments pour ou contre et de voter pour ces derniers. Des modérateurs (des fonctionnaires) vérifiaient le bon déroulement de l'opération. Cette plateforme d'intelligence collective a été un pas en avant, mais reste en deçà des besoins. Les outils d'écriture collaborative ainsi que ceux de suivi de version sont probablement des pistes d'amélioration du processus. La Fabrique de la Loi [63] qui permet de suivre l'évolution d'une loi entre le texte initial et la version finale, pourrait aussi être une source d'inspiration.

Quoi qu'il en soit, cette expérience a été perçue comme une réussite. Elle a permis de mettre les différentes parties prenantes, les citoyens, mais aussi les associations, autour d'un texte en devenir et de permettre à chacun d'avancer ses arguments en toute transparence. Le résultat de la consultation a donné lieu à l'ajout de 10 articles aux 30 du projet de loi initial et a intégré 70 modifications au texte initial. Il s'agit donc d'un progrès même si tout le monde n'a pas joué le jeu. Ainsi la proposition de lutte contre le *copyfraud* à savoir «*empêcher qu'une œuvre soit attaquée par d'autres droits alors que, de fait, celle-ci est dans le domaine public*» [64], a été combattue par les lobbies des droits d'auteur en dehors de la consultation, directement auprès du ministère de la Culture, lequel s'est farouchement opposé à cette proposition d'article et a ob-

62. Toujours visible sur `https://www.republique-numerique.fr/`

63. `https://www.lafabriquedelaloi.fr/`, un projet de l'association Regards Citoyens[g].

64. Comme l'a rappelé la députée Isabelle Attard.

tenu son retrait alors même qu'elle a fait l'objet d'un large consensus lors de la consultation.

L'étape suivante serait de permettre à une telle consultation de pouvoir proposer des amendements lors des débats à l'assemblée nationale, lui donner les pouvoirs d'un député. Bien sûr cela entre en contradiction avec le principe de démocratie représentative et rien ne dit que les citoyens dans leur ensemble soient plus résistants aux lobbies que les élus, mais la perte de confiance en la politique et les innovations techniques poussent vers une évolution du système.

La prise de décision collective

Après le débat, les choix. Dans ce domaine, force est de constater que tout est à faire du point de vue légal. Actuellement seul le référendum permet aux citoyens de donner leur avis sur un texte de loi ou un choix de société. Ce référendum est toujours à l'initiative du pouvoir, que ce soit le référendum classique à l'initiative de l'exécutif ou le référendum d'initiative partagée à l'initiative du législatif. Dans le second cas, il ne s'agit pas de voter pour un texte mais de voter pour qu'un texte soit étudié par le pouvoir ; il ne s'agit donc pas d'une réelle prise de décision collective.

Seul un référendum d'initiative citoyenne ou populaire pourrait permettre aux citoyens d'être à l'initiative d'un référendum classique. Réclamé par les Gilets jaunes et prévu dans la Constitution européenne que les Français ont rejetée, ce RIC ou RIP est dans les esprits depuis la Révolution, mais a toujours été refusé par le pouvoir et plus largement par les intellectuels. Le risque de laisser le peuple choisir leur semble trop grand, arguant que nombreux sont les domaines où la grande majorité des citoyens n'a ni les compétences, ni le temps pour faire un choix éclairé [65].

D'autre part, un référendum n'est pas un outil politique de précision. Le référendum pour la Constitution européenne en est un exemple flagrant. Dire oui ou non à un texte aussi complexe est

65. Surtout si on leur demande un avis tous les jours sur un nouveau sujet.

trop globalisant. Comment faire lorsqu'on juge un point totalement inacceptable alors qu'on approuve le reste ?

Une technique mise en avant dans la gouvernance ouverte est celle du consensus. Il s'agit de prendre en compte les divergences en amont pour arriver à un texte acceptable par tous. Cela veut dire sortir de la logique de la majorité qui l'emporte et ne pas passer en force contre une minorité braquée. À l'inverse, cela implique que la minorité laisse passer ce qui est désagréable, mais pas inacceptable pour elle. Cette recherche du consensus, en vogue dans les pays scandinaves et sur Internet, ne correspond pas à l'esprit français habitué la confrontation, mais elle paraît être une voie intéressante pour impliquer plus les citoyens dans la prise de décision. Cependant, même le consensus n'est pas exempt de reproches. Wikipédia qui défend ce principe a en fait un tout petit groupe de personnes hyper-actives qui contrôle l'élaboration des règles [66]. Il semble qu'en l'état, un travail pédagogique important doit être mené en amont, ainsi que le développement d'outils adaptés, pour espérer une prise de décision par consensus efficace.

Il est également possible d'imaginer de nouveaux modes opératoires. Un simple changement du système électoral peut changer radicalement la situation. Par exemple, plutôt que de voter pour un candidat ou une liste pour gérer la ville, on pourrait voter par thématiques [67] avec un système de délégation de voix où l'électeur donne sa voix à un proche, lequel peut donner sa voix ainsi que celles qu'il a reçues, etc [68]. Dans ce cas l'incompétence citée ci-dessus est largement réduite puisque les représentants sont choisis par thématique. Notons que ce système est implicitement appliqué par le parlement, avec des parlementaires spécialisés sur certains sujets qui travaillent en commissions préparatoires, leurs collègues suivant leurs avis.

66. Voir à ce sujet l'article de Gilles Sahut « La gouvernance de Wikipédia : élaboration de règles et théorie d'Ostrom », https://journals.openedition.org/ticetsociete/2426

67. Actuellement ces thématiques sont gérées par les adjoints au niveau des mairies, par les ministres au niveau de l'État.

68. Ce système cherche une ville pour être testé, cf. https://ricou.eu.org/e-politique.html#D%C3%A9mocratie%20permanente

Le tirage au sort d'un groupe de personnes représentatif à qui on donne le temps de pouvoir travailler sur un sujet est aussi une solution possible. La Convention citoyenne pour le climat [69] construite sur ce schéma a réuni 150 citoyens aidés tout le long des 6 mois par un comité de gouvernance composé de 13 experts dans les domaines du climat, de la démocratie participative, du champ économique et social auxquels se sont joints deux représentants du ministère de l'Écologie. De plus ces citoyens pouvaient s'appuyer sur un groupe de 14 experts pour répondre aux différentes questions et un groupe légistique pour aider à la rédaction. Enfin, ils ont pu interroger 150 personnes liées à la thématique, des personnalités scientifiques, politiques, syndicalistes et des entrepreneurs. Le résultat de ce travail collectif a abouti à 149 propositions dont 3 ont été bloquées par le président de la République [70]. Certaines mesures étaient déjà sur le point d'être prises, ce qui simplifie leur mise en application. D'autres devraient être soumises à un référendum, d'autres encore passeront par décrets. Enfin les dernières passeront par le parlement. Cette expérience intéressante ne mène pas encore à la décision puisqu'elle ne fait que soumettre des idées au parlement et au gouvernement alors que l'idée du tirage au sort est justement de prendre les décisions autrement.

On voit qu'il existe des pistes pour remplacer la démocratie représentative, mais ce n'est pas le but du gouvernement ouvert, tel qu'il est défini par les autorités. Pourtant, les taux d'abstention qui dépassent régulièrement les 50 % et les manifestations de mécontentement comme celles des gilets jaunes sont le signe d'un système à bout de souffle. Aussi, il est nécessaire de continuer à multiplier les expériences et les mener le plus loin possible pour trouver une solution de remplacement. Tester localement un système alternatif jusqu'à la prise de décision serait un premier pas.

69. `https://www.conventioncitoyennepourleclimat.fr/`

70. 110 km/h sur les autoroutes, taxer à 4 % les dividendes et introduire la supériorité de l'écologie dans la Constitution. Une 4e mesure est en train de disparaître, il s'agit du moratoire sur la 5G.

Chapitre 3

Participer

Ce chapitre entre dans les aspects pratiques : comment apporter sa pierre à ce mouvement de transparence numérique.

Contrairement aux idées reçues, il n'est pas nécessaire d'être informaticien pour participer. Entrer manuellement des données dans un tableur ou dans une page web est à la portée de tous. Certains jeux de données existent dans des administrations et ne demandent qu'à sortir pour peu qu'une personne les réclame assez fort. Parfois, la création de données est un peu technique, sans pour autant être inaccessible à une personne motivée. Le chapitre commence donc par la présentation des différentes façons de créer un jeu de données.

La seconde partie aborde l'analyse des données. On y étudie différents outils, du tableur aux langages de programmation permettant l'analyse des très grands jeux de données. La partie suivante se concentre sur des pièges qu'il convient de savoir éviter, tant pour comprendre les résultats que pour éviter une mésinterprétation. Quelques exemples sont présentés.

Enfin la partie 3.4 conclut ce chapitre sur les lanceurs d'alerte. C'est l'acte ultime d'ouverture de données pour dénoncer un comportement illégal ou amoral. Il est encadré par la loi et offre des protections à ceux qui prennent ce risque sans arrières pensés.

3.1
Créer des données

Avant d'analyser des données, il faut les assembler. La création de données peut se faire à partir de rien, en reprenant un ou

des jeux de données existants qu'on restructurera, ou en intégrant un projet existant pour compléter ou corriger ses jeux de données. C'est une tâche à la portée de tous, mais qui nécessite une mise en œuvre correcte si l'on veut un résultat valide.

Ex nihilo

Lorsqu'il s'agit d'un petit jeu de donnée, la liste des ministres avec une mise à jour régulière, on peut le faire à la main avec son éditeur favori ou un tableur [1]. Si on désire avoir la liste des 36 000 maires avec des informations les concernant, on sent que cela va être plus long, mais c'est encore possible avec de la volonté.

Pour un jeu de données plus gros, il faut s'organiser différemment et bien étudier tous les aspects : générer les données, les mettre en forme, les diffuser et choisir la licence à utiliser.

La génération de données est :

— soit manuelle, comme c'est le cas pour Wikipédia et beaucoup de données publiques,

— soit semi-automatique, comme pour OpenStreetMap,

— soit totalement automatique, pour les projets qui s'appuient sur des capteurs ou autres objets connectés.

Dans tous les cas, il est difficile de construire seul des données de taille importante. Il est préférable de constituer un groupe de passionnés en ligne en s'appuyant sur les nombreux outils disponibles sur Internet. Des rencontres physiques régulières permettront d'améliorer les relations humaines, les comportements sont bien meilleurs entre personnes qui se sont rencontrées et qui se reverront.

Le premier défi technique est de définir comment stocker les données. Il faut trouver un format assez souple pour s'adapter aux évolutions futures tout en étant assez performant pour être mis en production. Le bon format est rarement trouvé du premier coup, cela implique qu'il faut au minimum pouvoir évoluer d'un format à un autre. De plus, il faut choisir un format ouvert pour toucher le

1. `http://remaniement.fr` est un exemple de petite base faite à la main.

plus grand nombre possible d'utilisateurs et garantir la pérennité du projet.

Vient ensuite la partie centrale, la création des données. Elle est très variable suivant le type de projet. Dans certains cas, elle demande le développement d'une interface web pour que chacun puisse ajouter ses données en ligne. Le développement sera simple s'il ne s'agit que d'intégrer des données chiffrées qui respectent un format précis, mais il sera plus compliqué si les données ne sont pas structurées. Si le projet est une collection de données semblable à Wikipédia, alors il est possible d'utiliser son infrastructure basée sur MediaWiki. C'est un très gros logiciel, fruit d'un énorme travail collaboratif, heureusement sous licence ouverte et donc utilisable librement.

Une fois les données prêtes, comment les diffuser ? Pour un jeu de données archivé dans un fichier, il suffit de le déposer sur son site web ou sur `data.gouv.fr` qui est ouvert à tous [2]. Pour une jolie présentation des données, il est préférable de développer une interface web, ce qui implique un développement logiciel nettement plus conséquent. La diffusion doit aussi être référencée pour toucher un large public. En déposant ses données sur le portail `gouv.data.fr`, on bénéficie déjà de son audience. Pour encore plus de visibilité, il est conseillé d'ajouter des métadonnées [3] et des mots clefs décrivant les données. Cela aide les moteurs de recherche à faire ressortir le site web ainsi annoté.

Le choix de l'infrastructure informatique est plus délicat. Plus il y a de données, plus elles sont consultées et plus l'infrastructure informatique doit être solide. Elle peut être gérée en interne par les informaticiens du groupe ou s'appuyer sur le nuage et des services comme Google Drive. Dans le cas d'une gestion en interne, le coût du stockage et des serveurs est faible au début. Il peut être supporté par une seule personne, mais il augmentera avec la taille du projet. OpenStreetMap[g] a dépensé plus de 20 000 € en 2018 pour ses sites web et ses serveurs. Heureusement, lorsqu'on arrive au ni-

2. `https://doc.data.gouv.fr/jeux-de-donnees/publier-un-jeu-de-donnees/`

3. Métadonnées qui renseignent sur l'auteur, le format et la construction du jeu de données.

veau de notoriété d'OpenStreetMap, on trouve des financements : dons, adhérents, organisation de conférence...

Enfin vient le choix de la licence sous laquelle sont diffusées les données. Pour un projet altruiste ou qui ne vise pas à dégager de profits, la licence naturelle est une licence ouverte (cf. section 2.1.2). Il est important de noter qu'il est possible d'utiliser une licence ouverte tout en gardant pour soi la possibilité d'une exploitation commerciale. Ainsi le propriétaire des données peut créer une licence ouverte sans droit d'utilisation commerciale pour le grand public et une autre fermée avec droit d'exploitation commerciale pour la vente aux entreprises.

Aspirer Internet

Il est aussi possible de créer un jeu de données important, en aspirant des données présentes sur Internet. On utilise pour cela une API[g], cf. page 57, ou on prend des données sur des pages web. Dans ce dernier cas, les données ne sont pas structurées mais mises en pages pour être lues par un humain. Cela rend la récupération des données plus difficile, mais reste possible avec des bibliothèques [4], comme Beautiful Soup [5], qui permettent d'analyser la syntaxe des pages web pour en extraire les informations voulues. Ainsi, si on sait que telle page web présente en temps réel le nombre de places libres dans les parkings de la ville et que ces nombres sont rangés dans un tableau sur la page web [6], alors il suffit de demander à Beautiful Soup de ne récupérer que le tableau de la page web et d'ignorer le reste. En répétant régulièrement l'opération, par exemple toutes les 5 minutes, on peut créer l'historique du nombre de places dans les parkings de la ville.

Autre exemple : la SNCF considère qu'un train n'est en retard qu'à partir de 6 minutes et ainsi n'inclut pas les trains arrivés avec un retard inférieur. Cette option pénalise proportionnellement les banlieusards dont le trajet normal est en moyenne de 15 minutes.

4. En informatique les bibliothèques sont des extensions d'un langage.
5. Pour Python : `https://www.crummy.com/software/BeautifulSoup/`
6. Ce que fait la ville de Luxembourg, cf. `https://www.luxembourg-city.com/fr/planifier-votre-sejour/informations-voyageurs/parking`

On peut donc préférer interroger l'API de la SNCF toutes les minutes pour générer son jeu de données avec l'heure exacte d'arrivée de chaque train à chaque station. Calculer les retards, même d'une minute, devient alors possible.

Attention cependant, il est nécessaire de vérifier les aspects légaux pour savoir ce qu'on peut faire et ce qu'on peut diffuser.

Intégrer un projet

Une autre façon d'apporter sa pierre à l'édifice est d'intégrer un projet existant. C'est la façon la plus simple puisque tout est déjà en place. Corriger une faute d'orthographe sur Wikipédia est déjà une participation. Si on ajoute un paragraphe, c'est encore mieux et ainsi, de fil en aiguille, se construit une encyclopédie. On peut aussi être un participant régulier, voire participer au fonctionnement interne du projet. Les grands projets sont souvent portés par des associations ouvertes à tous.

Certains projets vont demander d'extraire des données pour les présenter dans un autre format ou de vérifier qu'il n'y a pas de manques ou d'incohérences. Ce n'est pas une tâche toujours attrayante mais elle est très importante. Un jeu de données comportant des manques ou des erreurs peut fausser une analyse. Parfois ce sont ces données bizarres cachées au milieu des autres qui permettent de soulever un lièvre. Des données manquantes sont aussi intéressantes lorsque qu'elles indiquent un dysfonctionnement.

Ainsi Anticor[g] demande à ses adhérents de vérifier les 5 000 déclarations de patrimoine et d'intérêts auprès de l'HATVP[g]. Chaque déclaration comportant plusieurs pages. On devine que toute aide est appréciée.

Transparency International[g] France a créé le projet *Visualiser la corruption*[7] pour combler le manque de données ouvertes en matière de justice. Il faut dire que ce ministère n'a toujours pas mis en œuvre la loi pour une République numérique[g] de 2016 et ne publie pas les comptes-rendus des décisions de justice, qui seraient pourtant bien utiles pour analyser notre société. Donc, en at-

7. http://www.visualiserlacorruption.fr/

tendant, Transparency International France fait appel aux bonnes volontés : « *nous avons besoin de vous pour connaître les condamnations qui n'ont pas fait l'objet d'un traitement dans la presse nationale ou qui n'apparaissent pas encore sur notre carte* ».

D'autres projets sont plus techniques, comme ceux qui utilisent des objets connectés. Dans ce cas, il faut faire fonctionner l'appareil, penser à la sécurité (changer le mot de passe par défaut au minimum) et le connecter au projet pour y transférer les données. Heureusement ces projets publient des guides et disposent de forums pour aider les débutants.

En général les projets ont besoin de contributeurs, aussi les nouveaux arrivants sont les bienvenus. Bien sûr, intégrer une structure demande de s'y adapter, ce qui n'est pas toujours simple. Un groupe constitué a ses habitudes qui peuvent dérouter le néophyte. Si l'intégration est trop difficile, il est toujours possible de rejoindre un autre projet. Une bonne volonté trouvera de quoi faire.

Créer des jeux de données dérivés

Une troisième façon de créer des données est de modifier l'existant. Un des principes fondamentaux du libre est de permettre de s'appuyer sur l'existant pour construire du nouveau. Cela a permis depuis des décennies de créer des logiciels de plus en plus complexes et des systèmes d'exploitation. C'est aussi un principe bien connu dans le monde de la culture même si l'aspect légal y est plus compliqué.

Dans le monde de la transparence et des données, créer un jeu de données dérivé peut simplifier, voire améliorer le travail d'analyse qui suit [8]. Ce travail demande des compétences plus ou moins importantes en informatique. Prendre un jeu de données et le sauver sous un autre format est relativement simple et suffit à aider d'autres personnes moins à l'aise.

En revanche, la manipulation de données à grande échelle est plus complexe. L'un des problèmes des jeux de données sur le site

8. Les données influencent notre manière de penser. Ajouter des données, les exprimer différemment peut faire apparaître de nouveaux éléments.

`data.gouv.fr` est que les administrations déposent souvent chaque année un nouveau fichier. Cela complique le suivi dans le temps puisqu'il faut charger plusieurs fichiers dont rien ne garantit qu'ils aient conservé le même format d'une année à l'autre. La création d'un gros fichier incluant toutes les données de toutes les années est utile, mais c'est un travail qui peut nécessiter des compétences avancées.

Il peut être intéressant de combiner des jeux de données. Si on désire poser des données sur une carte, mais que l'administration ne fournit qu'un fichier d'adresses postales, alors on peut le croiser avec la Base Adresse Nationale ce qui permet de générer un fichier comportant les coordonnées GPS des lieux choisis. D'autres bases possèdent les limites géographiques des communes, ce qui permet d'associer des couleurs aux communes en fonction de nos données.

Combiner des jeux de données revient parfois à en compléter un. Une enquête de l'observatoire opendata des territoires indique que le taux d'ouverture de leurs données par les organismes et collectivités concernés par la loi République numérique est de 14 %. Cela semble très peu. Heureusement l'observatoire ne se limite pas à donner des résultats statistiques ; il fournit aussi les données brutes [9]. On y découvre une colonne `pop-insee` indiquant la population de chaque collectivité territoriale. Il est ainsi possible de calculer quel pourcentage de la population de chaque commune, département ou région a accès à des données ouvertes [10]. On trouve ainsi que 52 % de la population française peut accéder librement aux données publiques locales (commune, communauté d'agglomération ou métropole), 61 % pour les départements et 99 % pour les régions [11] début 2022. La perception n'est plus du tout la même [12]. Si cette colonne population n'avait pas été présente, l'ajouter aurait amélioré significativement ce jeu de données. Cela étant, il ne faut

9. cf `https://www.observatoire-opendata.fr/les-donnees/`
10. La ligne de code qui permet cela est présentée page 96
11. Seule la Réunion a ouvert ses données parmi les régions ultramarines.
12. Mais on ne connait pas la quantité de données ouvertes. Si un seul compte rendu d'un vieux conseil municipal publié sur le site web de la mairie suffit pour être référencé, alors on est loin de l'esprit de la loi.

pas tomber non plus dans l'excès inverse et mettre toutes les données du monde dans un énorme fichier que personne ne pourra ouvrir sans saturer sa machine !

Enfin, simplement nettoyer un jeu de données est un travail utile, l'idéal étant de faire remonter les erreurs aux créateurs. De très nombreux jeux de données sont faits à la main ce qui rend les erreurs presque inévitables. Il peut s'agir d'erreurs simples à détecter comme un mot là où on attend un nombre, mais c'est parfois plus délicat, une valeur réelle ou négative dans la colonne `population` ou toute autre incohérence que la machine ne détectera pas au chargement. Même lorsque les données sont créées par des machines, comme dans le cas de l'Internet des objets, on peut trouver des erreurs, car les machines aussi ont des faiblesses. Aussi, prendre un jeu de données, le nettoyer sans rien modifier à sa structure et le sauver dans le même format est déjà un travail de grande valeur.

Demander des données

La dernière façon de créer des données consiste à engager ceux qui les possèdent à les diffuser. La procédure pour obtenir des données publiques consiste à contacter l'organisme qui les détient. Si cela ne donne rien [13], l'étape suivante est de faire appel à la CADA[g] dont la mission est d'aider le citoyen face à une administration récalcitrante (administration au sens large, cf. le chapitre 2.1.1) [14].

Deux associations ont mis en place des sites web pour aider le citoyen dans sa démarche.

Le site Ma Dada [15] de l'OKFN[g] France offre un formulaire pour contacter l'organisme qui doit avoir les données recherchées (le site propose plus de 50 000 organismes à contacter). L'avantage de

13. Si l'organisme ne bloque pas la demande pour des raisons politiques mais pour des raisons pratiques, on peut lui suggérer la lecture du livre de Jacques Priol *Le big data des territoires*, Fyp, 2017. Ce livre motive l'ouverture des données publiques et donne de bons conseils sur la manière de faire.

14. `https://www.cada.fr/particulier/quand-et-comment-saisir-la-cada`

15. `https://madada.fr`

passer par ce site est de rendre la demande publique, ce qui laisse une trace. Cela permet de voir la réactivité de l'administration et cela permet aux suivants de trouver la question et l'éventuelle réponse de l'administration. C'est utile lorsqu'on passe à la seconde étape, la demande d'intervention de la CADA[g].

L'association Ouvre-Boite [16] recueille les demandes dans son forum, puis choisit les cas qu'elle désire prendre en charge. À partir de là, l'association fait la demande des données auprès de l'organisme concerné, saisit la CADA et poursuit au tribunal administratif si nécessaire. Son site web offre un suivi très intéressant de ses actions.

3.2
Analyser les données

Une fois les données recueillies, l'analyse peut commencer. Le but de l'analyse est de comprendre ce qu'expriment les données. Cela peut se faire de différentes façons, à travers des calculs savants, en affichant les données sous forme de graphiques, avec des tableaux, en mélangeant les calculs et l'affichage, etc.

L'exemple récent le plus marquant est le site CovidTracker [17] de Guillaume Rozier [18]. Ce site aspire les données publiques liées à la pandémie et les présente sous forme de graphiques, qui permettent à chacun de mesurer l'évolution de la pandémie. Il y a aussi associé des indices pour synthétiser des données pas toujours simples à comprendre. Notons que ce projet personnel est en fait une œuvre collaborative dans le plus pur esprit des logiciels libres. Guillaume Rozier souligne lui-même avoir été aidé par plus de cent personnes pour résoudre des problèmes et améliorer le code source[g] de son site web, code diffusé sous licence libre.

16. `https://ouvre-boite.org/`
17. `https://covidtracker.fr/`
18. Guillaume Rozier, jeune ingénieur de 25 ans, a été fait chevalier de l'ordre national du Mérite à « titre exceptionnel » en mai 2021 pour ce site et son autre site *Vite ma dose*, `https://vitemadose.covidtracker.fr/`, qui centralise les crénaux disponibles pour se faire vacciner.

Pour commencer une analyse de données, il faut une méthode et de bonnes connaissances en mathématique, car il y a de nombreux pièges qu'on évoquera à la fin de ce chapitre [19]. Cela étant il est quand même possible d'utiliser son bon sens pour extraire des informations pertinentes.

Voyons les outils dont on dispose.

Les outils d'analyse de données

Voici quatre types d'outils largement utilisés pour manipuler des données et les analyser :

— les tableurs comme Excel, LibreOffice ou l'équivalent dans le nuage, pour stocker et manipuler des petits jeux de données,

— SQL pour créer et interroger une base de données conséquente,

— Tableau [20], un logiciel commercial à la mode pour analyser graphiquement des données. Autre logiciel, DataWrapper [21] est largement utilisé dans le monde du data journalisme pour générer des graphiques.

— les langages de programmation Python et R qui permettent aussi bien un simple affichage de données que des analyses très complexes.

L'analyse des données peut aussi être confiée à une intelligence artificielle (IA). Cela permet de trouver des informations enfouies dans les données qu'il serait difficile de trouver autrement, tant par la difficulté de repérer les signaux faibles, que par la quantité de données à traiter. Etalab[g] a ainsi crée le Lab IA [22] pour aider

19. Le lecteur désirant s'investir dans ce domaine peut s'inscrire à des cours en lignes sur de la plateforme FUN, `https://www.funmooc.fr/` ou sur Edx et Coursera en anglais (mot clef "Data analysis"). Citons sur FUN les cours « Analyse de données quantitatives en sciences humaines et sociales (ADSHS) » et, plus avancé, « Analyse des données multidimensionnelles ».

20. `https://www.tableau.com`

21. `https://www.datawrapper.de`

22. `https://www.etalab.gouv.fr/datasciences-et-intelligence-artificielle`

les administrations à utiliser l'IA. Elle a aussi créé un concours pour récompenser les projets les plus intéressants. On y trouve par exemple, l'analyse par une IA de données relatives aux centrales nucléaires pour améliorer la sécurité. D'autres projets ont pour but de construire des IA pour détecter des fraudes. Dans les entreprises, et en particulier chez les GAFAM[g], des IA choisissent les publicités à afficher en fonction de l'historique de l'internaute (ce qui fait énormément de données). Parfois l'analyse de toutes ces données fait qu'elles en savent trop sur nous. L'affaire Cambridge Analytica en est un exemple (cette entreprise a utilisé des données personnelles prises sur Facebook afin d'influencer les électeurs en faveur de Donald Trump lors de l'élection présidentielle américaine de 2016). Depuis la Commission Européenne a travaillé sur le sujet pour établir des règles qui doivent permettre d'éviter de tels débordements [23]. L'institut Ada Lovelace insiste, de son côté, sur le besoin d'impliquer la population lors de développement d'IA qui prennent des décisions à partir de données potentiellements personnelles [24]. Il s'agit, là encore, de trouver un équilibre entre les bénéfices pour la société et la protection des individus.

L'intelligence artificielle pose aussi un autre problème : on ne sait pas expliquer son fonctionnement. Outre l'embarras pour les scientifiques, cela soulève surtout un problème lorsque l'administration s'appuie sur une IA pour prendre une décision concernant un individu. Il lui est alors impossible d'expliquer la raison de la décision, ce qui va à l'encontre de l'ouverture des algorithmes.

Aujourd'hui l'IA est accessible au grand public avec l'arrivée de ChatGPT en 2022. Il est possible de l'utiliser sans rien connaitre à l'informatique, il suffit de lui parler. C'est une étape majeure dans l'IA, dans la science et probablement dans l'histoire de notre société. Dans notre cadre, cela va bien nous aider pour traiter des données.

Auparavant faisons un tour d'horizon sur les outils utilisés pour gérer et analyser les données.

---

23. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52021PC0206

24. https://www.adalovelaceinstitute.org/wp-content/uploads/2021/09/Participatory-data-stewardship_Final-report.pdf

Les tableurs

Les tableurs, en particulier Excel, sont des logiciels bien connus du grand public. Tout le monde n'en maitrise pas toutes les possibilités, mais nul n'est besoin d'être un expert pour ouvrir un fichier Excel, faire des calculs simples ou réarranger les données.

Un tableur peut stocker des données à 2 paramètres (ligne, colonne) dans un tableau. C'est le cas le plus simple à lire. Les données peuvent être les moyennes (1ère dimension) des élèves (2e dimension) par matière (3e dimension). Si l'on ajoute un 3e paramètre comme les trimestres les choses se compliquent. On va devoir choisir si l'on désire voir l'ensemble des matières par trimestre ou l'évolution trimestrielle au sein de chaque matière. Si les noms des élèves sont sur les lignes, il faudra des colonnes et des sous-colonnes (2e et 3e paramètres). L'utilisation d'onglets permet d'avoir une page par trimestre ou par matière et permet le retour à 2 paramètres et donc 3 dimensions. Graphiquement des données en 3 dimensions peuvent être des couleurs qui représente la note moyenne dans un tableau en 2 dimension. La représentation graphique avec une dimension de plus va être difficile et si on augmente encore la dimension, cela devient rapidement impossible.

Lorsqu'on passe à 5, 10 ou 20 dimensions, le tableur n'est plus adapté. La base de toutes les notes de tous les élèves de France pendant leur scolarité avec les noms des établissements et les noms des enseignants devient inexploitable avec un tableur.

Pourtant, un tel jeu de données pourrait permettre une analyse fine d'une partie du fonctionnement de l'éducation nationale. Pourrait-on y découvrir des paramètres qui influencent la réussite des élèves ? Que pourrait-on savoir sur les enseignants, sur les collèges, sur les départements ? Toutes ces questions et d'autres pourraient peut-être trouver des réponses intéressantes, mais avec un autre outil qu'un tableur.

SQL

SQL pour *Simple Query Langage* est un langage permettant de créer et d'exploiter une base de données. Il est la référence, relati-

vement facile à comprendre et contrairement à un tableur, il permet de gérer des données avec un grand nombre de dimensions ainsi que de très grands jeux de données.

Une base de données usuelle est composée de tables en 2 dimensions. Une technique classique de SQL consiste à utiliser des identifiants qui correspondent aux numéros de ligne de la table. Ainsi la table des notes comprend une note par ligne et une ligne contient plusieurs informations : l'identifiant de la note, sa valeur, la date, l'identifiant de l'élève concerné et la matière... cf. table 3.1. Pour obtenir des informations sur l'élève, on reporte son identifiant dans la table élèves. cf. table 3.2.

Bien sûr, il faut d'autres tables pour archiver toutes les notes des élèves avec les informations souhaitées. Définir les tables nécessaires et la façon de les relier par les identifiants est un exercice intéressant que ChatGPT sait très bien faire avec la requête suivante :

> *« Je désire construire une base SQL avec des tables pour stocker les élèves de l'école et leurs notes suivant les matières. Quelle architecture me proposez-vous ? »*

Suivi d'une seconde requête pour avoir le code et les champs en français (*« Donnez moi le code pour construire les tables avec les champs en français svp. »*).

Pour une architecture vraiment compliquée, il peut être bon d'avoir un architectes de bases de données mais on sent bien qu'il devient très accessible de construire soit même sa base de donnée (ChatGPT peut aussi expliquer comment installer et configurer les logiciels nécessaires).

Voici deux tables SQL qu'on va utiliser comme exemple :

```
id | valeur |  date  | eleve | matiere |
----------------------------------------
22 |   10.5 | 9/2/19 |   123 |       5 |
23 |   14.0 | 9/2/19 |   126 |       5 |
```

TABLE 3.1 – Table des notes

```
 id | prenom |     nom | naissance | sexe |
----------------------------------------
123 |  Julie | Breizou |  21/08/11 |    F |
124 | Pierre | Brejard |  01/01/10 |    M |
```

TABLE 3.2 – Table des élèves

Les requêtes SQL sont du type

```
SELECT * FROM notes WHERE eleve = 123 AND valeur < 10;
```

qui donne toutes les lignes de la table notes de l'élève 123 (Julie) avec une valeur en dessous de la moyenne (cf. tables 3.1 et 3.2 pour voir la structure des données.). Une requête plus compliquée va croiser les tables (la comparaison après le mot `WHERE`) :

```
SELECT valeur FROM notes,eleves
 WHERE notes.eleve = eleves.id AND eleves.prenom = "Julie";
```

Cette requête retourne toutes les notes de toutes les Julie. La demande suivante faite à ChatGPT 4o donne la bonne réponse [25] :

> *« J'ai deux tables SQL, la table des notes des élèves (colonnes : id, valeur, date, eleve, matiere) et la table des élèves (colonnes : id, prenom, nom, naissance, sexe). Quelle est la requête SQL pour avoir toutes les notes de toutes les Julie ? »*

Si on utilise les tables que ChatGPT nous a proposé, alors la première phrase de mise en contexte n'est pas nécessaire.

Il existe aussi des requêtes pour remplir et modifier la base.

SQL permet aussi de faire des calculs bien que ce ne soit pas sa mission première. Voici la moyenne générale de Julie pour la matière 5 (là encore ChatGPT produit la bonne requête) :

```
SELECT AVG(valeur) FROM notes WHERE eleves = 123 AND matiere = 5;
```

25. Une réponse plus moderne que celle de l'auteur, les 2 produisant le même résultat.

Cependant, un logiciel comme Tableau ou un langage comme Python sont plus adaptés pour faire des calculs. SQL est alors principalement utilisé pour accéder aux données voulues dans la base de données.

Tableau

Tableau est un logiciel commercial [26] qui permet de récupérer des données dans des bases de données et de générer des graphiques. Sa force vient de son interface ergonomique. Il est relativement simple de faire des opérations assez complexes sur les données puis de construire un tableau de bord avec différents types de graphiques pour présenter ce que l'on désire.

Ainsi la figure 3.1 utilise la base de données Weather Data qu'on a connectée comme indiqué en haut à gauche. Dans cette base on choisit les données sur les villes, en particulier la pluie, la température et la vitesse du vent. La première est une agrégation des quantités de pluie des différentes villes, les deux suivantes sont des moyennes. Cela donne les 3 courbes. Au sein de chaque courbe, on choisit une couleur différente pour le jour et pour la nuit. Pour comprendre comment on indique tout cela avec l'interface graphique de Tableau, le mieux est de regarder la vidéo de présentation [27].

L'étape suivante consiste à assembler des figures en un tableau de bord pour avoir une vision complète d'un problème. Lorsque les données sont régulièrement mises à jour, en utilisant un service derrière une API[g] par exemple, le tableau de bord est aussi mis à jour. Cela permet de suivre en direct la pollution au-dessus d'une région avec des alertes ou, comme dans la figure 3.2, l'évolution de la Covid-19. Le tableau de bord peut aussi être interactif. Dans l'exemple de la figure 3.2, il est possible de sélectionner un pays pour avoir des informations supplémentaires.

26. A partir de 35 $ par mois pour visualiser les données et 70 $ par mois pour travailler dessus.

27. Sur la page `https://www.tableau.com/products/desktop` ou directement à ce lien `https://cdnl.tblsft.com/sites/default/files/pages/getinsightsfast.gif`

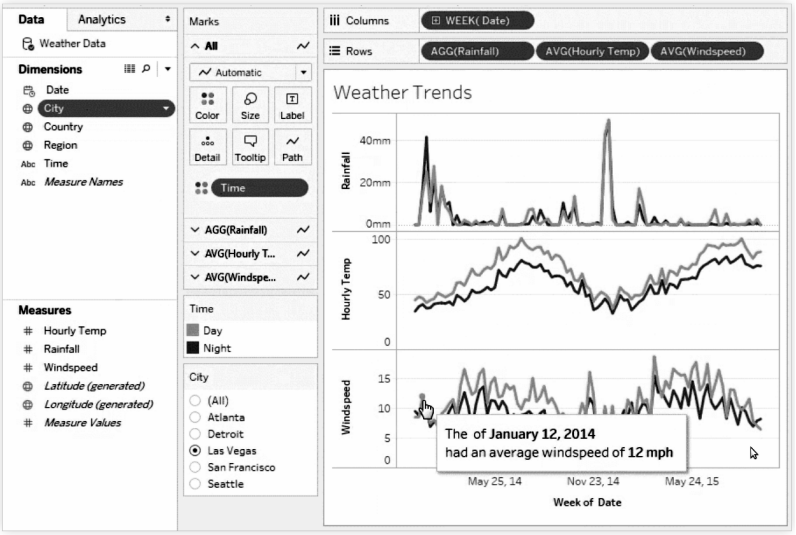

FIGURE 3.1 – Choix des données pour un graphique avec Tableau

Pour traiter des cas concrets, il faut savoir utiliser pleinement l'interface. Cela demande une étude approfondie de la documentation, ainsi que des passages par les forums de la communauté pour y lire des exemples et poser ses questions. Cependant, une interface reste limitée aux options prévues par ses développeurs. Tout n'est pas possible, surtout si on désire que l'interface reste conviviale. C'est pourquoi même si les possibilités de manipulation des données avec Tableau sont importantes, elles restent inférieures à celles d'un langage de programmation. De plus, avec l'arrivée des IA génératives, il est possible qu'un tel logiciel perde de son attrait pour favoriser les langages de programmation dont l'usage devient nettement plus simple, comme on a pu le voir avec SQL.

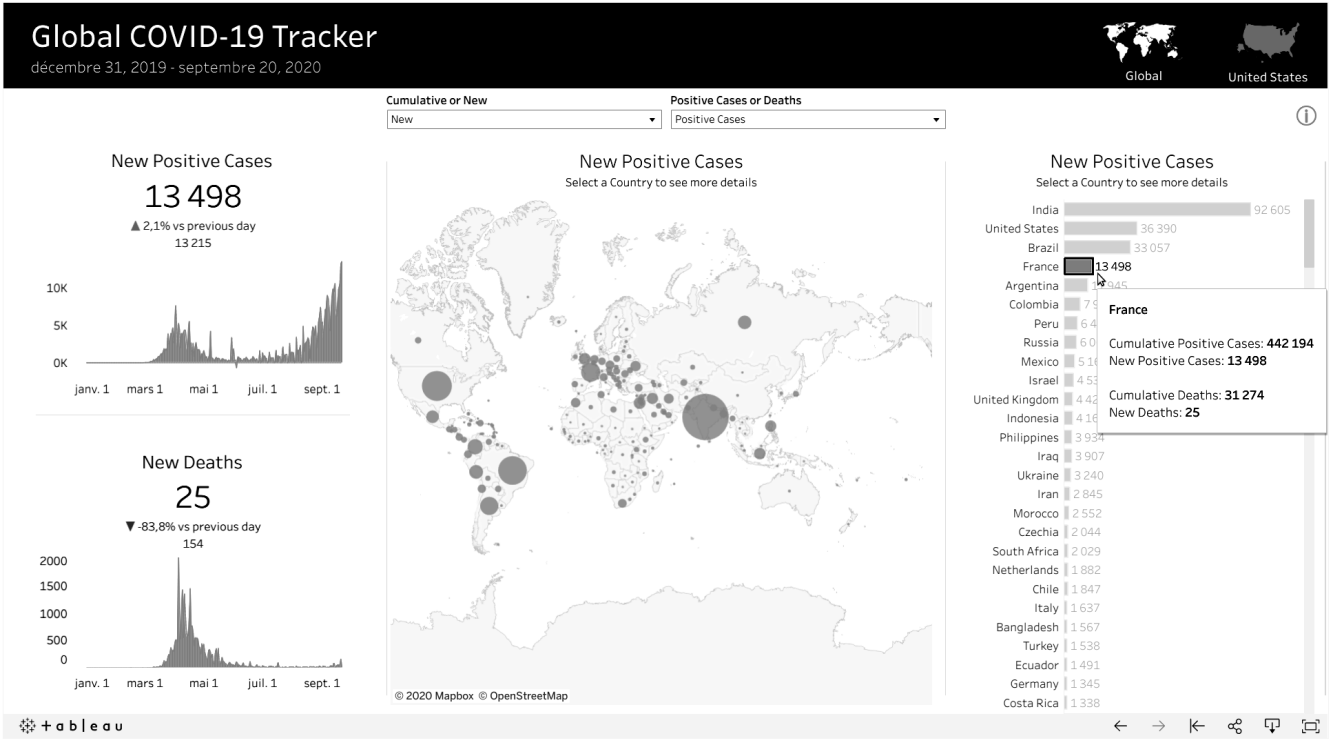

FIGURE 3.2 – Assemblage de graphiques pour créer un tableau de bord avec Tableau

https://public.tableau.com/app/profile/covid.19.data.resource.hub

Python + Pandas

Les deux langages informatiques de traitement de données les plus utilisés sont Python et R. Pour un scientifique motivé, l'apprentissage d'un tel langage et de ses bibliothèques peut se faire en une à deux semaines. Mais seule une longue pratique permet d'avoir un niveau d'aisance suffisant pour manipuler les données sans trop perdre de temps dans la documentation [28]. La bonne nouvelle est qu'avec ChatGPT et ses consœurs, il est possible de se faire aider et donc d'être productif bien plus rapidement.

Les bibliothèques de Python pour manipuler des tableaux de données sont Numpy et Pandas [29].

Une façon agréable de travailler est d'utiliser l'environnement Jupyter [30] qui permet de combiner du texte formaté, du code et des résultats graphiques dans une page visible et éditable depuis son navigateur. Il est ainsi possible de charger des données, de les regarder, de les organiser, de faire des calculs, de regarder les résultats bruts ou dans un graphique, puis de revenir pour corriger ses calculs, voir ce que cela change, etc. Cela permet de produire d'une page web digne d'un article scientifique (si on s'en donne la peine).

Regardons un exemple. On désire charger les données sur les musées de France pour avoir la liste des derniers musées agréés par l'État (ayant l'appellation « Musée de France »).

Le code de la figure 3.3 commence par charger la bibliothèque Pandas. On récupère ensuite sur Internet le fichier Excel listant les musées de France [31] puis on stocke le résultat dans la variable `musees`, que l'on affiche finalement (les figures ne présentent que

28. Un cours sur Python et ses bibliothèques pour faire de l'analyse de donnée est librement disponible sur `https://python3.mooc.lrde.epita.fr/`. Le projet Delta, lié à ce cours, présente des graphiques ainsi que le code source[g] les générant, `https://delta.lrde.epita.fr/`. Ces sites sont produits par l'auteur.

29. Numpy gère les tableaux en N dimensions de même type de données, Pandas accepte différents types, c'est un super tableur.
`https://numpy.org/` et `https://pandas.pydata.org/`

30. `https://jupyter.org/`

31. L'argument `parse-dates` permet d'indiquer les colonnes qui contiennent des dates afin de les interpréter au chargement.

les deux premières réponses, mais il y en a bien plus. De même il y a plus de colonnes que ce qu'on voit.)

```
import pandas as pd

url = "https://www.data.gouv.fr/fr/datasets/r/22df4a13-72d8-4b34-940e-8aec297b5ded"
musees = pd.read_excel(url, parse_dates=['DATE APPELLATION'])
musees
```

	NEW REGIONS	NOMDEP	Date retrait appellation par Haut Conseil	DATE APPELLATION	ID MUSEE	NOM DU MUSEE	ADR	CP	
0	AUVERGNE-RHÔNE-ALPES	AIN	NaT	2003-01-02	0105302	Musée Départemental des Pays De l'Ain	34, rue du Général Delestraint	1000.0	B BI
1	AUVERGNE-RHÔNE-ALPES	AIN	NaT	2003-01-02	0105301	Musée du Brou	Monastère Royal de Brou\n63, Boulevard de Brou	1000.0	B BI

FIGURE 3.3 – Python + Pandas – Charger un fichier Excel

Maintenant que le tableau est dans la variable `musees`, il est simple de faire des manipulations. Dans ce cas les données sont essentiellement des mots, ce qui ne permet pas de faire beaucoup de calculs, par contre il est possible d'appliquer un filtre qui restreint la liste aux musées qui ont obtenu leur appellation après le 1er janvier 2017 :

```
musees[musees['DATE APPELLATION'] > '01/01/2017']
```

	NEW REGIONS	NOMDEP	Date retrait appellation par Haut Conseil	DATE APPELLATION	ID MUSEE	NOM DU MUSEE	ADR	CP	
222	BOURGOGNE-FRANCHE-COMTE	SAONE ET LOIRE	NaT	2017-11-17	7144001	Musée de Bibracte	Mont Beuvray Col du Rebout	71990.0	S/
252	BRETAGNE	FINISTERE	NaT	2017-10-07	2910401	Musée de l'Ancienne Abbaye de Landévennec	Place Yann de Landévennec	29560.0	LA

FIGURE 3.4 – Python + Pandas – Filtrer

Là encore la requête pour limiter l'affichage aux valeurs qui nous intéressent est une opération très courte lorsqu'on en maîtrise la méthode.

On pourrait demander à ce que les lignes soient triées par date, restreinte à tel département, les plus proches de tel lieu, etc.

Autre exemple introduit page 83. Il s'agissait de connaître le pourcentage de la population pouvant accéder aux données publiques ouvertes selon son administration locale (commune, métropole, département, région). Pour cela, on commence par charger les données dans un tableau qu'on appelle `data`. Ensuite, on regroupe toutes les lignes de ce tableau ayant la même valeur `type` et on additionne leur population ; on divise par la taille de la population française (67 millions) et on multiplie par 100 pour obtenir un pourcentage. Pour plus de lisibilité, on limite les données sur lesquelles on travaille au type d'administration et à la population. La ligne de code qui permet tout cela est :

```
data[['type','pop-insee']].groupby('type').sum() / 67063700
```

Ces exemples montrent un aperçu de la puissance des langages informatiques dédiés au traitement des données. La concision des requêtes permet une manipulation aisée des données ce qui simplifie l'analyse. Le rendu graphique peut aussi être très riche.

Erreurs d'analyse

La difficulté principale en analyse de données est moins l'écriture d'une ligne de Python que de savoir ce qu'on cherche comme information, ce qui est pertinent et ce qui est révélateur. Ce n'est pas toujours simple. L'adage dit qu'on fait dire ce qu'on veut aux chiffres et ce n'est pas totalement faux. Il s'agit donc de faire le contraire, en trouvant la vérité objective cachée dans les chiffres et non pas notre vérité.

D'autre part l'analyse demande de comprendre les données et de savoir les visualiser. Voici l'exemple de deux erreurs possibles.

Échelle logarithmique

Prenons les courbes de la Covid-19. Imaginons qu'on ait 100 malades par jour et 200 un mois plus tard. Combien en aurons-nous un mois après si on garde la même progression ? Est-ce 300

ou 400 ? Si on considère que le nombre de malade a augmenté de 100, alors la réponse est 300. Si on considère que le nombre de malades a doublé (+ 100 %), alors la réponse est 400. Le principe de diffusion de la maladie est que les malades contaminent les autres, donc plus on a de malades, plus il y a de contaminations. Si on retient +100 cela veut dire que la progression est 100, 200, 300, 400 et on voit que le nombre de contaminés (100 par mois) ne dépend pas du nombre de malades. Il faut donc doubler le nombre de malades pour suivre la progression de notre maladie. Cela donne 100, 200, 400, 800. Cet aspect est en général bien compris, par contre sa traduction en courbe est quasiment toujours mal faite à savoir que les courbes de la Covid utilisent en général une échelle linéaire, c.à.d. avec un axe vertical qui progresse linéairement, – 100, 200, 300, 400 – au lieu d'utiliser une échelle logarithmique avec un axe qui progresse de façon exponentielle [32] – 100, 200, 400, 800. Le résultat est mauvais, les courbes sont trompeuses car elles exagèrent les grandes valeurs et elles ne sont pas lisibles pour les petites valeurs.

Ainsi sur l'échelle linéaire du graphique de gauche de la figure 3.5, il est difficile de dire quoi que ce soit à part que le nombre de cas testés positifs explose. On voit qu'il y avait une baisse en mai, mais il est difficile de dire à quel moment exact elle a pris fin.

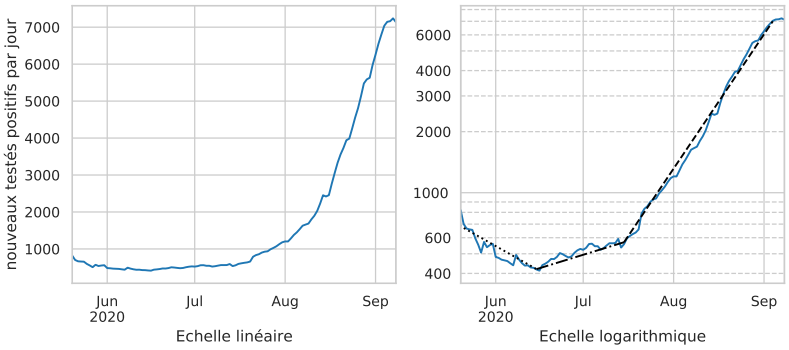

FIGURE 3.5 – Covid-19 – Cas testés positifs en France
Note : ces figures ont été générées avec Python + Matplotlib

32. cf. la présentation de l'échelle logarithmique de la Khan Academy `https://www.youtube.com/watch?v=HWEZZ7fk4JA`

Au contraire avec l'échelle logarithmique, graphique de droite, on peut lire facilement l'évolution de la maladie. On note la baisse du nombre de cas positifs jusqu'à la mi-juin, puis la nombre de cas repart à la hausse et accélère à la mi-juillet. Comme l'échelle est logarithmique, une progression $\times x$ par mois donne une droite (pour une échelle linéaire c'est $+ x$ par mois qui donne une droite). On peut ainsi voir les 3 rythmes de variation représentés par les droites. La troisième droite, la ligne de tirets, correspond à $\times 4$ par mois ou $\times 1,046$ par jour soit une progression de 4,6 % par jour.

L'échelle logarithme est la bonne façon de visualiser l'évolution de données chaque fois que les variations s'expriment en pourcentage. Les placements financiers en sont un autre exemple. Lorsqu'on place une somme d'argent dans un compte rémunéré, ce qui est important est son rendement en pourcentage. Si la banque indique un rendement de 200 € par an, alors je place 1 € et le banquier va regretter de ne pas s'être exprimé en pourcentage ! Il devrait en être de même pour la bourse, mais ce n'est qu'en partie le cas. Lorsqu'une action progresse, on indique son évolution en pourcentage, on ne dit pas qu'elle gagne 1 €, car un gain d'un euro si l'action en vaut 10 n'est pas la même chose que si l'action en vaut 100. Aussi la façon normale de regarder l'évolution d'un cours de bourse devrait être en pourcentage, c.a.d. en échelle logarithmique. Pourtant, ce n'est jamais le cas par défaut sur les sites web boursiers. Les courbes de cours de bourse sont toujours proposées par défaut avec une échelle linéaire. Peut-être est-ce la crainte d'effrayer le novice ? C'est regrettable, cela n'induit pas le bon réflexe.

Attention aux biais

Une autre erreur est de ne pas faire attention aux biais. La courbe de progression de la Covid de la figure 3.5 montre une croissance très importante des nouvelles personnes testées positives, mais cette courbe n'indique pas si le nombre de tests a augmenté durant la même période, ce qui biaiserait le jugement. En effet, si on double le nombre de tests dans une population qui a 1 % de nouveaux cas positifs par jour, on doit constater le doublement des cas positifs. Pourtant, le pourcentage de nouveaux cas est toujours

de 1 quel que soit le nombre de tests.

Les autorités et les journaux ont souligné ce point, mais ils ne l'ont pas souvent intégré dans leurs courbes, ce qui fait qu'on sait qu'il y a plus de tests, mais on ne sait pas quelle est leur part dans la progression du nombre de cas positifs. Pour bien comprendre, il convient de regarder l'évolution du nombre de cas positifs en pourcentage du nombre de personnes testées.

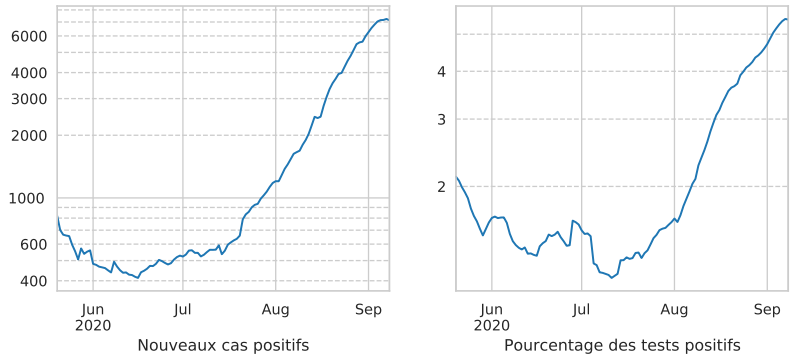

FIGURE 3.6 – Covid-19 – Cas testés positifs en France

La courbe qui présente le pourcentage de tests positifs, graphique de droite figure 3.6, montre que la baisse ne s'est pas arrêtée mi-juin, mais un mois après. Elle montre aussi que la pente est légèrement plus faible depuis mi-août, ce qu'on ne voyait pas sur le graphique de gauche, qui est certes en échelle logarithmique mais qui ne prend pas en compte le nombre de tests. L'augmentation du nombre de tests a donc faussé notre perception initiale, mais pas au point d'annuler la forte pente du mois d'août.

Si le pourcentage des tests positifs peut améliorer la compréhension, là encore il y a un biais : qui sont les personnes testées ? S'agit-il de personnes présentant des symptômes ou de personnes choisies aléatoirement dans la population ? Le résultat ne sera pas le même suivant le cas. Dans l'idéal il faudrait que cela soit le second cas pour bien évaluer la progression de la maladie au sein de la population. Cela ne signifie pas qu'il faille jeter nos courbes à la poubelle, mais on pourrait faire mieux en suivant un échan-

tillon de personnes représentatif, comme le fait l'INSEE[g] pour ses enquêtes.

Les biais sont une des sources principales de fausses conclusions. Tout résultat bizarre doit entraîner le doute et la recherche d'un biais qui aurait pu induire une erreur.

Ces deux exemples d'erreur ne sont malheureusement pas les seuls. Parfois, on utilise la moyenne alors que la médiane aurait plus de sens. Parfois, on oublie l'écart type alors que c'est l'information importante. Bref, analyser des données est un métier qu'on peut certes pratiquer en amateur, mais avec modestie. Il peut aussi être intéressant de présenter ses résultats à des experts en ligne ou dans des forums pour avoir des avis éclairés (par courriel, tweet ou dans un forum spécialisé comme StackOverFlow avec le mot clef *data-analysis* [33]).

Plus grave, certains utilisent les biais pour tromper le lecteur, en particulier à l'aide de représentations graphiques qui génèrent des « illusions d'optique ». Là encore, il est conseillé de lire le petit cours d'autodéfense intellectuelle [34].

3.3
Réutiliser les données

La réutilisation des données est probablement ce qui motive le plus le gouvernement lorsqu'il ouvre les données publiques. Cette réutilisation peut être personnelle ou d'intérêt général, une application ou un nouveau jeu de données, libre ou commerciale avec tous les intermédiaires possibles.

Parmi ces réalisations, les plus connues sont liées aux transports, que ce soit pour trouver le parcours optimal suivant le type de transport choisi, avoir les horaires, afficher les pistes cyclables, connaître la disponibilité des vélos en libre-service, trouver le parking qui a encore des places libres, etc. Quasiment toutes les villes qui ont ouvert leurs données ont vu des personnes ou des entre-

33. `https://stackoverflow.com/questions/tagged/data-analysis`
34. p.151-168 du *Petit cours d'autodéfense intellectuelle* cité p. 22.

prises développer de telles applications pour ordiphone. Si cela peut sembler anecdotique, il faut savoir que ces applications ont une valeur économique très importante. Elles permettent à chacun d'optimiser ses trajets et de gagner du temps, elles permettent aussi aux exploitants de réduire leur frais d'information auprès du grand public, puisqu'il leur suffit de transmettre les données qui seront reprises par les applications. Une étude de 2017, qui liste tous les bénéfices, a estimé ce gain entre 100 et 150 M€ par an pour Londres [35].

Une autre utilisation courante consiste à présenter les données dans une application (sur ordiphone ou site web) afin d'améliorer l'information. Certaines permettent de suivre le travail des administrations et, en cela, participent pleinement à la transparence. Une application donne le menu du jour de la cantine des enfants, une présente la qualité des cours d'eau, une autre décrypte le budget ou les marchés publics. Ces applications sont souvent citées sur les sites de données ouvertes des collectivités territoriales ou de l'État [36].

D'autres applications facilitent notre quotidien en poussant l'information dont on a besoin. Cela peut être les horaires de la piscine, des bibliothèques, de la déchetterie ou d'autres services publics. Cela peut être plus avancé, comme trouver la bibliothèque qui a un certain livre ou connaitre les aides sociales dont chacun peut disposer [37]. Certaines applications sont plus généralistes comme le site DataFrance qui intègre un grand nombre de données statistiques, poliques et pratiques sur une carte de la France [38].

Mais comment faire son application ? On a vu que Tableau permet de créer des graphiques complexes relativement facilement. Il est possible de les mettre en ligne sur son site ou sur le site de Tableau [39]. Les blogs sont une autre façon simple de mettre en

35. https://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf
36. https://www.data.gouv.fr/fr/reuses/?type=application
37. https://mes-aides.org/
38. http://map.datafrance.info/
39. Vous pouvez y apprécier la page sur la densité de la population française de Simon Lafosse https://public.tableau.com/app/profile/simon.lafosse/viz/PopulationDensity-France/FrenchPopulationDensity ainsi que

ligne des résultats. Le niveau suivant est de créer son site web, ce qui peut être fait relativement simplement avec un WCMS (*web content management system*) comme Joomal ou Drupal [40].

Programmer une application pour ordiphone est plus difficile et demande de réelles compétences en informatique. Le lecteur motivé trouvera tout ce qu'il faut sur Internet pour le guider pas à pas [41]. Pour les autres, il existe de nombreuses entreprises spécialisées dans le développement de telles applications.

3.4
Lancer une alerte

La forme la plus extrême de participation, lorsque les risques pour notre société sont avérés, est de lancer une alerte. Il s'agit alors de faire sortir de l'information, des données qui vont éclairer sur les agissements de certains.

L'alerte peut aussi être le résultat d'une analyse de données ouvertes qui laissent paraître un dysfonctionnement sévère. C'est possible, mais rare. Par exemple les statistiques d'un hôpital du Kentucky ont montré un taux anormal de réussites d'opérations cardiaques. Cela était dû à la mise en place d'un système d'opérations sur des patients qui n'en avaient pas besoin [42]. Un travail sur les données de cet hôpital aurait pu permettre de détecter la triche, mais ce sont des médecins dont les patients ont été traités dans cet hôpital qui l'ont dénoncée en premier.

Les années 2010 ont vu une explosion d'alertes d'échelle mondiale, essentiellement des alertes liées à l'évasion et l'optimisation fiscales. La plus retentissante est l'alerte d'Edward Snowden qui a expliqué, preuves à l'appui, le système de surveillance généralisée

d'autres.

40. L'hébergeur OVH explique comment mettre en place un WCMS sur ses serveurs, cf. `https://www.ovhcloud.com/fr/web-hosting/uc-website`.

41. Le site `https://developer.android.com/` est la référence pour Android et Kotlin le langage de développement qui a le vent en poupe. Pour iOS le site de référence est `https://developer.apple.com/`.

42. `http://www.khpi.org/blog/serious-accusations-of-medical-overtreatment-made-against-responders-to-uofls-partner-search/`

mis en place par la National Security Agency (NSA)[g], y compris contre ses propres citoyens, ce qui est illégal.

Si la société apprécie ces alertes, les personnes visées apprécient moins et font tout pour punir le lanceur d'alerte. Rui Pinto, qui a dénoncé les magouilles fiscales au sein du monde du football, a passé un an en prison avant que son procès commence en septembre 2020, devant répondre de 90 chefs d'accusation. Edward Snowden est toujours exilé en Russie. Ceux qui ont finalement été reconnus comme lanceurs d'alerte ont souvent été persécutés par la justice avant d'être acquittés. Denis Robert qui a révélé l'affaire Clearstream a été poursuivi pendant 10 ans pour calomnie, diffamation, recel de vol et abus de confiance. Comme l'indique la Maison des lanceurs d'alerte (MLA)[g] sur son site :

> *« L'alerte est un mécanisme vital pour le bon fonctionnement des démocraties et la sauvegarde de l'intérêt public mais elle expose souvent les lanceurs d'alerte à des risques de représailles une fois que celle-ci est lancée. Parce que ceux qui sont visés par l'alerte chercheront souvent à faire taire ceux qui signalent ou révèlent des violations de l'intérêt public plutôt que de remédier à celles-ci, il est de la plus haute importance de se préparer et de s'informer avant d'agir. »*
>
> https://mlalerte.org/procedure-aide-pour-lancer-l-alerte/

Même si la loi protège mieux qu'avant, il peut être plus prudent de se cacher derrière un organisme. En France, le Canard enchaîné a longtemps été le principal relais pour dénoncer les scandales. Dans le cas des Panama papers et des Paradise papers, les lanceurs d'alerte sont restés anonymes et se sont reposés sur l'ICIJ[g] pour analyser et diffuser l'information. La MLA[g] propose une liste de plateformes pour lancer une alerte [43]. Elle même peut aider le lanceur d'alerte.

43. https://mlalerte.org/plateformes-de-lancement-dalerte/

La loi

La France a attendu 2007 pour protéger les employés du privé qui signalent des faits de corruption, puis 2013 et le scandale du Mediator, pour élargir la protection des lanceurs d'alerte au secteur public, mais en limitant son champ au médical et à l'environnement. C'est seulement en 2016, avec la loi Sapin II, que les cas particuliers ont laissé la place à une loi généraliste de protection des lanceurs d'alerte. Cette loi a évolué pour prendre en compte les demandes de la directive européenne « sur la protection des personnes qui signalent des violations du droit de l'Union »[44]. Les évolutions sont indiquées dans la loi Waserman du 21 mars 2022.

L'article 6 de loi la Sapin II actualisée indique :

> *« Un lanceur d'alerte est une personne physique qui signale ou divulgue, sans contrepartie financière directe et de bonne foi, des informations portant sur un crime, un délit, une menace ou un préjudice pour l'intérêt général, une violation ou une tentative de dissimulation d'une violation d'un engagement international régulièrement ratifié ou approuvé par la France, d'un acte unilatéral d'une organisation internationale pris sur le fondement d'un tel engagement, du droit de l'Union européenne, de la loi ou du règlement. Lorsque les informations n'ont pas été obtenues dans le cadre des activités professionnelles mentionnées au I de l'article 8, le lanceur d'alerte doit en avoir eu personnellement connaissance.*
>
> *Les faits, informations ou documents, quel que soit leur forme ou leur support, couverts par le secret de la défense nationale, le secret médical, le secret des délibérations judiciaires, le secret de l'enquête ou de l'instruction judiciaires ou le secret professionnel de l'avocat sont exclus du régime de l'alerte défini au présent chapitre, sous réserve des dérogations prévues par la loi. »*

44. La directive 2019/1937 est présentée sur `https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L1937`

La procédure pour faire remonter une alerte peut se faire en interne, au sein de l'entreprise, ou en externe auprès de l'autorité compétente, du défenseur des droits, de la justice ou d'un organe européen [45].

La diffusion publique de l'information et des données qui l'appuient peut être faites :

— en absence de traitement à la suite d'un signalement externe dans un certain délai ;
— en cas de risque de représailles ou si le signalement n'a aucune chance d'aboutir ;
— en cas de « danger grave et imminent » ou pour les informations obtenues dans un cadre professionnel, en cas de « danger imminent ou manifeste pour l'intérêt général ».

La protection accordée protège contre le licenciement et accorde l'anonymat lors de la diffusion de l'information à la justice. Elle offre aussi la protection des sources et une certaine protection pour les personnes qui aident le lanceur d'alerte. Il est aussi prévu des sanctions contre ceux qui feraient obstacle à la diffusion de l'alerte et contre ceux qui intenteraient une procédure abusive en diffamation.

De plus différentes lois depuis 2017 permettent à l'administration fiscale à indemniser toute personne étrangère aux administrations publiques [46].

Le site de la MLA[g] [47] présente en détail cette loi et d'autres qui ont aussi un impact sur les lanceurs d'alerte. Il s'agit d'un site à bien lire si on désire lancer une alerte.

À titre de comparaison, les États-Unis disposent du False Claims

45. Un décret précisera la liste des autorités compétentes pour recueillir et traiter les alertes externes, parmi les autorités administratives ou indépendantes, les ordres professionnels... Ce décret fixera les conditions et délais dans lesquels elles devront accuser réception des signalements (sept jours maximum) et fournir un retour d'information aux lanceurs d'alerte (trois mois ou six mois si cela est justifié).

46. https://www.legavox.fr/blog/jean-claude-carra/quelle-definition-role-aviseur-fiscal-34079.htm

47. https://mlalerte.org/

Act [48] qui récompense les lanceurs d'alerte qui dénoncent les individus ou les entreprises qui fraudent. La récompense est une partie de l'amende exigée, c'est le *qui tam* [49]. En 1978 le Civil Service Reform Act définit la protection des lanceurs d'alerte du service public et génère la création de l'Office of Special Counsel vers lequel les lanceurs doivent se retourner pour transmettre leurs informations. Puis des lois comme le Whistleblower Protection Act de 1989 ont renforcé les protections des lanceurs d'alerte. Malheureusement ces protections sont limitées par de nombreux textes qui restreignent le statut de lanceur d'alerte. Ainsi Edward Snowden est poursuivi en application de l'Espionnage Act (une loi de 1917).

Et demain ?

Internet est un formidable outil de transparence comme a pu le montrer Wikileaks. Même sans pratiquer le piratage ou sans contacts internes, les informations librement disponibles sont tellement nombreuses qu'elles ouvrent les portes à des secrets bien cachés. Bellingcat [50] est un groupe de passionnés qui enquête sur des événements tragiques à partir de sources publiées sur Internet. Un de leur exploit a été de retrouver qui a abattu l'avion de la Malaysia Airlines en Ukraine en 2014, à partir de vidéos et de photos postées sur les réseaux sociaux. Ils ont ainsi retrouvé le véhicule lance-missile responsable ainsi que son trajet depuis Koursk en Russie. Des messages des soldats et de leurs familles sur les réseaux sociaux russes leur ont même permis de connaitre la liste des soldats du bataillon présent en Ukraine au moment du tir. Toujours en Ukraine, durant l'invasion russe de 2022, les vidéos de Xavier Tytelman offrent des analyses très précises sur la situation,

48. Le False Claims Act date de 1863. Il a été modifié en 1986 et 2008.

49. Les médecins qui ont dénoncé l'hôpital qui faisait des opérations inutiles ont reçu 800 000 $ chacun, cf `https://www.justice.gov/usao-edky/pr/saint-joseph-london-hospital-pay-165-million-settle-false-claims-act-allegations`

50. `https://www.bellingcat.com/`

en s'appuyant là encore sur le ROSO [51]. L'État aussi utilise Internet pour enquêter. Le ministère des Finances a déclaré utiliser les réseaux sociaux pour détecter des fraudes fiscales, en particulier de contribuables ayant un niveau de vie nettement supérieur à leur déclaration.

Avec l'ouverture des données et leur analyse par les citoyens, une nouvelle étape est franchie. Les incohérences révélatrices de situations anormales ou simplement de faiblesses à améliorer devraient être détectées plus rapidement. Les succès locaux ont aussi plus de chances d'être détectés et diffusés. L'Internet des objets améliore notre vision du monde et permet de suivre les différentes formes de pollution ou de noter des problèmes de flux dans les transports. Ainsi tout analyste de données a la matière pour devenir un journaliste d'investigation, voire un lanceur d'alerte [52].

Se pose alors la question de la rémunération. Analyser des données est un long travail qui demande des compétences élevées. Certains journalistes travaillant les données, les *data journalistes* puisque le terme existe déjà, sont financés par leur journal et/ou en écrivant des livres. Comment financer les autres, les analystes indépendants ? Le système américain verse une partie des sommes recouvertes par l'État ou de condamnations aux personnes ayant permis de découvrir le scandale. Dans le milieu informatique, les informaticiens trouvant des bogues critiques dans des logiciels sont de plus en plus souvent rémunérés par les propriétaires de ces logiciels. Mais la rémunération fait penser aux chasseurs de prime avec tout l'imaginaire négatif associé. En France, elle est officiellement rejetée, le lanceur d'alerte doit être un chevalier blanc. Les raisons sont probablement autant historiques, collaboration durant la seconde guerre mondiale, que le reflet de notre société : trop de personnes trichent et ne désirent pas que cela se sache, y compris chez les élus. Pourtant, ces blocages pénalisent l'intérêt gé-

51. Le Renseignement d'Origine Sources Ouvertes ou OSINT *Open Source Intelligencedes*, utilise tout ce qui est disponible sur Internet pour faire du renseignement. Là encore les nombreuses vidéos postées en temps réel sur les réseaux sociaux fournissent des informations précieuses.

52. Le site `https://datajournos.fr/` réunit des journalistes et autres analystes de données. Il offre des liens et la possibilité de discuter.

néral. La France aurait beaucoup à gagner [53] à avoir des analystes d'investigation indépendants qui aident à détecter les fraudes et autres malversations.

53. 100 milliards d'euros par an rien qu'en redressement des fraudes fiscales.

Conclusion

Notre société est de plus en plus gourmande de données. Les GAFAM[g] et d'autres entreprises du secteur les exploitent déjà largement. Elles échangent des services contre nos données personnelles, sans vraiment nous demander notre accord. Ce sont ces données qui ont fait de Google et de Facebook des entreprises si puissantes [54], peut-être trop. Aussi les gouvernements écrivent des lois pour éviter certains abus, mais en prenant soin de ne pas aller trop loin, pour ne pas casser ces entreprises dont on apprécie les services, en particulier en période de confinement.

Si ces entreprises aspirent nos données, elles en sont aussi de grands fournisseurs. Dans le domaine de la cartographie, les internautes peuvent apprécier Google Maps et Google Earth. Les conducteurs se font guider gratuitement par Google Maps ou l'application Waze. C'est donc un troc de données pour le bonheur de tous, même s'il n'est pas certain que les conducteurs qui laissent à Google toutes les informations sur leur parcours et leur façon de conduire en soient parfaitement conscients.

Nous sommes nous-mêmes devenus dépendants aux données. Nous voulons des services gratuits que l'on payait autrefois. Aujourd'hui l'IGN[g] offre cartes et données qu'il vendait avant que Google et OpenStreetMap ne le forcent à se réinventer.

Cet exemple fait ressortir des éléments vertueux. Google a démocratisé l'accès gratuit à une cartographie mondiale de qualité. Ce faisant, il a poussé l'IGN à changer de mission pour proposer une offre gratuite encore plus riche, mais limitée à la France. De son côté le projet OpenStreetMap a pris conscience de l'impor-

54. Leur valorisation en 2021 les classe dans les 10 premières entreprises à travers le monde, elles sont utilisées par des milliards d'individus tous les jours.

tance de la cartographie et a choisi d'en faire un commun [55]. Probablement inspiré par cet exemple, l'IGN étudie maintenant la mise dans les communs de certaines de ses données.

La bonne nouvelle, comme on l'a vu, est qu'il ne s'agit pas d'un cas isolé. Les données publiques s'ouvrent. Le paradigme a changé, l'État retourne aux citoyens les données non sensibles plutôt que de les garder ou de les vendre.

Cela ne veut pas dire que le combat est fini. Météo France continue de faire payer l'accès à ses données brutes alors qu'il est financé à plus de 90 % par des subventions publiques et une redevance. Même ce qui est gratuit et important, comme les alertes météo, n'est pas toujours à jour. Ce problème des mises à jour se retrouve dans de nombreux jeux de données disponibles sur `data.gouv.fr` (c'est à se demander si les données sont générées ou validées à la main).

Autre exemple, durant la pandémie de la Covid, le cabinet du ministre de la Santé a proposé de transmettre des données au site CovidTracker sans les déposer sur des plateformes ouvertes. Il a fallu que l'auteur de ce site refuse de les utiliser tant qu'elles ne seraient pas diffusées publiquement pour que le ministère le fasse. Par quel réflexe le ministre de la Santé a-t-il décidé de faire de la rétention d'information, quand le président de la République annonçait une transparence exemplaire [56] ?

Cependant, la situation progresse et il y a des raisons d'être optimiste. Certes le chemin va être long et cahoteux, mais il en vaut la peine !

— La transparence est fondamentalement bonne pour réduire les injustices et améliorer l'égalité des chances.

— L'expérience de l'ouverture a déjà été menée avec succès pour

55. L'IGN avait beau être financé à 80 % par les citoyens, ses cartes papier étaient au même prix que les cartes Michelin. De fait l'IGN travaillait pour l'État et non pour les citoyens. Notons que les États-Unis ont toujours laissé un libre accès à leurs données cartographiques et autres, suivant le principe que ce qui est financé par les citoyens doit être dans le domaine public.

56. L'article du Monde du 22/01/21 « *Guillaume Rozier, prodige des data sur la piste du Covid-19* » laisse deviner que cela pourrait être lié à la qualité des données remontées.

les logiciels libres ainsi que pour les protocoles d'Internet [57].

— Si l'État n'aide pas à la création de solutions libres alternatives, l'emprise des GAFAM[g] & Co sur notre société va devenir de plus en plus problématique.

Enfin, n'oublions pas que l'intelligence artificielle se nourrit de données. Cette technologie vit une renaissance depuis une décennie et ses progrès sont impressionnants, à tel point qu'elle peut changer radicalement notre économie [58]. Là encore l'ouverture de données est indispensable pour permettre des développements de solutions locales, sinon seuls ceux qui peuvent récupérer des données par eux-mêmes peuvent réussir. Sans surprise, les GAFAM sont à la pointe de la recherche en IA.

La transparence numérique va donc bien plus loin que la surveillance des actions de l'État. Il s'agit d'un changement qui touche aussi l'économie, notre vie au quotidien, notre façon d'appréhender les événements, notre engagement citoyen.

Nous avons un rôle important à jouer, de notre côté ainsi qu'en interagissant avec les services publics. Il s'agit de montrer que l'ouverture en cours des données est réellement une source de progrès ; que le travail des fonctionnaires pour les mettre en ligne n'est pas vain ; que l'exemple de Guillaume Rozier n'est pas une exception, mais l'annonce de ce changement ; que de plus en plus de citoyens, d'associations et de professionnels participeront, chacun à son niveau.

57. Les aspects techniques d'Internet sont ouverts depuis toujours. Les normes, les RFCs, sont ouvertes ainsi que les groupes de travail qui les rédigent.

58. Pas seulement, les armées intègrent l'IA et certains pays développent des robots tueurs autonomes.

GLOSSAIRE

A

AFA Agence française anticorruption
https://www.agence-francaise-anticorruption.gouv.fr/. Pages : 11, 12

AGD Administrateur Général des Données
Il coordonne l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données par les administrations
https://agd.data.gouv.org/. Page : 27

AMF Autorité des marchés financiers
https://www.amf-france.org/. Page : 34

Anticor Anticor est une association française anticorruption
https://www.anticor.org/. Pages : 11, 81

API Application Programming Interface.
Interface de programmation composée de methodes permettant interroger un autre programme informatique, un site web, le logiciel d'un capteur... Pages : 57, 80, 91

ARCEP Autorité de régulation des communications électroniques, des postes et de la distribution de la presse
Elle définie les règles et supervise le déploiement des connexions à Internet fixes et mobiles. Ses rapports et données sont librement accessibles.
https://www.arcep.fr. Page : 70

B

BSD La licence BSD est une licence ouverte écrite pour le système BSD (Berkeley Software Distribution). Il s'agit d'une licence

très permissive qui permet de réutiliser le code dans un but commercial.
https://fr.Wikipédia.org/wiki/Licence_BSD. Page : 47

C

CADA Commission d'accès aux documents administratifs
https://www.cada.fr/. Pages : 50, 84, 85

code source Ensemble d'instructions exécutées par un ordinateur pour produire un résultat à partir de données. Dans le principe, un code source est l'équivalent d'une recette en cuisine. Pages : 47, 48, 85, 94

CRPA Code des relations entre le public et l'administration,
https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000031366350. Pages : 39, 50

D

DINUM La Direction interministérielle du numérique conseille le gouvernement, accompagne les ministères dans leur transformation numérique et développe des services et ressources partagées.
https://www.numerique.gouv.fr/dinum/. Page : 43

E

Etalab Département de la DINUM chargé de coordonner la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée.
https://www.etalab.gouv.fr/. Pages : 28, 30, 47, 50, 52, 86

F

Fing Un groupe de réfléxion sur les transformations numériques. Il s'est fortement impliqué pour l'ouverture des données.
https://www.fing.org/. Page : 31

FMI Fond Monétaire International
https://www.imf.org/external/french/index.htm. Pages : 10, 12

FSF Free Software Foundation
https://www.fsf.org/. Pages : 47, 49

G

GAFAM est un sigle pour Google Apple Facebook Amazon Microsoft, les géants de l'Internet. Pages : 23, 34, 87, 109, 111

GitHub est un site web qui héberge le code source d'applications souvent libres. Fin 2021 il avait 73 millions d'utilisateurs et plus de 200 millions de projets (dépots). Il a été racheté par Microsoft en 2018.
https://github.com/. Pages : 26, 67

GPL GNU General Public License
Licence ouverte adaptée aux logiciels. Elle est virale à savoir que les codes dérivés doivent aussi être sous la licence GPL.
http://www.gnu.org/licenses/. Page : 47

Greco Groupe d'États contre la corruption
https://www.coe.int/en/web/greco. Page : 11

H

HATVP Haute autorité pour la transparence de la vie publique
https://www.hatvp.fr/. Pages : 12, 44, 81

I

ICIJ International Consortium of Investigative Journalists
https://www.icij.org/. Pages : 13, 103

IGN Institut national de l'information géographique et forestière
https://www.ign.fr/. Pages : 30, 46, 54, 70, 109

INSEE Institut national de la statistique et des études économiques
https://www.insee.fr/. Pages : 30, 53, 100

L

loi CADA loi n° 78-753 du 17 juillet 1978, dite loi CADA
https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000339241. Page : 39

loi pour une République numérique loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/. Pages : 28, 35, 40, 64, 68, 69, 73, 81

M

MLA Maison des lanceurs d'alerte
https://mlalerte.org/. Pages : 103, 105

MOOC Cours en lignes ouverts. Les principaux acteurs sont EdX, Coursera (en anglais) et FUN (en français) pour le supérieur, la Khan Académie (français et anglais) pour le primaire et secondaire. Page : 18

N

NSA National Security Agency
Agence états-unienne en charge de la surveillance des communications électromagnétique et d'Internet.
https://www.nsa.gov/. Page : 103

O

OKFN Open Knowledge Foundation,
https://okfn.org/. Pages : 49, 54, 84

OMS Organisation Mondiale de la Santé,
https://www.who.int/fr. Page : 25

open data Terme anglais pour désigner les données ouvertes, à savoir leur mise à disposition gratuite et sans restriction technique, juridique ou financière d'utilisation. Souvent la seule contrainte est de citer la source. Pages : 5, 10, 27, 35, 51, 65

OpenStreetMap est une base de données géographique ouverte construite par des bénévoles. La carte peut être consultée comme Google Maps ou servir de base à des applications libres ou commerciales.
https://www.openstreetmap.org. Pages : 30, 59, 79

OSI Open Source Initiative,
https://opensource.org/. Page : 48

R

Rapport de la mission Bothorel Pour une politique publique de la donnée – Décembre 2020
https://www.mission-open-data.fr/. Pages : 24, 27, 55

Reddit est un site de forums essentiellement anglophone avec un système de notation des textes publiés.
https://www.reddit.com/. Page : 22

Regards Citoyens est une association qui simplifie et enrichit l'accès à l'information politique
https://www.regardscitoyens.org/. Pages : 44, 54, 73

RGPD Règlement européen de 2018 sur protection de la vie privée sur Internet
https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32016R0679. Page : 63

S

SIE Système d'information sur l'eau, un service de Eau France lui-même rattaché à l'Office français de la biodiversité.
https://www.eaufrance.fr/le-systeme-dinformation-sur-leau-sie. Page : 53

T

Transparency International est une association internationale de lutte contre la corruption
https://www.transparency.org/. Pages : 11, 12, 81

trolls Provocateurs qui génèrent des polémiques sur Internet, en particulier dans les forums. Page : 22

Données, Transparence et Démocratie

Bienvenue dans l'âge des données. Nos actions sur Internet sont enregistrées au profit d'entreprises qui valorisent ces données et offrent des services en échange. Pouvons-nous faire de même ? Pouvons-nous utiliser les données de l'État pour améliorer notre démocratie ?

Depuis 2016, les données publiques doivent être ouvertes à tous. Les citoyens peuvent les analyser pour mesurer l'efficacité de l'action publique ou pour leur compte personnel. Les data journalistes les utilisent pour nous éclairer, les chercheurs pour comprendre. Ainsi la transparence permet de lutter contre la corruption et les intox, tout comme elle est source de progrès.

Ce changement de paradigme, l'accès aux données de l'État, est surtout une opportunité pour participer. Noter un dysfonctionnement permet de suggérer une amélioration, un manque peut être une opportunité économique, même un jeu de données incomplet est une occasion pour tisser des liens entre l'administration, les associations et les citoyens.

À travers cet essai, l'auteur nous propose un voyage optimiste dans le monde des données. Chemin faisant, les liens entre ces données et la transparence ouvrent la voie vers une démocratie plus ouverte, plus interactive et donc plus juste.

Olivier Ricou est enseignant-chercheur à l'EPITA. Auteur d'un guide sur Internet en 1992, fondateur puis président de l'Association des Utilisateurs d'Internet, il est un observateur averti de l'impact d'Internet sur notre société.

opendata.ricou.eu.org
5 € TTC

9 782958 187309