Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data

Samuel Goeta

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2016-ENST-0045

EDITE ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech
Spécialité “ Sociologie ”

présentée et soutenue publiquement par

Samuel GOËTA

le 8 septembre 2016

Instaurer des données, instaurer des publics
Une enquête sociologique dans les coulisses de l’open data

Directeur de thèse : **Jérôme DENIS**

Jury

Jérôme Denis — Maître assistant habilité à diriger les recherches au Centre de Sociologie de l’Innovation (Mines ParisTech) — Directeur de thèse
Emmanuel Didier — Chargé de recherche CNRS, directeur adjoint d’Epidapo Lab, Institute for Society and Genetics (Université de Californie à Los Angeles) — Examinateur
Paul N. Edwards - Professeur à la School of Information et au Department of History (Université du Michigan) - Examinateur
Sophie Pène — Professeure au Centre de Recherche Interdisciplinaire (Université Paris Descartes) — Rapporteure
Serge Proulx — Professeur émérite à l’Ecole des Médias (Université du Québec à Montréal) — Examinateur
Valérie Schafer — Chargée de recherche CNRS habilitée à diriger les recherches, à l’Institut des Sciences de la Communication (CNRS/Paris-Sorbonne/UPMC) — Rapporteure

T H È S E

Télécom ParisTech
école de l’Institut Mines Télécom – membre de ParisTech
46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Remerciements

Mes premiers remerciements vont à Jérôme Denis pour ces quatre années de travail ensemble qui ont été si enrichissantes pour moi. Je lui suis extrêmement reconnaissant de m'avoir proposé de conduire cette enquête sur l'open data, de m'avoir accompagné avec sa grande pédagogie et de m'avoir permis d'écrire un certain genre de thèse dans lequel je me reconnais. Je mesure ma chance d'avoir eu un directeur de thèse présent à chaque moment, qui a su relire et améliorer les nombreuses versions de ce manuscrit à travers plusieurs centaines de commentaires, parfois hilarants, en marge du texte. J'espère que notre dialogue se poursuivra dans l'écriture commune de prochains articles. Je suis fier d'être le premier doctorant d'une lignée qui sera longue et riche en recherches passionnantes.

Mes remerciements vont ensuite à Sophie Pène et à Valérie Schafer qui ont accepté d'être rapporteures de ce travail ainsi qu'à Emmanuel Didier, Paul Edwards et Serge Proulx d'avoir accepté de prendre part à ce jury. Je tiens par ailleurs à remercier en particulier Emmanuel Didier et David Pontille pour leurs conseils avisés lors de la soutenance à mi-parcours de cette thèse. Ils m'ont permis d'approfondir certaines questions que j'effleurais tout juste alors qu'elles sont devenues, grâce à eux, des aspects essentiels de cette thèse.

Je suis aussi très reconnaissant à Florence Millerand, Guillaume Latzko-Toth et Serge Proulx de m'avoir accueilli très chaleureusement à l'Université du Québec à Montréal où j'ai pu, pendant quatre mois, rencontrer une formidable équipe de recherche et prendre de la distance pour commencer à écrire cette thèse dans le bureau du LabCMO avec sa vue imprenable sur le Mont Royal.

Cette thèse n'aurait pas pu exister sans l'ouverture des personnes interrogées travaillant au sein des administrations françaises où j'ai pu conduire mon enquête. Je pense aussi aux « informateurs » qui ont su me guider vers les bonnes personnes. Leur ayant promis de garantir leur anonymat, je ne les citerai pas, mais je tiens à exprimer ici ma gratitude pour leur disponibilité et leur accueil favorable.

Cette thèse doit aussi beaucoup à l'environnement de travail très favorable que j'ai pu trouver à Télécom ParisTech. Un grand merci aux enseignants-chercheurs et doctorants du département SES pour leur accueil et les discussions passionnantes que nous avons pu avoir. Je pense en particulier à mes voisines de bureau, Mohini Vanhille, Marine Jouan et Chloé Lebail, pour tous les bons moments passés ensemble et à deux larrons, Clément Marquet et François Huguet, pour leurs bons conseils, leurs relectures et leur bonne humeur. Je remercie aussi Marie-Josée Vatin et Florence Besnard pour leur aide précieuse et leur disponibilité. Une pensée particulière pour Nicolas Auray avec qui j'aurais aimé discuter des résultats de cette thèse.

Cette thèse est aussi le fruit de très nombreuses rencontres académiques. Je ne vais pas, là encore, les citer tous, mais je tiens à remercier quelques personnes en particulier. Je pense à deux voisins de bureau pendant un temps : Michael Bourgatte que je remercie pour son humour et sa complicité, Benjamin Loveluck pour ses bons conseils et la grande richesse intellectuelle

de nos discussions. Je remercie aussi Clément Mabi et Romain Badouard pour nos échanges passionnants, leur enthousiasme et leur envie de creuser ensemble certains aspects de cette thèse. Je tiens aussi à remercier Sylvain Parasie de m'avoir proposé au début de cette thèse de venir présenter mes travaux au séminaire W2S.

Je suis reconnaissant à celles et ceux qui m'ont permis d'enseigner : à Régine Serra, Tomasso Vitale de l'école urbaine de Sciences Po et à Antoine Jardin, François Briatte, Joël Gombin avec qui nous avons donné le cours « Open Data for Urban Research », à Étienne Candel pour m'avoir permis d'enseigner dans mon ancien master au Celsa avec Thomas Grignon, à Magali Nonjon que je remercie de m'avoir donné la chance de retourner à Sciences Po Aix pour enseigner et de m'avoir fait confiance pour y développer la cartographie de controverse avec le beau projet du magazine Chicane. Je remercie les étudiants avec qui j'ai pu avoir des discussions enrichissantes et stimulantes.

Pour leur aide dans ce travail, je tiens à remercier les membres de l'Open Knowledge Foundation à l'international et en France, à mes anciens collègues de la Netscouade qui ont eu la bonne idée de m'envoyer à Helsinki à l'Open Knowledge Festival, à mon oncle Alain pour son appartement et à mes collègues de Yelloworking, en particulier Lilian, Marie et Victor qui m'ont fait confiance à mon arrivée à Aix, et à celles et ceux avec qui nous avons monté la belle maison coworking dans laquelle j'ai fini cette thèse.

Enfin, je suis profondément reconnaissant à mes amis et à ma famille pour leur soutien et la joie qu'ils m'apportent au quotidien. Je les remercie aussi pour leur patience, je sais que ce travail m'a parfois isolé d'eux. Je ne peux pas les citer tous et dire à quel point leurs bonnes ondes ont compté. Et je tiens enfin à remercier infiniment Camille, mon épouse, qui chaque jour a su me supporter, m'encourager, me remotiver et me faire prendre de la hauteur. Il y a une bonne dose de son énergie et de son enthousiasme dans cette thèse.

Résumé et mots clés

Alors que plus de cinquante pays dans le monde ont entrepris une démarche d'ouverture des données publiques, la thèse enquête sur l'émergence et la mise en œuvre des politiques d'*open data*. Elle repose sur l'analyse de sources publiques et sur une enquête ethnographique conduite dans sept collectivités locales et institutions françaises. Revenant sur six moments de définition de grands « principes » de l'*open data* et leur traduction en politique publique par une institution française, Etalab, ce travail montre comment la catégorisation par l'open data a porté l'attention sur les données, en particulier sous leur forme « brute », considérées comme une ressource inexploitée, le « nouveau pétrole » gisant sous les organisations.

L'enquête montre que le processus de l'ouverture débute généralement par une phase d'identification marquée par des explorations progressives et incertaines. Elle permet de comprendre que l'identification constitue un geste d'instauration qui transforme progressivement les fichiers de gestion de l'administration en données. Leur mise en circulation provoque des frictions : pour sortir des réseaux sociotechniques de l'organisation, les données doivent généralement passer à travers des circuits de validation et des chaînes de traitement. Par ailleurs, les données doivent souvent subir d'importantes transformations avant leur ouverture pour devenir intelligibles à la fois par les machines et par les humains. Cette thèse montre enfin que l'instauration concerne aussi les publics dont il est attendu qu'ils visualisent, inspectent et exploitent les données ouvertes. L'instauration des publics par des instruments très divers constitue un autre pan du travail invisible des politiques d'*open data*.

Il ressort enfin de cette thèse que l'obligation à l'ouverture des données publiques, une suite possible des politiques d'*open data*, pose de manière saillante une question fondamentale « qu'est-ce qu'une donnée ? » Plutôt que de réduire la donnée à une catégorie relative, qui s'appliquerait à toutes sortes de matériaux informationnels, les cas étudiés montrent qu'elle est généralement attribuée dès lors que les données sont le point de départ de réseaux sociotechniques dédiés à leur circulation, leur exploitation et leur mise en visibilité.

Mots clés : open data, ouverture des données publiques, données, infrastructure studies, Etalab, standards, CSV, GTFS, raw data, data frictions, instauration, publics, concours, hackathons

Abstract and keywords

Instantiate data, instantiate publics: a sociological inquiry in the backrooms of open data

As more than fifty countries have launched an open data policy, this doctoral dissertation investigates on the emergence and implementation of such policies. It is based on the analysis of public sources and an ethnographic inquiry conducted in seven French local authorities and institutions. By retracing six moments of definitions of the “open data principles” and their implementation by a French institution, Etalab, this work shows how open data has brought attention to data, particularly in their raw form, considered as an untapped resource, the “new oil” lying under the organisations.

The inquiry shows that the process of opening generally begins by a phase of identification marked by progressive and uncertain explorations. It allows to understand that data are progressively instantiated from management files into data. Their circulation provoke frictions: to leave the sociotechnical network of organisations, data generally go through validation circuits and chains of treatment. Besides, data must often undergo important treatments before their opening in order to become intelligible by machines as well as humans. This thesis shows eventually that data publics are also instantiated as they are expected to visualize, inspect and process the data. Data publics are instantiated through various tools, which compose another area of the invisible work of open data projects.

Finally, it appears from this work that the possible legal requirement to open data asks a fundamental question, “what is data?” Instead of reducing data to a relational category, which would apply to any informational material, studied cases show that they generally are applied when data are a starting point of sociotechnical networks dedicated to their circulation, their exploitation and their visibility.

Keywords: open data, data, infrastructure studies, Etalab, standards, CSV, GTFS, raw data, data frictions, data publics, open data contests, hackathons

Glossaire des abréviations

AGD : Administrateur Général des Données

API : Application Programming Interface

APIE : Agence du Patrimoine Immatériel de l'État

CADA : Commission d'Accès aux Documents Administratifs

CC : Creative Commons

CSV : Comma Separated Values

DbHD : Database Hugging Disorder

DSI : Direction des Systèmes d'Information

FING : Fondation Internet Nouvelle Génération

FOIA : Freedom of Information Act

GTFS : General Transit Feed Specification

HATVP : Haute Autorité pour la Transparence de la Vie Publique

IETF : Internet Engineering TaskForce

NOTRe : Nouvelle Organisation des Territoires de la République

ODbL : Open Database License

OGP : Open Government Partnership

OKFN : Open Knowledge Foundation

OSI : Open Source Initiative

PSI : Public Sector Information

RFC : Request for Comments

STS : Science and Technology Studies

W3C : World Wide Web Consortium

Sommaire

Introduction

Depuis maintenant presque une dizaine d'années, les politiques d'ouverture des données publiques (*open data*) ont conduit à la publication de jeux de données extrêmement variés par des États, des villes, des institutions internationales, et parfois des entreprises. Ces données, librement réutilisables par tous et généralement gratuites, sont souvent présentées comme la source d'un renouvellement de la transparence et les instruments d'une nouvelle *accountability* institutionnelle. Pour certains, elles pourraient aussi devenir des ressources majeures pour l'innovation ; par exemple, le cabinet McKinsey a estimé les retombées d'une ouverture des données généralisées à trois milliards de dollars par an dans le monde[1]. Enfin, pour d'autres, ces données pourraient servir de vecteur à une transformation des pratiques administratives. En 2013, les dirigeants des huit pays les plus riches du monde, lors de la réunion du G8 en Irlande du Nord, ont adopté une charte sur l'*open data* dans laquelle ils s'engagent à ce que l'ouverture des données devienne la pratique par défaut des administrations des pays signataires. Aujourd'hui, selon la dernière version du classement Open Data Barometer de la Web Foundation[2], 51 gouvernements dans le monde ont adopté une politique d'*open data*. Les attentes suscitées par l'ouverture des données publiques sont donc très fortes et plusieurs gouvernements se sont engagés en faveur de sa généralisation.

D'un point de vue plus personnel, j'ai commencé à travailler sur le sujet en 2010 dans le cadre d'un mémoire de fin d'études au Celsa, l'école de communication de la Sorbonne. Le 29 novembre 2010, je me rendais à Rennes pour assister à une rencontre internationale qui se tenait dans l'hémicycle de la Métropole, intitulée « *Open Data and Reuse: what is happening at local levels in Europe?* » La conférence était organisée par ePSI Platform, un organisme créé par la Commission européenne pour promouvoir la réutilisation des données publiques et la Fondation Internet Nouvelle Génération (FING), un acteur associatif majeur sur les sujets numériques. Daniel Kaplan, le directeur de la FING, avait clôturé la rencontre en faisant un bilan de la journée et en invitant les participants, pour la plupart travaillant dans

[1] McKinsey Global Institute, « Open data: Unlocking innovation and performance with liquid information », http://www.mckinsey.com/insights/businesstechnology/opendataunlockinginnovationandperformancewithliquid_information, consulté le 31 octobre 2013.
[2] Open Data Barometer, « ODB Global Report Third Edition », http://opendatabarometer.org, consulté le 21 avril 2016.

des collectivités locales, à ouvrir leurs données. J'avais été à l'époque très intrigué par ses propos dont j'ai pu retrouver un enregistrement[3].

> Ce n'est pas souvent que les acteurs publics se retrouvent face à une opportunité pareille. Vous êtes assis sur… enfin vous avez déjà un stock de données que vous produisez de fait parce que vous faites votre boulot, vous prenez des décisions, vous menez des études, vous représentez graphiquement ou informatiquement votre territoire, parce que vous coordonnez un certain nombre d'activités ou de services… Donc elles sont là et il se trouve qu'il y a des gens qui vous les demandent. Il se trouve qu'en les donnant, ça peut produire un peu de croissance ou de meilleurs services, ça peut répondre à un certain nombre d'attentes citoyennes ou d'associations ou de médias. Ça peut produire des connaissances auxquelles personne n'avait accès. C'est quand même cadeau cette histoire. Alors c'est évidemment un peu moins simple au quotidien, mais c'est quand même cadeau. Vous n'avez pas autant d'opportunités que ça qui vous passent sous la main et qui, somme toute, vous ne coûtent pas si cher que ça dans la période actuelle. Et ça, je pense que c'est quand même le point de départ. Ensuite, quand on va commencer à s'y coller, ce sera quand même moins rigolo. On va devoir parler de licence, il va falloir discuter avec des informaticiens, mais le point de départ c'est quand même une opportunité extraordinaire et ce n'est pas si compliqué que ça et pas si coûteux que ça.

Ce discours offre un excellent point de départ pour mener une enquête sur les politiques d'*open data* qui semblent se répandre partout dans le monde et concerner des administrations à tous les niveaux, mais sur lesquelles on ne sait finalement que peu de choses sur leurs origines et les conditions de leur mise en œuvre concrète dans les administrations[4].

---

[3] Vimeo, « Conclusion du forum ePSI par Daniel Kaplan », http://vimeo.com/m/18081080, consulté le 1 juillet 2016.

[4] Les études existantes se sont surtout concentrées sur les usages des données et les conséquences potentielles de l'ouverture, en particulier pour la transparence de l'action publique. Ces travaux ont montré que les données numériques sont souvent associées à un idéal d'objectivité et d'immédiateté (Birchall, 2014) qui occulte les dispositifs par lesquels la transparence se réalise (Mazzarella, 2006 ; Hansen & Flyverbom, 2014). Dans les pays en développement, l'ouverture de données sur la propriété des terres a créé de nouvelles inégalités (Raman, 2012 ; Donovan, 2012 ; Johnson, 2013) et n'est pas parvenue à installer la transparence dans les routines des administrations (Raman, 2012). Enfin, d'autres insistent sur le fait que le renouveau de la transparence souvent associé à l'ouverture des données publiques dépend en grande partie de la capacité des publics à analyser les données mises à disposition (Gurstein, 2011 ; Ruppert, 2012 ; Peixoto, 2013 ; Birchall, 2015). Bien qu'ils apportent un premier regard critique sur les projets d'open data, ces travaux ne prennent pas en considération les conditions concrètes de l'ouverture.

Premièrement, Daniel Kaplan évoque dans son discours que l'ouverture des données est « une opportunité extraordinaire » qui permettrait de produire de la croissance, des services ou de nouvelles connaissances. Si, à travers son discours, on aperçoit que les bénéfices possibles de l'ouverture des données semblent clairement établis, on a plus de mal à voir quels en sont ses grands principes. Daniel Kaplan avait résumé les demandes essentielles des revendications de l'ouverture des données sur InternetActu.net, le site de prospective de la FING, dans un article qui tente d'imaginer les conséquences possibles d'une généralisation de l'*open data* à l'ensemble des administrations publiques : « Imaginons que nous avons gagné : une part très significative des "données de service public" sont désormais accessibles et réutilisables, brutes, en un format lisible par des machines, à un coût faible, voire (le plus souvent) nul. »[5] On voit à travers cet extrait que l'ouverture ne signifie pas seulement la publication des données ; pour certains, ces données doivent être « brutes », « de service public », « lisible par les machines » ou à « faible coût. » Mais existe-t-il vraiment un consensus autour d'une définition de l'ouverture des données ? Qui sont les acteurs qui ont formulé ces revendications ? Et comment les ont-ils formulées ? Leurs demandes se rejoignent-elles ou voit-on apparaitre des lignes de tension entre ces acteurs ? Et, par ailleurs, comment ces demandes ont-elles été traduites en politiques publiques ? Quelles transformations ces revendications ont-elles provoquées dans les pratiques de publication des informations publiques ? Et, plus fondamentalement, qu'est-ce qui change lorsqu'on réclame des données ? Dans cette enquête, je vais d'abord tenter de retracer la généalogie de ce que les acteurs qualifiaient de « principes de l'*open data* » et de reconstituer comment ils ont été traduits plus localement en France dans des politiques publiques.

Pour répondre à cette première série de questions, je me suis appuyé sur des sources publiques en ligne, issues en particulier des archives du web (Schafer & Thierry, 2015) et des listes de diffusion (Akrich, 2012) dans lesquelles ont été débattues les définitions de l'*open data*. En plus de rejoindre des convictions personnelles, ma « participation observante » à l'Open Knowledge Foundation, a aussi constitué une source précieuse d'informations sur les pratiques émergentes de l'*open data*. En effet, au début de cette thèse,

[5] InternetActu.net, « L'ouverture des données publiques, et après ? », http://www.internetactu.net/2010/11/09/louverture-des-donnees-publiques-et-apres/, consulté le 9 janvier 2011.

j'ai choisi de participer avec quelques-uns à la création du chapitre français de l'Open Knowledge Foundation[6], une organisation qui, nous le verrons, a joué un rôle majeur dans la définition et le développement de l'*open data*. Par ailleurs, je me suis appuyé sur certains auteurs qui ont retracé l'émergence d'un droit de réutilisation des informations publiques (Ronai, 1997 ; Boustany, 2013 ; Trojette, 2013), l'histoire des mouvements se revendiquant de l'ouverture (Kelty, 2008 ; Strasser, 2011 ; Tkacz, 2012 ; Russell, 2014) et l'apparition de la notion d'*Open Governement Data* (Yu & Robinson, 2012).

Dans le premier chapitre, je retrace les origines des grands « principes » de l'ouverture des données en revenant sur six moments, sous la forme de six épisodes, lors desquels des manifestes, des outils de *benchmarking* (Bruno & Didier, 2013) et une déclaration diplomatique ont été élaborés. À travers ces épisodes qui montrent que l'histoire de l'*open data* n'est pas linéaire et qu'il n'existe pas de consensus complet sur sa définition, nous verrons se tisser un réseau d'acteurs autour de l'ouverture des données et émerger des lignes de tension parmi leurs revendications. Il montre qu'au-delà de grands principes dont la définition est régulièrement débattue et renégociée, ces acteurs ont répandu une nouvelle catégorisation (Cefaï, 1996) des politiques de diffusion de l'information publique fondée sur la donnée et se sont intéressés en particulier aux données brutes, au matériau de l'information avant son traitement, pour formuler la promesse d'une réduction des asymétries d'information et d'une décentralisation des lieux de calcul.

Dans le deuxième chapitre, je montre comment, en France, ces grands principes ont été traduits en politiques publiques. Dans une approche monographique, je reconstruis une

[6] Lors de l'Open Knowledge Festival à Helsinki en septembre 2011, quelques jours avant mon arrivée à Telecom ParisTech, nous étions quelques participants français à regretter l'absence de représentation officielle de la France au sein du réseau international des groupes locaux. Regards Citoyens, une association sur laquelle je reviendrai à plusieurs reprises, assurait cette représentation de manière officieuse. Ses membres ont bien voulu accompagner notre projet et rejoindre en tant que membre notre association, nous accompagnant dans la conduite de bon nombre de nos projets. Par rapport à cette thèse, mon implication au sein de l'Open Knowledge Foundation m'a donné un accès facilité à des « informateurs » même si, tout au long de ce travail, je me suis attaché, tant que possible, à clarifier mon statut auprès de mes interlocuteurs, à bien distinguer la « casquette » de doctorant de celle d'acteur associatif. Je ne vais pas revenir dans ce mémoire sur les projets que l'association a conduits et son activité fluctuante au cours de ces quatre dernières années. J'évoquerai seulement dans le récit mon implication personnelle pour clarifier mon rôle, notamment dans le premier chapitre à propos de l'Open Data Index, un outil de classement développé par le réseau international pour lequel j'ai assuré une partie du travail d'évaluation et de communication pour la France.

« histoire » (Grosjean & Lacoste, 1999), celle d'Etalab, le service en charge de l'ouverture des données de l'État français, afin de saisir les formes concrètes d'engagement des acteurs dans la dynamique des projets. La trajectoire de cette institution entre 2011 et 2016 révèle comment l'attention portée aux données a conduit à la création de structures dans l'organisation dédiée à leur exploitation et à leur circulation. En considérant la donnée comme une ressource inexploitée, le « nouveau pétrole » gisant sous les organisations, l'ouverture ne constitue qu'un des niveaux d'une politique plus large favorisant la circulation et l'exploitation des données.

Deuxièmement, dans son discours, Daniel Kaplan considère que les données sont « là », sous la main et à disposition des agents et que leur ouverture « ne coûte pas si cher » et n'est pas « si compliquée. » Pourtant, il consent que « c'est évidemment un peu moins simple au quotidien » et « quand on va commencer à s'y coller, ce sera quand même moins rigolo. » Mais, avec plus de recul[7], peut-on vraiment dire que l'ouverture des données est « cadeau » ? Que se passe-t-il concrètement dans les administrations quand les projets d'*open data* sont mis en œuvre ? Est-ce que les données sont vraiment « là », prêtes à leur ouverture ? Est-ce que l'ouverture est si peu coûteuse ? Qu'est-ce qui se passe « quand on s'y colle » ? Pourquoi faut-il discuter de la licence et négocier avec les informaticiens ? Entre quelles « petites mains » (Denis & Pontille, 2012) les données passent-elles avant d'être libérées ? Et quelles formes d'invisibilité ce travail subit-il ? À prendre pour acquise la circulation même des données, on laisse de côté les processus concrets qui amènent à leur ouverture même. Ce point est d'autant plus sensible qu'il est courant dans les projets d'*open data* d'emprunter des métaphores au domaine des ressources naturelles, qui font de la donnée un allant de soi, un « pétrole », une entité naturellement disponible qu'il suffirait de libérer pour produire de la transparence et favoriser l'innovation.

Pour travailler ces questions, je me suis essentiellement appuyé sur des études qui sont allées voir du côté des sciences où les pratiques de partage des données sont monnaie courante depuis plusieurs années. En biologie, en astronomie, en biomédecine, en

[7] Replaçons le discours de Daniel Kaplan dans son contexte. En France, seule la ville de Rennes avait ouvert des données à l'époque, il était très difficile de prendre la mesure du travail qu'allaient occasionner les projets d'open data. Le cas de Rennes a d'ailleurs servi de « laboratoire » pour la FING pour élaborer le guide de l'ouverture des données publiques que l'association a publié en janvier 2011 à destination des administrations.

géophysique, en cristallographie, en botanique, les exemples sont innombrables. Ces programmes sont en quelque sorte les ancêtres des politiques d'ouverture de données publiques : ils visent à faciliter la mise à disposition de données pour des équipes scientifiques internationales et interdisciplinaires, afin que les chercheurs puissent d'une part se focaliser sur l'analyse plutôt que sur la récolte de données et d'autre part collaborer à une échelle jusqu'ici inédite (Bowker, 2000 ; Hine, 2008). Essentiellement centrées sur l'étude des activités scientifiques et leurs transformations récentes, les *Science and Technology Studies* (STS) offrent un cadre précieux pour comprendre sous un angle jusqu'ici inédit ce qui est en jeu avec l'ouverture des données publiques. Ces programmes, et plus généralement la mise en place de plateforme de données collectives, ont été étudiés par des chercheurs issus des STS qui se sont rassemblés sous le label d'*Infrastructure Studies*. Les travaux en STS dégagent deux pistes principales pour cette enquête : l'épaisseur du travail de production des données et les frictions qu'implique leur circulation. Les premières ethnographies de laboratoires (Knorr-Cetina, 1981 ; Latour & Woolgar, 2000 ; Lynch, 1985) ont exploré méthodiquement les espaces de travail des scientifiques et ont apporté une compréhension fine des opérations successives qui sont nécessaires à la production des résultats scientifiques. Ces opérations constituent des coulisses au sens de Goffman (1973) : leur accès est généralement réservé à quelques spécialistes et très peu d'entre elles sont rendues publiques. Les ethnographies de laboratoires ont permis de mettre en lumière la chaîne des transformations mises en œuvre dans la « réduction » des données scientifiques et la fabrication des résultats destinés à circuler après qu'ils aient été figés par la publication, devenus des « mobiles immuables » (Latour, 2006). En montrant que les données n'étaient pas passivement récoltées, mais soigneusement façonnées, ces enquêtes ont mené à la remise en cause de la notion de données brutes (Gitelman, 2013) et ont en quelque sorte mis en avant le coût de production des données. Deuxièmement, les travaux en STS ont montré que les résultats scientifiques et les données sont toujours ancrés dans des écologies pratiques, orientés vers des problèmes particuliers. Leur intelligibilité est intrinsèquement liée aux conditions locales de leur production et de leur usage. À la vision de données transparentes qu'il suffirait de formater selon des standards adéquats pour assurer la collaboration, les STS ont montré que la circulation des données produit ce que Edwards (2010) appelle des « frictions ». Ce point reboucle sur le premier : si les données donnent lieu à des frictions, il faut reconnaître le travail supplémentaire que supposent leurs échanges et le coût de la fabrique et de l'entretien des métadonnées dédiées à la

« fluidification » de leur circulation (Baker & Bowker, 2007 ; Edwards et al, 2011). Au fil de ce mémoire, nous verrons ce que les *Science and Technology Studies*, et notamment les *Infrastructure Studies*, peuvent apporter à l'étude de l'*open data*, mais aussi ce que les pratiques d'ouverture des données dans les administrations peuvent apporter à la compréhension des infrastructures informationnelles contemporaines.

Pour cette partie de l'enquête, le matériau est composé d'entretiens, d'observations et de documents internes collectés principalement dans sept institutions afin de varier les configurations administratives et politiques. J'ai suivi les bases de l'analyse sociologique qualitative (Glaser & Strauss, 1967) en mettant en œuvre une ethnographie des infrastructures telle que Star l'a décrite (1999), qui fait varier les configurations institutionnelles afin de développer une recherche multisite (Marcus, 1995). Elle s'appuie sur des études de cas dont l'objectif est, non pas de produire une « représentativité » qui n'aurait aucun sens statistique, mais les conditions d'une comparaison raisonnée permettant d'appréhender la variabilité des programmes d'ouverture des données. Il s'agit donc de mettre en place une « jurisprudence de cas » (Dodier & Bazanger, 1997) qui élargisse au maximum l'éventail des éléments de compréhension. Concrètement, ces études de cas ont pris deux formes : un suivi en temps réel de certains projets d'*open data* et une enquête a posteriori pour d'autres. Je me suis appuyé sur la conduite d'entretiens semi-directifs approfondis avec des responsables de projet d'*open data* et des agents administratifs en charge de données ouvertes, sur des séances d'observation participative via le suivi et l'accompagnement de réunions et d'échanges en situation et sur l'analyse de documents de nature variée (comptes rendus, documents internes, formulaires, etc.)

La majeure partie de ce matériau a été collecté dans des collectivités locales françaises. Avec l'aide de Simon Chignard, auteur d'un ouvrage sur l'*open data* (Chignard, 2012), un premier terrain a été ouvert à Rennes Métropole. Après lui avoir présenté le projet dès son acceptation par la Fondation Mines-Télécom qui l'a financé, Simon Chignard m'a orienté vers une dizaine de contacts qu'il a identifiée dans le cadre de ses activités associatives et lors de la rédaction de son livre. À Montpellier, j'ai pris contact avec l'équipe en charge du projet d'*open data* qui, notamment du fait de liens avec Télécom ParisTech et les institutions de recherche en général, a bien voulu accepter que je rencontre les agents administratifs avec lesquels elle échangeait. Au sein de la ville de Paris, l'entrée sur le terrain a fait l'objet

de quelques brèves négociations. Avec Jérôme Denis, nous avons convaincu l'équipe en charge du projet d'*open data* de me laisser observer trois réunions de pilotage auxquelles devaient assister les correspondants d'*open data* de la municipalité. Il m'a aussi été permis de conduire des entretiens auprès de ces correspondants et de certains agents identifiés pour leur rôle dans l'ouverture de données. En échange, il était prévu que je fasse une restitution informelle auprès des correspondants, elle a été finalement faite en privé. Enfin, j'ai eu la chance de voir un ami proche embauché au conseil régional d'Ile-De-France. Avec l'accord de sa hiérarchie, il m'a raconté son travail dans des entretiens enregistrés, approfondis et non directifs et m'a donné accès à une réunion de l'ensemble des agents qui ont contribué à l'ouverture des données de la région.

En dehors des collectivités locales, mon enquête m'a aussi mené dans des institutions nationales et internationales. Sans que cela ne fasse l'objet d'un accord formel avec la direction, j'ai effectué trois entretiens approfondis au sein de la mission Etalab et avec deux de ses correspondants dans les ministères. Par ailleurs, j'ai contacté une grande entreprise de service française[8] qui, avec l'aval de son directeur de la communication et après signature d'un accord de confidentialité, a accepté que j'assiste aux réunions du comité de pilotage mensuel du projet *open data*, réunions que j'ai pu enregistrer. Enfin, après plusieurs mois de négociation, un contrat de recherche a été signé avec une organisation internationale pour l'accompagnement de son programme d'*open data*. Le contrat, conduit au cours de l'année 2013 en partenariat avec le département sciences économiques et sociales de Telecom ParisTech et soumis lui aussi à un accord de confidentialité, consistait à conduire un benchmark mesurant l'avancement d'autres projets d'*open data*. Même si je n'ai pas pu exploiter directement les entretiens et les observations dans l'organisation internationale pour la thèse, ils m'ont toutefois permis de mieux comprendre certains enjeux, de produire des effets de contraste entre les institutions et d'approfondir le rôle de certains « principes fondateurs » de l'*open data*. Côtoyer cette organisation, participer à ses réunions et contribuer à l'avancement de son projet m'a mis en position d'observateur privilégié de la naissance, de la négociation et de l'évolution d'un projet d'*open data*. En tout, j'ai conduit 47 entretiens et observations, principalement entre septembre 2012 et juillet 2014. Ce matériau a été anonymisé pour garantir la confidentialité des agents qui ont accepté de

[8] Il a été convenu que mes travaux ne permettraient pas d'identifier le nom de cette entreprise ou son contexte d'action.

répondre à mes questions. Dans certains cas, j'ai aussi anonymisé le service et/ou l'organisation concernée. Pour les chefs de projet d'*open data*, l'anonymat n'est pas garanti puisqu'ils peuvent être facilement retrouvés, étant souvent seuls à tenir ce poste dans les organisations. En revanche, le pseudonymat permet d'éviter que leurs propos soient trouvés directement dans un moteur de recherche (Jounin, 2014).

Dans le troisième chapitre, je m'intéresse à la question de l'identification des données. Contrairement à certaines injonctions qui considèrent que les données sont disponibles et connues de l'administration, l'enquête révèle qu'elles sont identifiées au prix d'un travail important d'investigation qui se nourrit d'explorations progressives et incertaines. L'identification « travaille l'organisation » (Cochoy, Garel & de Terssac, 1998) par la désignation de lieux et de personnes responsables de l'ouverture. Je montre que l'identification constitue un geste d'instauration à part entière (Souriau, 2009 ; Latour, 2015). Elle engendre un périmètre de données qui sont instaurées non seulement comme « ouvertes » ou « brutes », mais aussi comme « données » tout court.

Le quatrième chapitre s'intéresse aux frictions (Edwards, 2010) qui, après l'identification, peuvent empêcher l'ouverture des données. Plutôt que de balayer de la main les résistances exprimées par les agents, ce chapitre prend au sérieux les « bonnes raisons organisationnelles » (Garfinkel & Bittner, 1967) qu'ils invoquent. Que ce soit du fait des difficultés d'extraction de données gérées dans des systèmes d'information, de la crainte d'une « mauvaise qualité » de données jamais sorties de l'organisation ou à cause des risques pour la carrière des agents de l'ouverture de données « sensibles » sans l'approbation de la hiérarchie, les cas étudiés rappellent que la circulation des données provoque des « frictions ». Elles doivent passer à travers des circuits de validation plus ou moins formalisés et des chaînes de traitement plus ou moins abouties.

Le cinquième chapitre poursuit l'exploration du processus de l'ouverture et décrit les transformations que peuvent subir les données avant leur ouverture. Nous verrons que ces transformations sont souvent opérées par des standards de données. Je m'intéresse ici à deux formats de fichier : le CSV qui définit de grands principes d'organisation des données et un standard émergent dans le domaine des transports, le GTFS qui impose des définitions aux catégories et une organisation très précise des données. Nous verrons à

travers ces cas que les standards demandent souvent des transformations importantes et coûteuses, voire des réorganisations du travail. Du point de vue des agents, ces transformations constituent un investissement (Thévenot, 1986) dans la lisibilité des données par les machines. Enfin, des opérations d'édition (Desrosières, 2005), parfois qualifiées de nettoyages (Walford, 2013), peuvent aussi intervenir sur le contenu même des données pour atténuer certaines des sources de friction identifiées dans le chapitre 4. Nous verrons que ces cas sont guidés par une double exigence d'intelligibilité, pour les humains et pour les machines. Ils interrogent la notion de données brutes dont l'intelligibilité résulte souvent des transformations évoquées précédemment.

Troisièmement, dans son discours, Daniel Kaplan évoque des « gens qui demandent » les données, des « attentes citoyennes ou d'associations ou de médias. » Mais, en pratique, cette demande de données est-elle aussi évidente ? Comment se révèlent ces publics ? Ces publics réutilisent-ils les données par eux-mêmes ? Concrètement, comment sont produits la croissance, les services ou les connaissances que promet Daniel Kaplan, comme beaucoup d'autres, comme les bénéfices des politiques d'*open data* ? Dans le dernier chapitre, j'explore un autre pan du travail invisible des politiques d'*open data* en m'intéressant aux instruments (Lascoumes & Le Gales, 2005) qui contribuent à instaurer les publics des données ouvertes. À l'opposé d'une vision qui considérerait que les réutilisations apparaitraient d'elles-mêmes, mon enquête montre que des instruments divers, chacun à leur manière, instaurent les publics de l'*open data*. Ruppert (2012) a montré que les politiques d'*open data* « imaginent » des *data publics* qui visualisent et créent les interfaces qui rendent les données intelligibles à un public plus large et permettent de promettre un renouveau de la transparence publique. Mais elle ne se préoccupe pas de savoir si ces publics peuvent juste rester un produit de l'imagination des promoteurs des projets d'*open data*. Or l'enquête révèle que l'absence de publics de données peut constituer un problème important pour les responsables de projets d'*open data* qui ne sont pas seulement évalués en fonction du nombre ou de la qualité des données ouvertes, mais aussi des réutilisations qui en sont faites par le public. Considérer l'ouverture des données comme une « politique de l'offre de participation » (Gourgues, 2012) nous permettra de voir comment sont instaurés les publics de données qui constituent le postulat essentiel des projets d'*open data*.

Dans le sixième chapitre, je m'intéresse à trois instruments en particulier qui contribuent à instaurer les publics de données : les métadonnées, la visualisation des données et les concours de réutilisation. En restant dans l'environnement des portails, nous verrons dans quelle mesure les métadonnées peuvent atténuer les frictions que provoque la réutilisation des données. En étudiant le cas du portail de la région Ile-de-France, nous verrons comment les fonctionnalités de visualisation permettent d'élargir la cible des données ouvertes en ne demandant pas de télécharger et d'ouvrir les données, réclament de nouvelles transformations des fichiers. Enfin, j'analyse deux concours de réutilisation de données ouvertes mis en place dans le cadre des projets d'*open data* de Montpellier et de Rennes. Je les présente comme des dispositifs d'« intéressement » au sens de Callon (1986) qui, en se plaçant entre les données et les publics, créent des assemblages sociotechniques temporaires qui peuvent servir à justifier l'existence d'un public pour les données ouvertes.

Enfin, en lisant le discours de Daniel Kaplan, une question encore plus fondamentale émerge : que sont les données ? Qu'est-ce qui les différencie des informations, des documents ou encore des fichiers que traitent et gèrent au quotidien les agents dans les administrations ? Nous verrons qu'il est frappant de ne trouver nulle part une définition claire et stable de ce qu'est (ou n'est pas) une donnée dans les textes et dans les discours qui fondent les politiques d'*open data*. Or, cette question a des implications très pratiques pour le travail des agents au moment où certains responsables politiques envisagent d'imposer l'ouverture des données comme la norme dans les administrations. Avec la question des données brutes, ce sera un des fils rouges de cette thèse pour lesquels je tenterai d'apporter des réponses en conclusion.

Chapitre 1

L'invention de l'open data : retour sur six moments de définition

Le 27 novembre 2013, le service d'innovation numérique de la région Île-de-France et la Fonderie, son agence de développement de l'économie numérique, organisent une réunion à l'Institut d'Aménagement et d'Urbanisme (IAU) de Paris. Une quarantaine d'agents de la région assistent à cet évènement intitulé « *Open Data Bootcamp* ». Les organisateurs me présentent comme « observateur » tout comme Simon Chignard, auteur d'un ouvrage sur l'ouverture des données (Chignard, 2012) et un de mes « informateurs » à Rennes. Les organisateurs diffusent une vidéo de l'association nantaise Libertic intitulée « L'*open data*, on a tous à y gagner ». En voici un court extrait.

> L'Open Data est une démarche qui vise à rendre des données numériques accessibles et utilisables par tous. Pour les collectivités et les organismes publics, l'Open Data consiste à publier sur une plateforme ouverte des informations : statistiques, cartographiques, des horaires, des données économiques et financières sur les territoires... La mise à disposition des données publiques est une obligation légale. Un cadre juridique strict définit les informations qui peuvent être rendues publiques et celles qui ne le peuvent pas. Les données sensibles et à caractère personnel sont exclues, de fait, de la démarche Open Data[9].

Après la diffusion de ce clip promotionnel de deux minutes, les organisateurs distribuent un document d'une dizaine de pages. Intitulé « vadémécum de l'ouverture des données de la région », il s'inspire d'une brochure produite par Etalab, la mission en charge de l'ouverture des données du gouvernement français, à l'attention des gestionnaires de données (j'y reviendrai dans le chapitre suivant) et se présente sous la forme de questions-réponses. Laurent[10], un des animateurs de la réunion, prend le micro et invite les participants assis face à lui à répondre à un quizz qui décline le vadémécum. Il explique que ce quizz vise à « mettre en discussion un certain nombre de sujets qui tournent autour de la

[9] Libertic, « L'Open Data à la Loupe », https://www.youtube.com/watch?v=aHxv_2BMJfw, consulté le 1 juillet 2016.

[10] Les prénoms ont été changés à des fins d'anonymisation.

problématique *open data*. » Il projette une présentation sur l'écran et s'arrête sur la première question : « en quoi consiste une démarche d'ouverture et de partage des données publiques ? Moi j'ai une réponse, je ne vous la donne pas. » Laurent demande si Simon veut répondre, il hésite, mais finalement répond : « une démarche d'ouverture et de partage des données publiques, c'est de mettre en ligne des données d'une manière qui facilite leur réutilisation par des tiers. » Laurent demande si quelqu'un aurait une autre définition. Pas de réponse dans la salle, il donne donc celle qui figure dans le vadémécum : « c'est mettre à disposition sur Internet toutes les données brutes qui ont vocation à être librement accessibles et réutilisables. »

Arrêtons-nous à cet instant dans le récit de cet évènement, j'y reviendrai en ouverture des prochains chapitres. Ici, les organisateurs définissent l'*open data* en s'appuyant sur des ressources : la vidéo de Libertic, le vadémécum de la région et la contribution d'un expert. La définition varie pour chaque version : Libertic insiste sur l'obligation légale d'ouverture, Simon Chignard sur les possibilités de réutilisation et Laurent sur le caractère brut des données publiées. Malgré ces différences, les personnes présentes ne débattent pas de ces définitions. Chacune des formulations contient donc un élément relativement admis par les participants. Ce chapitre retracera les lignes principales de ces définitions, encore mouvantes et débattues, de l'ouverture des données en suivant la trajectoire des acteurs qui l'ont définie, leurs revendications et leurs moyens d'action. Cette exploration des origines de l'*open data* révèlera des ressources essentielles qui fondent les politiques publiques d'ouverture de données. Pour reconstituer la genèse de ce qu'on appelle aujourd'hui l'*open data* sans prétendre retracer une histoire exhaustive des mouvements qui s'en réclament, je vais me limiter ici à des sources de seconde main, ainsi qu'aux nombreuses ressources disponibles publiquement sur le web : des pages, des articles, des listes de diffusion, des wikis ou encore des enregistrements réalisés par le site archive.org.

Je propose ici d'isoler six moments de définition de ce qui est devenu « l'*open data* ». J'ai sélectionné ces épisodes, car on y voit progressivement se consolider de grands principes qui vont porter sur le processus de l'ouverture des données, les politiques publiques qui vont le définir et le cadre juridique de leur réutilisation. D'autre part, ils révèlent un réseau d'acteurs qui reprennent, reformulent ou contredisent les définitions et les critères exposés précédemment. Ces épisodes, délimités de manière chronologique, soulignent la diversité

des acteurs qui ont formulé des définitions de l'*open data*. Une décennie après leur formulation, on peut délimiter trois demandes essentielles dans leurs revendications : la diffusion volontaire et proactive des données produites par les agents de l'État ; leur ouverture juridique et technique ; leur publication sous leur forme la plus « brute ». Formant un cadre juridique et technique très particulier, ces demandes orientent les politiques de diffusion de l'information publique vers les données de l'État dont la réutilisation permettrait de renouveler la transparence, de nourrir l'innovation et de transformer les pratiques de travail des administrations.

Le premier épisode de cette généalogie de l'*open data* débute en 2005 avec la rédaction de l'*Open Definition* par l'Open Knowledge Foundation. Fondé sur la définition de l'*open source*, ce texte propose des critères essentiellement juridiques qui décrivent les droits des usagers d'un savoir ouvert. Notons que cette définition s'attache au savoir en général et ne formule pas de revendication pour faire évoluer les politiques publiques de diffusion de l'information. A l'inverse, le deuxième moment que je retrace, la réunion de Sebastopol en Californie en 2007, a défini des principes de l'ouverture des données gouvernementales. Ses protagonistes ont espéré que leurs revendications soient adoptées par le futur président des États-Unis. Nous verrons dans quelle mesure leur ambition a été satisfaite avec la signature par Barack Obama d'un mémorandum sur l'*Open Government* à son entrée à la Maison-Blanche. L'inventeur du web, Tim Berners-Lee joue le premier rôle des deux épisodes suivants. En 2009, il donnait une conférence restée célèbre lors de laquelle il réclamait l'ouverture des données brutes. En 2010, son modèle en cinq étoiles proposait une approche progressive pour que les gouvernements adoptent des standards ouverts de données. Tim Berners-Lee a suivi l'application de ses préconisations en conseillant le gouvernement britannique dans sa politique d'*open data*. Ensuite, nous retournons auprès de l'Open Knowledge Foundation lorsqu'en 2012, elle créait un outil de *benchmarking*, l'Open Data Index qui classe les États selon la publication d'une sélection de données « essentielles. » Enfin, je reviens sur l'adoption en 2013 d'une charte par les chefs d'État du G8 qui ont déclaré vouloir faire de l'*open data* la pratique par défaut des administrations qu'ils dirigent et reprennent en partie le travail des groupes d'intérêt évoqués précédemment. On le voit à travers ce résumé, ces six épisodes retracent la trajectoire d'acteurs et de projets très différents. Ils montrent comment un vocabulaire, des revendications, des pratiques ont accompagné l'élaboration des politiques d'ouverture de données.

Episode 1, "Open Definition" : des droits de l'usager d'un savoir ouvert

En mai 2004, Rufus Pollock, chercheur en économie à l'université de Cambridge, annonçait la création de l'Open Knowledge Foundation (OKFN), une organisation à but non commercial visant à « promouvoir l'ouverture de toutes les formes de savoir [...] information, données et tous les termes synonymes[11]. » La thèse de Rufus Pollock portait sur la valeur économique du domaine public pour les œuvres culturelles dont le copyright a expiré. L'OKFN a pour principe la discussion ouverte[12] ; ses échanges se déroulent sur des listes de diffusion publiques et archivées qui constituent un matériau très riche pour restituer les débats au sein du mouvement (Akrich, 2012).

Peu après la création de l'organisation, en août 2005, Pollock invitait les premiers membres de l'OKFN et son réseau de partenaires à adopter collectivement une définition du savoir ouvert. Dans son appel à commentaire (*Request for Comments*), Pollock souhaitait décliner une série de conditions essentiellement juridiques permettant d'établir qu'un savoir est ouvert[13]. La définition devait aussi servir à énumérer les licences ouvertes spécifiques au savoir et à fédérer des disciplines éparses.

> *Below is a first draft of an open knowledge definition. The intent is to get down in a simple but clear way what open knowledge means and the principles that open knowledge licenses should embody. The concept of openness has already started to spread rapidly beyond its original roots in software with "open access" journals, open genetics, open geodata, open content etc. However just as with software we can expect (or are already) seeing a proliferation of licenses and a potential blurring of what is open and what is not. A good definition will serve to promote compatibility, guard against dilution and provide a common thread to diverse projects across a multiplicity of disciplines. This is a first draft and all comments and corrections will be much appreciated.*

Pollock n'employait pas le terme « *open data* » dans son message et dans la définition, mais il signalait une prolifération de mouvements se revendiquant de l'ouverture. Sa définition

[11] OKFN, « Open Knowledge Foundation Launched », http://blog.okfn.org/2004/05/24/open-knowledge-foundation-launched/, consulté le 15 avril 2015.

[12] OKFN, « Governance », https://web.archive.org/web/20050311010327/http://www.okfn.org/about.html, consulté le 15 avril 2015.

[13] Open Knowledge Definition mailing list, « [okd-discuss] RFC: Open Knowledge Definition v0.1. » https://lists.okfn.org/pipermail/okfn-discuss/2005-August/005233.html, consulté le 15 avril 2015.

se fondait directement de l'expérience de l'*open source*, une généalogie clairement affirmée dans le premier brouillon du texte. Pollock y créditait l'*Open Source Definition* comme la ressource essentielle qui a servi à la rédaction de la définition, mais aussi à forger l'idée même d'ouverture.

> *Acknowledgement: The idea of openness and its specific expression here owe a huge debt to Free and Open Source software movement. In particular much of the below draws directly from the Open Source Definition available at: http://www.opensource.org/docs/definition.php*

A la lecture de cet extrait, il semblerait donc que la demande d'ouverture trouve une source commune dans les mouvements de l'*open source* et du logiciel libre (Tkacz, 2012). Kelty (2008) a montré que la demande d'ouverture est intervenue dans une lutte entre les acteurs de l'industrie informatique entre 1980 et 1993 pour standardiser les systèmes d'exploitation et les standards de télécommunication afin de permettre l'interopérabilité des ordinateurs. Dans *Open Standards and the Digital Age*, Andrew Russell (2014) de remonter plus loin pour retrouver les origines de la demande d'ouverture. Il invite à relire l'histoire des standards de télécommunications pour comprendre le foisonnement contemporain de mouvements qui se revendiquent de l'ouverture. Russell considère que l'idée d'ouverture émerge au milieu du XXe siècle dans les écrits de philosophes et de théoriciens libéraux en Europe et aux États-Unis. Il cite, en particulier, Karl Popper qui, dans *The Open Society and Its Ennemies* publié en 1945, dénonçait les « sociétés fermées » fondées sur des vérités incontestables (Tkacz, 2012 ; Russell, 2014). Il évoque aussi la cybernétique de Wiener et les théoriciens des systèmes. Après la Seconde Guerre mondiale, ces derniers ont défendu le modèle des systèmes ouverts dans lequel l'information circule et fait reculer l'entropie, la menace inéluctable du chaos annoncé par la thermodynamique (Breton, 2004 ; Lafontaine, 2004 ; Triclot, 2008 ; Turner, 2008). C'est aussi après la Seconde Guerre mondiale, du fait notamment de l'influence de la théorie cybernétique, qu'a émergé aux États Unis la notion d'*Open Government* pour réclamer la révélation des secrets de l'État (Yu & Robinson, 2012). Dans un tout autre contexte, des ingénieurs ont affirmé la supériorité des systèmes ouverts lors de la conception des standards mondiaux de télécommunication. En particulier, en août 1977, l'*International Organisation for Standardization* (ISO) a créé un groupe de travail intitulé *Open Systems Interconnection* (OSI) pour concevoir des standards ouverts d'interconnexion des réseaux et des terminaux. Bien que l'OSI a échoué dans sa mission, la métaphore de

l'ouverture s'est imposée auprès d'ingénieurs issus de domaines très variés comme une critique de l'ordre établi et un cri de ralliement. Les recherches de Russell révèlent une filiation qui dépasse celle du mouvement de l'*open source* mais, en résumant l'ouverture à une « idéologie » aux caractéristiques précises, elles tendent à gommer les différences qui peuvent exister dans et entre les groupes qui réclament l'ouverture.

On peut repérer dans les sciences une autre origine de la demande d'ouverture. Selon Yu et Robinson (2012), le terme « *open data* » est apparu pour la première fois dans les accords qu'a signés la NASA avec des pays partenaires en vue du partage de données satellitaires. En étudiant le cas des premières collections de données sur le génome, Strasser (2011 ; 2012) montre que le succès de GenBank (aujourd'hui la plus grande base de données en génétique au monde) a reposé sur la valorisation de l'ouverture du projet en opposition à ses concurrents. Dans la correspondance de Walter Goad, le concepteur de GenBank, Strasser révèle que l'ouverture a joué une fonction rhétorique essentielle dans l'attribution du contrat du National Institute of Health (NIH) des États-Unis pour la création d'une base de données génétique nationale en 1982. Dès 1979, Goad a présenté GenBank comme un projet ouvert dont les données étaient accessibles librement et gratuitement en utilisant le réseau Arpanet. Strasser montre aussi que l'expérience de GenBank a servi de modèle pour le mouvement qui réclame le partage libre et gratuit des publications scientifiques (2011). Nommé *open access* à partir d'un colloque en mai 2000, ce mouvement émerge simultanément dans deux milieux : d'une part, les chercheurs qui, en archivant en ligne leurs publications indépendamment des éditeurs, ont subi des poursuites ; d'autre part, les bibliothécaires qui ont fait face à une forte augmentation des frais d'abonnement aux revues (Pontille & Torny, à paraître). Depuis près d'une décennie, des groupes très divers se sont aussi ralliés autour de la bannière de l'ouverture avec le concept d'*open science* pour réclamer un grand nombre de mutations des sciences : la libération du savoir, la participation des citoyens à la recherche, l'évolution de ses infrastructures, la collaboration entre disciplines ou encore des méthodes alternatives d'évaluation (Fecher & Friesike, 2014). Les origines de la métaphore de l'ouverture sont donc multiples au croisement notamment de la philosophie, de la cybernétique, des standards de télécommunication et des sciences.

A la lecture de l'extrait de Pollock évoqué précédemment, on pourrait aussi croire que libre et ouvert étaient deux synonymes pour désigner les mouvements revendiquant le droit

d'inspecter et de partager le code des logiciels. Or, le fondateur de l'OKFN occultait ici les tensions et les controverses qui ont animé ces mouvements. Pour revenir sur le mouvement du logiciel libre, il faut revenir sur la figure de Richard Stallman[14]. Ancien ingénieur du MIT se revendiquant comme un hacker, Richard Stallman a créé le mouvement du logiciel libre en 1983. Il en a défini les principes en 1986 en déclinant les quatre libertés fondamentales du logiciel libre : l'utilisation, la modification, la copie et la redistribution. En 1989, il a créé la licence GPL (*General Public License*) qui instaurait le principe du *copyleft*, imposant aux utilisateurs de conserver les quatre libertés fondamentales lorsqu'ils copient, repartagent et modifient le code (Kelty, 2008). En opposition aux revendications politiques et morales de liberté de Stallman et sa Free Software Foundation, Eric Raymond et Bruce Perens, ont proposé le terme *open source* en 1998 pour « se débarrasser de l'attitude moralisatrice et belliqueuse qui avait été associée au logiciel libre par le passé, et en promouvoir l'idée uniquement sur une base pragmatique et par un raisonnement économique. »[15] La définition de l'*Open Source*, que Pollock crédite, a remis en cause le principe du *copyleft* : elle n'exigeait pas que les logiciels dérivés le respectent alors que la Free Software Foundation de Stallman en ont fait une condition essentielle de liberté (Broca, 2013).

Je ne vais pas m'étendre sur les lignes de tensions entre logiciel libre et *open source*, ce n'est pas essentiel pour comprendre les débuts de l'*open data*. Mais une comparaison des deux textes montre que l'*Open Knowledge Definition* a emprunté la majorité de son contenu à l'Open Source Definition. Pour la résumer en quelques mots, l'*Open Knowledge Definition* (devenue Open Definition quelques années après sa publication) décline les conditions de l'ouverture du savoir. Cette définition utilise la notion de savoir pour désigner un domaine très large, qui rassemble des objets informationnels très différents (donnée, document, contenu, œuvre, article…) Sans entrer dans le détail de chacune des clauses, l'Open Definition exclut les licences qui « discriminent » selon les types d'usagers ou la finalité de la réutilisation. Elle demande d'accorder trois droits fondamentaux (utiliser, réutiliser,

[14] Voir notamment sur le logiciel libre et l'open source : Auray, N. (2010) *Politique de l'informatique et de l'information*. Thèse de sociologie dirigée par Laurent Thévenot, Paris, École des hautes études en sciences sociales ; Broca, S. (2013). *Utopie du logiciel libre*, Neuvy-en-Champagne, éditions Le Passage Clandestin ; Coleman, G. (2013). *Coding Freedom. The Ethics and Aesthetics of Hacking*, Princeton, Princeton University Press ; Kelty, C. (2008), *Two Bits. The Cultural Significance of Free Software*, Durham, Duke University Press ; Weber, S. (2004), *The Success of Open Source*, Cambridge, Harvard University Press.

[15] OSI, « History of the OSI », http://www.opensource.org/history, extrait traduit par Broca (2013).

redistribuer) et autorise à contraindre les réutilisateurs à deux exigences possibles : la citation de la source et le partage des modifications de l'œuvre avec la même licence (clause de *share alike*). Les débats qui ont suivi la diffusion de ce premier brouillon n'ont pas conduit à une réécriture importante de l'Open Definition. En effet, la version définitive reprend légèrement la formulation de certains critères sans remettre en cause leurs fondements.

Les discussions sur les listes de diffusion que j'ai pu analyser révèlent un point de controverse qui n'est toujours pas refermé : l'exclusion des œuvres adossée à une clause non commerciale. Ce débat porte particulièrement sur le cas des licences Creative Commons. Créées en 2001 par Lawrence Lessig, un juriste de Harvard et militant de la culture libre, elles proposent une alternative au copyright en permettant aux créateurs de conserver certains droits et partager gratuitement des œuvres (Kelty, 2008). Chaque licence Creative Commons comporte une ou plusieurs clauses standardisées. Elles peuvent être combinées pour exiger de l'usager de citer la source (*Attribution - BY*), interdire les œuvres dérivées (*Non Derivates - ND*), les usages commerciaux (*Non Commercial - NC*) et demander le partage avec la même licence (*Share Alike - SA*). Or, l'Open Definition stipule dans son huitième article[16] que « la licence ne peut exclure l'utilisation de l'œuvre dans un domaine spécifique. Elle ne peut par exemple interdire l'utilisation de l'œuvre dans le domaine commercial. » En août 2005, Cory Doctorow, auteur de fiction et militant de la réforme du copyright, avait averti Rufus Pollock à ce sujet sur la liste de discussion de l'*Open Knowledge Definition*. Pour Doctorow, la définition classait les œuvres avec une clause non commerciale comme « fermées » au même titre qu'une œuvre publiée selon les règles classiques du droit d'auteur. En réponse, Pollock évoquait la possibilité de compléter l'*Open Knowledge Definition* par la distinction d'œuvres « faiblement ouvertes » pour désigner les œuvres libérées avec la clause non-commercial (NC). Cette proposition n'a pas figuré dans la version finale de la définition. Pollock a dû, par la suite, préciser sa position par rapport aux licences Creative Commons dans un billet de blog[17]. Il y suggérait que l'*Open Knowledge Definition* complète Creative Commons par des principes qui s'assurent de la compatibilité des œuvres entre elles : « *Any CC non-commercial license is incompatible with the CC Attribution-ShareAlike (by-sa) license. By contrast one would hope and expect that any license which is conformant with the Open*

[16] Open Definition, « Définition du Savoir Libre v.1.0. » http://www.opendefinition.org/okd/francais/, consulté le 20 avril 2015.

[17] OKFN, « The Open Definition and Creative Commons », http://blog.okfn.org/2007/10/23/the-open-definition-and-creative-commons/, consulté le 25 avril 2015.

Knowledge/Data Definition would be compatible with any other such license." Bien que la définition est toujours débattue et négociée parmi les acteurs qui se revendiquent de l'ouverture, elle a joué un rôle crucial dans le développement de l'Open Knowledge Foundation. Elle a permis de fédérer les participants à des projets qui produisent ou réutilisent un savoir ouvert. L'un d'entre eux, l'Open Data Index sera l'objet du cinquième épisode de ce chapitre.

En posant la base d'un élargissement de l'*open source* au savoir, l'Open Definition a constitué une ressource précieuse pour l'ouverture des données publiques. Elle a établi des critères essentiellement juridiques qui caractérisent l'ouverture en termes de droits des usagers sans préjuger du type de savoir concerné. Cet effort de définition s'est inscrit aussi dans le prolongement du travail de Creative Commons qui a défini une série de licences assorties à des droits et devoirs des usagers d'un savoir ouvert.

L'Open Definition a fourni une définition et des critères sans formuler de revendications ou d'exigences à l'égard des décideurs politiques. En 2007, quelques mois avant l'élection de Barack Obama, des militants de l'ouverture du savoir se sont réunis pour faire part de leurs demandes au futur président des États-Unis. Portant spécifiquement sur les données gouvernementales, ils ont réclamé une évolution radicale de leurs procédures de diffusion et des conditions de leur réutilisation.

Episode 2, "Sebastopol" : l'ouverture exhaustive des données primaires

Le 22 octobre 2007, Carl Malamud envoyait une invitation en vue de l'organisation d'une rencontre de l'« *Open Government Working Group* »[18] les 7 et 8 décembre 2007[19]. L'évènement s'est tenu à Sebastopol en Californie au sein des locaux de la maison d'édition que dirige Tim O'Reilly, l'autre organisateur de la rencontre. Après avoir fondé une des premières radios en ligne, Carl Malamud a créé le site associatif PublicRessource.org pour partager des données que le gouvernement des États-Unis refusait de diffuser librement sur le web. Malamud s'est fait connaitre en 1995 après avoir forcé un organisme fédéral, la *US Securities and Exchange Commission (SEC)*, de fournir un accès libre aux données sur les entreprises

[18] Malgré de nombreuses recherches, je n'ai pas trouvé l'origine du Working Group. David Orban dans sa biographie se revendique comme le fondateur du groupe mais je n'ai pas pu recouper cette information.
[19] Open Government Working Group, http://public.resource.org/opengovernmentmeeting.html, consulté le 10 avril 2015.

qu'elle collecte. Tim O'Reilly, quant à lui, dirigeait une maison d'édition spécialisée dans les sujets technologiques et s'est fait connaitre pour avoir popularisé l'expression « web 2.0. » Malamud et O'Reilly ont obtenu un financement pour l'organisation de cet évènement par la Sunlight Foundation, une ONG qui défend la transparence et deux grandes entreprises du numérique, Google et Yahoo. Je n'ai trouvé aucune information sur les montants et les objectifs de ces partenariats.

Dans le texte de l'invitation, les deux organisateurs se sont fixés pour ambition de lister dix principes de l'*Open Government*. Ils espéraient que les candidats à l'élection du président des États-Unis suivraient leurs recommandations : « *can the group devise a list of 10 principles of Open Government? […] The hope is to be able to publish these principles and perhaps even get candidates in the upcoming U.S. elections to adopt them.*" L'invitation suggérait un programme composé de sessions plénières et de groupes de travail dont les participants devront déterminer le contenu. Lawrence Lessig semble avoir joué un rôle moteur dans l'animation de l'évènement et dans la conduite des groupes de travail. Un autre participant, David Orban, a couvert l'évènement en publiant des photos et des vidéos[20] sur son compte Flickr, un matériau particulièrement riche pour retracer la généalogie de ces principes. Au terme des deux jours, il interrogeait Lessig dans une vidéo[21] publiée sur YouTube sur les ambitions de cet évènement et ses résultats.

> *David Orban: We are here in Sebastopol having just completed a two-day session for defining the Open Government Data principles. I have Larry Lessig here with me and, Larry, I would like to have your comments: what did we talk about? And what did we decide to release?*
>
> *Lawrence Lessig: The objective of these two days was to find simple ways to express values that a bunch of us I think agrees are pretty common. And these are values about how government could make its data available in a way that enables a wide arrange of people to make the government function better. That means more transparency about what the government is doing and more opportunity for people to leverage government data to produce insights or other great business models. So we came up with a set of principles that articulate the components to program what would qualify as open government data. In just the way the Open Source*

[20] David Orban on Flickr, « Open Government », https://www.flickr.com/photos/davidorban/sets/72157603410393877, consulté le 9 avril 2015.

[21] YouTube, « Larry Lessig on Open Government Data Principles. », http://www.youtube.com/watch?v=AmlzW980i5A, consulté le 5 avril 2015.

> *movement set out the Definition of Open Source Software, we want to set a definition of what Open Government data would look like and that's what I think we accomplished.*

Arrêtons-nous sur les propos de Lessig avant de présenter plus en détail les participants de ces deux journées de décembre 2007 et leurs réalisations. Lessig soulignait essentiellement deux bénéfices qui pourraient découler aux principes qu'ils ont édictés. L'*Open Governement Data* pourrait renouveler la participation de la société civile tout en créant des opportunités économiques pour les entrepreneurs. Notons aussi que, comme Rufus Pollock, ils se sont inspirés de l'*Open Source Definition* pour édicter les principes de l'*Open Government Data*. Enfin, on remarque un basculement sémantique entre le moment de l'invitation et de la clôture de la réunion. Les principes ne portaient plus sur l'*Open Government*, mais définissaient la notion d'*Open Government Data* créée par la même occasion. L'*Open Government*, autrefois synonyme de révélation des secrets étatiques, vise à faciliter la réutilisation de données déjà disponibles (Yu & Robinson, 2012).

Figure 1. Les participants à la réunion à Sebastopol de l'*Open Government Data working group*. Image : David Orban sur Flickr.

Trente participants (figure 1) sélectionnés par les organisateurs, une seule femme, ont accepté l'invitation de Malamud et O'Reilly. Sans revenir en détail sur le parcours de

chacun, les participants[22] ont été sélectionnés en fonction de leur affiliation à une organisation qui exige, ouvre ou réutilise des données publiques. D'autres ont été invités pour leur implication dans des projets très variés relatifs à la participation des citoyens. Trois participants ont représenté les sponsors : Google, la Sunlight Foundation et Yahoo. Selon Micah Sifry, créateur des conférences Personal Democracy Forum, le dénominateur commun de ce groupe résidait dans la défense de la liberté de l'information et dans le potentiel démocratique d'Internet.

> *The common denominator of this group of non-profit and for-profit social entrepreneurs is the conviction that freedom of information is a cornerstone of democracy, and that the Internet is the most powerful system ever invented for expanding public information and participation in the decisions that affect our lives. Thus just about everyone in attendance is actively involved in projects that take*

[22] Micah Sifry, consultant de la Sunlight Foundation et fondateur de la conférence Public Democracy Forum sur les apports des technologies numériques en politique, a publié un billet de blog dans lequel il présente les participants et leurs affiliations.

« *In attendance were Adrian Holovaty and Daniel O'Neil of the soon-to-be-unveiled EveryBlock; Michal Mugurski and Eric Rodenbeck of Stamen Design, which does amazing work with data visualization; Josh Tauberer of GovTrack.us, which makes Thomas useful and amazes the rest of us with his efficiency; Lawrence Lessig of Stanford, who's focusing his prodigious energies on the problem of corruption; Dan Newman of MAPLight.org, which is doing path-breaking work connecting money, legislators, votes and power; John Geraci of outside.in, which is localizing the blogosphere down the neighborhood level; Ed Bender of the Institute for Money in State Politics, which has state-of-the-art APIs for mashing up state-level campaign finance data; Tom Steinberg of mySociety.org, probably the world's leader in pro-democracy web services (see TheyWorkForYou.com); David Moore and Donny Shaw of OpenCongress, which brings social wisdom to unveil what's really going inside Congress now; JL Needham of Google, you've probably heard of them; Ethan Zuckerman of the Berkman Center, who has more accomplishments in the geek-to-social-good sector than anyone I know (and he's only 34!!); Greg Palmer, whose stepping down as Congressman Henry Waxman's tech director soon to venture into some exciting projects in the private sector; Jamie Taylor of Metaweb, which is building a powerful platform called Freebase for public information sharing; Bradley Horowitz of Yahoo!, you've probably heard of them too; Zack Exley of the New Organizing Institute, whose one of my favorite progressive agitators; Michael Dale of Metavid, which is bringing transparency and interactivity to Congressional video; Joseph Lorenzo Hall of UC Berkeley, one of the world's experts on e-voting; Marcia Hoffman, a staff attorney for the Electronic Frontier Foundation, which I am a proud member of; David Orban of Metasocial Web, who is exploring the frontier of networked politics; Will Fitzpatrick of Omidyar Network, which is moving toward embracing transparency as a top priority; Aaron Swartz of Open Library, which is working on creating a wiki page for every book in the world; and myself and Greg Elin of the Sunlight Labs*. »

in Techpresident, « Open Govt Data Geeks Unite, and the Rise of 3-D Journalism » http://www.techpresident.com/node/15170, consulté le 9 avril 2015.

> *publicly available data and, using all kinds of new software, make it dramatically more meaningful and engaging*[23].

La journée du 7 décembre a débuté par une session plénière lors de laquelle plusieurs participants ont présenté leurs projets comme GovTrack.us de Justin Tauberer qui extrait le site du Congrès pour suivre l'activité des parlementaires. Ces démonstrations ont servi à détailler les difficultés pour obtenir et exploiter des données publiques. Ces débats ont fait émerger les premières demandes qui vont fonder les principes de l'*Open Government Data*. Dans l'après-midi, les participants se sont réunis en groupes pour approfondir les thèmes qui ont émergé lors de la matinée. En fin de journée, le groupe a décliné une série de principes : *non proprietary*, *machine processable*, *timely*, *complete*, *free*, *non discriminable access*, *accessible*, *primary* et *reviewable* (figure 2). L'ordre et la définition des principes ont évolué avec la suite des débats.

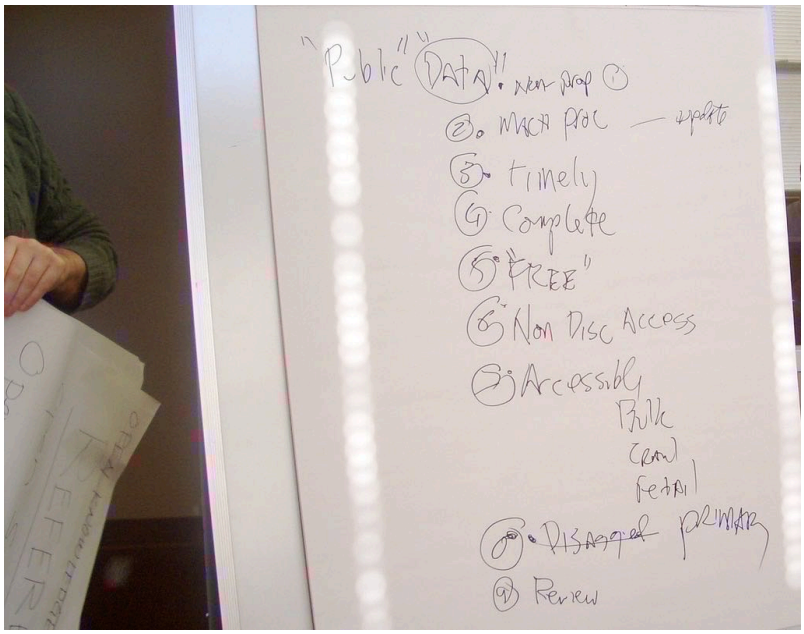


Figure 2. Document de travail de la réunion de Sebastopol : un tableau blanc avec une première liste de principes. Image : David Orban sur Flickr.

Lawrence Lessig a publié une page éditable par tous sur son wiki dans laquelle il reprenait ces principes. En une heure, Lawrence Lessig et Aaron Swartz ont édité les articles et présenté une première version des principes dans la salle où se sont tenues les séances plénières (figure 3). À la fin de la journée du 7 décembre, les principes et leur ordre ont été

[23] Ibid.

redéfinis. Par exemple, *free* est devenu *unlicensed* et les critères ont porté sur des « données publiques primaires » (*primary public data*) qui n'ont été ni agrégées ni transformées.

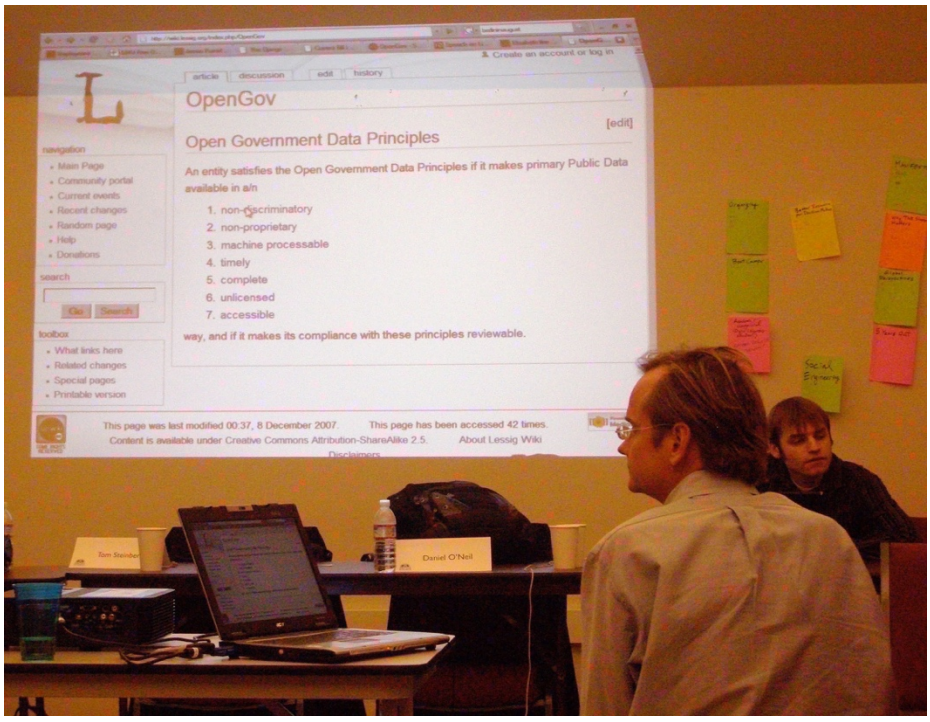


Figure 3. Le 7 décembre au soir, Lawrence Lessig présentant la première version des principes de Sebastopol. Image : David Orban sur Flickr.

L'édition des principes a repris sur le wiki à 11 h le lendemain, le 8 décembre 2007, et s'est terminée par une déclaration finale publiée le jour même. Il serait laborieux de revenir en détail sur chacune des modifications étant donné que la page wiki a été éditée soixante fois avant d'obtenir la version finale publiée au terme de la réunion. Arrêtons-nous toutefois sur trois moments pour mieux comprendre les conditions d'écriture de ces principes. Premièrement, une modification réalisée par JL Needham de Google à 11 h 16 indiquait que le groupe a choisi de décliner les principes selon une séquence en trois temps[24]. D'abord les deux premiers ont porté sur les données elles-mêmes (*complete* et *primary* dans la version finale), ensuite sur les conditions d'accès (*accessible* et *machine-processable*) et enfin sur leurs conditions d'utilisation (*non-proprietary* et *unlicensed*). Deuxièmement, les participants se sont accordés rapidement sur l'ordre des principes et leurs intitulés. Dans une modification réalisée par Aaron Swartz à 11 h 28, on retrouve l'ordre final des principes ainsi que leurs

[24] Wiki Lessig, « Open Gov », http://wiki.lessig.org/mw/index.php?title=OpenGov&direction=next&oldid=1978, consulté le 9 avril 2015

intitulés, mise à part *unlicensed* qui deviendra *license-free*[25]. Troisièmement, l'historique des modifications indique que les principaux débats ont porté sur la forme de la déclaration finale, son introduction et sur la définition qui accompagnait chaque principe. Ces modifications réalisées en séance plénière (figure 4) ont duré près de trois heures de 11 h 28 à 14 h 46.

Figure 4. Débats en séance plénière lors de la réunion de Sebastopol. Image : David Orban sur Flickr.

Le communiqué final a pris la forme d'une *Request For Comments* publiée sur publicressource.org. L'introduction insistait sur les bénéfices de l'*Open Government Data* pour la démocratie, l'innovation et l'amélioration du service public.

> *December 7–8, 2007—This weekend, 30 Open Government advocates gathered to develop a set of principles of Open Government data. The meeting, held in Sebastopol, California, was designed to develop a more robust understanding of why open government data is essential to democracy.*
>
> *The Internet is the public space of the modern world, and through it governments now have the opportunity to better understand the needs of their citizens and citizens may participate more fully in their government. Information becomes more valuable as it is shared, less valuable as it is hoarded. Open data promotes increased civil discourse, improved public welfare, and a more efficient use of public resources.*
>
> *The group is offering a set of fundamental principles for open government data. By embracing the eight principles, governments of the world can become more effective, transparent, and relevant to our lives.*

Le texte a défini une série de huit critères pour que des données gouvernementales soient considérées comme ouvertes. Les données doivent être complètes (toutes les données

[25] Wiki Lessig, « Open Gov », http://wiki.lessig.org/mw/index.php?title=OpenGov&oldid=1989, consulté le 9 avril 2015.

publiques doivent être rendues disponibles dans les limites légales) et primaires (telles que collectées à la source, non-agrégées avec le plus haut niveau de granularité). Trois critères s'appliquent au processus de diffusion : les données doivent être disponibles dès que possible (*timely*), accessibles pour le plus grand nombre d'usagers potentiels et structurées pour permettre leur traitement par des machines (*machine-processable*). Enfin, les données peuvent être utilisées par tous sans enregistrement préalable (*non-discriminatory*), dans un format ouvert (*non-proprietary*) et sans que le droit d'auteur ne s'applique (*license-free*). Sur ce dernier point, l'*Open Definition* était plus précise en attribuant une série de conditions alors que les principes de l'*Open Government Data* autorisent des « privilèges » et des « restrictions raisonnables » en matière de vie privée ou de sécurité. Enfin, le texte se termine en définissant trois termes : *public*, *data* et *reviewable*. Je reviendrai à la suite en détail sur la notion de données publiques, notons seulement que les principes définis à Sebastopol considèrent une donnée comme toute information ou tout enregistrement stocké de manière électronique. *Reviewable* signifie qu'une personne de contact est désignée pour répondre aux usagers et en cas de « violations » de ces principes. Une autorité administrative ou judiciaire doit pouvoir vérifier l'application de ces principes.

Maintenant que nous avons aperçu le contenu de ces deux journées et leur résultat, revenons-en au texte de l'invitation : les participants ont-ils rempli leur objectif, à savoir l'adoption de ces principes par le futur président des États-Unis ? Le 21 janvier 2009, jour de son investiture à la Maison-Blanche, Barack Obama a signé deux mémorandums sur l'*Open Government*. Le premier exigeait une plus grande coopération des agences gouvernementales aux procédures du *Freedom of Information Act* (FOIA). Le second réclamait que les agences gouvernementales mettent en œuvre des politiques en faveur de la transparence, la collaboration avec la société civile et la participation des citoyens.

> *We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.*

Ce mémorandum a donné un nouveau sens à l'*Open Government* qui, en plus d'être synonyme de transparence et d'*accountability*, a été associé à des politiques de participation des citoyens et de collaboration avec la société civile. Il a établi trois principes : le premier, « *Government should be transparent* », demandait que l'information fédérale soit valorisée

comme un actif stratégique (*national asset*) ; le second, « *Government should be participatory* », encourageait les politiques de participation des citoyens à la vie publique ; le troisième, « *Government should be collaborative* », recommandait des méthodes innovantes et invitait à la collaboration avec les associations (*non-profit*) et les entreprises. L'administration Obama a repris ici, en partie, certains des travaux de l'*Open Government Working Group*. Lors de sa campagne, il avait promis de restaurer la confiance par un gouvernement plus ouvert et transparent. Un groupe, nommé TIGR comme *Technology, Innovation & Government Reform Policy*, en charge de définir le programme en matière d'innovation numérique du futur président, a joué le rôle de passeur des principes établis à Sebastopol (Yu & Robinson, 2012). Parmi les membres de ce groupe de travail, deux personnes ont été nommées pour mettre en œuvre la politique d'*Open Government* du président Obama. Vivek Kundra, ancien *Chief Technology Officer* (CTO) de la ville de Washington est devenu *Chief Innovation Officer* et Beth Noveck, professeure de droit, *Deputy Chief Technology Officer*. En mars 2009, Vivek Kundra a annoncé la création de data.gov inspiré par l'expérience de Washington qu'il a pilotée. Pour inciter les agences gouvernementales à y diffuser leurs données, la Maison-Blanche a publié en décembre 2009 la directive *Open Government*. Elle stipulait que les administrations devaient fournir un plan d'action en vue de l'ouverture de nouvelles données. La directive a repris ou reformulé certains des principes de Sebastopol ; par exemple, elle demandait la publication volontaire des données publiques dès que possible (« *timely publication of information is an essential component of transparency* ») dans des formats ouverts. On retrouve encore la trace des principes de Sebastopol en 2011 dans l'*Open Government Declaration* signé par les États-Unis et sept autres pays qui a donné naissance à l'*Open Government Partnership*. Les gouvernements s'y sont engagés à fournir des informations de haute valeur, dont des données brutes (« *including raw data* »), « *in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse*. »

Ce renouveau de l'*Open Government* pourrait être considéré comme une consécration des principes établis à Sebastopol. Mais, sans entrer dans une évaluation des politiques publiques d'*Open Government* qui nous emmènerait très loin du questionnement de départ, notons que plusieurs militants, dont des participants à la réunion de Sebastopol, ont exprimé leur déception quant aux résultats de leur action. Beth Noveck a écrit un billet de blog dans lequel elle regrettait la confusion qui s'est créée autour de la notion d'*Open Government*.

> *In retrospect, "open government" was a bad choice. It has generated too much confusion. Many people, even in the White House, still assume that open government means transparency about government… The aim of open government is to take advantage of the know-how and entrepreneurial spirit of those outside government institutions to work together with those inside government to solve problems.*

Lawrence Lessig, quant à lui, s'est exprimé dans un article sur les périls de la transparence et les risques de l'utilisation « aveugle » des données publiques numériques (Lessig, 2009). La Sunlight Foundation, sponsor de la rencontre de Sebastopol, a actualisé les critères établis en 2007. En introduction de ses dix principes pour ouvrir des données gouvernementales, John Wonderlich, *Policy Director* de la fondation, a réalisé un bilan critique de l'application des principes de l'*Open Government Data*.

> *In October 2007, 30 open government advocates met in Sebastopol, California to discuss how government could open up electronically-stored government data for public use. Up until that point, the federal and state governments had made some data available to the public, usually inconsistently and incompletely, which had whetted the advocates' appetites for more and better data. The conference, led by Carl Malamud and Tim O'Reilly and funded by a grant from the Sunlight Foundation, resulted in eight principles that, if implemented, would empower the public's use of government-held data. We have updated and expanded upon the Sebastopol list and identified ten principles that provide a lens to evaluate the extent to which government data is open and accessible to the public. The list is not exhaustive, and each principle exists along a continuum of openness. The principles are completeness, primacy, timeliness, ease of physical and electronic access, machine readability, non-discrimination, use of commonly owned standards, licensing, permanence and usage costs*[26]*.*

La Sunlight Fondation admettait que les principes ont été suivis partiellement par les gouvernements. En conséquence, la fondation a amendé certains critères ou rajouté de nouveaux. Le critère de *completeness* y est plus développé, il demande la diffusion des données brutes et réclame des métadonnées complètes pour comprendre comment les données sont produites et agrégées. Contrairement à l'*Open Definition*, la Sunlight Foundation demande des conditions juridiques d'utilisation les plus permissives possible et

[26] Sunlight Foundation, « Ten Principles for Opening Up Government Data », http://sunlightfoundation.com/policy/documents/ten-open-data-principles/, consulté le 9 avril 2015.

déconseille les clauses d'attribution à la source pour faciliter la « dissémination » de l'information. Deux principes sont rajoutés : la permanence des données (leur maintien en ligne même après une mise à jour ou un changement) ainsi que leur gratuité.

On le voit, les principes de l'*open data* sont régulièrement renégociés et redéfinis. Aucun « standard » ne s'est imposé, mais une pluralité de principes qui s'entrecoupent et se contredisent pour orienter l'ouverture des données publiques. Dans un billet de blog[27], Rufus Pollock explique le rapport entre l'*Open Definition* et deux autres textes de référence qui ont établi des principes pour l'ouverture des données publiques : les huit principes de Sebastopol suivis des dix principes de la Sunlight Foundation et l'échelle en 5 étoiles de Tim Berners-Lee.

> *The Open Definition doesn't provide in-depth guidance for those publishing information in specific areas […] The [Open Government Data] principles share many of the key aspects of the Open Definition, but include additional requirements and guidance specific to government information and the ways it is published and used. […] In 2010, web Inventor Tim Berners-Lee created his 5 Stars for Linked Data, which aims to encourage more people to publish as Linked Data—that is using a particular set of technical standards and technologies for making information interoperable and interlinked. The 5 stars have been influential in various parts of the open data community, especially those interested in the semantic web and the vision of a web of data, although there are many other ways to connect data together.*

En quoi consiste cette échelle en cinq étoiles ? Comment l'inventeur du web s'est-il impliqué dans l'ouverture des données publiques ? En quoi ce texte diffère-t-il des principes précédents ? Pour répondre à ces questions, il nous faut revenir sur deux épisodes qui concernent Tim Berners-Lee : la conférence TED lors de laquelle il a formulé son appel à l'ouverture des données brutes et la publication de son modèle « en cinq étoiles » de l'*open data*.

Episode 3, "Raw Data Now" : l'entrée en politique des données « brutes »

Tim Berners-Lee a formulé son appel à l'ouverture des données brutes le 4 février 2009 à Long Beach en Californie lors d'une conférence TED. TED est un réseau de conférences retransmises gratuitement sur le web qui vise à présenter simplement des idées parfois complexes et à convaincre l'audience de s'impliquer. Le tout en moins de quinze minutes.

[27] OKFN Blog, « The Open Definition in context: putting open into practice. » http://blog.okfn.org/2013/10/16/open-definition-in-context/, consulté le 20 avril 2015.

Tim Berners-Lee[28] s'y est présenté comme l'inventeur du web. Il a raconté d'abord son parcours au sein du CERN, l'accélérateur de particules situé à la frontière franco-suisse. Il a expliqué que le web partait de sa frustration de ne pas pouvoir accéder aux documents produits dans son laboratoire. Berners-Lee a dit ressentir la même frustration aujourd'hui avec les données. Pour lui, les données sont invisibles, mais elles déterminent une grande partie de nos vies. Il s'est félicité de l'apparition de l'*Open Government Data* et des engagements du président Obama (son discours est intervenu deux mois après la signature des mémorandums).

Mais il estimait que l'ouverture des données implique aussi de transformer les attitudes des administrations. Il expliquait que, très souvent, les agents publics sont tentés de garder leurs données et trouvent une multitude de raisons pour ne pas les diffuser et permettre leur réutilisation. Dans sa présentation, Berners-Lee a fait référence au médecin suédois Hans Rosling qui, avec son outil Gapminder, a contesté des mythes répandus sur le développement des populations dans le monde. Pour produire cet outil, Rosling a dû exiger des données à une multitude d'institutions internationales. Avec l'expression « *database hugging* », Rosling avait proposé une métaphore dans laquelle les agents s'accrochent à leurs données au point de les « câliner ». Berners-Lee a repris cette métaphore et l'a mimée sur la scène de TED (figure 5).

[28] TED, « The next web. Présenté à TED Talk. », http://www.ted.com/talks/timbernersleeonthenextweb.html, consulté le 19 avril 2015

Figure 5. Tim Berners-Lee, lors de sa conférence TED de 2009, mimant le *database hugging*, l'attitude des administrations qui « s'accrochent » à leurs données.

Il a expliqué que les administrations n'arrêtent le *database hugging* qu'à partir du moment où elles ont présenté leurs données sur un beau site web. Il a demandé d'inverser cette logique et d'abord de fournir les données.

> *If you know about some data in a government department, often you find that these people, they're very tempted to keep it—Hans calls it database hugging. You hug your database, you don't want to let it go until you've made a beautiful website for it. Well, I'd like to suggest that rather—yes, make a beautiful website, who am I to say don't make a beautiful website? Make a beautiful website, but first give us the unadulterated data, we want the data. We want unadulterated data.*

Tim Berners-Lee a réclamé que les administrations fournissent d'abord les données à leur état « pur » et non modifié (*unadultarated*). Cette idée, il l'a empruntée directement à Rufus Pollock. En 2007, celui-ci a publié un billet sur le blog de l'OKFN intitulé « *Give us the data raw, Give us the data now* »[29] (il figurait dans les crédits de la conférence TED de Tim Berners-Lee[30]). Pollock y dénonçait l'attitude des administrations qui créent de belles interfaces (*shinny front end*) rapidement obsolètes alors que les données dans des standards ouverts ne vieillissent pas. Le fondateur de l'OKFN considérait que les interfaces coutent cher et détournent les administrations de leur « tâche centrale » : la publication des données. Selon lui, cette approche *interface-centric* empêche leur diffusion : "*because the interface is taken as primary, the data does not get released until the interface has been developed. This can cause significant delay in getting access to that data.*" Lorsqu'elles acceptent de publier les données, les administrations vont souvent vouloir les nettoyer et les préparer pour enlever leur complexité. Pollock suggérait alors de demander les données brutes pour les obtenir tout de suite : « *we should reply: "No, we want the data raw, and we want the data now"*. » Deux ans plus tard, Tim Berners-Lee a repris l'argumentaire de Pollock dans sa conférence TED. À la onzième minute, il demandait au public de crier « *Raw data now!* » à l'attention des administrations (figure 6).

[29] OKFN blog, « Give Us the Data Raw, and Give it to Us Now », http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/, consulté le 12 avril 2015.

[30] Berners-Lee, T. (2009). Linked Data (34) (34). Consulté 24 mars 2014, à l'adresse http://www.w3.org/2009/Talks/0204-ted-tbl/#(34)

Figure 6. Tim Berners-Lee appelle le public à crier « *raw data now* ».

Extrait de la transcription de la conférence :

« *- TBL: OK, we have to ask for raw data now. And I'm going to ask you to practice that, OK? Can you say 'raw'?*

— Audience: Raw.

— Tim Berners-Lee: Can you say 'data'?

— Audience: Data.

— TBL: Can you say 'now'?

— Audience: Now!

— TBL: Alright, 'raw data now'!

— Audience: Raw data now!"

Pourquoi revenir sur cette conférence dans cette section qui remonte aux origines de ce qui est aujourd'hui considéré comme les « principes » de l'*open data* ? D'une part, ce discours a imposé la demande de données brutes comme un aspect essentiel de l'*open data*. Par la simplicité de son message et l'influence de l'inventeur du web, cette conférence a été très visionnée et citée (le compteur de vues de TED dépasse le million). La demande de Tim Berners-Lee était facilement mémorable : ouvrez les données brutes maintenant. Il est resté très lacunaire sur le sens de cette expression évoquant seulement des données « à l'état pur » (*unadultarated*). On peut pourtant se demander à quel moment une donnée est brute, comment elle peut perdre cette qualité et s'interroger les raisons qui le poussent à réclamer

leur ouverture. Gardons en tête ces questions, car nous verrons l'implication concrète de ces revendications pour les travailleurs des données.

Enfin, cette conférence a eu un impact considérable dans la trajectoire de Tim Berners-Lee et dans la création de data.gov.uk. Son appel à l'ouverture des données brutes a été entendu au 10 Downing Street. Jusqu'en 2009, la demande d'ouverture des données publiques au Royaume-Uni était portée par une multitude d'acteurs. En particulier, la commission *Power of Information Taskforce* avait été créée pour ouvrir de nouvelles données, mais s'était heurtée à de fortes résistances. Le Guardian avec sa campagne *Free Our Data* avait structuré un réseau d'acteurs qui exigeait la diffusion gratuite des données publiques[31]. Mais l'ouverture des données de l'État britannique s'est accélérée en juin 2009 lorsque le Premier ministre, Gordon Brown, a nommé Berners-Lee *information advisor* du gouvernement aux côtés de Nigel Shadbolt, un professeur d'informatique connu pour avoir fondé la « science du web ».[32] Berners-Lee a raconté cette nomination à Charles Arthur, le journaliste du *Guardian* à l'origine de la campagne *Free Our Data*[33]. Selon l'inventeur du web, le Premier ministre voulait prendre une initiative dans le domaine des technologies, il lui a proposé « *Just put all the government's data on [the Internet]*. » Gordon Brown lui a répondu, à sa grande surprise, d'un bref « *OK, let's do it*. » Durant l'été 2009, Berners-Lee a obtenu la diffusion des données sur les accidents cyclistes. En quelques heures, des citoyens en ont produit une carte. Considérée comme une démonstration du potentiel de l'ouverture des données, elle aurait convaincu le gouvernement d'approuver la création de data.gov.uk selon l'article du Guardian cité précédemment. Dans l'été 2009, ils se sont rendus à la Maison-Blanche où ils ont rencontré l'équipe en charge de data.gov. Ils ont déclaré au *Guardian* avoir été

[31] Cette campagne a débuté en mars 2006 par la publication d'une tribune rédigée par deux journalistes du quotidien londonien. Ils protestent contre les nombreuses redevances à payer pour accéder à certaines données publiques essentielles, en particulier celles de l'Ordnance Survey, l'institut géographique du Royaume-Uni. Pour eux, les données publiques sont « les joyaux modernes de la couronne » : elles doivent être rendues au peuple pour stimuler l'innovation par leur diffusion gratuite [*]. Par la suite, la campagne s'est poursuivie avec la publication régulière d'articles sur les potentiels de l'ouverture des données et le journal est devenu un important bastion du journalisme de données.

[] Arthur, C., & Cross, M. (2006, 9 mars). Give us back our crown jewels. The Guardian. Consulté à l'adresse http://www.theguardian.com/technology/2006/mar/09/education.epublic

[32] BBC, « Web creator job "beyond politics." », http://news.bbc.co.uk/2/hi/technology/8096273.stm, consulté le 14 avril 2015.

[33] The Guardian, « 'OK, let's do it': How Britain's official data was freed. », http://www.theguardian.com/technology/2010/jan/21/how-official-data-freed, consulté le 26 juillet 2015.

convaincus par l'approche des États-Unis : un portail unique qui renvoie vers les données publiques librement réutilisables. En septembre 2009, Tim Berners-Lee et Nigel Shadbolt ont été reçus au 10 Downing Street par Gordon Brown pour faire part de leurs avancées (figure 7).

Figure 7. Tim Berners-Lee et Nigel Shadbolt reçus par Gordon Brown au 10 Downing Street le 15 septembre 2009. Source : Number10.gov.uk[34].

Lors de cette entrevue, les deux *information advisers* auraient convaincu le Premier ministre de suivre l'exemple de data.gov aux États-Unis en proposant un portail central pour les données publiques britanniques, le futur data.gov.uk.

> *Sir Tim Berners-Lee told Cabinet about the goal of delivering a single online access point to Government information, similar to the one introduced by the Obama administration in the US. […] After the update from Sir Tim and Professor Shadbolt, The Prime Minister confirmed his full support for the next phase of their work.*

Le 21 janvier 2010, data.gov.uk a été lancé en version beta. À cette occasion, Berners-Lee et Shadbolt ont publié un manifeste pour les données gouvernementales dans les colonnes du *Guardian*. Ils y ont expliqué que 2400 développeurs ont été consultés lors de la conception de data.gov.uk et que le portail est un catalogue des données publiques disponibles sur les sites web des départements du gouvernement : « *we have created a single online place where*

[34] Number10.gov.uk, « PM welcomes Sir Tim Berners-Lee to Downing Street. », http://webarchive.nationalarchives.gov.uk/20091005122636/http://www.number10.gov.uk/Page20595, consulté le 26 juillet 2015.

those looking for government data can go to find it, without having to know which department holds what and where it is. »[35] En avril 2010, Berners-Lee et Shadbolt ont annoncé avoir obtenu la publication de certaines données de l'institut géographique britannique *Ordnance Survey*, une demande qui était à l'origine de la campagne du Guardian en 2006[36]. Le 6 mai 2010, David Cameron entrait en fonction au 10 Downing Street, il avait fait de l'*open data* une des composantes essentielles de son projet de *Big Society* (Chrzanowski, 2011). Arrivant à la fin de leur mandat d'un an, Berners-Lee et Shadbolt sont restés conseillers du gouvernement en matière d'*open data*. Ils ont rejoint une nouvelle instance intitulée *Public Transparency Board* chargée de contrôler l'application des projets de transparence publique. Rufus Pollock y a siégé aux côtés notamment de Tom Steinberg. Ce dernier a participé à la réunion de Sebastopol en 2007 et a fondé MySociety, une organisation britannique connue notamment pour WhatDoTheyKnow, un site qui permet de demander aux administrations des informations publiques.

La conférence TED de Tim Berners-Lee a donc imposé la demande de données brutes comme une composante essentielle de l'*open data*. Elle a mené l'inventeur du web au cœur de la politique britannique d'ouverture des données. Un an après avoir appelé à l'ouverture des données brutes, Tim Berners-Lee a tenté de nouveau d'influencer les politiques d'*open data* en proposant un classement des formats de données qui incitait les administrations à lier leurs données et à les décrire par des nomenclatures partagées. Ce classement, prenant la forme d'un « modèle en cinq étoiles », a placé l'utilisation de formats ouverts comme une revendication essentielle de l'*open data*.

Episode 4, "5-star model" : des formats ouverts et lisibles par les machines

Après le lancement du site data.gov.uk, Tim Berners-Lee a continué de conseiller les gouvernements dans l'ouverture des données publiques. Après avoir exigé l'ouverture des données brute et déterminé la politique d'*open data* du Royaume-Uni, il a appelé ici à l'utilisation de formats ouverts de données. En 2010, il proposait un outil d'évaluation sur la page dédiée au Linked Data de son site web. Il y a proposé un nouveau concept, le Linked

[35] The Guardian, « Tim Berners-Lee and Nigel Shadbolt: our manifesto for government data. », http://www.theguardian.com/news/datablog/2010/jan/21/timbernerslee-government-data, consulté le 26 juillet 2015.

[36] BBC, « Ordnance Survey offers free data », http://news.bbc.co.uk/2/hi/technology/8597779.stm, consulté le 26 juillet 2015.

Open Data. Pour pouvoir prétendre à ce label, les données liées doivent être diffusées selon une licence ouverte.

> *Linked Open Data (LOD) is Linked Data which is released under an open licence, which does not impede its reuse for free. Creative Commons CC-BY is an example open licence, as is the UK's Open Government Licence. Linked Data does not, of course, in general have to be open—there is a lot of important use of Linked data internally, and for personal and group-wide data. You can have 5-star Linked Data without it being open. However, if it claims to be Linked Open Data then it does have to be open, to get any star at all*[37].

Berners-Lee ne s'est pas attardé sur les critères juridiques contrairement à l'*Open Definition*, il a défini une licence ouverte par la réutilisation gratuite des données qu'elle accorde. Il citait comme exemples les licences Creative Commons (CC-BY en particulier) et celle adoptée par le gouvernement britannique. Il proposait un modèle en cinq étapes, une hiérarchie de la première à la cinquième étoile qui, à la manière de la classification des hôtels, permet aux réutilisateurs de distinguer la qualité des données. Ce modèle s'adressait particulièrement aux gouvernements pour les encourager à adopter le Linked Data pour ouvrir leurs données. Sur la boutique en ligne du W3C, le consortium en charge des standards du web, Berners-Lee vend même des tasses sur lesquelles figure son modèle en cinq étoiles. Il a déclaré espérer que la circulation de ces tasses dans les bureaux inciterait à ouvrir et lier toujours plus de données (figure 8).

[37] W3C, « Linked Data - Design Issues », http://www.w3.org/DesignIssues/LinkedData.html, consulté le 28 juillet 2015.

Figure 8. Tasse du W3C reprenant le modèle en cinq étoiles de Tim Berners-Lee. Source : w3.org.

Dans la hiérarchie de Tim Berners-Lee, les données sont ouvertes dès la validation du premier critère. Il a considéré que, plus une donnée obtient d'étoiles, plus elle sera simple à utiliser. La première étoile demande la publication sur le web des données, quel que soit leur format avec une licence ouverte. La deuxième étoile exigeant que les données publiées sur le web sous une licence ouverte soient lisibles par les machines et structurées. En plus des deux autres critères précédents, l'obtention de la troisième étoile réclame la publication des données dans un format non propriétaire. Pour obtenir la quatrième étoile, les données doivent être publiées dans les standards ouverts du W3C (RDF et SPARQL) qui imposent que les objets contenus dans les données soient décrits. Enfin, la cinquième étoile demande qu'elles soient liées à d'autres données publiées sur le web.

Dans les administrations que j'ai étudiées dans mon enquête, le modèle de Tim Berners-Lee a été employé essentiellement pour inciter les agents à ouvrir les données dans des formats ouverts comme le CSV plutôt que d'utiliser le format Excel. L'utilisation de formats sémantiques, les deux derniers niveaux du modèle, réclame un travail trop important de transformation des données au regard des moyens alloués aux projets d'*open data* que j'ai pu étudier. Un des responsables de data.gov.uk m'a ainsi expliqué que son équipe a visé l'utilisation des standards du Linked Data, mais qu'il peinait à convaincre les

administrations de l'intérêt de ce format. Pour l'instant, il vise les trois étoiles pour l'ensemble des jeux de données. En France, dans l'équipe d'Etalab, le constat était similaire : demander aux agents de publier leurs données en tant que Linked Data imposait trop de contraintes alors que la simple publication des données posait déjà problème. Toutefois, un agent d'Etalab interrogé m'a expliqué que, comme au Royaume-Uni, ils essayaient de publier le plus de données possible en CSV. Comme la publication des données dans les formats du Linked Data n'était pas requise par les agents en charge de la mise en œuvre des politiques d'*open data*, je n'ai pas pu suivre au cours de mon enquête de cas de publication de données dans ces standards.

Retenons donc du classement en cinq étoiles qu'il a suggéré aux administrations d'ouvrir les données de manière progressive. En quelque sorte, il leur propose une marche à suivre : d'abord publier les données sur le web avec une licence ouverte, ensuite avec des formats lisibles par les machines puis dans des formats ouverts et enfin éventuellement selon les standards du Linked Data. Tim Berners-Lee s'est servi de ce modèle pour composer un outil d'audit de l'ouverture des données, l'Open Data Barometer. En novembre 2009, il a créé la Web Foundation, un organisme à but non lucratif qui vise à maintenir le web libre et accessible. Elle publie depuis 2013 l'Open Data Barometer, un classement des pays selon leurs politiques d'*open data*. Il évalue les politiques d'accès à l'information, les initiatives publiques en matière d'*open data* et l'ouverture d'une liste de données publiques jugées essentielles[38]. Je ne vais pas décrire ici le fonctionnement de ce classement, car sa méthodologie s'étend sur près d'une vingtaine de pages et les articles de presse n'en retiennent essentiellement que les scores. Je vais plutôt m'intéresser à un autre classement, l'Open Data Index de l'OKFN, car il a établi neuf critères de l'ouverture d'un jeu de données. Dans mon enquête, ce classement a joué un rôle déterminant en faveur de l'ouverture de certaines données. Surtout, ses critères ont consolidé la définition d'une donnée ouverte et marqué une rupture avec l'ouverture complète revendiquée dans les principes de Sebastopol en délimitant un périmètre de données essentielles, à ouvrir en priorité.

Episode 5, « Open Data Index » : un score d'ouverture et des données « essentielles »

[38] J'ai participé en ce projet en tant que lead researcher en 2013 pour deux pays : la Belgique et la Tunisie.

Le 17 avril 2012, un billet sur le blog de l'OKFN présentait le nouveau projet de l'organisation, l'Open Data Index[39]. Selon Rufus Pollock, l'*open data* s'est propagé partout dans le monde, mais les données n'ont pas toujours été publiées de la bonne manière : « *simply putting a few spreadsheets online under an open license is obviously not enough. Doing open government data well depends on releasing key datasets in the right way*.»[40] L'Open Data Index est un outil de *benchmarking* qui évalue le niveau d'ouverture de données jugées essentielles. L'Index ne s'intéresse pas aux politiques publiques d'*open data* et à leur mise à œuvre, il analyse seulement le niveau d'ouverture des données publiées par les gouvernements : « *focussing on data will also allow us to keep the census very concrete. Analysing policy or even law is a complex process ; whether a dataset is 'open' or not is usually a clear yes or no answer. In this Census, we are interested in the current status of data: is it open, is it accessible, can I use it now*[41]*?*"

L'Open Data Index reprend les trois formes principales des outils de *benchmarking* (Bruno & Didier, 2013). Premièrement, c'est un tableau de bord, qui indique en un coup d'œil le niveau d'ouverture d'une donnée. Les résultats de l'Index sont présentés dans un tableau dans lequel chaque ligne concerne un pays. Chaque cellule indique le niveau d'ouverture d'une donnée essentielle avec neuf barres. Un code couleur confirme le respect de chaque critère (vert : ouvert/rouge : fermé/bleu : incertain). Deuxièmement, l'Index est un baromètre mis à jour chaque année qui valorise la progression des « bons élèves ». Enfin, c'est un palmarès dans lequel les pays sont classés en totalisant les scores. Cette présentation vise à identifier en un coup d'œil les pays en tête et les données les plus ouvertes (figure 9).

[39] Au départ, le projet était nommé Open Data Census puis il est devenu « Open Data Index » en 2013. Pour faciliter la lecture, je vais utiliser uniquement sa dénomination actuelle, Open Data Index.

[40] OKFN Blog, « The Open Data Census – Tracking the State of Open Data Around the World », http://blog.okfn.org/2013/02/20/open-data-census-tracking-the-state-of-open-data-around-the-world/, consulté le 26 juillet 2015.

[41] OKFN Blog, « Launching the Open Data Census 2012! », http://blog.okfn.org/2012/04/17/launching-the-open-data-census-2012/, consulté le 30 juillet 2015.

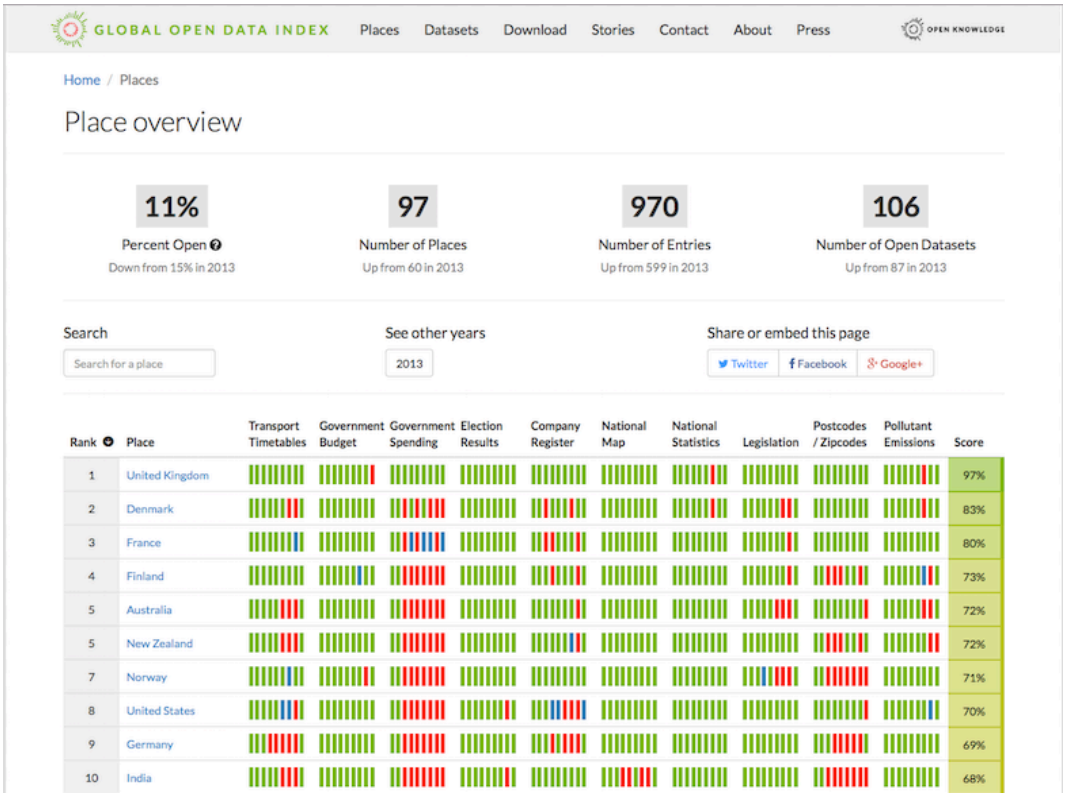


Figure 9. Résultats de l'Open Data Index 2014 présentés par pays.

Dans son billet de 2013, Rufus Pollock indiquait que trois principes guidaient l'Open Data Index : la comparabilité, l'importance et le classement. Comparables, les données ont été soumises à des critères uniformisés pour établir un score d'ouverture quelque soit le contexte juridique ou politique. Pollock soulignait que le but de l'Index était d'évaluer la « qualité » des données plutôt que leur quantité. Deuxièmement, les données de l'Open Data Index ont été sélectionnées en fonction d'un critère d'importance. Pollock considérait que certains jeux de données doivent être ouverts en priorité : "*We want to know whether governments around the world are releasing key datasets, for example critical information about public finances, locations and public transport rather than less critical information such as the location of park benches or the number of streetlights per capita.*" Un gouvernement pourrait ainsi arriver en tête de l'Open Data Index en publiant uniquement les données essentielles. Il pourrait même ne pas avoir de politique de transparence ou de droit à l'information. Enfin, troisième principe, le classement : il est fondé sur la somme des scores obtenus par chaque pays pour chacun des dix jeux de données clés. Pollock indiquait que l'Open Knowledge Foundation s'est inspirée de plusieurs classements relatifs à la transparence des

gouvernements publiés par des ONG[42] dont il a repris la mécanique. L'Open Data Index repose sur des contributeurs bénévoles qui évaluent le niveau de conformité des données essentielles par rapport aux critères. En 2014, une nouvelle procédure a exigé la validation des contributions par un *reviewer* bénévole, sélectionné pour son expertise ou son implication dans l'OKFN.

L'OKFN a établi une liste des jeux de données essentiels en 2012, après une première proposition de Rufus Pollock sur les listes de diffusion de l'OKFN. De 2012 à 2014, ces jeux de données étaient définis par les intitulés suivants :

- les résultats des élections nationales au niveau de la circonscription ;
- les registres des entreprises (sans les informations financières sur le bilan notamment) ;
- les données géographiques sur le territoire national à l'échelle 1:250000 ;
- les dépenses de l'État au niveau transactionnel ;
- le budget de l'État ;
- la législation (lois et décrets) ;
- les données statistiques nationales sur l'économie et la démographie ;
- la base de données des codes postaux avec leur géolocalisation ;
- les horaires des transports publics nationaux ;
- les données environnementales sur les principaux polluants.

L'OKFN a précisé brièvement comment cette sélection s'est opérée : « *the datasets have been chosen for their breadth and relevance. We have attempted to select data which most governments could reasonably be expected to collect*." Par rapport à l'*Open Definition*, aux principes de Sebastopol et aux travaux de Tim Berners-Lee, l'Index a marqué plusieurs ruptures importantes. Plutôt que de réclamer l'ouverture complète des données sans évaluer préalablement leurs conditions concrètes de production et de réutilisation, il a délimité une sélection de données essentielles à ouvrir prioritairement et partout dans le monde. Cette sélection s'est effectuée en fonction de l'offre, l'OKFN estimant que ces données sont

[42] Pollock cite en particulier trois classement relatifs à la transparence des gouvernements et à la perception du niveau de corruption des agents publics. Il indique que ces classements les ont inspirés à agréger les résultats dans un seul classement et à les présenter sous la forme d'un tableau de bord : « *Inspired by work such as Open Budget Index from the International Budget Partnership, the Aid Transparency Index from Publish What You Fund, the Corruption Perception Index from Transparency International and many more, we felt a key aspect is to distill the results into a single overall ranking and present this clearly.* »

produites dans chaque pays et de la demande, ces données sont considérées comme « pertinentes » pour les réutilisateurs, essentielles à la transparence et à la création de services aux citoyens. L'existence d'une demande n'y était pas considérée comme un préalable mais comme une variable dépendant du contenu même des données. D'autre part, ce classement a fragilisé les définitions de l'ouverture des données en atomisant les critères et en établissant une hiérarchie parmi ces derniers. En effet, une pondération est attribuée à chacune des questions qui permettent d'établir le score d'ouverture : les critères techniques (six premières questions) portent sur cinquante points, autant que les critères légaux (figure 10).

Icône	Question	Score
	1. Does the data exist?	5
	2. Is data in digital form?	5
	3. Publicly available?	5
	4. Is the data available for free?	15
	5. Is the data available online?	5
	6. Is the data machine-readable?	15
	7. Available in bulk?	10
	8. Openly licensed?	30
	9. Is the data provided in a timely and up to date basis?	10

Figure 10. Les neuf critères de l'ouverture d'une donnée selon l'Open Data Index avec leur pondération. Adapté de la page Methodology de l'Open Data Index : index.okfn.org/methodology.

Pour l'OKFN en particulier, l'Index a complété son texte fondateur, l'*Open Definition* sur deux points en particulier : la gratuité et l'utilisation de formats ouverts. En 2005, la première version de la définition demandait que les données soient diffusées à « un prix ne dépassant pas un coût raisonnable de reproduction » alors que le quatrième critère de l'Index exige la gratuité des données. Le sixième critère demande des données *machine-readable* alors que

l'*Open Definition* exige l'ouverture « dans un format qui ne présente pas d'obstacles techniques ».[43] Notons que l'Index n'a pas exigé l'utilisation d'un format ouvert, contrairement aux principes de Sebastopol ou aux recommandations de Tim Berners-Lee. On voit donc bien que les acteurs se revendiquant de l'*open data* n'aboutissent pas à un consensus sur les conditions de l'ouverture d'une donnée qui sont sans cesse débattus et remises en cause.

Par rapport aux moments de définition précédents, l'Index a imposé une priorité dans l'ouverture de certaines données jugées plus essentielles que d'autres et a introduit une évaluation du contenu des données et de leur demande de réutilisation. L'Open Data Index a créé les conditions d'une concurrence entre les pays par la définition de critères mesurables du niveau d'ouverture d'une donnée. À travers ces cinq épisodes, nous avons abordé des initiatives locales et des instruments dont on a du mal à mesurer la portée concrète au niveau international. Mais en 2013, l'*open data* est apparu à l'agenda des discussions des chefs d'État participants à la réunion du G8 en Irlande du Nord. Comment la charte qui en résulte a-t-elle traduit ces initiatives dans le langage de la diplomatie et des engagements internationaux des États ?

Episode 6, « G8 » : la reconnaissance de données à forte valeur

Les 17 et 18 juin 2013 à Loughe-Erne en Irlande du Nord, le Premier ministre britannique, David Cameron, accueillait la réunion du G8, la rencontre de huit chefs d'État parmi les plus grandes puissances économiques mondiales (Allemagne, Canada, États-Unis d'Amérique, France, Royaume-Uni, Italie, Japon, Russie). Les journalistes en ont essentiellement retenu les déclarations autour de la Syrie et de la lutte contre l'évasion fiscale. David Cameron entendait pourtant faire de Loughe-Erne le « sommet de la transparence. » L'agenda comportait une session (figure 11) sur la publication d'information sur les industries extractives, la transparence de la propriété des terres et l'adoption d'une charte sur l'*open data*.

[43] Open Definition, « Définition du Savoir Libre v.1.0. » http://www.opendefinition.org/okd/francais/, consulté le 20 avril 2015.

Figure 11. Une session de travail des chefs d'État lors du G8 de 2013[44].

Contrairement aux cas précédents, je n'ai pas eu accès à des informations sur les coulisses de ce sommet et de la rédaction de la charte. Toutefois, j'ai été impliqué au sein de l'OKFN dans la préparation de l'Open Data Index en vue du G8. L'OKFN était un des experts techniques sollicités dans la préparation du G8. La charte sur l'*open data* du G8 se compose d'une série de cinq principes et trois annexes. La charte part du constat que l'*open data* (nommé comme tel[45]) est au cœur d'« un mouvement mondial » facilité par la technologie, les médias sociaux et l'information qui pourra créer de la croissance économique et rendre les gouvernements plus redevables (*accountable*) et efficaces. Le préambule détaille les bénéfices de l'*open data* : création de services, transparence de l'action publique, meilleure gouvernance, amélioration du débat public, lutte contre la corruption, soutien à l'innovation des entreprises et de la société civile, prospérité renouvelée… Pour éviter que l'*open data* ne soit une « opportunité manquée », les chefs d'État du G8 ont décidé de l'adoption de cinq principes pour régir l'accès aux données.

Les trois premiers principes établissent les conditions d'ouverture des données puis les deux derniers fixent deux objectifs : l'amélioration de la gouvernance et le soutien à l'innovation. Le premier point de la charte annonce que les pays signataires s'engagent à faire de l'*open data* la pratique par défaut des administrations pour les données publiques tout en

[44] G8UK sur FlickR, https://www.flickr.com/photos/g8uk/, consulté le 1 aout 2016.

[45] L'Elysée a traduit open data en accessibilité des données dans la version complète de la déclaration du G8 de Lough Erne sur le site de l'Elysée : http://www.elysee.fr/communiques-de-presse/article/communique-final-du-g/. Mais Etalab a publié une version non officielle avec le gouvernement canadien qui traduit open data en Ouverture des Données Publiques :

respectant les législations en vigueur sur la propriété intellectuelle et la vie privée. Cette annonce n'engage toutefois pas les gouvernements qui doivent chacun préciser, d'une part, leur stratégie d'ouverture des données publiques et, d'autre part, publier un plan d'action pour mettre en œuvre la charte du G8. L'annexe technique précise que les gouvernements sont encouragés à publier les données sur un portail national unique où elles ne pourront être « retirées sans préavis. »[46] Dans son deuxième principe, la charte promet la publication de données de qualité. Partant du constat que la préparation de données exige du temps pour les administrations, elle propose que les gouvernements se concertent avec des représentants d'utilisateurs pour définir les données à améliorer en priorité. Les gouvernements s'engagent à publier les données dès que possible, « sous leur forme originale et non modifiée, et au plus fin niveau de granularité disponible ». Cette dernière demande se rapproche des exigences de données « primaires » des principes de l'*Open Government Data* ou de données brutes selon Tim Berners-Lee. Selon le troisième principe, les données doivent être publiées dans des portails uniques par pays qui n'exigent pas l'enregistrement des utilisateurs. Elles doivent aussi être gratuites et « dans des formats ouverts chaque fois que possible. » Dans le quatrième principe, les États du G8 s'engagent à partager leur expertise technique avec les pays du monde entier, notamment au sein d'initiatives multilatérales telles que l'Open Government Partnership. Ils déclarent vouloir identifier les jeux de données à ouvrir en priorité avec les organisations de la société civile. Dans le cinquième principe, les gouvernements s'engagent à développer la culture de l'ouverture des données (« *increase open data litteracy* ») et à encourager les organisations de la société civile qui promeuvent l'*open data*. Dans l'annexe technique, la charte soutient la publication de données avec une licence libre, mais n'en fait pas une exigence. Pourtant, tous les acteurs évoqués dans les épisodes précédents la placent comme une condition essentielle de l'ouverture des données.

À travers ce résumé, on voit donc que la charte du G8 s'est inscrite dans la continuité des définitions de l'*open data* évoquées précédemment. En particulier, elle a repris la majeure partie des principes de l'*Open Government Data* établis à Sebastopol. Comme le montre le tableau suivant (Figure 12), plusieurs critères ont été repris quasiment à l'identique dans le texte de la charte du G8.

[46] La formulation reprend le principe de permanence établi par la Sunlight Foundation dans ces dix principes dérivés de l'Open Government Data.

Principe de Sebastopol	Définition du principe de Sebastopol	Principe de la charte du G8	Extrait de la charte semblable au principe de Sebastopol
1. Complete	*All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.*	1. Open data by default	Annexe technique: *We will establish an expectation that all government data be published openly by default.*
2. Primary	*Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms*.	2. Quality and quantity	*To the extent possible, data will be in their original, unmodified form and at the finest level of granularity available;*
3. Timely	*Data is made available as quickly as necessary to preserve the value of the data.*	2. Quality and Quantity	*We will: release high-quality open data that are timely, comprehensive, and accurate, […] release data as early as possible, allow users to provide feedback, and then continue to make revisions to ensure the highest standards of open data quality are met*[47]*.*
4. Accessible	*Data is available to the widest range of users for the widest range of purposes.*	3. Usable by all	*the data are available to the widest range of users for the widest range of purposes*[48]
5. Machine-processable	*Data is reasonably structured to allow automated processing.*	5. Releasing Data for Innovation	*We will […] empower a future generation of data innovators by providing data in machine-readable formats*. Annexe technique : *We will […] ensure data are machine readable in bulk*[49] *by providing data that are well structured to allow automated processing*
6. Non-discriminatory	*Data is available to anyone, with no requirement of registration.*	3. Usable by All	*We agree that when open data are released, it should be done without bureaucratic or administrative barriers, such as registration requirements, which can deter people from accessing the data.*

[47] « Public dissemination will allow users to verify that information was collected properly and recorded accurately. »

[48] Notons ici que, mise à part le pluriel à data, la formulation est identique de celles des principes de Sebastopol.

[49] On retrouve ici un des critères de l'Open Data Census.

7. Non-proprietary	*Data is available in a format over which no entity has exclusive control.*	3. Usable by all	*We will release data in open formats wherever possible, ensuring that the data are available to the widest range of users for the widest range of purposes* Annexe technique : *We will make data available in convenient open formats to ensure files can be easily retrieved, downloaded, indexed, and searched by all commonly used Web search applications. Open formats, for example non-proprietary CSV files, are ones where the specification for the format is available to anyone for free, thereby allowing the data contained in a file to be opened by different software programmes.*
8. Licence-free	*Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.*	5. Releasing data for innovation	Annexe technique : *We will support the release of data using open licences or other relevant instruments—while respecting intellectual property rights—so that no restrictions or charges are placed on the re-use of the information for non-commercial or commercial purposes, save for exceptional circumstances*

Figure 12. Tableau comparatif des principes de Sebastopol et de la charte du G8 de 2013 sur l'*open data*. Les formulations similaires dans les deux textes sont soulignées.

Néanmoins, les promesses de la charte diffèrent sur deux aspects essentiels. Premièrement, alors que les principes de l'*Open Government Data* demandent l'ouverture complète des données publiques, la charte la conditionne à l'existence d'un public qui réclame les données : « nous souhaitons que le public attende désormais de l'État qu'il publie librement l'ensemble de ses données par défaut. » Deuxièmement, concernant les conditions juridiques de réutilisation, les chefs d'État ont soutenu la publication de données avec une licence ouverte sans que cela soit formulé comme un engagement.

D'autre part, la charte s'est inspirée de l'Open Data Index en sélectionnant des données à forte valeur ajoutée devant être ouvertes en priorité (high value datasets). L'ensemble des jeux de données essentiels sélectionnés par l'OKFN figure dans les exemples cités dans l'annexe de la charte du G8 (figure 13).

Catégories de données	Exemples de jeux de données
Criminalité et justice	Statistiques sur la criminalité, sécurité
Développement mondial	Aide au développement, sécurité alimentaire, industries extractives, terres
Données géospatiales	Topographie, codes postaux, cartes nationales ou locales
Éducation	Liste des écoles, valeur ajoutée, compétences numériques
Entreprises	Registre des entreprises
Environnement	Niveaux de pollution, consommation énergétique
Finances et marchés	Valeur des transactions, marchés publics attribués ou à venir, budget local ou national (prévu et exécuté)
Mobilité et protection sociales	Logement, prestations sociales, assurance-maladie et assurance-chômage
Observation de la Terre	Conditions météorologiques, agriculture, foresterie, pêche et chasse
Responsabilisation des gouvernements et démocratie	Guichets et points de contact des administrations, résultats des élections, lois et règlements, salaires (échelles salariales), dons
Santé	Données issues de prescriptions, données de performance
Science et recherche	Données relatives au génome humain, recherche et activités pédagogiques, résultats d'expérience
Statistiques	Statistiques nationales, recensements, infrastructure, statistiques économiques et éducatives
Transport et infrastructure	Horaires des transports publics, services à large bande

Figure 13. Liste des données à forte valeur ajoutée sélectionnées dans la charte du G8 (traduction : Etalab). Les données correspondant à la sélection de l'Open Data Index sont soulignées.

Dans son communiqué de presse publié après le sommet de Lough-Erne, l'OKFN s'est félicitée que les chefs d'État du G8 reconnaissent une liste de jeux de données à forte valeur.

L'ONG a trouvé « décevant » que les États ne s'engagent pas à les publier : « *it is therefore good to see that the Charter recognizes a list of "high value datasets" which should be prioritized for release, though it is disappointing that there are no explicit commitments to release the types of data mentioned*. » Dans la déclaration finale du sommet, ce sont ces jeux de données essentielles qui résument la charte plutôt que l'engagement d'ouverture par défaut des données publiques.

> *10. Governments should publish information on laws, budgets, spending, national statistics, elections and government contracts in a way that is easy to read and re-use, so that citizens can hold them to account*[50].

Les chefs d'État du G8 ont considéré que l'ouverture des données devra se faire de manière progressive en se concentrant d'abord sur les données « à forte valeur ajoutée », celles pour lesquelles une multitude d'usages potentiels est projetée.

Au final, sept des huit états ont publié leur plan d'action, y compris la Russie en préparation du sommet qui devait se tenir à Sotchi en 2014. Mais ce dernier a été annulé après l'annexion de la Crimée et l'exclusion de la Russie du G8. Pour inciter les gouvernements à honorer leurs promesses, des groupes d'intérêt ont suivi sa mise en pratique par la méthode du *benchmarking*. La Sunlight Foundation a publié une feuille de calcul dans laquelle elle a évalué le contenu des plans d'action et a attribué un score maximum aux gouvernements qui s'engagent dans des actions concrètes[51]. Un autre groupe d'intérêt basé à Washington, le Center for Data Innovation, a évalué les plans d'action en attribuant des scores aux pays selon leur conformité avec les promesses du G8[52]. Le suivi de la charte sur l'*open data* ne figure plus au programme des sommets. Le G7 de juin 2014 était dédié à la crise en Crimée et celui de 2015 en Allemagne n'a pas évoqué la charte, le pays hôte n'ayant pas publié son plan d'action. Bien qu'elle ne soit plus au programme des débats du G7, la charte a été reprise lors d'une rencontre en marge de la conférence internationale sur l'*open data* qui

[50] Gov.uk, « G8 Lough Erne Declaration HTML version », https://www.gov.uk/government/publications/g8-lough-erne-declaration/g8-lough-erne-declaration-html-version, consulté le 3 aout 2016.
[51] European Public Sector Information Platform, « Comparing the G7 countries' Open Data Action Plans », http://www.epsiplatform.eu/content/comparing-g7-countries-open-data-action-plans, consulté le 3 aout 2016.
[52] NextGov, « UK, US Most Committed in G8 to Unleashing Data. » http://www.nextgov.com/technology-news/2015/03/uk-leads-g8-commitment-open-data-charter-russia-dead-last/107742/, consulté le 4 aout 2016.

s'est tenue à Ottawa en mai 2015. La réunion visait à créer une charte de l'*open data* au-delà des seuls pays du G8. Elle regroupait des représentants gouvernementaux, des organisations de la société civile, de plusieurs institutions internationales et des chercheurs (figure 14).

Figure 14. Réunion à Ottowa du groupe de travail en charge de la définition d'une charte internationale de l'*open data*.

La nouvelle charte est encore en débat au moment de l'écriture de ces lignes, une consultation est prévue en vue de sa finalisation[53]. Elle reprend de la charte du G8 le préambule, l'intitulé des trois premiers principes et une grande partie du contenu. Parmi les multiples modifications de la charte de 2013, elle demande aux gouvernements de s'engager à publier des inventaires des données produites pour faciliter la sélection de celles à ouvrir en priorité (une revendication déjà formulée par la Sunlight Foundation en réponse à la charte du G8[54]). Encore une fois, ce groupe tente d'établir des principes de l'*open data* reconnus et appliqués par les gouvernements du monde entier.

Conclusion

Nous arrivons au terme de cette petite généalogie de l'*open data*. Lors de ces six épisodes, nous avons fait la navette entre les États-Unis et la Grande-Bretagne à la rencontre d'acteurs, majoritairement masculins, aux profils très divers : des informaticiens, des juristes, des experts de l'innovation, des diplomates ou encore des gouvernants. Tout d'abord, les trois premiers épisodes ont montré comment de grands principes de l'ouverture des données ont

[53] Une version préliminaire de la charte est en ligne sur le site opendatacharter.net

[54] Sunlight Foundation, « G8 Open Data Charter Action Plan: Open data by default, but you may have to pay. », http://sunlightfoundation.com/blog/2014/07/28/g8-open-data-charter-action-plan-open-data-by-default-but-you-may-have-to-pay-for-it/, consulté le 3 aout 2016.

été définis. L'*Open Definition* a décliné des droits de l'usager du savoir ouvert, dans la lignée des revendications des mouvements de l'*open source* et de l'*open access*. De leur côté, les participants de la réunion de Sebastopol se sont intéressés spécifiquement aux données de l'État à travers des principes adressés aux candidats à l'élection présidentielle états-unienne. Dans leur conception, l'*Open Definition* et les principes de l'*Open Government Data* ont suivi un modèle bien particulier, celui des *Requests For Comments* (RFC), une procédure de discussion et de délibération qui remonte à l'élaboration des standards ouverts de l'Internet[55]. À l'opposé de ce modèle fondé sur le consensus d'une communauté, Tim Berners-Lee a formulé son appel à l'ouverture des données brutes sous la forme d'un manifeste fondé sur sa renommée en tant qu'inventeur du web. Après cette phase de définition, les premières initiatives gouvernementales d'*open data* aux États-Unis et au Royaume-Uni ont traduit certains de ces principes en politiques publiques. Les quatrièmes et cinquièmes épisodes ont montré comment des outils d'évaluation ont été élaborés pour agir sur la mise en œuvre des politiques d'*open data*. Se concentrant sur les standards de données, le modèle en cinq étoiles de Tim Berners-Lee a proposé une « marche à suivre » et a décliné l'ouverture sous la forme d'un gradient qui contraste avec l'*Open Definition* et les principes de l'*Open Government Data* dans lesquels c'est une variable binaire. De son côté, l'Open Data Index décline neuf critères de l'ouverture d'une donnée, reprenant en partie certains des principes élaborés à Sebastopol. Pour influencer les politiques d'*open data*, le modèle en cinq étoiles de Tim Berners-Lee et l'Open Data Index investissent dans la force performative des outils de notation et de classement dont Espeland et Stevens (2007) ont montré, au sujet des classements des écoles de droit aux États-Unis, comment la mise en concurrence par des indicateurs publics et commensurables provoquait généralement la réaction des acteurs mesurés. Enfin, la charte du G8, objet du sixième épisode, a tenté de synthétiser ces principes dans une déclaration diplomatique signée par les chefs des huit États membres dans le but affiché d'institutionnaliser l'ouverture des données et de devenir la pratique par défaut des administrations des huit pays signataires. Sa mise en œuvre repose

[55] En 1969, un étudiant de UCLA, Steve Crocker, publiait la première RFC au sein du Network Working Group, un groupe relativement informel d'étudiants et d'enseignants où étaient définis les protocoles d'Internet. Par la suite, les RFC sont devenus le standard pour l'élaboration des protocoles d'Internet. Ce mode de discussion, particulièrement répandu dans les communautés du logiciel libre, implique que chacun puisse critiquer les propositions ou en soumettre de nouvelles. Il vise à l'élaboration de solutions techniquement efficaces dont l'adoption s'établit par un consensus, assumé comme approximatif, entre les membres du collectif qui délibère (Flichy, 2001 ; Kelty, 2008 ; Loveluck, 2012 ; Russel, 2014).

sur la méthode du *benchmarking* (Bruno & Didier, 2013) : les pays signataires doivent publier des plans d'action dont la mise en œuvre est évaluée publiquement[56].

Tout au long de ce chapitre, j'ai souligné les contradictions et les divergences entre les acteurs qui montrent que l'histoire de l'ouverture des données n'est pas linéaire et qu'il n'existe pas de consensus complet sur la définition de l'*open data*. En particulier, on peut souligner une tension importante entre deux modèles, l'un qui prône une ouverture exhaustive, l'autre qui insiste sur la mise à disposition prioritaire de données jugées essentielles. D'un côté, les principes de Sebastopol proposent un modèle fondé sur la *completeness*, l'ouverture exhaustive de toutes les données publiques sous leur forme primaire qu'on retrouve aussi dans l'appel de Tim Berners-Lee à la publication de l'ensemble des données brutes. De l'autre côté, l'Open Data Index introduit une sélection de données essentielles qui doivent être ouvertes en priorité en fonction de la demande et du potentiel de la réutilisation. Le premier modèle propose des critères de l'ouverture qui visent à réduire les frictions de la réutilisation des données par les machines. Ces critères sont à la fois juridiques en réclamant l'utilisation de licences standardisées et techniques en demandant l'ouverture des données dans des standards aux spécifications ouvertes et lisibles par les machines. Ces revendications s'inscrivent dans la lignée des mouvements du logiciel libre qui réclament ce que Kelty (2008) qualifie de « *modifiability* », la capacité de transformer un objet, de le réutiliser dans un contexte différent ou encore de l'améliorer[57]. On retrouve essentiellement ces critères dans l'Open Data Index, mais ce deuxième modèle apporte une nouvelle dimension puisqu'il porte sur le contenu même des données. La charte du G8 se situe en quelque sorte sur une ligne de partage en demandant à la fois

[56] D'autres initiatives intergouvernementales telles que l'ePSI Scorecard de la Commission européenne, l'Open Data Readiness Assessment de la Banque Mondiale ou les plans d'action de l'Open Government Partnership ont aussi recours au benchmarking pour institutionnaliser l'ouverture des données.

[57] Dans *Two Bits*, Kelty considère que la demande de modifiabilité se retrouve aussi au sein de Creative Commons, un projet dont on a vu les liens avec l'ouverture des données dans le premier épisode : « *Modifiability includes the ability not only to access—that is, to reuse in the trivial sense of using something without restrictions—but to transform it for use in new contexts, to different ends, or in order to participate directly in its improvement and to redistribute or re- circulate those improvements within the same infrastructures while securing the same rights for everyone else. In fact, the core practice of Free Software is the practice of reuse and modification of software source code. Reuse and modification are also the key ideas that projects modeled on Free Software (such as Connexions and Creative Commons) see as their goa*l. » (p.11)

l'ouverture de l'ensemble des données par défaut dès leur production et en désignant des « *high-quality data* », des données essentielles à ouvrir en priorité.

Mais fondamentalement, il apparait en filigrane que tous ces acteurs portent une même revendication, font part d'une même demande, celle d'obtenir des données. Si cette demande semble aller de soi tout au long des six épisodes, du point de vue des administrations, elle reconfigure en profondeur les politiques de diffusion de l'information publique. En effet, les données brutes de l'administration ne correspondent pas aux objets informationnels dont la circulation est prévue par la loi. Elles ne sont pas des statistiques, des informations quantitatives produites par un organisme spécialisé et visant à fournir une information « agrégée de portée générale » (Desrosières, 2005). Comme elles sont gérées et diffusées directement par les agents qui les produisent, elles ne sont pas non plus des archives, des documents transmis à une institution dédiée à sa préservation. On pourrait toutefois les faire entrer dans une catégorie plus générale, celle de document administratif, sur laquelle se fonde la transparence étatique en France. Sur ce point, Kafka (2012) a montré l'importance de l'article 15 de la Déclaration des droits de l'homme et du citoyen de 1789 qui stipule que « la Société a le droit de demander compte à tout Agent public de son administration. » Porté par l'abbé Sieyès, cet article a créé un « nouvel éthos documentaire » dans lequel toute action réalisée au nom de l'État doit être documentée sous forme écrite et archivée en anticipation d'un contrôle. Tout en soulignant la fragilité des écrits et les débordements qu'ils entrainent, Kafka a montré comment la « paperasse » est devenue une technologie de la représentation politique. Vismann (2008) et Weller (2012) ont montré comment la transparence de l'État s'est aussi matérialisée de manière très concrète dans l'organisation spatiale du traitement de l'information à travers le rangement des dossiers, la luminosité des bureaux ou encore l'architecture des bâtiments. En France, jusqu'à la seconde moitié du 20e siècle, l'obligation de redevabilité des agents publics reposait uniquement sur des institutions dédiées, telle que la Cour des Comptes, qui accédaient aux documents administratifs pour exercer le contrôle de l'action publique au nom des citoyens. Depuis l'apparition d'un « droit de savoir » devenu progressivement la norme dans les démocraties, les documents administratifs peuvent être réclamés directement par les citoyens[58]. Adoptée le 17 juillet 1978, la loi CADA a accordé un droit d'accès puis de

[58] Pour la Sunlight Foundation dans son document Open Data Policy Guidelines, l'enjeu de l'ouverture des données consiste à passer d'un système réactif de révélation dans lequel les agents publics répondent à

réutilisation des documents administratifs. La France était alors le troisième pays à se doter d'une telle législation, après la Suède en 1766 et les États-Unis en 1966 avec le *Freedom of Information Act* (Boustany, 2013). La loi CADA, du nom de la Commission d'Accès aux Documents Administratifs en charge d'arbitrer les demandes, donne une définition particulièrement large des documents administratifs : « tous les documents produits ou reçus par l'administration qu'ils se présentent sous forme écrite (dossiers, rapports, études, comptes rendus, procès-verbaux, statistiques, directives, instructions, circulaires...), sous forme d'enregistrement sonore ou visuel ou sous forme numérique ou informatique. Sont également concernées les informations contenues dans des fichiers informatiques et qui peuvent en être extraites par un traitement automatisé d'usage courant[59]. » La loi CADA s'applique aux administrations d'État, aux collectivités territoriales, aux établissements publics, mais aussi aux organismes privés chargés d'une mission de service public. Elle exclut du droit d'accès les informations nominatives ou personnelles et les documents préparatoires.

Bien qu'elle délimite un premier périmètre parmi l'ensemble des informations produites par les administrations, la notion de document administratif ne différencie pas les données, sous leur forme brute en particulier, des autres objets informationnels. Des juristes ont ainsi montré que la notion de données n'a pas de fondement juridique qui la distinguerait des documents administratifs définis par la loi française ou des informations du secteur public dans la législation européenne (Boustany, 2013 ; Trojette, 2013)[60]. Pour tenter de

une demande à un système proactif dans lequel les administrations publient directement leurs données : « *Most public records systems, including the Freedom of Information Act itself, are systems of reactive disclosure -- meaning that a question has to be asked before an answer given; public information requested, before it is disclosed. Proactive disclosure is the opposite. Proactive disclosure is the release of public information -- online and in open formats -- before it is asked for. This is no simple task, but, in a way, it's what all "open data" is aiming to accomplish*. »

[59] CADA, « La notion de document administratif », http://www.cada.fr/la-notion-de-document-administratif,56.html, consulté le 30 avril 2015.

[60] Dans son rapport au Premier ministre, Mohammed Adnène Trojette, magistrat à la Cour des comptes, évoque la constitution d'un groupe de travail de l'Observatoire juridique des technologies de l'information qui s'est interrogé, à la demande du Secrétaire général du Gouvernement (SGG), sur la signification juridique de la notion de données publiques. Ce groupe de travail a souligné l'ambiguïté juridique de la notion de données et en a conclu qu'elle était synonyme d'information.
Dans un article dressant un état des lieux de l'accès et de la réutilisation des données publiques depuis 1970, Boustany a souligné que plusieurs expressions étaient utilisées dans les textes de loi et les publications officielles pour désigner indifféremment les documents, les informations et les données produites par les autorités publiques.

comprendre ce qui distingue les données, retournons alors dans le troisième épisode quand Tim Berners-Lee tente une définition des données lors de sa conférence TED de 2009.

> *There is still a huge frustration that people have because we haven't got data on the web as data. What do you mean, "data"? What's the difference—documents, data? Well, documents you read, OK? More or less, you read them, you can follow links from them, and that's it. Data—you can do all kinds of stuff with a computer.*

La définition des données que donne Tim Berners-Lee reste assez obscure, mais l'inventeur du web marque une séparation intéressante entre les documents, intelligibles par les humains, et les données, intelligibles par les machines. Son expression en rend bien compte lorsqu'il estime que les données sur le web ne sont pas encore des données, car il considère qu'elles ne sont pleinement utilisables par des machines. Gardons en mémoire cet enjeu d'intelligibilité des données par les humains et par les machines que nous aurons l'occasion d'aborder plus en détail. La question de la définition des données reste donc ouverte, elle sera un des fils rouges de ce mémoire. Mais, à la lecture de ces épisodes, nous sommes en mesure de mieux caractériser les données brutes. Dans le deuxième épisode, le deuxième principe de l'*Open Government Data* n'évoque pas des données brutes, mais des données primaires : « *data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms*. » De son côté, Tim Berners-Lee donne une définition très élusive des données brutes, évoquant des données à l'état pur (*unadultarated*) : « *make a beautiful website, but first give us the unadulterated data, we want the data. We want unadulterated data. OK, we have to ask for raw data now*." Si l'on s'en tient à ces deux extraits, les données brutes constitueraient en quelque sorte une matière première « pure » n'ayant pas subi de modifications ou d'agrégation. Dans la lignée de la métaphore des données comme le nouveau pétrole, comme une *commodity* qui s'échange de manière fluide (Ribes & Jackson, 2013), les données brutes seraient en quelque sorte le matériau brut, non traité, dont émanent des objets informationnels transformés, aboutis et intelligibles. Dans l'horizon d'un renouveau de la transparence par les données (Birchall, 2014), l'ouverture des données sous leur forme brute permettrait de réduire les asymétries d'information entre les gouvernants et les administrés, chacun disposant de données avec le même niveau de granularité et de détail. Leur exploitation par des activistes (Bruno, Didier & Prévieux, 2014), des journalistes et tout acteur disposant des compétences suffisantes formule la promesse

d'une décentralisation des centres de calcul, d'un déplacement des lieux vers lesquels convergent les réseaux de données qui sont mobilisés dans le débat public.

> La mobilisation des bases de données par les journalistes et les activistes peut être identifiée à un processus de décentralisation des centres de calcul équipant le débat public. Jusqu'à une date récente, seules les grandes institutions publiques et privées étaient en mesure de constituer de grandes bases de données et d'élaborer des indicateurs statistiques permettant d'alimenter le débat public (Desrosières, 1993, pp. 397-400). Or des organisations de presse et des associations militantes considèrent aujourd'hui les bases de données comme des formes permettant de gagner en marge de manœuvres à l'égard des institutions. À travers la structuration et le traitement des données, un ensemble de réalités peuvent ainsi être offertes aux citoyens et en quelque sorte proposées à leur indignation collective. Si, comme nous l'avons vu, des limites importantes pèsent sur la capacité des journalistes et des activistes à mobiliser un public tout en mettant à distance les institutions productrices de données, on peut néanmoins conclure à un déplacement. (Parasie, 2013)

J'aurai, à plusieurs reprises dans ce mémoire, l'occasion de rediscuter les implications de la demande de données brutes pour la transparence publique et le travail des agents administratifs. Mais avant cela, dans le chapitre suivant, je vais resserrer la focale et me concentrer sur la France pour retracer brièvement la trajectoire d'une organisation en particulier, Etalab. À travers ses nombreuses mutations en cinq ans, nous verrons comment les grands principes de l'ouverture des données ont donné lieu à de multiples traductions et adaptations dans les politiques publiques. Par ailleurs, qu'est-ce qui change lorsque les politiques informationnelles sont catégorisées par la donnée ?

Chapitre 2

Vers une administration des données : la trajectoire d'Etalab

Comme en ouverture du chapitre précédent, nous retournons à la réunion Open Data Bootcamp de la région Ile-de-France de novembre 2013. Xavier Crouan, directeur de la communication de la région, ouvre l'évènement. Il prend le micro et affirme que « nous n'avons pas à rougir » de la démarche *open data* de la région, « une des collectivités les plus dynamiques » en la matière et souligne qu'il faut continuer l'ouverture de nouvelles. Il annonce qu'il y aura prochainement une obligation à ouvrir les données dans la loi de décentralisation et que la région va prochainement adhérer à l'association Open Data France « pour travailler de concert avec les autres collectivités. » Enfin, il déclare que l'*open data* est un chantier prioritaire du plan de communication de la région qu'il va présenter à la suite au cabinet du président de région : « l'ouverture de la donnée vient désormais au cœur de notre stratégie d'information et désormais pour chaque information qu'on met sur notre site, on cherche à avoir de la donnée. Et la donnée constitue aussi aussi un système d'information à part entière. On a vraiment une politique assez forte, assez volontariste et ça fait partie je le dis, des chantiers structurants de la direction de la communication et plus largement de la région. »

Faisons un pas de côté, je reviendrai dans les prochains chapitres sur cet évènement. Le discours du directeur de la communication montre l'importance stratégique qui est désormais attachée à la politique d'ouverture des données de la région. Mais comment une telle politique publique est-elle entrée dans les priorités des collectivités locales et des institutions ? Comment les grands principes de l'*open data* ont-ils été importés en France ? Quels en ont été les passeurs ? Pour répondre à ces questions, je vais revenir sur la trajectoire de la mission Etalab en charge de la mise en œuvre de l'*open data* pour le gouvernement français. J'ai choisi dans ce chapitre une approche monographique pour décrire en détail l'évolution de cette organisation qui a joué un rôle déterminant dans l'institutionnalisation des politiques d'*open data* au niveau local et national en France. En effet, Etalab a créé un réseau au sein de l'administration et tissé des liens avec la plupart des équipes de projets

open data que j'ai pu rencontrer. La trame de cette histoire est constituée principalement par l'analyse d'un corpus d'entretiens et de documents. J'ai divisé les histoires en phases qui correspondent aux moments d'évolution majeure du projet, c'est-à-dire quand l'ouverture des données est remise en cause ou connait une nouvelle impulsion. La sélection des différentes étapes de la narration de ces histoires s'inspire de la méthode proposée par Grosjean et Lacoste (1999) dans leur ethnographie du travail à l'hôpital : « l'histoire est une structure de traitement des données qui rend compte du caractère arborescent des interactions dans les services hospitaliers : une même histoire est comme le furet dans la chanson : elle court, passe de l'un à l'autre ; on la croit réglée en un endroit, elle ressurgit d'ailleurs, donnant un autre sens aux paroles échangées précédemment. »

Cette histoire d'Etalab débute par la création de l'APIE, une entité dédiée à la valorisation financière des données du « patrimoine informationnel » de l'État. L'APIE a été progressivement dessaisie des données à partir du décret qui a créé Etalab et marqué l'arrêt du développement des redevances pour les données publiques. Ce revirement s'explique par le lancement de data.gov aux États-Unis suite à l'élection du président Obama et par la volonté du gouvernement d'afficher sa transparence par le lancement d'un portail *open data* quelques mois avant la campagne présidentielle. Le projet s'est matérialisé par la création de la mission Etalab, en charge de mettre en œuvre l'ouverture des données par la création d'un portail, la rédaction d'une licence ouverte, la mobilisation de réseaux favorables à l'*open data* et le recensement des données publiques. Suite à l'alternance de 2012, l'existence de cette mission et le maintien de la politique d'*open data* sont, pendant un moment, remis en question, mais l'arrivée d'une nouvelle direction et l'adoption d'une feuille de route sur l'*open data* par le gouvernement ont attribué de nouveau un caractère prioritaire à l'ouverture des données. Au moment de l'écriture de ces lignes, Etalab est encore une structure administrative récente en mouvement permanent. J'observe très régulièrement des changements dans son action, ses priorités, ses équipes ou encore ses attaches politiques et administratives. Aujourd'hui, Etalab, service d'une vingtaine d'agents désormais rattaché au Secretariat Général pour la Modernisation de l'Action Publique (SGMAP), est reconnu comme l'organisme gouvernemental en charge de l'ouverture des données publiques. Avec la création du poste d'administrateur général des données nommé par le Premier ministre, la mission a progressivement intégré une nouvelle compétence, l'expérimentation en

matière de science des données, et, au-delà de l'*open data*, est chargée d'assurer la circulation des données dans l'État.

Le renvoi de l'APIE : un virage de la politique gouvernementale en faveur de la gratuité

En octobre 2008, le gouvernement présentait un plan de développement de l'économie numérique pour le quinquennat, intitulé « France Numérique 2012 », comprenant un volet sur la transformation du service public. Il proposait, dans son action 39, de « favoriser le développement de nouveaux produits et services par la création d'un portail unique d'accès aux données publiques dont la conception sera pilotée par l'APIE[61]. » L'Agence pour le Patrimoine Immatériel de l'État (APIE) avait été créée à la suite du rapport Lévy-Jouyet de décembre 2006 qui considérait les données publiques comme un actif à valoriser dans le « patrimoine immatériel de l'État notamment par la création de nouvelles redevances. En 2008, l'APIE était chargée de la création d'un portail unique pour les données publiques. Au-delà de la valorisation économique, cette initiative s'inscrivait dans le cadre de la transposition d'une directive européenne de 2003 dite PSI (Public Sector Information) qui demande aux États membres de rendre leurs documents administratifs réutilisables, si possible sous forme électronique. Dans son article 9, la directive demande aux administrations d'établir des listes de documents administratifs réutilisables agrégées par les gouvernements nationaux dans des portails qui facilitent la recherche des informations publiques. C'est dans ce cadre que l'APIE avait commencé l'élaboration d'un portail pour les données publiques françaises.

Dans un rapport de 2010, le nom « État lab » a été proposé pour dénommer le portail que devait concevoir l'APIE. Le gouvernement avait demandé à huit experts de faire des propositions sur l'« amélioration de la relation numérique à l'usager[62] ». La proposition 22 (figure 15) du rapport publié le 12 février 2010 et dirigé par le député Franck Riester appelait à « créer une plateforme d'innovation de services "État lab" pour permettre aux acteurs tiers de développer des services innovants à partir de données publiques ».

[61] Secrétariat d'État chargé de la prospective, de l'évaluation des politiques publiques et du développement de l'économie numérique, « France Numérique 2012 - Plan de développement de l'économie numérique », http://francenumerique2012.fr/pdf/081020FRANCENUMERIQUE_2012.pdf, consulté le 13 février 2015.

[62] EPSIPlatform, « Digital Experts call for French State Lab to Develop Services from Government Data », http://www.epsiplatform.eu/content/digital-experts-call-french-state-lab-develop-services-government-data, consulté le 27 novembre 2014.

➲ Innover

PROPOSITION N°22 : Créer une plateforme d'innovation de services « Etat Lab » permettant aux acteurs tiers de développer des services innovants à partir de données publiques

Enjeux

Contexte :

- Le succès de l'Iphone repose sur la boutique d'applications téléchargeables en ligne, l'Appstore, riche de plus de 100 000 applications. Plus de 125 000 développeurs ont travaillé sur ces applications et près de 100 nouvelles applications sortent chaque jour ;
- En France, une application Iphone permet de connaître le lieu le plus proche pour trouver un Vélib, les autoroutes communiquent, par application, le prix de l'essence dans les stations services.

Problématique :

- Faciliter la co-création de contenus par les usagers permet de proposer des services au plus près des besoins et d'en démultiplier rapidement le nombre en recourant à des développeurs externes à l'Etat ;
- Le développement d'applications pour les services administratifs nécessite éventuellement la mise à disposition de données publiques à des développeurs privés.

Solution proposée

- **Créer une plateforme d'innovation de services en prenant appui sur le futur portail de l'APIE (Agence du Patrimoine Immatériel de l'Etat) de mise à disposition des données publiques ;**
- **Proposer aux tiers de tester des services développés à partir des données publiques ;**
- **Lancer des API et banques de données (databank).**

Meilleures pratiques des administrations

- En France, l'APIE conduit une politique d'accès aux données publiques de l'Etat
- Aux USA lancement d'un AppStore gouvernemental ; certaines villes américaines ont lancées leurs propres « databanks ».

Retour d'expériences des experts

- Permettre à certaines applications d'offrir des interfaces ouvertes pour permettre la création d'un marché d'applications construites par les citoyens, qui peuvent ensuite être mis à la disposition de tous.

Objectif cible

- **D'ici fin 2010 :** initialisation de la plateforme d'un Etat Lab

Figure 15. Proposition 22 du rapport Riester.

Le rapport soulignait le succès de l'App Store d'Apple, deux ans après son lancement, avec 100 000 applications disponibles pour iPhone. Il citait des applications pratiques pour trouver un Vélib ou connaitre les prix de l'essence dans les stations-service, postulant que de nombreux services pourraient apparaitre si les données publiques étaient librement utilisables. Trois solutions étaient évoquées pour permettre aux usagers de « proposer des services au plus près des besoins » : « créer une plateforme de mise à disposition des données publiques intitulée « État lab », « proposer aux tiers de tester des services développés à partir des données publiques » et de « lancer des API et banques de données ». Le rapport recommandait de prendre appui sur le travail de l'APIE qui défendait une valorisation financière des données par des redevances[63]. On remarque dans ce rapport que « l'*open*

[63] L'APIE fait suite au rapport Lévy-Jouyet de décembre 2006□1 consacré à « l'économie de l'immatériel » qui encourage l'État à trouver de nouvelles sources de revenus dans la vente et l'exploitation de ses actifs

data » n'était pas évoqué et que la gratuité des données ne fait pas partie des propositions. Malgré ce rapport qui plaçait l'agence parmi les bonnes pratiques des administrations, l'APIE est progressivement dessaisie du dossier au profit d'une nouvelle structure dédiée à la diffusion gratuite des données publiques. Comment expliquer ce revirement ? Plusieurs causes peuvent être identifiées : la volonté du gouvernement d'imiter les États-Unis avec data.gov, le retard de l'APIE dans la conception du portail, l'émergence de licences « libres » pour les données et le contexte politique de l'élection présidentielle.

Tout d'abord, le lancement de data.gov en 2009 et la politique d'*open data* de l'administration Obama ont provoqué une redéfinition du projet et ont joué un rôle déclencheur dans le lancement de la mission Etalab. Les conseillers numériques de l'exécutif, Nicolas Princen pour le président et Séverin Naudet pour le Premier ministre, ainsi que la secrétaire d'État en charge des questions numériques ont demandé le renvoi de l'APIE et la redéfinition du projet à la suite d'un voyage à Washington.

> NKM[64] avait fait un gros lobbying puisqu'elle était allée rencontrer Vivek Kundera, le CTO des États-Unis[65]. Il l'avait super impressionné et, du coup, elle appuyait fortement le projet relatif à data gouv. Après tu regardes le benchmark : les États-Unis l'avaient déjà fait et le Royaume-Uni aussi. On s'était dit, on commence à être un peu à la traine quoi. Les collectivités s'étaient déjà lancées, Rennes, Paris… On s'est dit que c'était le sens de l'histoire, il faut qu'on le fasse quoi. Donc voilà on s'est lancé dans ce gros projet.
> (T.Y., un agent de la mission Etalab)

L'expérience de l'administration Obama a contribué à justifier l'abandon du modèle de redevances défendu par l'APIE, la gratuité des données étant considérée comme une des conditions du « succès » de data.gov. Au terme de ce voyage, les représentants du gouvernement ont considéré la France en retard en matière d'*open data*, son portail n'était pas lancé alors que les initiatives se multipliaient au niveau local comme international. Un

immatériels. L'APIE est créée le 23 avril 2007 et rattachée au Ministère de l'Économie et des Finances. L'agence a pour mission d'assister les ministères pour valoriser financièrement leurs actifs immatériels. Les données publiques font partie de ces actifs qui sont considérés comme une source potentielle de recettes considérables pour les administrations.

[64] Nathalie Kosciusko-Morizet (NKM) a été secrétaire d'État en charge des questions numériques avant d'être ministre de l'écologie des gouvernements Fillon.

[65] Le Chief Technology Officer nommé par Barack Obama à son entrée en fonction est en charge d'assurer l'interopérabilité des données et des systèmes d'information de l'État.

rapport publié par Sofrecom, une filiale de conseil et de prospective d'Orange-France Télécom, indiquait que la conception du portail de l'APIE avait pris un retard considérable par rapport au calendrier initial : « à date, ce portail n'a pas été lancé et est désormais annoncé pour 2011. Les informations disponibles semblent indiquer qu'il s'agirait d'un simple agrégateur de liens, renvoyant vers des sources de données déjà existantes, et non d'un portail "one stop" sur le modèle britannique ou états-unien » (Peugeot, Duprat & Tramblay, 2010). Le cahier des charges conçu par l'APIE envisageait la création d'un catalogue qui renverrait vers les sites des institutions disposant de données réutilisables. Or, les deux sites de référence dans ce rapport, data.gov et data.gov.uk, ont choisi un accès direct et ont imposé leur gratuité, à l'opposé du modèle de redevances de l'APIE. Pour les émissaires du gouvernement à Washington, les choix de l'agence apparaissaient alors en porte-à-faux par rapport aux initiatives américaines et britanniques qu'ils vantaient au Premier ministre et au Président.

La position de l'APIE en faveur des redevances a été aussi remise en cause en mai 2010 suite à une initiative d'un agent du ministère de la Justice qui a ouvert une brèche en faveur de l'usage de licences dites « libres » pour les données. Thomas Saint-Aubin, chargé d'enseignement à l'université Paris I et chef du bureau de la stratégie éditoriale du ministère de la Justice, a publié un article le 6 avril 2010 sur le site village-justice.com[66]. Dans son article, il annonçait la création d'une licence « IP » comme Information Publique compatible[67] avec les licences Creative Commons et les licences libres. L'association Regards Citoyens s'est félicitée de la création de cette licence : « l'initiative du Ministère de la Justice est salutaire, elle démontre que la notion de licence libre pour les contenus (documents ou données) est bel et bien compatible avec le droit français et ses obligations réglementaires[68]. » Cette initiative au sein du ministère de la justice a créé une alternative à l'approche de l'APIE : il était désormais possible, en droit français, d'ouvrir des données gratuitement en utilisant une licence « libre » et compatible avec les standards

[66] Village Justice,. « Peut-on diffuser des données publiques sous licences libres et ouvertes ? » , http://www.village-justice.com/articles/diffuser-donnees-publiques,7658.html#lrxr6dLRgRb65vQv.99, consulté le 1 décembre 2014.

[67] La compatibilité signifie que la licence est mise à disposition «selon les termes de la licence Creative Commons Paternité-Partage des Conditions Initiales à l'Identique 2.0 France» (CC by-sa).

[68] Regards Citoyens, « Licence « Information Publique » : un grand pas pour la France ? ». http://www.regardscitoyens.org/licence-%C2%AB-information-publique-%C2%BB-un-grand-pas-pour-la-france/, consulté le 3 décembre 2014.

internationaux. À cette initiative, s'ajoutait celle de la ville de Paris qui a demandé à une association de défense du logiciel libre, Veni Vedi Libri, de traduire la licence ODbL (*Open Database Licence*) de l'Open Knowledge Foundation. Cette licence s'inspirait des principes du copyleft dans le logiciel libre et des biens communs de la connaissance[69]. Elle impose de partager les données avec la même licence en cas de réutilisation publique qu'elle soit commerciale ou non. En décembre 2010, les conseillers de Paris ont adopté la licence ODbL pour les données publiées sur le portail opendata.paris.fr lancé fin janvier 2011. L'intégration des licences Information Publique et ODbL a contribué à discréditer l'approche choisie par l'APIE et a proposé des alternatives libres et ouvertes, conformes aux exigences de l'Open Definition, pour la diffusion de données publiques. L'adoption de ces licences ouvertes par les collectivités locales a créé un précédent juridique en faveur de la gratuité des données gouvernementales.

Enfin, le contexte politique de l'époque a contribué en faveur de l'abandon de la stratégie de l'APIE. Avant le lancement de data.gouv.fr, essentiellement des collectivités locales dirigées par des élus socialistes ont mis en œuvre des politiques d'*open data*. Plusieurs acteurs interrogés ont interprété l'accélération du développement de data.gouv.fr comme une volonté pour la majorité de rattraper son « retard » face aux collectivités locales d'opposition. D'autre part, de manière plus officieuse, Etalab aurait été créée en vue de la campagne présidentielle pour renforcer le bilan du président sortant. Le portail data.gouv.fr avait été conçu comme un symbole de la transparence et de l'innovation du candidat : « le projet était très politique : ouvrir le site avant la campagne électorale, c'était pour montrer le bilan du candidat Sarkozy, tout était calculé dans cet esprit-là. » (Q.H., Correspondant du réseau Etalab, ministère). La création de la mission Etalab s'est en effet précipitée quelques mois avant le début de la campagne officielle. Sa mise en œuvre a été portée par un proche de l'exécutif, Séverin Naudet, le conseiller du Premier ministre sur les questions numériques.

À la fin de l'année 2010, l'APIE a été progressivement désinvestie du projet de création d'un portail unique. Ce projet a évolué pour s'inscrire dans la filiation du portail data.gov dont il

[69] La licence a été créée dans le cadre du projet Open Data Commons de l'Open Knowledge Foundation qui vise à inscrire les données dans le mouvement des biens communs de la connaissance par la création d'outils et de licences.

a repris le nom avec data.gouv.fr. Le mimétisme à l'égard des initiatives anglo-saxonnes a transformé le projet en l'orientant vers la gratuite des données. Le contexte politique de l'élection présidentielle a aussi entrainé une accélération de la conception du portail qui est devenue une priorité des équipes gouvernementales.

Etalab : un engagement affiché en faveur de « l'open data »

Le 30 juin 2010, le conseil de modernisation des politiques publiques, un organe interministériel en charge de la révision générale des politiques publiques (RGPP), a décidé de la création d'un « État lab », un « portail Internet recensant les données existantes et permettant leur réutilisation ». L'APIE a publié un communiqué de presse dans lequel elle indiquait que le portail s'inspire « des initiatives engagées en Grande-Bretagne et aux États-Unis[70] ». Le conseil des ministres du 24 novembre 2010 a confirmé la décision de juin et annoncé la mise en ligne de ce portail avant la fin de l'année 2011 : « un portail unique des données publiques, intitulé "Etalab" sera créé. Il favorisera la réutilisation des données publiques par des acteurs privés. Un directeur de projet sera prochainement nommé afin de piloter la mise en ligne de ce portail d'accès aux données publiques d'ici fin 2011[71] ». À partir du conseil des ministres de novembre 2010, l'APIE a été dessaisie du dossier. « État lab » ne désigne plus le portail, devenu data.gouv.fr, mais la mission en charge de sa création. Sa direction a été attribuée à Séverin Naudet, ancien vice-président du site de partage de vidéos Dailymotion nommé en 2007 « conseiller spécial sur Internet et le multimédia » de François Fillon. Le 21 février 2011, le décret 2011-194 a créé la mission « Etalab » placée sous l'autorité du Premier ministre et rattachée au secrétaire général du Gouvernement. Le décret donnait à Etalab pour missions de créer le portail data.gouv.fr et de coordonner les actions des administrations pour faciliter la réutilisation des informations publiques. Séverin Naudet avait alors pour tache de constituer l'équipe d'Etalab. Il cherchait des personnes avec une expérience dans le secteur numérique capables de travailler avec un calendrier serré. L'équipe d'Etalab (figure 16) a dû travailler dans l'urgence avec pour impératif de lancer le site en décembre 2011, avant le début de la campagne présidentielle officielle pendant laquelle l'administration n'a plus le droit de lancer de nouveaux projets. Leurs réalisations

[70] Economie.gouv.fr, « Conseil des ministres du 30 juin 2010 : revue générale des politiques publiques », http://www.economie.gouv.fr/apie/2010-07-conseil-des-ministres-30-juin-2010-revue-generale-des-politiques-publiques, consulté le 3 décembre 2014.

[71] Vie Publique, « Conseil des ministres du 24 novembre 2010. L'administration électronique. » http://discours.vie-publique.fr/notices/1060002518.html, consulté le 3 décembre 2014.

étaient suivies directement par les services du Premier ministre qui exigeaient un avancement rapide des travaux.

> On avait un calendrier politique extrêmement serré. On arrive en février et il fallait, en décembre, lancer une plateforme des données publiques de l'État. Et, en 2012, on avait la présidentielle donc c'était une grosse pression, il ne fallait pas qu'on se foire. (T.Y., un agent de la mission Etalab)

Les services du Premier ministre ont placé l'équipe d'Etalab, dans leurs locaux, rue de Babylone à Paris, ce qui témoignait de la priorité du projet pour l'exécutif, de l'avis de plusieurs personnes interrogées.

Figure 16. L'équipe d'Etalab au lancement du portail[72].

Pendant la période qui a précédé le lancement de data.gouv.fr, l'équipe d'Etalab a organisé une série de rencontres pour mieux connaitre les acteurs et les pratiques de l'*open data*. Le 15 mars 2011, l'équipe a rencontré le directeur technique de la ville d'Edmonton au Canada qui lancé un portail en janvier 2010. Le 17 mars, elle rencontrait le professeur Nigel Shadbolt pour discuter du cas de la création de data.gov.uk. Le 23 mars 2011, la mission organisait une rencontre avec des représentants de l'Open Knowledge Foundation, la Mairie de Paris et Regards Citoyens (figure 17). À partir du mois de mai, Etalab a aussi organisé quatre

[72] Le blog de la mission Etalab, « L'équipe d'Etalab », https://www.etalab.gouv.fr/lequipedetalab, consulté le 3 décembre 2014.

réunions de travail publiques, ouvertes à tous, lors desquelles certains des aspects du portail data.gouv.fr ont été débattus.

Figure 17. Rencontre du 23 mars 2011[73].

À travers ces rencontres, l'équipe d'Etalab s'attachait à comprendre les normes et les pratiques de l'*open data* pour les refléter dans ses choix et éviter d'être associée aux initiatives de l'APIE. En effet, l'approche de commercialisation des données préconisée par l'APIE était considérée comme antinomique de l'*open data* pour une association comme Regards Citoyens qui a incité Etalab à revoir plusieurs de ses positions. La réunion entre cette association et Etalab (figure 17) s'est tenue, selon T.Y., dans un climat tendu, car Regards Citoyens était radicalement opposée à la politique de vente des données de l'APIE et considérait qu'Etalab s'inscrivait dans cette filiation. D'autre part, Regards Citoyens exigeait l'ouverture de données sur la transparence de l'État, des informations que l'APIE n'avait pas inclus dans son périmètre, l'agence ayant pour mission de générer des revenus et de soutenir la création économique.

> À l'époque, il y avait l'APIE qui avait fait travail de préfiguration sur l'***open data***. Nous on arrivait en essayant de marquer une sorte de rupture par rapport au positionnement de l'APIE en insistant sur le principe de gratuité par défaut. [Regards Citoyens], ils nous ont vu arriver, ils nous ont dit « non, mais attendez les gars on

[73] Etalab, « Etalab Participe À Une Rencontre OKFN / Mairie de Paris / Regards Citoyens, », https://www.etalab.gouv.fr/etalabparticipeaunerencontreokfnmairiedeparisregardscitoyens, consulté le 3 décembre 2014.

> nous l'a déjà fait, on a vu ce que l'État avait fait en la matière, il y a du boulot, on est loin de la transparence de l'action publique. » Quand on s'est vu, c'était le tout début de la mission : la plateforme n'était pas encore montée, on n'avait pas de données, on venait de finaliser la rédaction de la circulaire du 26 mai 2011, les correspondants n'étaient pas encore nommés, il y avait que dalle quoi. Et donc ils nous disaient, « voilà nous ce qu'on attend, c'est davantage de transparence avec plus de données en matière de dépenses publiques, au niveau des élections… » Bon après je vais ne pas citer des propos qui ne sont peut être pas les bons… Les gros problèmes qu'ils ont soulevés, c'est « on manque de données, le travail qui avait été mené auparavant [par l'APIE] n'était pas en adéquation avec nos attentes, c'était plus de la valorisation du patrimoine immatériel que de l'***open data***. »
> (T.Y., un agent de la mission Etalab)

Ces rencontres ne se résument pas qu'au simple échange de « bonnes pratiques », elles ont été un des lieux où ont été débattus et négociés quelques-uns des aspects essentiels de la mise en œuvre de la politique *open data* du gouvernement. Dans le cas précédent, Regards Citoyens a demandé l'ouverture de données sur la transparence de l'État, la gratuité des données publiques et l'utilisation de standards ouverts. Etalab a marqué une rupture avec l'APIE et satisfait une des revendications de Regards Citoyens le 17 octobre 2011 lorsque la mission a publié la Licence Ouverte qui acte de la gratuité des données publiées, autorise les usages commerciaux et impose aux réutilisateurs de citer la source. Regards Citoyens a salué sa publication en annonçant que « la guerre française des licences s'achève », car la Licence Ouverte « répond globalement aux attentes et aux demandes de la communauté des réutilisateurs »[74] notamment par sa compatibilité avec les licences Creative Commons et ODbL. Cela signifie que les clauses juridiques de ces licences, comme l'attribution à la source, sont standardisées et équivalentes entre elles. Le billet de l'association demandait qu'Etalab se concentre « sur l'autre enjeu crucial de l'Open Data : les formats » en réclamant l'ouverture des données dans « des formats ouverts et structurés clairement définis et reconnus » l'abandon du format XLS propriété de Microsoft. À travers ces exigences juridiques et techniques, Regards Citoyens a tenté de définir ce qui relevait ou non de l'*open data* comme on le voit avec le cas de l'APIE dans l'extrait d'entretien précédent. Ce cas montre aussi l'importance des « entrepreneurs de cause » (Cobb & Elder, 1972), des groupes ayant la capacité de publiciser leurs analyses et de mobiliser les acteurs politiques autour de

[74] Regards Citoyens, « Open data & Etalab : la guerre française des licences s'achève », http://www.regardscitoyens.org/opendata-etalab-la-guerre-francaise-des-licences-sacheve, consulté le 3 décembre 2014.

leurs revendications, dans la mise en œuvre de la politique d'ouverture des données du gouvernement.

En juin, un prototype a été lancé, mais data.gouv.fr n'était alors qu'une coquille vide sans les données des administrations. L'équipe d'Etalab s'est alors orientée vers une autre priorité : l'identification des données publiques. Le 26 mai 2011, le Premier ministre, François Fillon, a publié un décret[75] et une circulaire[76] adressés aux ministres, secrétaires d'État et préfets. Le décret marquait l'arrêt du développement des redevances pour les données publiques. Il imposait que la liste des données publiques soumises à redevance soit arrêtée par décret. Hors de cette liste maintenue par Etalab, les données publiques dans leur ensemble sont devenues gratuites à partir du 1er juillet 2012. La circulaire établissait comme « priorité dans la politique gouvernementale de modernisation de l'État et de développement de l'économie » l'accès et la réutilisation des données publiques. Du point de vue des acteurs interrogés, le point important de cette circulaire se situait dans la quatrième annexe. Elle imposait dans un délai de dix jours à chaque ministère de désigner un interlocuteur unique pour Etalab, responsable du recensement des informations publiques, de la mise en place d'une procédure pour transmettre les données et de la réponse aux demandes des réutilisateurs.

La circulaire a permis à Etalab de tisser un réseau dans l'administration à travers les secrétaires généraux qui chapeautent les administrations des ministères. Ils se sont appuyés aussi sur le soutien politique des cabinets ministériels qui ont appuyé leurs demandes du fait de la priorité assignée au projet par le Premier ministre. Après la constitution de ce réseau, le recensement des données publiques a débuté par l'organisation de réunions lors desquelles ont été discutées les données à ouvrir au lancement de data.gouv.fr. La circulaire de mai 2011 exigeait, dans un délai d'un mois, une rencontre « bilatérale » avec les représentants de chaque ministère lors de laquelle ont été fixés des objectifs quantitatifs et qualitatifs de publication de données avec des dates de livraison. Les ministères étaient

[75] Premier ministre. Décret n° 2011-57 du 26 mai 2011 relatif à la réutilisation des informations publiques détenues par l'État et ses établissements publics administratifs. Journal Officiel de la République Française. 27 mai 2011.

[76] Premier ministre, Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'État « data.gouv.fr » par la mission « Etalab » et l'application du droit de réutilisation des informations publiques. Journal Officiel de la République Française. 27 mai 2011.

représentés lors de ces réunions bilatérales par des correspondants et des conseillers du cabinet du ministère. Ils y découvraient bien souvent en quoi consistait la politique d'*open data* engagée par le gouvernement. Dans ces rencontres, Etalab réclamait l'obtention d'un maximum de données. En effet, les services du Premier ministre ont assigné comme objectif à Etalab de dépasser le nombre de jeux de données publiés sur data.gov et data.gov.uk. Les agents de la mission ont donc encouragé les correspondants à obtenir des données dans les différents services des ministères pour diversifier les thématiques et accroitre la quantité de données publiées.

L'équipe d'Etalab contraignait souvent les agents à l'ouverture des données à leur disposition. Séverin Naudet s'appuyait sur son rattachement direct au Premier ministre pour obtenir le soutien des cabinets ministériels. Un correspondant d'Etalab (T.Y.) expliquait que « quand il arrivait dans un cabinet, c'était le représentant de Fillon politiquement donc il arrivait avec cette légitimité. » Passant dans le registre de l'injonction, le directeur de la mission négligeait les réticences des producteurs et déconsidérait parfois la fonction publique dans son ensemble. Des tensions ont émergé progressivement entre l'équipe d'Etalab et leurs interlocuteurs dans les ministères.

> Séverin Naudet avait quand même l'arrogance du gars qui vient du privé, qui méprise un petit peu l'administration. […] Sur le point de vue relationnel, il était épouvantable et il nous a maltraités si je puis dire. Je pense qu'il avait un profond mépris pour l'administration : « l'administration, c'est des feignants, il faut les pousser, y a toujours des marges de manœuvre, il faut toujours en tirer quelque chose. » Même si ça a quand même son efficacité […] il faut quand même motiver les administrations, les associer et non pas leur mettre le pistolet sur la tempe. Il nous a traités comme de la merde entre guillemets.
> (Q.H., Correspondant du réseau Etalab, ministère)

L'ouverture des données s'est souvent faite selon des procédures exceptionnelles qui dérogeaient avec les circuits habituels et les délais d'exécution de l'administration. Les demandes d'Etalab perturbaient parfois les relations des correspondants avec les agents de leur ministère : « sa démarche, à court terme, elle était viable, mais heureusement qu'il est parti parce qu'à long terme, ça devenait tendu avec les administrations. » (Q.H.) Les correspondants se sont réunis une fois de manière informelle pour discuter « de la démarche autocratique de Séverin Naudet » (Q.H.). Ces tensions ne relèvent pas uniquement de la

personnalité de son directeur, elles témoignent aussi du faible ancrage de l'ouverture des données dans les pratiques de l'administration. Le faible enracinement de la mission dans les pratiques étatiques était particulièrement palpable dans la période qui a suivi la défaite de Nicolas Sarkozy. Pendant plusieurs mois, l'activité d'Etalab a été interrompue, car la mission était associée à ses attaches politiques.

L'alternance : Etalab sur la sellette

La première version de data.gouv.fr a été lancée le 5 décembre 2011, conformément au calendrier du projet. Le site (figure 18) était présenté comme la « plateforme française d'ouverture des données publiques (*open data*) ». Résultat du recensement des données publiques, Etalab communiquait particulièrement sur le nombre de données publiées sur le site (352 000) excédant les chiffres annoncés pour data.gov aux États-Unis et data.gov.uk au Royaume-Uni.

Figure 18. Capture d'écran du site data.gouv.fr.(version du 28 décembre 2011, source : Archive.org)

Après le lancement de data.gouv.fr, Etalab a poursuivi ses activités malgré la campagne pour l'élection présidentielle qui a interrompu bon nombre de projets gouvernementaux. En février 2011, la mission annonçait le lancement de Dataconnexions, un concours pour valoriser les meilleures réutilisations des données ouvertes. Une dizaine de partenaires privés[77] ont financé son organisation et la tenue de remises de prix deux fois par an. En avril 2012, l'équipe d'Etalab publiait une nouvelle version de data.gouv.fr permettant de lancer un sujet de discussion autour d'un jeu de données et de demander de nouvelles données dans une « boite à idées ». La poursuite de l'activité d'Etalab était pourtant contestée : les services administratifs étaient tenus par leur hiérarchie et la jurisprudence de ne pas communiquer pendant la campagne officielle. Certains correspondants considéraient le fait de publier une donnée sur le site comme une action de communication, mais Séverin Naudet et son équipe continuaient néanmoins d'exiger l'ouverture de nouvelles données. Pendant la campagne, Séverin Naudet a pris parti en faveur de Nicolas Sarkozy et a communiqué publiquement son militantisme bien que statutairement il était soumis à un devoir de réserve. Un correspondant craignait alors que le projet d'ouverture des données ne soit abandonné suite à la défaite du président sortant en mai 2012.

> Il n'a pas du tout respecté son devoir de réserve. Il était quand même directeur d'administration centrale. Même pour le candidat…des mecs qui veulent faire des histoires, un directeur d'administration centrale s'exprime. Donc ce qui m'a un peu inquiété, il colore trop ce projet comme un projet de droite […] J'avais peur qu'il soit entaché en projet Sarkozy et, à ce titre là, on fait le ménage et on passe en phase de purgatoire, la période a duré presque un an jusqu'à maintenant. Dans mon ministère, au niveau du cabinet, c'était silence radio.
> (Q.H., Correspondant du réseau Etalab, ministère)

Pour ce correspondant, la défaite de Nicolas Sarkozy et le changement de majorité ont bloqué l'ouverture des données pendant l'année suivant les élections, le projet restant associé à la majorité précédente. En l'absence de soutien politique dans le nouveau cabinet, il peinait à convaincre les agents d'ouvrir volontairement leurs données : « on marchait un peu sans tête, on faisait notre travail, mais c'était le coordinateur qui a repris finalement le projet à lui seul. » À la suite de l'élection présidentielle, plusieurs personnes interrogées ont

[77] Google, Microsoft, Orange, La poste, SNCF, et Dassault Systèmes via sa filiale Exalead étaient partenaires de Dataconnexions à son lancement.

douté de l'avenir de la mission Etalab et du maintien en ligne de data.gouv.fr du fait de la dépendance du projet au pouvoir en place à Matignon. J'ai pu le constater lors du comité de pilotage du projet d'*open data* d'une entreprise publique. Il y était question de l'éventuelle dé-publication d'un jeu de données qui posait problème pour des raisons de qualité. Les agents d'Etalab avaient fortement insisté pour obtenir ce fichier sur data.gouv.fr. La personne qui plaide pour la dé-publication du fichier, en relation régulière avec Etalab, suggérait de profiter de la période pour se désengager : « en même temps, ils sont morts. Justement il y a peut-être une fenêtre de tir puisqu'il y a un flottement, ils n'ont plus de patron. » (comité de pilotage, 27 novembre 2012, entreprise publique). Séverin Naudet est resté en fonction jusqu'en octobre, mais il s'est exprimé rarement de manière publique depuis le changement de majorité. L'équipe a connu une vague de départ : seuls deux de ses membres sont restés dans l'année qui a suivi l'élection présidentielle.

L'incertitude sur la politique d'*open data* a été exacerbée par un article publié dans Les Échos en octobre 2012 qui annonçait que certaines administrations pourraient commercialiser leurs données publiques[78]. Les membres du gouvernement avaient pourtant signé une charte de déontologie qui disposait que le gouvernement « mène une action déterminée pour la mise à disposition gratuite et commode sur Internet d'un grand nombre de données publiques[79]. » L'équipe d'Etalab ne parvenait toutefois plus à ouvrir de nouvelles données tant qu'un nouveau directeur n'a pas été annoncé et soutenu par les nouveaux cabinets ministériels. Certains correspondants et producteurs de données ont refusé de prendre part à une politique à l'avenir incertain.

> Séverin a mis du temps avant de se faire virer. Il a été viré en septembre-octobre, les élections avaient eu lieu en mai donc nous on était un peu dans l'expectative, c'était un peu le ralenti. Les administrations, les correspondants n'avaient pas bougé, ils savaient très bien qu'il y avait une instabilité politique donc on ne pouvait plus jouer sur les cabinets. Les mecs attendaient un petit peu de savoir qui allait être désigné comme nouveau responsable d'Etalab.
> (Q.H., Correspondant du réseau Etalab, ministère)

[78] Les Echos, « Open Data : l'État pourrait renoncer à la gratuité de certaines données publiques », http://www.lesechos.fr/journal20121017/lec2hightechetmedias/0202329690871-open-data-l-etat-pourrait-renoncer-a-la-gratuite-des-donnees-publiques-501147.php, consulté le 15 décembre 2014.

[79] Numérama, « Internet et l'Open Data dans la déontologie du gouvernement Ayrault », http://numerama.com/magazine/22534-internet-et-l-open-data-dans-la-deontologie-du-gouvernement-ayrault.html, consulté le 15 décembre 2014.

Le récit de la période d'incertitude suite à l'alternance met en exergue le caractère fragile de la politique d'ouverture des données publiques. Sa mise en œuvre par Etalab s'est reposée en grande partie sur les attaches politiques de la mission auprès du gouvernement et des cabinets ministériels. En ouvrant des données, les agents ont souvent répondu à une injonction politique avant de participer à une pratique normale de l'administration. La nouvelle équipe d'Etalab a en partie repensé les pratiques et les instruments de l'ouverture des données publiques : comment a-t-elle procédé pour accroitre l'ancrage de son action dans les pratiques de l'administration ? La nouvelle majorité a-t-elle proposé une politique différente de ses prédécesseurs ? Si oui, quelles en ont été les conséquences pour la mission Etalab ?

La refonte de data.gouv.fr : « faire vivre » les données

En décembre 2012, les services du Premier ministre ont annoncé la nomination d'Henri Verdier à la tête d'Etalab. Ancien directeur de Cap Digital, pôle de compétitivité des entreprises du secteur numérique en Ile-de-France, il a aussi fondé MFG Labs, une entreprise spécialisée dans l'exploitation de données massives. Le 18 décembre 2012, le Premier ministre convoquait le premier Comité interministériel pour la modernisation de l'action publique[80] (Cimap) dont l'un des cinq axes abordait « l'administration numérique[81]. » À cette occasion, Jean-Marc Ayrault annonçait « réaffirmer le principe de gratuité de la réutilisation des données publiques » et la publication d'une « feuille de route » d'Etalab lors du séminaire gouvernemental sur le numérique en février 2013. Ces annonces, en replaçant l'ouverture des données comme une priorité gouvernementale, ont dissipé les doutes sur l'éventuel abandon de la politique d'*open data*. Au cours de l'année 2013, de nouvelles orientations ont été données à la mission Etalab qui a changé de rattachement pour rejoindre le SGMAP (Secrétariat Général à la Modernisation de l'Action Publique). Ce changement de rattachement a eu des conséquences sur le discours des agents d'Etalab. Henri Verdier a présenté régulièrement l'ouverture des données publiques comme une manière de transformer l'action de l'État avant d'insister sur son potentiel économique comme le faisait son prédécesseur.

[80] La modernisation de l'action publique (MAP) fait suite à la Révision Générale des Politiques Publiques (RGPP).

[81] PCinpact, « Le gouvernement réaffirme le principe de gratuité des données publiques ». http://www.pcinpact.com/news/76176-le-gouvernement-reaffirme-principe-gratuite-donnees-publiques.htm, consulté le 15 décembre 2014.

La feuille de route du gouvernement publiée en février 2013[82] a donné les nouvelles orientations de la mission Etalab. Le premier changement portait sur les objectifs assignés à la mission. Plutôt que d'être évaluée sur la quantité de données publiées selon un objectif d'exhaustivité proche des préconisations des principes de Sebastopol, Etalab doit désormais publier des jeux de données « stratégiques » qui seront identifiés lors de six débats thématiques avec la société civile. Le modèle retenu par le gouvernement se rapproche donc plus des approches de l'Open Data Index ou de la charte du G8 avec la sélection de données à forte valeur devant être ouvertes en priorité. Évaluée sur sa capacité à publier des jeux de données stratégiques, l'équipe d'Etalab s'est assurée que la France allait améliorer son classement dans l'Open Data Index après avoir obtenu une 16e place en 2013. Pour remonter dans le classement, ses agents ont inspecté les critères et la méthode d'évaluation et ont obtenu la diffusion sur data.gouv.fr de plusieurs données « essentielles » : la base de données des lois, décrets et ordonnances et l'ouverture du fichier des codes postaux par la Poste[83]. Le gouvernement a alors engagé une campagne de communication autour de ce résultat (figure 19).

Figure 19. Illustration diffusée sur le compte Twitter officiel du gouvernement français suite à la publication des résultats de l'Open Data Index.

En outre, la feuille de route annonçait le lancement d'une nouvelle version du portail data.gouv.fr. Henri Verdier reprochait en particulier la difficulté de trouver un jeu de

[82] Etalab, « La feuille de route du Gouvernement en matière d'ouverture et de partage des données publiques », http://www.etalab.gouv.fr/article-la-feuille-de-route-du-gouvernement-en-matiere-d-ouverture-et-de-partage-des-donnees-publiques-115767801.html, consulté le 12 décembre 2014.

[83] Etalab, « La base officielle des codes postaux est disponible sur data.gouv.fr », http://www.etalab.gouv.fr/la-base-officielle-des-codes-postaux-est-disponible-sur-data-gouv-fr, consulté le 14 décembre 2014.

données sur le portail suite à une recherche. Pour sa refonte, il déclarait s'inspirer des principes proposés dans le livre qu'il a coécrit avec Nicolas Colin (Verdier & Colin, 2013), *L'Âge de la multitude*. Selon eux, la richesse dans l'économie numérique dépend de capacité d'un acteur à capter la valeur de la « multitude. » Pour repenser data.gouv.fr, il a nommé un réseau d'experts pour conseiller Etalab et lancé une consultation. Verdier a aussi décidé que le développement de la nouvelle version de data.gouv.fr sera assuré en interne plutôt que par des prestataires. Deux développeurs et un chef de produit ont été recrutés pour sa réalisation. Le portail a été remis à plat à l'issue d'une consultation lors de laquelle « l'écosystème » était invité à répondre à un questionnaire en ligne et à participer à des ateliers de « codesign. » Pour un agent d'Etalab, le dialogue a mis en pratique les principes de l'ouverture des données.

> Ensuite, il y a eu la partie codesign, on a travaillé avec tout l'écosystème. Henri a voulu mettre en place un comité d'experts avec des mecs avec lesquels il a été amené à bosser dans le passé pour les solliciter sur un certain nombre de sujets liés au droit de la donnée publique, à la plateforme… On a essayé de s'entourer au maximum parce qu'on sera toujours plus intéressant avec les personnes qui ont une vraie expérience sur le sujet plutôt que de fonctionner en vase clos. Donc on a vraiment mis en pratique les fondamentaux de l'***open data*** finalement. Qu'on prétendait mettre en pratique auparavant, mais qu'on ne faisait pas réellement dans la pratique.
> (T.Y., un agent de la mission Etalab)

Au terme de la consultation, Etalab annonçait reconsidérer certains des principes qui ont présidé l'ouverture des données. Une des premières conclusions de la consultation publiée sur le blog d'Etalab[84] consistait à « contester les grands dogmes de l'*open data* ».

> Le mouvement ***open data***, tout jeune qu'il soit, est déjà traversé de grands dogmes : il s'agirait de diffuser des données brutes, dans l'état où elles sont produites par l'administration, il ne faudrait pas essayer d'interpréter ces données, l'indexation serait une question centrale…

En pratique, l'équipe d'Etalab a imposé moins de contraintes techniques aux agents. Cela s'est matérialisé dans le nouveau portail qui laissait la possibilité aux agents de publier

[84] Etalab, « Les premiers enseignements de l'opération CoDesign » http://www.etalab.gouv.fr/article-les-premiers-enseignements-de-l-operation-codesign-119057937.html, consulté le 14 décembre 2014.

plusieurs fichiers pour un seul jeu de données. Les visiteurs du portail pouvaient aussi republier un fichier dans un nouveau format après l'avoir traité afin que l'administration puisse bénéficier du travail de la « multitude ». Le nouveau data.gouv.fr était présenté comme un espace de mise en relation entre l'administration et les développeurs. Il a été lancé le 18 décembre 2013 à Matignon par le Premier ministre en présence d'une centaine d'invités. J'y intervenais, au nom de mon association, pour saluer l'utilisation de CKAN, le portail *open source* développé par l'Open Knowledge Foundation. La communication autour du nouveau site insistait sur son caractère interactif et participatif (figure 20). Cette nouvelle version du portail était présentée comme une « plateforme ouverte » et les données publiques comme un « bien commun informationnel » édifié avec les citoyens.

- Pour atteindre ces objectifs, il faut une plus grande quantité de données : le site a été profondément refondé pour y contribuer et permettre ce passage à l'échelle
    - Dépôt facilité des données : expérience simplifiée et référencement vers les sites des producteurs
    - Dynamique vertueuse pour « faire vivre » les données et les améliorer grâce aux contributions extérieures
    - Accueil de nombreux fournisseurs de données : recherche scientifique, établissements publics, collectivités. Ainsi, les grandes collectivités locales engagées dans l'open data ont-elles déclaré l'ensemble des données disponibles sur leurs portails sur cette adresse unique

- Le nouveau data.gouv.fr a été développé avec trois exigences :
    - Permettre à l'utilisateur d'accéder facilement aux données les plus pertinentes pour répondre aux questions qu'il se pose ;
    - Permettre à tout détenteur de données publiques de les partager en une minute ;
    - Enrichir les données publiques grâce aux améliorations ou aux interprétations des réutilisateurs.

- Cette nouvelle version de data.gouv.fr ouvre une nouvelle dimension de l'open data, qui devient plus collaboratif et s'enrichit des données et des contributions de la société civile.

- Ainsi, le nouveau data.gouv.fr n'est pas seulement le portail de diffusion des données du service public : c'est un outil collectif pour co-construire, avec les citoyens, un bien commun informationnel pour l'ensemble de la nation

Figure 20. Extrait des éléments de langage rédigés pour le lancement de la deuxième version de data.gouv.fr.

La communication d'Etalab présentait le nouveau data.gouv.fr comme une plateforme qui, selon les éléments de langage, « permet de "faire vivre" les données et de rencontrer des innovateurs permettant de faire naître de nouveaux services. » Dans sa communication, Etalab insistait tout autant sur la valeur des données publiées sur data.gouv.fr que sur celle de la « communauté » qui s'y active. Cela s'est traduit dans les objectifs opérationnels assignés à la mission Etalab qui doit accroitre le nombre d'utilisateurs actifs de la plateforme et de réutilisations recensées sur data.gouv.fr sur son site, comme on peut le voir dans le projet de loi de finances 2015 (figure 21).

INDICATEUR 7.2 : Ouverture et diffusion des données publiques

(du point de vue du citoyen)

	Unité	2012 Réalisation	2013 Réalisation	2014 Prévision PAP 2014	2014 Prévision actualisée	2015 Prévision	2017 Cible
Nombre de ressources en open data (site "data.gouv.fr")	Nombre	SO	SO	SO	36 000	37 000	40 000
Nombre de contributeurs actifs (site "data.gouv.fr")	Nb de comptes actifs	SO	SO	SO	3 500	4 000	10 000
Nombre de réutilisations (site "data.gouv.fr")	Nb	SO	SO	SO	1 400	2 000	5 000

Figure 21. Indicateurs de performances de l'ouverture des données publiques dans le projet de loi de finances 2015.

Avec la refonte du portail, Etalab s'est donc vue attribuer de nouveaux objectifs qui ne consistent pas uniquement à ouvrir les données, mais à les « faire vivre » en s'assurant de leur réutilisation et de l'existence d'une communauté autour des données. Cela apparait encore plus clairement dans le détail des objectifs qui figurent dans le projet de loi de finances 2015.

> La mission Etalab assure la promotion de la réutilisation des données publiques par des acteurs de l'économie réelle, à cette fin elle multiplie les démarches afin d'intéresser et de fédérer une communauté d'utilisateurs actifs qui partagent des données ou des projets sur le site « data.gouv.fr ». […] Le site « data.gouv.fr » permet aux réutilisateurs de publier et de partager avec la communauté les réutilisations faites à partir des données. La mesure du nombre de ces réutilisations est effectuée sur le site. Il est un indice de l'utilité des données partagées, et démontre la vitalité de la communauté des réutilisateurs des données des administrations.

Les équipes en charge de projet d'*open data* n'ont souvent pas pour seule mission de publier des données. Ces agents doivent aussi s'assurer que ces dernières sont effectivement utilisées, quitte à encourager des publics de s'en saisir. Ce sera précisément l'objet du sixième chapitre dans lequel j'aborderai les instruments qui font exister des publics de données. « Faire vivre les données », cet objectif a aussi conduit Etalab à reconsidérer son rôle dans l'État. Au départ, la mission devait se charger de l'ouverture au public des données. Avec la nomination d'un administrateur général des données rattaché à la mission Etalab, cette institution se charge aussi de coordonner la circulation et l'exploitation des données dans l'État. Après s'être concentré sur l'ouverture des données, Etalab s'intéresse désormais aux données tout court.

L'administrateur général des données : de l'ouverture à la « gouvernance » des données

Au cours de l'année 2015, le cadre juridique de l'ouverture des données a connu de nombreuses évolutions au point que la Gazette des Communes a dû éditer un poster récapitulatif pour tenter de dénouer un « écheveau législatif ».[85] En effet, en 2015, trois lois ont prévu des dispositions relatives à l'ouverture des données. Je vais en aborder très brièvement le contenu étant donné qu'elles n'interviennent pas sur les missions d'Etalab. La première, la loi relative à la gratuité et aux modalités de réutilisation des informations du secteur public, a transposé la révision de 2013 de la directive européenne PSI et a inscrit le principe de gratuité des données dans la loi. Par ailleurs, le projet de loi « NOTRe (Nouvelle Organisation des Territoires de la République) » a disposé que les collectivités locales de plus de 3500 habitants doivent désormais publier volontairement les documents administratifs relevant de la loi CADA. Cette disposition doit être précisée par décret au cours de l'année 2016. Enfin, les premiers articles de la loi pour une République numérique, en cours de discussion au moment de l'écriture de ces lignes, devraient porter sur l'ouverture des données publiques. Entre autres, les dispositions actuellement débattues évoquent des évolutions du droit d'accès auquel les administrations pourraient avoir recours pour réclamer les données d'un autre service et une obligation de publier les documents après l'acceptation d'une demande en vertu de la loi CADA.

Au-delà de ces enjeux législatifs, une évolution importante pour Etalab a été l'entrée de la France dans le Partenariat pour le Gouvernement Ouvert (Open Government Partnership). En avril 2014, lors de la conférence de Paris sur le gouvernement ouvert, Henri Verdier a annoncé que la France allait rejoindre cette organisation internationale créée en 2009 par le président Obama, pour promouvoir les « bonnes pratiques » en matière de transparence des États, la participation des citoyens et la collaboration avec la société civile. Par la suite, la France a intensifié son engagement dans le Partenariat en rejoignant son comité directeur puis en prenant sa présidence en 2016. En décembre 2016, la ville de Paris accueillera son sommet annuel. En plus de leurs missions liées à l'ouverture des données, les agents d'Etalab assurent une grande partie de la représentation de la France au sein de cette institution internationale et élaborent, en concertation avec la société civile, les engagements que prend le gouvernement dans le cadre de son plan d'action. Ces nouvelles missions liées

[85] La Gazette des Communes, « [Poster] Open data : démêler l'écheveau législatif », http://www.lagazettedescommunes.com/397545/poster-open-data-un-echeveau-legislatif/, consulté en octobre 2015.

à des enjeux de participation sont critiquées en particulier par l'association Regards Citoyens qui considèrent qu'elles détournent Etalab de sa mission : « il est temps qu'Etalab se recentre sur les actions concrètes simples et rapides au cœur de ses missions pour vraiment faire avancer la transparence et l'Open Data. »[86]

Du point de vue des données, la mutation la plus importante d'Etalab concerne la création d'une fonction d'Administrateur Général des Données (AGD) annoncée le 21 mai 2014 en conseil des ministres. Henri Verdier a été nommé à cette fonction par décret le 19 septembre 2014. Rattaché au SGMAP tout comme Etalab, l'AGD a pour mission de coordonner l'action des administrations en matière d'« inventaire, de gouvernance, de production, de circulation et d'exploitation des données. » Dans le cadre de ses missions, il peut être saisi par tout citoyen ou toute personne morale et doit remettre chaque année un rapport au Premier ministre sur la gouvernance des données. Au cours de l'année 2015, l'équipe de l'AGD[87] a conduit plusieurs expérimentations avec des administrations notamment sur des domaines variés tels que la consommation électrique de l'État, les vols de voiture ou l'emploi. Les cinq *data scientists* rattachés à l'AGD ont exploité et croisé des données publiques pour créer des services de prédiction et d'aide à la décision.

Le premier rapport de l'AGD a été publié en décembre 2015[88], son contenu est très riche par rapport aux questionnements qui animent ce travail. Pour résumer brièvement la cinquantaine de pages de ce rapport, l'administrateur général des données présente la « révolution des données » et la science des données comme des « leviers de transformation de l'action publique. » Dans sa première partie, il relit l'histoire de l'action publique à l'aune des données : « La construction progressive de l'État moderne s'accompagne de celle d'un ensemble de données de références nécessaires à son organisation et à son fonctionnement, ainsi que de données indispensables au bon fonctionnement de l'économie et de la société. » En particulier, l'AGD replace son action dans la continuité de l'histoire de la statistique et

[86] Regards Citoyens, « La France presidera-t-elle l'Open Communication Partnership? », https://www.regardscitoyens.org/la-france-presidera-t-elle-lopen-communication-partnership/, consulté en juin 2015.

[87] Administrateur Général des Données, « L'équipe », https://agd.data.gouv.fr/lequipe/, consulté en juin 2016.

[88] Administrateur Général des Données, « Rapport annuel 2015: mettre les données au service de la transformation de l'action publique », https://agd.data.gouv.fr/2016/01/21/rapport-annuel-2015-mettre-les-donnees-au-service-de-la-transformation-de-laction-publique/, consulté en mars 2016.

de l'information publiques. Dans la deuxième partie, le rapport s'intéresse aux « freins » qui réduisent le potentiel des données. Le premier « frein » signalé par l'AGD porte sur l'absence de vision centrale sur les données publiques : « nul n'est aujourd'hui en mesure de connaitre avec précision l'étendue des données que l'administration publique possède. » Pour expliquer cette situation, l'équipe d'Henri Verdier regrette que de nombreuses administrations produisent des données « sans les considérer telles quelles. » En particulier, l'AGD invite à s'intéresser aux « données de gestion » produites dans les systèmes d'information de l'État.

> La plupart des données existantes sont aujourd'hui produites dans de grands systèmes de gestion informatisés, et ne sont pas connues ni repérées comme telles. Une histoire connue dans les communautés open data concerne cette grande municipalité qui souhaitait ouvrir son portail d'open data et recherchait dans ce but des données concernant les pratiques culturelles. Il lui fallut près d'un an pour réaliser que l'application de gestion des bibliothèques municipales recelait un trésor : la liste des ouvrages empruntés quotidiennement dessinait une sociologie des pratiques culturelles, permettait de comprendre la saisonnalité des pratiques, d'identifier des corrélations inédites entre types d'ouvrages, de recommander des livres à emprunter, etc. […] De telles données, issues des grands systèmes de gestion, représentent aujourd'hui un sujet central de la gouvernance de la donnée.

Cette description des données de gestion correspond, à peu de choses près, aux « sources administratives » de la statistique publique évoquées par Desrosières (2005) lorsqu'il distingue deux sources de la statistique publique : d'une part, les « enquêtes » produites spécifiquement par des institutions dédiées selon des normes scientifiques et d'autre part, les « sources administratives » issues de services « dont les activités de gestion impliquent la tenue, selon des règles générales, de fichiers ou de registres individuels, dont l'agrégation n'est qu'un sous-produit, alors que les informations individuelles en sont l'élément important, notamment pour les individus ou les entreprises concernés. » Tout comme les données brutes qu'il suffirait de réclamer pour obtenir leur diffusion, les fichiers de gestion des administrations sont souvent présentés comme un matériau dont l'ouverture ne coûterait rien, « comme s'"il n'y avait qu'à se baisser pour les cueillir" » (Desrosières, 2005). Les autres « freins » à l'exploitation des données publiques signalés par l'AGD portent sur la conception et la gestion des systèmes d'informations de l'État, Henri Verdier étant directeur interministériel du numérique et du système d'information et de communication

de l'État depuis septembre 2015[89]. L'AGD déplore que les systèmes d'informations de l'État n'aient pas été prévus pour extraire les données et que le recours à la sous-traitance « sans équipe interne qui garde la main sur les données » ait porté atteinte à « l'autonomie de l'État. » Sur les aspects organisationnels, l'AGD 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. » Enfin, dans sa troisième partie, le rapport de l'AGD préconise notamment au Premier ministre la poursuite des expérimentations et la conduite de nouveaux projets grâce à un marché d'appui à la *data science*, la cartographie des données de l'État dans les ministères volontaires et le renforcement de l'expertise juridique et technique en matière d'anonymisation des données.

En parcourant ces quelques jalons dans l'histoire récente d'Etalab, nous avons pu voir que les missions et le cadre législatif et administratif de cette institution connaissent des mouvements constants. L'adhésion et l'engagement croissant de la France au sein du Partenariat contribuent à orienter cette institution vers des enjeux de participation. Cette nouvelle mission attribuée à Etalab s'inscrit dans la lignée des objectifs fixés par le gouvernement dans le projet de loi de finances qui considèrent que ses agents ne doivent pas uniquement ouvrir des données, mais les « faire vivre », leur trouver un public et les inscrire dans de nouveaux réseaux sociotechniques. C'est aussi le sens de l'action de l'Administrateur Général des Données qui oriente les données vers un autre public qu'Etalab, celui des usagers internes des administrations. À travers ses premières expérimentations et son rapport, il porte l'attention sur une ressource « inexploitée » et brute, les données de gestion des administrations.

Conclusion

Avant de tirer des conclusions de la trajectoire d'Etalab que nous venons de parcourir dans ce chapitre, il me faut d'abord la resituer dans une histoire plus longue, celle de la réutilisation des informations publiques. Aux États-Unis, cette question avait déjà fait l'objet d'une mobilisation qui a mené à l'adoption d'un principe de réutilisation libre et gratuite

[89] Henri Verdier a quitté la direction d'Etalab. Laure Lucchesi, auparavant direction adjointe d'Etalab, l'a remplacé à la tête de la mission.

des informations publiques. À la fin des années 1970, les industriels de l'information avaient mis en place une action de lobbying, sous l'égide de l'*Information Industry Association* (IIA), pour obtenir l'adoption d'un principe général selon lequel le secteur public ne doit pas concurrencer les entreprises privées dans le domaine de l'information (Ronai, 1994). Cette campagne d'influence a abouti à l'adoption du *Paperwork Reduction Act* en 1985 et du *Copyright Act* en 1988 qui excluent les informations publiques du droit d'auteur et permettent aux entreprises leur exploitation sans restrictions. En France, le principe de réutilisation libre et gratuite des informations publiques est entré plus tardivement dans la réglementation. Les différences de la France ont souvent été expliquées comme étant liées au modèle économique du Minitel qui, jusqu'à la fin des années 1990, a assuré des revenus importants aux administrations par la revente de leurs informations (Ronai, 1996 ; Boustany, 2013 ; Trojette, 2013). C'est essentiellement par l'Union européenne que le droit à la réutilisation s'est imposé en France. Dans les années 2000, la Commission européenne s'est inspirée du cadre réglementaire états-unien et a multiplié les études sur le potentiel économique de la réutilisation de l'information publique, évaluant jusqu'à 200 milliards d'euros par an la valeur de leur circulation optimale dans les pays de l'Union (Vickery, 2011). Sans imposer la gratuité, elle a fixé des règles limitant le montant des redevances et demandant l'utilisation de licences standardisées. En 2007, la directive européenne dite « INSPIRE » a imposé l'ouverture proactive et gratuite des données géographiques pour favoriser la protection de l'environnement.

Dans le chapitre précédent, je me demandais comment la mise en lumière des données a changé les politiques informationnelles de l'État. Nous avons ici un cas précis de catégorisation[90] en termes d'*open data* et de données du problème de la réutilisation de l'information publique. En faisant démarrer cette petite histoire d'Etalab à la création d'un portail par l'APIE, nous avons pu voir comment cette catégorisation a conduit à une

[90] En sociologie de l'action publique, la notion de catégorisation a été développée en particulier par Cefaï (1996) qui a montré que l'apparition de nouvelles catégories accompagnent souvent l'émergence d'arènes publiques dans lesquelles les problèmes publics sont discutés : « Nommer et narrer, c'est déjà catégoriser, faire advenir à l'existence et rendre digne de préoccupation, qu'il s'agisse de « nouvelle pauvreté » ou d' « avortement volontaire », de « malaise des banlieues » ou de « commerce d'enfants. » […] Inscrire le problème public dans un contexte d'interprétation, d'explication et de jugement, ce n'est pas seulement le désigner comme un référent objectif ; c'est aussi le faire advenir en tant que problème. Dans la lignée pragmatiste de J. Dewey, le problème public est plus que le produit d'un « étiquetage collectif », c'est une « activité collective » en train de se faire. «

redéfinition de la politique gouvernementale. C'est à la suite d'un voyage d'études de plusieurs émissaires du gouvernement à Washington, à la rencontre de l'équipe derrière data.gov, et d'une volonté du président sortant d'afficher sa transparence, qu'Etalab a été créé pour mettre en œuvre une politique d'*open data*. Ces circonstances singulières, qui rappellent que les politiques publiques prennent la forme d'expérimentations marquées par l'incertitude (Dewey, 2010) et naissent de circonstances inédites très éloignées du mythe de la décision « rationnelle » et « visionnaire » (Sfez, 1976 ; Lascoumes & Le Galès, 2007 ; Hassenteufel, 2011), ont conduit à ce que le gouvernement désinvestisse l'APIE de la mission de création d'un portail pour les données publiques. Les agents d'Etalab ont alors tenté de traduire les grands principes de l'ouverture des données à travers plusieurs mesures comme la gratuité des données publiées sur le portail, le choix d'une licence compatible avec les standards internationaux, la mise en place d'un portail unique avec data.gouv.fr et l'affichage d'un objectif de lisibilité des données par les machines. Dans ce cadrage par l'*open data*, les agents d'Etalab se sont intéressés en particulier aux données brutes pour lesquels ils ont organisé un réseau de correspondants dans l'État dédié à leur recensement et leur ouverture. En attribuant des responsabilités inédites à des agents et en créant de nouvelles procédures, Etalab a « travaillé l'organisation » de l'État (Cochoy, Garel & de Terssac, 1998) pour permettre l'ouverture des données brutes.

La trajectoire d'Etalab au cours de ces cinq dernières années souligne l'instabilité et la fragilité de la politique publique d'ouverture des données. En effet, comme on a pu le voir à travers l'alternance, l'évolution de son contexte légal et l'attribution régulière de nouveaux objectifs en matière d'*Open Government* ou de réutilisation des données, cette structure reste très liée à ses attaches politiques et à un contexte législatif mouvant. Les missions pourraient donc évoluer en 2017 après l'élection présidentielle. Néanmoins, au-delà de la trajectoire d'Etalab, on peut retenir de l'histoire de cette structure qu'elle a contribué à porter l'attention sur les données publiques, bien au-delà de la question de leur ouverture. En considérant la donnée comme une ressource inexploitée, le « nouveau pétrole » gisant sous les organisations, l'ouverture ne constitue qu'un des niveaux d'une politique plus large de « gouvernance des données. » Cela s'est traduit par la distinction dans l'organisation de l'État entre Etalab chargé de l'ouverture au public et l'Administrateur Général des Données organisant leur circulation dans l'État et leur exploitation. Une fois passés les moments de déploiement des politiques d'*open data*, si l'on suit les projets en train de se faire, nous

verrons que la circulation fluide des données n'a rien d'évident. Dans le chapitre suivant, nous allons voir que cela débute par une étape d'identification lors de laquelle les données brutes doivent être localisées parmi la masse d'objets informationnels produits quotidiennement par les administrations. Au-delà des enjeux organisationnels que soulève ce travail d'identification, il permettra d'apporter de nouvelles réponses à la question la plus générale de cette thèse : « qu'est-ce qu'une donnée ? »

Chapitre 3

L'identification : la découverte progressive et collective des données

Revenons-en à la réunion de novembre 2013 de la région Ile-de-France. L'invitation annonce que les participants doivent « prendre le temps d'identifier d'éventuelles données avant la tenue de l'évènement. » Au début de la rencontre, Laurent, un des organisateurs, présente le programme de la demi-journée et annonce la tenue d'« un jeu de rôle » qui va servir à identifier de nouvelles données.

> Ce que nous allons faire, c'est une session d'animation dans laquelle nous allons chercher à simuler l'enrichissement du portail, une sorte de jeu de rôle où on va se mettre en situation. Vous voyez de l'autre côté, il y a un grand tableau avec les thématiques de classement du portail ***open data***. Ici, on a des fiches d'identification qui représentent les jeux de données. On va essayer de jouer l'enrichissement du portail ***open data***. On va imaginer quelles seraient les données qui seraient susceptibles d'enrichir les jeux de données ***open data*** déjà publiés. […] La première étape, c'est l'identification. C'est avec vous qu'on va identifier les jeux de données qui seront libérables avec une prise en considération des aspects juridiques et techniques pour évaluer le degré de facilité avec lequel on va ouvrir cette donnée.

Figure 22. Photo de la salle de l'évènement Open Data Bootcamp de la région Ile-de-France avec le tableau représentant les catégories du portail. Image communiquée par un des organisateurs.

Derrière les participants, un grand tableau représente les catégories[91] qui organisent les jeux de données sur le portail *open data* (figure 22). Pendant la moitié du temps de l'évènement, les participants simulent l'enrichissement du portail en remplissant des fiches qui représentent un nouveau jeu de données et en les positionnant sur le tableau. Un organisateur explique les « règles du jeu » et distribue les fiches avec lesquelles les participants vont décrire les jeux de données qu'ils ont préalablement identifiés.

> Je prends ma fiche, ma bombe de colle repositionnable et je regarde la nomenclature de classement. Et donc je me dis que ça s'est plutôt « enjeux économiques et innovation ». Et l'objectif, ça va être de simuler l'enrichissement du portail à partir de ce que vous allez pouvoir référencer. Donc je vais distribuer les fiches. Ceux qui savent qu'ils ont des jeux de données dynamiques, demandez-moi une fiche rouge, on n'en a imprimé moins. Il y a des bases de données dynamiques ? Qui veut du rouge ? Du vert ? Du rouge ?

L'audience se répartit en petits groupes souvent composés d'agents du même service du conseil régional. Les organisateurs aident les participants à remplir les fiches et discutent de l'ouverture de certaines données. Dans un groupe, ils convainquent un agent d'un service spécialisé dans l'agriculture et l'alimentation d'ouvrir une base de données sur les restaurants dont elle ne voyait pas l'intérêt pour les citoyens. Dans d'autres groupes, les organisateurs se réjouissent en voyant les fiches de données qui sont remplies. Pour un fichier dynamique, je les entends dire « la fiche est rouge, bien ! Woo ! ça veut dire que c'est un fichier dynamique. » Pour un jeu de données sur les logements sociaux, « ça, c'est formidable », des données sur les transports « c'est un gros morceau », la liste des offices de tourisme « ah ça c'est bien » ou des données budgétaires « ça, c'est intéressant. » Leurs commentaires distinguent certaines données pour leur intérêt et encouragent les participants à remplir plus de fiches pour enrichir la simulation. Les fiches sont restituées aux animateurs qui lisent la fiche, sélectionnent avec les participants une catégorie et la collent sur le tableau qui représente le portail.

[91] Les catégories affichées au fond de la salle étaient les suivantes, placées de gauche à droite : aménagement du territoire ; bâtiment, équipements ; logement, santé, social ; déplacements, transports ; vie sociale ; emploi ; assemblée régionale ; justice ; vie culturelle : cadre de vie, environnement ; enseignement, formation, recherche ; vie économique, innovation ; administration ; finances publiques ; vie urbaine ; sport, tourisme, loisirs.

Un mois plus tard, je revois l'un des animateurs en entretien. Je lui demande de m'expliquer les raisons de l'organisation de cet évènement et l'intérêt d'avoir fait remplir des fiches par les participants lors de l'atelier que je viens de décrire. Ce qui était annoncé comme un « jeu de rôle » s'est avéré être un très bon moyen d'identifier de nouvelles données. Alors que le portail *open data* avait été lancé il y a près de six mois, les responsables du projet ont identifié de nouvelles données dont ils ignoraient l'existence. L'organisation de cette rencontre a permis aussi de formaliser un réseau de producteurs de données ouvertes, une « communauté » sur laquelle le projet est bâti.

> On voulait profiter qu'ils soient tous là pour faire émerger de nouvelles données, c'est ce qu'on a fait avec les fiches. Donc on a eu cinquante jeux de données exclusifs dont on n'avait quasiment jamais entendu parler, ça, c'était vraiment très bien. Et troisièmement, on voulait fédérer la communauté, ça fait partie de la carotte qu'on veut leur donner, c'est-à-dire qu'ils font partie d'une sorte de service invisible dans la région. […] Là ils étaient tous ensemble, ils ont tous craché leurs données. Si j'étais allé les voir un par un, on n'aurait pas eu de résultats comme ça. C'était genre la communauté, c'est toujours plus fort quoi.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

Cet épisode est riche en enseignements quant aux enjeux des opérations d'identification des données. En suivant cette réunion, on voit dans un temps court quelques-unes des questions qui se posent lors de l'identification des données. On peut déjà tirer un premier enseignement très général de ce court récit : le processus d'ouverture débute par une phase d'identification lors de laquelle des données vont être localisées, certaines sélectionnées pour leur ouverture en fonction de critères plus ou moins précis. Pour étudier le travail d'identification, je vais m'appuyer dans ce chapitre sur l'analyse du corpus d'entretiens, d'observations et de documents constitué dans les six terrains de mon enquête.

Contrairement à ce qu'affirment certains des acteurs à l'origine de l'*open data* que nous avons rencontrés dans le premier chapitre, les données brutes ne sont pas disponibles au sein des administrations, prêtes à leur ouverture immédiate. Ici, les données ne sont pas des entités à portée de main des responsables du projet *open data*, elles sont progressivement localisées, négociées et caractérisées en vue de leur ouverture. Nous verrons, dans un premier temps, que là aussi contrairement à certaines idées reçues, les données ne sont pas répertoriées dans des catalogues ou des inventaires dans lesquels les responsables de projet *open data* n'auraient qu'à faire leur sélection pour déclencher une démarche d'ouverture des

données. Il ressort de mon enquête que les inventaires de données existants se révèlent nécessairement partiels et localisés et ne suffisent pas à guider l'identification. Les responsables de projet *open data* doivent donc partir à la découverte des données produites par les services de l'organisation. Le dernier extrait d'entretien, conduit un mois après l'évènement, montre bien que le « jeu de rôle » n'était pas une simulation factice pour l'équipe en charge de l'Open data, mais prenait bien part au processus d'identification des données et qu'elle poursuivait plusieurs mois de travail à la recherche de nouvelles données à ouvrir. Nous verrons donc, dans un deuxième temps, que l'identification prend la forme d'une exploration, d'une découverte progressive des données lors de laquelle les responsables du projet *open data* vont arpenter les bureaux et rencontrer dans les services pour localiser et sélectionner les données à ouvrir. Ces explorations sont guidées par une multiplicité des pistes qui se présentent à eux au cours de l'identification, mais aussi par les objectifs qualitatifs et quantitatifs qui sont fixés aux responsables de projet *open data*. Au cours de ces explorations, les données sont progressivement caractérisées et sélectionnées. Cela ressort clairement du récit précédent, les participants à la réunion attribuent une catégorie, une description, des mots clés aux données. De leur côté, les responsables de projet *open data* soulignent sans cesse l'intérêt de certaines données. Nous verrons que c'est au terme d'une préfiguration des usages, d'une évaluation informelle du potentiel de réutilisation des données, que certaines sont caractérisées comme intéressantes. Cette réunion, qui officialise la création d'une « communauté » de l'*open data* régional, montre que des rôles et des responsabilités sont désignés à des agents qui renseignent l'équipe en charge de l'*open data* sur les pratiques de production de données et étendent le travail d'identification au-delà des acteurs qu'ils ont déjà sollicités. Dans un troisième temps, nous verrons que l'identification attribue des rôles et crée de nouveaux réseaux dédiés à la circulation des données au sein de l'organisation. Enfin, un dernier point ressort de cette réunion, mais je ne vais pas me concentrer dessus ici : l'identification révèle des difficultés et des contraintes qui préviennent leur ouverture. J'aborderai ce point en détail dans le chapitre suivant afin de me concentrer sur l'objet de ce chapitre : comprendre comment des données sont identifiées et sélectionnées parmi l'ensemble des informations produites par les administrations.

L'utopie de l'inventaire exhaustif

Dans la plupart des projets étudiés, les personnes en charge de la mise en œuvre des projets d'*open data* ne sont pas productrices de données. Elles découvrent parfois même le fonctionnement de l'institution dont elles sont en charge d'ouvrir les données. Elles se

trouvent donc, au démarrage du projet, en situation d'exploration, enquêtant à la recherche de données candidates à l'ouverture. Cette exploration peut viser à déboucher sur un inventaire. Idéalement, aux yeux de certains, cet inventaire devrait prendre la forme d'un catalogue exhaustif qui recenserait, non seulement les données à ouvrir, mais « toutes » les données produites par l'institution. Par exemple, la Sunlight Foundation, dans ses guidelines pour une politique d'*open data*, réclame la production d'inventaires complets des données publiques et exige leurs publications : « *Government bodies often do not know what information they have. Open data policies should require a full public listing of government information*." À partir de cet inventaire, les équipes en charge de projets d'*open data* pourraient concentrer leurs efforts sur certaines données en fonction des priorités de l'organisation, des demandes des usagers ou encore de l'actualité. Or, dans mon enquête, les responsables de projet *open data* qui pensaient s'appuyer sur un inventaire pour sélectionner les données à ouvrir se rendent compte que sa production n'a jamais été entreprise. En l'absence d'un tel outil, ils donc doivent partir à la recherche d'entités dont les contours et la situation dans l'organisation ne font l'objet d'aucune vision centrale.

> On aimerait bien dans l'idéal avoir une liste absolue, exhaustive de toutes les données que produit chaque service de chaque collectivité publique, toutes les données qu'ils manipulent finalement. […] S'il pouvait y avoir un annuaire complet, c'est un peu utopique, mais ça serait génial. Parmi toutes ces données, on identifierait celles qui peuvent être ouvertes, celles qui ne peuvent pas parce qu'il y a des restrictions, les données personnelles, les sensibles, etc. Avoir une liste, ça permettrait de se dire « voilà, celles-là on pourrait les ouvrir, celles-là on peut pas parce qu'il y a des problèmes techniques, celles-là on n'a pas envie »
> (L.K., responsable projet ***open data***, Rennes)

> Moi c'est la politique de l'iceberg, on sait très bien qu'on ne va jamais sortir tout le glaçon de l'eau, mais au moins il faut qu'on sache ce qu'il y a sous l'eau. C'est ce que je dis à tout le monde, je leur dis « écoutez c'est hyper net qu'on va pas tout sortir. Par contre, je veux savoir tout ce qu'il y a. » […] Parce que sinon tu es en mode boite noire, tu sors plein de trucs, tu ne sais pas si c'est les bons trucs ou si tu devrais plus t'acharner à sortir ça au lieu de sortir des trucs simples.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

Dans les premiers temps du projet, la production de données constitue une « boite noire » pour les responsables de projet *open data*. Même s'ils constatent rapidement que l'effort de constitution d'un tel inventaire exhaustif n'a jamais été entrepris, les responsables de projet

open data peuvent toutefois s'appuyer sur des inventaires partiels qui sont constitués par les directions des systèmes d'information (DSI). Mais toutes les directions n'ont pas recours à ses services et les données ne sont pas toujours stockées dans les bases de données qu'elle gère.

> En fait, là on est rattaché à la DSI. Et donc, il y a des serveurs de données qui sont mutualisés entre plusieurs services. Ce qui n'est pas le cas de tous les services, il n'y a pas une grosse base de données qui comporte toutes les données de la ville et dans cette base de données, moi, j'y ai accès et donc je vois ce qu'il y a, à qui ça appartient. Donc, ça m'a donné une idée de quel service contacter. Quand je rencontre le service, par exemple, je vais voir le génie urbain, je les rencontre parce que je veux les trottoirs. Et moi, je sais tout ce qu'ils ont.
> (Un chef de projet *open data*)

Lorsqu'un inventaire a été entrepris par la DSI, cela peut révéler aux responsables de projet *open data* les données qui sont gérées par ce service. Ils peuvent alors réclamer l'ouverture en disposant d'informations précises sur les données que produisent les agents. Dans certains services, les gestionnaires des systèmes d'information ont déjà constitué un inventaire qui recense les données produites par le service. Dans le cas de la fiche des données du génie urbain ci-dessous (figure 23), il est amendé dans les trois dernières colonnes pour déterminer les conditions de l'ouverture des données.

Données Patrimoine Exploitation DGU

Objet	Producteur	Format	Disponibilité	Année de dernière mise à jour	Diffusable	Commentaires diffusion	Editeur MLD
Cadastre	DGI	ESRI	OK	2010	Voir SIG		DGI/SIG
Cadastre communes limitrophes	DGI	ESRI	OK	2010	Voir SIG		DGI/SIG
Plan Ville	SIG	ESRI	OK	janvier 2010	Voir SIG		
Photo aérienne	SIG	ESRI	OK	2009	Voir SIG		
Filaire et adresse	DEP	ESRI	OK	MAJ continue	Voir DEP		
Stations	CAM	ESRI	OK	2009	Voir CAM		
Réseau	CAM	ESRI	OK	juin 2010	Voir CAM		
Eclairage public	EP	ESRI	OK	En continu. 90%	Oui	Les mats et lanternes, sans descriptif	Imagis
Réseaux d'eau pluviale	Hydraulique urbaine	Dwg ESRI	OK	Existant 30%, 100% en 2013.	Non	Données pas complètes	Imagis
Ruisseau	Hydraulique urbaine	Dwg ESRI	OK		Oui		HU
Bassins versants	Hydraulique urbaine	Dwg ESRI	OK		Oui		HU
Assainissement	CAM	ESRI	OK	2010	Voir CAM		Imagis
Eau potable	CAM	ESRI	OK	2007. Avoir les feeders.	Voir CAM		Imagis
Eau brute	BRL	ESRI	OK	2008 (DIPAN)	Non		
Points bas et htes eaux	HU	Dwg ESRI	A valider		Non	Pas validées	Pas validées
Comptages	DO	ESRI	OK	Sept 2010.	Non	Sensible	
Sens de circulation	DO	ESRI	OK	MAJ continue	Oui	Avec filaire de voie (DEP)	Imagis

Figure 23. Extrait de l'inventaire des données du service du génie urbain d'une ville. Les trois dernières colonnes à droite (en vert) ont été rajoutées pour le projet *open data*.

Ici, l'inventaire a servi à qualifier les données en fonction de plusieurs critères qui permettent aux agents du service du génie urbain de déterminer si une donnée peut être diffusée ou non. En effet, toutes les données inventoriées ne sont pas toutes diffusables en tant que telles. D'abord, les agents du service ne peuvent pas décider seuls de l'ouverture de certaines données, c'est le cas pour les six premiers jeux de données. Pour ouvrir ces données qui font partie du « patrimoine » du service, il faudra que les agents en charge de l'*open data* obtiennent de nouvelles validations. D'autres données ne sont pas considérées comme diffusables pour des raisons variées telles que la sensibilité ou l'absence de validation. Je ne m'étends pas sur ce point, j'aurai l'occasion d'aborder en détail dans le prochain chapitre les raisons pour lesquelles certaines données ne sont pas ouvertes. Enfin, dans d'autres cas, l'ouverture est conditionnée à la sélection de certaines informations : « les mats et lanternes, sans descriptif » ou encore, hors de l'extrait ci-dessus, uniquement les arrêtés signés ou seulement les informations relatives à l'accessibilité. Toutes les données recensées ne sont pas « ouvrables » telles quelles. Dans de nombreux cas, il faudra discuter des conditions de leur ouverture.

Dans d'autres configurations, l'ouverture des données peut servir à relancer des projets d'inventaire ou de cartographie des systèmes d'information régulièrement conduits par les DSI. Lors du comité de pilotage du projet d'*open data* d'une entreprise, la mise en place d'un inventaire des données a été proposée pour accompagner le projet d'*open data*. S'inspirant d'une expérience locale de cartographie dans l'entreprise et des préconisations de l'administration américaine dans un mémo de 2013[92], un agent de la DSI proposait lors

[92] Le 9 mai 2013, le président Obama a signé un executive order accompagné d'un mémorandum intitulé « Open Data Policy-Managing Information as an Asset » qui exige que chaque agence fédérale complète leur inventaire des ressources informationnelles, une obligation légale, pour inclure les données utilisées dans ses systèmes d'information. Le mémo de la Maison Blanche conçoit l'inventaire comme une démarche progressive dont l'exhaustivité est un objectif lointain : « *The inventory will be built out over time, with the ultimate goal of including all agency datasets, to the extent practicable. The inventory will indicate, as appropriate, if the agency has determined that the individual datasets may be made publicly available (i.e., release is permitted by law, subject to all privacy, confidentiality, security, and other valid requirements) and whether they are currently available to the public*. »
in White House (2013, mai 9). Obama Administration Releases Historic Open Data Rules to Enhance Government Efficiency and Fuel Economic Growth. Consulté 25 mars 2014, à l'adresse

de la réunion de rencontrer les responsables des principales applications « métier » pour évaluer l'opportunité de l'ouverture de certaines données. Selon lui, l'*open data* constitue « un bon alibi » pour déployer une démarche d'inventaire et travailler sur la qualité des données.

> Un des leviers que permet l'open data coté SI [systèmes d'information], c'est d'expliquer en quoi les données sont utiles. […] On commence à regarder la non-qualité des données, on voit des chaines de traitement de l'info avec des processus de ressaisie, on se rend compte que finalement ça coute potentiellement une fortune. […] La problématique qu'on a aujourd'hui c'est que c'est compliqué, on ne peut pas décréter comme ça de cartographier les données, on est tout de suite face à une problématique de retour sur investissement. Et je trouve que l'open data, c'est un bon alibi pour adosser cette démarche à la constitution de catalogues de données. Je pense que ça permet aussi d'apprendre l'argumentaire de l'open data en interne.
> (F.T., responsable des systèmes d'information d'une des branches de l'entreprise)

De ce point de vue, l'inventaire constitue une fin et non seulement un moyen pour identifier les données à ouvrir. Il permet aussi de présenter aux agents le projet d'*open data* et, éventuellement, de les convaincre d'ouvrir leurs données. Lors de la réunion du comité de pilotage, cet agent de la DSI a fait circuler un document qui décrit les objectifs et la procédure de la démarche d'inventaire qu'il propose. Il y formule une proposition de critères qui pourront figurer dans le tableau d'inventaire (figure 24). Même si « *open data* » se trouve dans l'intitulé, le document évoque plutôt une démarche de « partage des données » pour intégrer la possibilité de diffuser les données en interne et auprès de publics ciblés. Dans la lignée des projets de la DSI, l'inventaire cible d'abord le « public interne » pour permettre aux employés de l'entreprise de localiser et d'utiliser les données produites par d'autres services. L'inventaire est aussi conçu comme un outil pour la DSI pour nourrir ses projets de cartographie des systèmes d'information et d'amélioration de la qualité des données. L'ouverture des données est abordée dans la deuxième partie du document, envisagée comme une seconde phase de l'inventaire.

---

http://www.whitehouse.gov/the-press-office/2013/05/09/obama-administration-releases-historic-open-data-rules-enhance-government

Direction de la communication
Open Data : inventaire des données

Maintien d'un inventaire des données de l'organisation
La première phase d'une démarche de partage des données repose sur la création et le maintien d'un inventaire des données sous forme de tableau. Dans l'idéal, cet inventaire regroupera l'ensemble des jeux de données et APIs qui sont présents dans l'organisation. Même les données qui n'ont pas vocation à être partagées peuvent être intégrées dans cet inventaire. L'objectif de cette première phase est d'avoir un aperçu général du patrimoine informationnel de l'organisation. À ce stade, la qualité des données ne doit pas être un critère déterminant de l'ajout dans l'inventaire, bien qu'il soit nécessaire de renseigner le document sur la qualité des données.

Ce tableau peut inclure les entrées suivantes :

Présentation générale
Titre
Code d'identification
Description
Tags

Responsable
Service en charge du maintien
Personne en charge du maintien
Mail de la personne en charge du maintien

Présentation technique
Support d'exposition (fichier ou API)
Lien d'accès
Format des données

Date
Date de création initiale
Date de mise à jour
Fréquence de mise à jour
Période couverte

Qualité des données
Données brutes
Données complètes
Données à jour
Données structurées

Cette première étape vise à cartographier les données de l'organisation en identifiant les données de faible qualité. Cette étape est fondamentale pour l'établissement de la feuille de route Open Data, même s'il est difficile de compléter l'ensemble du tableau. À la rigueur, le tableau pourra être complété en plusieurs vagues à mesure que l'investigation sur les données progresse.

Cette étape est aussi l'occasion de convaincre le public interne sur l'enjeu de cartographier les données. C'est un levier pour mobiliser le soutien des équipes et ne pas les laisser avec un sentiment de dépossession. La participation la plus large possible permettra d'enrichir toujours plus ce tableau.

Dans un second temps, les personnes en charge du SI pourront indiquer la localisation de chacun des jeux de données sur le schéma d'urbanisation du SI afin d'en avoir une représentation visuelle et d'identifier les enjeux métier et business du partage de données.

À terme, cet inventaire peut aussi se transformer en outils de recherche de données pour le public interne. Avec un système de gestion de base de données, cet inventaire pourra être questionné sur la base de critères pertinents dans le cadre d'une démarche de partage des données.

Évaluation de la pertinence d'un partage des données

La deuxième phase d'une démarche de partage des données consiste à évaluer la pertinence d'ouvrir les jeux de données et APIs catalogués sur la base des trois questions auxquelles on associe une note de 0 à 4 selon la pertinence de l'ouverture (4 étant la note la plus importante) :
– Les données ont-elles une opportunité (business, communication) ?
– Les données ont-elles un intérêt pour l'écosystème Open Data ?
– Les données sont-elles de qualité (brutes, complètes, à jour, structurées) ?

Une dernière question discriminante porte sur les enjeux stratégiques des données. Cette question sera volontairement traitée à la fin de l'évaluation afin de ne pas tronquer dès le départ l'évaluation des données.
– Le partage des données représente-t-il une menace (concurrence, confidentialité) ?

Figure 24. Ébauche d'un inventaire des données. Document distribué lors d'une réunion de pilotage du projet *open data* d'une entreprise.

Tout comme pour les deux chefs de projet *open data* évoqués plus haut, l'exhaustivité de cet inventaire est considérée comme un idéal. Prenant la forme d'une « investigation » progressive, conduite en plusieurs phases pour obtenir un « aperçu du patrimoine informationnel », la démarche proposée considère l'inventaire comme une découverte des données, une exploration progressive qui va permettre de les qualifier et d'identifier le réseau sociotechnique qui se tisse autour d'elles. Pour qualifier et catégoriser les données, une première liste de critères à remplir lors de l'inventaire est proposée dans le document. Même si la définition des critères n'est ni détaillée ni arrêtée, on peut déjà comprendre que l'inventaire pourrait attribuer des responsabilités nouvelles aux services et aux agents qui gèrent les données. Il pourrait aussi stabiliser certaines caractéristiques des données en demandant aux agents de renseigner un format, une fréquence de mise à jour et leur qualité à travers quatre critères particulièrement intéressants pour mon enquête (brut, complet, à

jour, structuré). Je n'ai malheureusement pas pu suivre l'évolution de ce projet d'inventaire pour observer si ces critères de qualité ont pu être détaillés et appliqués. En tout cas, la qualité des données ressort comme une préoccupation essentielle de l'entreprise dans ce document. À sa lecture, l'inventaire semble tout autant être l'opérateur d'une amélioration de la qualité des données que de leur ouverture. Cette dernière est déterminée, dans une deuxième phase de l'inventaire, à travers quatre questions évaluant l'opportunité, l'intérêt, la qualité et les risques de l'ouverture. Selon cette procédure qui, rappelons-le, n'était qu'une proposition à ce stade, les données ne peuvent être ouvertes qu'après un passage au crible des risques et des opportunités de leur ouverture.

Les responsables de projet d'*open data* n'ont donc pas à leur disposition des inventaires exhaustifs ou des catalogues à partir desquels ils pourraient sélectionner les données qui iront remplir les portails *open data*. Toutefois, certains services peuvent produire des inventaires partiels et localisés, généralement avec l'aide des directions des systèmes d'information. Dans d'autres cas, des projets d'*open data* donnent effectivement lieu à la réalisation d'inventaires. Dans une démarche d'exploration et d'enquête conçue comme progressive et non exhaustive, l'inventaire produit une première qualification des données, leur définit de nouvelles caractéristiques et attribue des responsabilités à des individus en charge de leur gestion. Il participe de la prise de connaissance des données et de leurs réseaux sociotechniques qui accompagne les projets d'*open data*. Cette démarche d'inventaire est plus généralement envisagée sur le long terme dans le cadre de grands projets de transformation des systèmes d'information. Elle intervient rarement au lancement d'un projet *open data* où l'identification prend plutôt la forme d'explorations que d'un recensement systématique des données.

L'exploration de l'organisation

Au commencement d'un projet d'*open data*, l'équipe qui en a la responsabilité explore généralement l'organisation à la recherche de données pouvant être ouvertes, parcourant l'organigramme et sillonnant les bureaux. Ils peuvent démarrer leurs explorations par une tournée des responsables de service pour présenter la démarche et obtenir l'ouverture des premiers jeux de données. En sensibilisant la hiérarchie à l'ouverture des données, ils pourront identifier des interlocuteurs et rapidement les premières données pouvant être ouvertes.

> [L'identification] s'est faite par beaucoup de rencontres des services, on a commencé par des réunions de présentation du projet à l'ensemble des directeurs, des chefs de service. Ensuite [H.B] est allé voir chaque service pour voir avec eux ce qui était disponible, comment ça pouvait être mis à disposition, les freins qu'il pouvait y avoir à publier telle ou telle donnée parce que c'est bien chaque service qui est responsable de ces données, propriétaires de ces données et en aucun cas la DSI.
> (G.H., Directeur des systèmes d'information, Montpellier)

L'identification se réalise souvent de manière graduelle : les chefs de service mettent en contact l'équipe de l'*open data* avec des producteurs de données qui vont eux aussi suggérer de nouveaux interlocuteurs. Par exemple, dans le cas de l'ouverture des données à Montpellier, le chef de projet présentait d'abord la démarche d'ouverture des données aux agents et les invitait à suggérer des données à ouvrir. Chaque rencontre, lors de laquelle les agents expliquaient leurs missions et présentaient les informations qu'ils produisent, donnait lieu à la découverte de nouvelles données pouvant potentiellement être ouvertes.

> Les six premiers mois, j'allais service après service, rencontrer, expliquer la démarche et demander « qu'est-ce que vous avez comme données ? Sur quoi on pourrait travailler ? Comment ça se passe ? Quelle est la licence ? » […] C'est quand on a identifié une donnée dans un service, on va dans le service et on demande ce qu'ils ont d'autre comme données. Et c'est là que la discussion s'amorce « Nous, on travaille sur ça ça ça. Moi je fais ça, ça pourrait être intéressant, ça, c'est pas intéressant ». Généralement, le contact sur le service se fait via une ou deux entrées. Et c'est après que je découvre qu'il y en a cinq, dix, quinze données en gestion dans le service.
> (H.B., Chef de projet ***open data***, Montpellier)

Le processus d'identification se nourrit lui-même, il fait émerger de nouvelles pistes au fur et à mesure de l'enquête et fait découvrir les méandres de l'institution aux personnes en charge du projet d'*open data*. Ce qui semble partagé, dans les cas étudiés, c'est le caractère progressif de l'enquête qui amène les agents à découvrir des jeux de données autant qu'à déplier l'organisation elle-même. C'est d'autant plus prégnant à l'échelle d'une région ou d'un État où parcourir l'organigramme peut donner l'impression de s'engager dans un dédale.

> L'idée c'est que je pars des unités, ensuite direction et ensuite les services. Ensuite, dans les services, tu as les producteurs de données […] Et en gros, on descend,

> jusqu'au plus petit dénominateur commun pour qu'on identifie toutes les données quoi. Et, ce qui est fou, c'est qu'à chaque fois que je fais un rendez-vous, je suis reparti avec 30 rendez-vous. À partir de ces 30 rendez-vous, ils m'identifient cinq autres personnes qu'il faudrait que je voie. Donc en gros c'est exponentiel. […] Tu vas voir tout le monde, c'est long et c'est fastidieux.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

> J'ai procédé par structure, je prenais le ministère ou les établissements publics sous tutelle, certains sont particulièrement connus comme l'IGN, Meteo France, le BRGM. J'en ai découvert certains autres qui étaient particulièrement intéressants et qui pourraient être des acteurs assez importants de l'ouverture des données. Ensuite je prenais la structuration ministérielle : direction, service, bureau. […] Si je devais faire une roadmap de mon job, j'en ai au moins pour 10 ans si je voulais tendre à l'exhaustivité du sujet.
> (T.Y., un agent de la mission Etalab)

Tout comme l'idée d'un inventaire exhaustif, les responsables du projet considèrent une exploration intégrale de l'organisation comme utopique. Loin du modèle de l'ouverture « complète » des données publiques réclamé dans les principes de Sebastopol, les responsables du projet *open data* peuvent tenter d'obtenir l'ouverture d'un « échantillon », un aperçu des données produites par chaque service.

> La démarche, ça a été de recenser ce qui pouvait être donné rapidement parce qu'on avait toujours cette contrainte de temps. Et en même temps, petit à petit, je me suis aperçu qu'il fallait donner un échantillon de toutes nos missions dans les différents secteurs d'interventions du ministère. Que l'échantillon que l'on donne soit représentatif de l'activité, ça, c'est le deuxième objectif qu'on s'est assigné.
> (Q.H., Correspondant du réseau Etalab, ministère)

Dans d'autres cas, les responsables de projet *open data* peuvent guider leurs explorations en observant les données publiées par d'autres organisations ayant mis en place un projet d'*open data*. À défaut d'inventaire, ces ressources peuvent les mener à localiser des services et à envisager l'existence de données déjà ouvertes dans d'autres contextes.

> Je suis tout simplement allé sur les portails quand on s'est lancé, en France c'était Rennes et Paris en France et après, je suis allé sur le portail de New York, de Chicago et j'ai regardé les listes de données. […] On a fait une réunion et on a hiérarchisé : « essaie de travailler sur ça ça ça. Ça serait vraiment intéressant, quel service peut l'avoir ? »
> (H.B., Chef de projet ***open data***, Montpellier)

> Quand on est allé voir le service de prestations à la population, on savait déjà que ce qui serait facile pour eux à ouvrir et ce qui serait intéressant, c'étaient les données concernant les prénoms des enfants nés à Rennes. C'est une chose qui avait déjà été faite à Paris et à Nantes donc on savait que ça pouvait être intéressant de les sortir aussi.
> (L.K., responsable projet ***open data***, Rennes)

Une grande multiplicité de pistes se présente aux responsables de l'identification. Pour assurer le suivi des explorations et établir les priorités, des tableaux de bord équipent souvent le travail d'identification et permettent parfois aux responsables de l'identification de rendre des comptes de leurs explorations. Ce document peut même intervenir pour obtenir certaines données dont les producteurs refusent l'ouverture.

> J'ai travaillé à la réalisation de tableaux de bord par ministère, c'était un petit peu dur parce que je me suis fait tacler à plusieurs reprises. J'ai fait des codes couleur : vert, orange et rouge. Je n'ai pas tilté parce que ma culture de l'administration était nulle, mais les mecs ont vu le spectre de la RGPP [Revue Générale des Politiques Publiques] revenir en pleine tronche. Du coup, je me suis pris des scuds. Lors des réunions à Matignon avec les correspondants ***open data*** et les secrétaires généraux, ils avaient mon tableau de bord sous les yeux. Moi j'avais mis du vert sur ce qui était déjà accessible, de l'orange pour ce qui était accessible, pas encore en ligne, mais qu'ils n'y avaient pas de points très bloquants. Et le rouge c'était symbolique, c'étaient les jeux de données qui étaient accessibles, mais qui ne l'étaient pas pour des raisons politiques ou par un manque de volonté de l'administration. On se focalisait pas mal sur le rouge, ça nous a permis de lever pas mal de loups.
> (T.Y., un agent de la mission Etalab)

L'identification ne consiste pas à recueillir des entités disponibles, prêtes à être diffusées. Les responsables de projet *open data* doivent faire œuvre de persuasion et parfois avoir recours à la contrainte hiérarchique pour obtenir l'ouverture des données. Le tableau de bord, outil conçu au départ pour le suivi de l'identification, peut alors être perçu comme un objet d'évaluation voire de coercition lorsque l'équipe en charge du projet fait intervenir la hiérarchie. Au fur et à mesure des rencontres, l'identification révèle des résistances, des difficultés ou des contraintes. Dans le tableau de bord ci-dessous (figure 25), les données dont l'ouverture pose problème sont mises en attente dans un espace dédié en bleu.

Avancement des contacts Open Data.

Légende :

Rendez-vous effectué

Rendez-vous pris

Pas de rendez-vous

Nom	Service	Données ciblées	Contact	Rendez-vous (date)	Données libérées	Données en attente
	DGU	Relative à la voirie : chaussée & trottoirs, accidentologie, comptage		Ok (sept 2011)	Eclairage, ruisseaux, bassin, sens circ, filaire, passa piét, limite, ouvrage art, arbres	
	Plan climat	Climat, biodiversité		Ok (sept 2011)	non	
	Zoo	Inventaire +géoloc		Ok (sept 2011)	Inventaires animaux	
	Mandarine, annonce verte, OSM	Relatives à l'handicap		Ok (sept 2011)	Collaboration sur fichier ERP, sur problématique handicap	
	budget	Relatives aux finances		Oui effectué par (oct 2011)	dotation générale de l'état	
	DAP	Socio démo + espace urbain		Ok (sept 2011)	Pistes cyclables, parkings vélos, parking ouvrage	
	Parking	Parking surface		Ok (sept 2011)	Via prestataire extérieur emplacement horodateurs, places réservées, infraction places GIG	
				Non	Non	
				Non (sept 2011)	non	
				Non	Non	
	SIG	Plan ville		Ok (sept 2011)	Filaire des voies, cours d'eau, espaces verts, bâtiments, bassin versants, plan ville, quartiers, sous-quartiers	
	DAI	Handicap		Non	Non	

Figure 25. Tableau de bord d'un responsable de projet *open data* d'une ville.

Dans ce cas, lorsque le chargé de projet *open data* partait à la recherche des données sur les parkings, il a découvert de nouvelles données telles que les infractions aux places handicapées ou l'emplacement des horodateurs. De même pour le plan de la ville qui a révélé un jeu de données sur les cours d'eau. À l'inverse, un rendez-vous où sont ciblées les données sur le climat peut ne révéler aucun fichier pouvant être ouvert. À travers ce tableau de bord et les différents cas évoqués précédemment, on comprend que l'identification se nourrit d'explorations progressives et incertaines lors desquelles chaque rencontre peut mener à de nouvelles pistes comme déboucher sur des impasses. Mais l'exploration progressive de l'organisation n'est pas la seule méthode par laquelle les responsables de projet *open data* parviennent à identifier les données. L'identification peut aussi partir des usages de données ouvertes repérés dans d'autres organisations ou cibler des usagers potentiels.

Le ciblage des usages

Les explorations qui nourrissent le travail d'identification ne se résument pas à jalonner les services de l'organisation à la recherche de données. L'identification peut aussi cibler des usagers en fonction d'une demande de réutilisation avérée ou préfigurée. Sans même avoir

à sortir du périmètre de l'organisation, des publics internes peuvent faire part de leurs demandes et orienter les explorations. Dans un cas, le responsable du projet *open data* s'est rapproché de l'équipe éditoriale rattachée à sa direction, celle de la communication, pour cibler certaines données en fonction des sujets choisis en comité de rédaction.

> L'idée c'est que je sois inclus dans les comités de rédaction [des magazines et sites édités par la région] et que, par exemple, sur les sujets qui ont déjà été choisis, je vienne appuyer les journalistes en proposant une infographie sur la base de tel ou tel data set. Et inversement, parce que moi je récupère un data set exclusif super intéressant, je fais en sorte qu'il y ait sur le sujet de ces données-là un article beaucoup plus large.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

Dans d'autres cas, la désignation de données à ouvrir prend sa source dans des usages déjà existants et constatés. En regardant les applications et les services réutilisant les données ouvertes par d'autres organisations, les responsables de projet *open data* peuvent identifier des fichiers et des bases de données auxquels ils n'auraient pas pensé. Ces cas d'usage peuvent aussi prouver l'existence d'une demande de réutilisation des données et montrer le potentiel de l'*open data* dans les discussions avec les producteurs de données. Dans les premiers moments du projet d'*open data* de Montpellier, les idées de données à collecter étaient recueillies dans un document (figure 26). Les pistes de données étaient ensuite discutées au sein de l'équipe du projet pour établir des priorités dans l'identification et remplissaient le tableau de bord de l'identification.

Données	Où la trouver	Pourquoi faire		Qui fait déjà ça
Enregistrements audio de tous les audio-guides de la ville	OT	Appli complète à tous les sites décrit par audio guide		Zevisit.com
Relatives au climat	Voir mail Plan climat			
~~les documents budgétaires de la ville de Montpellier (Budgets primitif et supplémentaire, compte administratif…)~~				~~Rennes~~
~~(Animaux présents, localisation, plan du zoo…)~~	~~Zoo~~	~~aplication pour téléphone mobile afin de se localiser pendant une visite dans le zoo~~		
Parking	Ville de Mtp (Police) et agglo	Connaître les lieux les plus occupés (à éviter) et les plus libre (pour trouver rapidement sa place)		

Figure 26. Document « idées de données à collecter », ville de Montpellier.

Ici, les idées de données à ouvrir émergent en imaginant des usages et en repérant des services qui pourraient être créés. Ces services pourraient apparaitre si des développeurs venaient à les réutiliser, spontanément ou lors de l'appel à projets, un concours régulier organisé par la ville de Montpellier qui prend part au projet d'*open data* pour encourager à la réutilisation des données (j'aurai l'occasion d'aborder en détail ce type de dispositifs dans le dernier chapitre). Ce document et cette démarche d'exploration, qui part des usages pour localiser des données, ciblent un type bien précis d'usagers : les développeurs qui ont les compétences techniques suffisantes pour créer des applications mobiles et le profil entrepreneurial pour mener un tel projet. En effet, comme on l'a vu dans le premier chapitre, un des objectifs très généraux de l'ouverture des données porte sur l'innovation et la création de valeur économique qui pourrait découler de l'ouverture des données. On retrouve ce principe, par exemple, dans la charte du G8 : « *Freely-available government data can be used in innovative ways to create useful tools and products that help people navigate modern life more easily. Used in this way, open data are a catalyst for innovation in the private sector, supporting the creation of new markets, businesses, and jobs*.” Pour certains projets d'*open data*, la réutilisation des données figure même dans les indicateurs de performance de l'ouverture

des données, c'est le cas de la mission Etalab qui est évaluée en fonction du nombre de réutilisations répertoriées sur data.gouv.fr (figure 21).

Pour favoriser la réutilisation et cibler des publics capables de créer des services à partir des données ouvertes, des responsables de projet *open data* peuvent sélectionnant les données à ouvrir prioritairement en appréciant la valeur économique ou le potentiel de réutilisation. Dans un cas, le responsable de l'identification s'est rapproché d'amis diplômés d'une école d'ingénieurs, dont certains sont développeurs ou porteurs de projets, pour déceler des problèmes et tenter d'identifier les données qui pourraient servir à les résoudre à travers des applications et des services.

> Je parlais beaucoup avec les gens de mon entourage, des anciens potes de promo de l'école, pour savoir « qu'est-ce qui vous intéresserait ? Quelles données vous plairaient ? » […] Je prenais la position suivante « qu'est ce qu'aujourd'hui potentiellement une administration pourrait t'apporter comme données ? » Au moins, j'essaie de partir des problèmes identifiés par des amis. Et en quoi on peut trouver des solutions à travers le spectre de l'administration, c'est ça l'approche que j'avais.
> (T.Y., un agent de la mission Etalab)

Les développeurs et les porteurs de projet sont donc souvent une cible prioritaire du travail d'identification. Mais, en présumant de compétences techniques avancées et en déterminant la sélection des données par le critère de la création de valeur économique, un tel ciblage peut donner une orientation bien particulière au projet d'*open data* et restreindre fortement le périmètre des données concerné par l'ouverture. Si le projet d'*open data* est conçu, pas uniquement pour soutenir l'innovation ou la croissance, mais pour atteindre aussi d'autres objectifs tels que la transparence de l'action publique ou la modernisation de l'administration, les responsables de l'identification doivent orienter leurs explorations vers d'autres données dont ils ne perçoivent pas nécessairement la valeur économique. Des données peuvent être ciblées pour l'intérêt qu'elles peuvent présenter pour des citoyens qui les consultent, sans nécessairement construire un service ou une application à partir d'elles.

> Quand on a travaillé sur les caméras de vidéos surveillance, on ne s'attend pas à ce qu'il y ait une appli qui soit créée, mais là on est sur la transparence de l'action publique. On montre qu'on donne l'emplacement de nos 116 caméras de surveillance, c'est quelque chose qui parle aux gens de suite. […] Quand tu fais de

> l'*open data*, tu dois à la fois satisfaire des développeurs geek qui eux en ont que faire de ta description de tes tarifs, de tes subventions, de ton budget. Ils veulent de la donnée brute, qu'ils peuvent attaquer, qu'ils peuvent injecter dans leur application. Et, d'un autre côté, tu dois satisfaire la transparence de ton action publique, quelque chose de plus vulgarisé pour que le commun des mortels comprenne. Il ne faut pas que travailler sur des données vraiment très brutes. Nous on se permet de travailler aussi sur une donnée qui a un peu moins de valeur pour les développeurs, mais qui a plus de valeurs en termes de compréhension de son territoire pour le citoyen.
> (H.B., Chef de projet *open data*, Montpellier)

Arrêtons-nous sur un point important de cet extrait : il y aurait donc des données plus brutes que d'autres, comme une sorte de gradient dans le caractère brut des données. Pour caractériser ces données plus brutes que d'autres, il évoque des données qui pourraient directement être « injectées » dans des applications, « attaquées » par les développeurs et leurs scripts, des données qui au fond seraient lisibles par les machines et réutilisables sans frictions. À l'opposé, les données des subventions, des tarifs ou la liste des caméras de surveillance seraient moins brutes car les développeurs ne peuvent pas les utiliser sans avoir à les transformer. Ces données seraient plus locales, mais aussi plus lisibles pour les usagers sans compétences techniques avancées et habitant dans la commune. Au-delà des questions qui sont posées par la demande de données brutes, que cela nous apprend-il sur le travail d'identification ? Les responsables de projet *open data* ne ciblent pas uniquement des profils types d'usagers tels que des développeurs, des porteurs de projet ou les rédacteurs du service de communication lorsqu'ils sélectionnent les données à ouvrir. Ils visent, et parfois font même des choix entre des données configurées pour l'utilisation par les machines et d'autres orientées pour leur utilisation par des humains. Ces choix, qui se posent notamment lors de l'identification, mais pas uniquement à ce stade du processus d'ouverture des données, comme nous le verrons plus loin, orientent les données vers des publics différents et façonnent progressivement les conditions de leur réutilisation. Je ne vais pas déplier plus encore cette question essentielle ici, car j'aurai l'occasion d'y revenir plus en détail dans le cinquième et le sixième chapitre lorsque j'évoquerai la question des formats de données.

À travers ces cas, on commence ici à cerner ce que les responsables du projet *open data* de la région Ile-de-France qualifiaient de donnée « intéressante » dans la réunion racontée en introduction de ce chapitre. Il s'agit vraisemblablement d'une donnée dont l'équipe en

charge de l'*open data* perçoit le potentiel de réutilisation et parvient à imaginer comment elle s'inscrira dans de nouveaux réseaux sociotechniques une fois son ouverture effectuée. Ces cas montrent aussi que l'ouverture des données ne constitue pas uniquement une politique de l'offre où des données sont identifiées et ouvertes sans prévoir leur éventuelle réutilisation. Le travail d'identification est aussi guidé par la demande et par l'évaluation du potentiel de réutilisation des données. Les responsables de l'identification, dans leurs explorations, tentent donc aussi de cibler des usages en sélectionnant des données qui ont été réclamées par un public, dont ils imaginent les cas d'usages ou qui, dans d'autres contextes, ont permis de créer des applications et des services.

L'organisation d'un réseau

Au-delà des pistes qui peuvent se présenter en inventoriant les données, en parcourant l'organisation ou en repérant d'éventuelles demandes de réutilisation des données, les explorations qui constituent le travail d'identification sont aussi déterminées par un objectif quantitatif. Très souvent, les actions de communication qui accompagnent les projets d'*open data* insistent sur le nombre de données publiées, parfois en considérant cet indicateur comme un benchmark pour se comparer à des institutions de taille similaire. Par exemple, Etalab avait comme objectif, au lancement de data.gouv.fr, de dépasser le nombre de jeux de données de data.gov. Aujourd'hui, nous l'avons vu précédemment, le projet de loi de finances comprend un objectif quantitatif du nombre de données publiées sur le portail data.gouv.fr et du nombre de réutilisations (figure 21). Dans des collectivités locales comme la ville de Montpellier, des objectifs quantitatifs guident aussi le travail d'identification.

> On a des objectifs. Quand je suis arrivé, on m'a dit « ça serait bien que d'ici la fin de l'année, on soit à soixante jeux de données et quatre applications créées. » Je ne vais pas trouver de la donnée pour arriver à soixante, pour faire du chiffre, mais ça nous donne un objectif, une ligne de conduite et moi, j'essaie de suivre.
> (H.B., Chef de projet ***open data***, Montpellier)

D'autres projets d'*open data* sont aussi guidés par des objectifs quantitatifs d'ouverture de données. Au début de la réunion Open Data Bootcamp de la région Ile-de-France, le directeur de la communication se fonde sur le nombre de données publiées pour affirmer que la région est « une des collectivités les plus dynamiques » en matière d'ouverture des données. À Rennes, le nombre de jeux de données publiées figure dans l'en-tête même du portail *open data* (figure 27).

Figure 27. En-tête du site data.rennes-metropole.fr (novembre 2013).

Les équipes en charge de l'identification, lorsqu'elles doivent répondre à des objectifs quantitatifs portant sur le nombre de données et de réutilisations, sont donc soumises à une double contrainte. D'une part, elles doivent trouver des données « intéressantes » dont elles reconnaissent un potentiel de réutilisation ou qui répondent à une demande, comme nous avons pu le voir précédemment. D'autre part, elles doivent aussi obtenir l'ouverture d'un grand nombre de jeux de données. Ces objectifs quantitatifs incitent les responsables de l'identification à maintenir leurs efforts et à travailler sans cesse à l'ouverture de nouvelles données.

Pour ancrer l'ouverture dans les réseaux sociotechniques de l'organisation, les responsables de projet *open data* tentent de stabiliser des circuits de diffusion, non seulement en désignant des types spécifiques de données à ouvrir, mais également en établissant des lieux dans l'institution et des personnes qui sont instituées en responsables des données et de leur circulation. Les équipes en charge de l'*open data* se sont souvent inspirées du réseau organisé par la mission Etalab pour assurer le recensement des données publiques. Dès la circulaire qui a créé la mission Etalab, il était prévu la mise en place d'un réseau de correspondants jouant le rôle de courroie de transmission des demandes de données dans les services et les bureaux de chaque ministère. Ce réseau s'est progressivement renforcé par la désignation d'interlocuteurs dans les services, en complément des correspondants *open data* de chaque ministère et la mise en place de réunions régulière. Trois organisations dans mon enquête ont formalisé un tel réseau (la région Ile-de-France, la ville de Paris, Etalab) qui repose sur la désignation de correspondants par les secrétaires généraux. Dans d'autres cas, ce réseau existe de manière informelle. Des interlocuteurs privilégiés relaient les demandes des responsables du projet sans que cela donne lieu à une désignation formelle.

> Je sais qu'il y a beaucoup de collectivités qui fonctionnent avec des référents officiels dans les services. Nous, on n'a pas du tout fonctionné comme ça. J'ai pris mon bâton de pèlerin, je suis allé faire service après service et ça c'est fait plutôt comme

> ça au contact humain. Donc, après dans les services, maintenant j'ai entre guillemets des référents, enfin, des contacts privilégiés. Donc, c'est eux que je vais voir quand j'ai besoin d'une donnée. C'est eux qui m'introduisent au directeur du service si besoin est. C'est eux qui me contactent quand il y a une mise à jour. Donc voilà, ça se fait. Ils n'ont pas le statut officiel, mais au final c'est un peu… le rôle qu'ils endossent.
> (H.B., Chef de projet *open data*, Montpellier)

Quand ce rôle fait l'objet d'une nomination, la désignation des correspondants par les secrétaires généraux ne répond pas à un critère établi à l'avance. Dans le cadre du projet *open data*, il est demandé que ce correspondant soit placé sous l'autorité du secrétaire général. Il faut aussi que cet interlocuteur se situe à un niveau hiérarchique ou dans un service qui travaille avec toutes les branches de l'organisation. Au sein de la ville de Paris, les correspondants, un public essentiellement masculin comme j'ai pu le constater lors des réunions auxquelles j'ai assisté, sont choisis au sein de services dits « support », informatique, DSI et communication essentiellement, ceux qui assistent les autres composantes de l'organisation dans la mise en œuvre des politiques publiques. Le degré d'implication des correspondants dans le projet varie selon leur engagement personnel en faveur de l'*open data* et du niveau de soutien dont dispose le projet dans l'organisation. Dans certains cas, l'ouverture des données devient une mission officielle des correspondants entrant parfois dans leur fiche de poste et dans le titre qui désigne leur rôle dans l'organisation. L'ouverture des données s'ancre donc à des niveaux différents dans le fonctionnement des services, selon les priorités qui sont établies pour le projet et le volume de travail qui est demandé aux correspondants. Dans certaines configurations, la structuration en réseau s'étend au sein même de la division que chapeaute le correspondant. Certains ont désigné des interlocuteurs et constitué eux aussi un réseau. Dans un ministère, le sous-réseau de correspondant s'est notamment matérialisé par la mise en place d'une lettre d'information qui tient les interlocuteurs au courant des dernières ouvertures de données et de l'évolution du contexte juridique national de l'*open data*.

La stabilisation du réseau repose parfois sur l'organisation régulière de réunions avec les correspondants. Ces dernières débutent généralement par une présentation des actions de l'équipe en charge du projet *open data*. : derniers développements du portail, bilan des concours de réutilisation, évolutions juridiques ou politiques de l'ouverture des données… Il y succède généralement un « tour de table » lors duquel les correspondants évoquent les

données qui pourraient être ouvertes, les mises à jour à effectuer et font part des difficultés auxquelles ils font face. Lors du tour de table, les responsables de projet *open data* encouragent les correspondants à ouvrir de nouvelles données pour tenir leurs objectifs quantitatifs d'ouverture. Dans l'entreprise que j'ai observée, l'ouverture de nouveaux jeux de données était rarement au cœur des discussions, ce sujet étant plutôt géré directement par les représentants de chaque branche. Mais, lors d'une réunion, le chef de projet a tenu à inscrire ce sujet à l'ordre du jour, car de nouvelles données n'avaient pas été publiées depuis longtemps. Il craignait que l'absence de mise à jour du portail affecte l'image de l'entreprise et remette en cause l'avenir du projet : « on a été moteur sur l'ouverture des données en France, mais je sens une perte de vitesse qu'il faut combler. Si on veut garder intact l'esprit du projet, il faut absolument qu'on ouvre de nouvelles données, sachant qu'on a tous un aperçu de ce qui pourrait être ouvert. » Lors du tour de table, il a incité les représentants de chaque branche à ouvrir des données même si l'intérêt de leur ouverture est contesté.

> Chef de projet ***open data*** : Si on peut faire un tour de table sur les projets d'ouverture et puis on va terminer avec David. Bon on commence avec Corinne.
> Corinne : On a toujours le temps réel avec la problématique juridique et stratégique. Ce n'est toujours pas tranché, je n'en sais pas plus aujourd'hui. Pierre voulait savoir si ça ne posait pas de soucis au niveau stratégique. […]
> Chef de projet ***open data*** : Et, en termes de fichiers un peu moins stratégiques, des fichiers Excel téléchargeables ?
> Corinne : après nous on a en perspective le prochain hackathon programmé en juin, mais finalement reporté à la rentrée, on va sortir tout un tas de données sur l'accessibilité
> Chef de projet ***open data*** : je comprends l'enjeu de communication, mais ça peut être intéressant d'ouvrir déjà même un jeu de données pas très stratégique pour garder le rythme, un jeu de données de transparence qui peut être intéressant.
> (Extrait de la retranscription d'une réunion de pilotage du projet ***open data*** d'une ville)

On le voit à travers cet extrait, ces tours de table prolongent le travail d'identification des données que conduisent les responsables de projet *open data*. Les responsables politiques et administratifs des projets *open data* se servent souvent de ces réunions pour contraindre les correspondants et les gestionnaires à ouvrir leurs données. En effet, la présence d'élus et de la hiérarchie administrative, renforcent les demandes des responsables du projet *open data* pour inciter voire contraindre les correspondants à ouvrir de nouvelles données.

D'autre part, la publicité des discussions entre chaque service et l'équipe en charge de l'*open data* renforce l'injonction à ouvrir qui est parfois adressée aux gestionnaires de données. Enfin, les présentations successives de chaque service révèlent les différences d'implication entre les correspondants, ceux qui ne rendent pas compte d'avancées jugées suffisantes peuvent se voir réprimandés publiquement. Les réunions du réseau des correspondants peuvent ainsi servir aux responsables de projet *open data* à contraindre les agents à ouvrir leurs données et à désigner des « mauvais élèves » parmi les correspondants.

> TY : Henri [Verdier] m'a demandé de mettre en place des comités réguliers avec tout le réseau, c'était vraiment extrêmement rare que je rassemble tous les correspondants ensemble. Là on a vraiment instauré une régularité avec des réunions tous les deux mois.
> SG : et quel est l'ordre du jour de ces comités ?
> TY : On parle de la plateforme, des prochaines fonctionnalités, des attentes des producteurs… On fait le point sur la mise à disposition de leurs données, s'il y a eu des décisions qui concernent la mise à disposition d'un certain nombre de leurs données. On essaie de savoir quel calendrier est prévu pour la mise à disposition de leurs données […]
> SG : et le fait de les faire en public avec tous les correspondants, ça change quelque chose ?
> TY : Bah ça montre les bons élèves et les mauvais élèves en fait. Tu vois le fait de confronter les différents acteurs, leur montrer qu'il y a certains ministères qui jouent le jeu à fond. On leur dit « mais pourquoi vous le faites pas quoi ? Vous avez l'air de gros cons, vous êtes tous seuls, vous ne balancez rien »
> (T.Y., un agent de la mission Etalab)

Par leur travail de mobilisation d'un réseau, qui peut parfois prendre la forme d'injonctions, les équipes en charge des projets d'*open data* tentent de distribuer la responsabilité de l'ouverture des données dans l'organisation. Pour ancrer le projet et stabiliser des circuits de diffusion des données, des lieux et des personnes responsables de l'ouverture des données sont désignés dans l'organisation. Tout comme on avait pu le voir précédemment avec Etalab, les projets d'*open data* « travaillent l'organisation » (Cochoy, Garel & de Terssac, 1998), ils redistribuent certaines cartes attribuent des rôles nouveaux et des responsabilités inédites.

Conclusion

Dans l'extrait que j'ai présenté en introduction, Daniel Kaplan disait des données publiques : « elles sont là. » À travers ce chapitre, nous avons pu voir que les données ne

sont pas des ressources disponibles, sous la main des chefs de projet d'*open data*. Au contraire, l'identification prend la forme d'une exploration, une découverte collective et négociée d'entités qui jusqu'alors étaient les outils quotidiens du travail des agents administratifs. On a pu voir que, dans cette exploration des méandres de l'administration, les responsables de projet d'*open data* ne disposaient pas d'inventaires exhaustifs à partir desquels ils auraient pu orienter leurs recherches. Et même si des projets d'inventaire de données sont parfois conduits dans les administrations, en particulier par les DSI, ces derniers se limitent à un périmètre circonscrit et sont conçus comme une investigation progressive qui permet une première qualification des données. Au lieu d'une récolte d'entités déjà reconnues et localisées, l'identification prend la forme d'une exploration progressive des services dans lesquels les responsables de projets d'*open data* découvrent et négocient l'ouverture des fichiers ou des bases de données que les agents gèrent au quotidien. Le travail d'identification se nourrit ainsi d'explorations progressives et incertaines de pistes de données à ouvrir. Nous avons vu que ces dernières peuvent émerger lors de rencontres avec les agents, par l'observation des usages des données ouvertes dans d'autres organisations ou encore par la mise en place d'un réseau de correspondants qui distribue le travail d'identification à travers les entités de l'organisation. L'identification ne permet donc pas uniquement de qualifier les données, de manière plus ou moins formelle, elle « travaille l'organisation » (Cochoy, Garel & de Terssac, 1998) en attribuant des rôles et des responsabilités inédites à des agents. En plus de la dénomination de correspondants qui intègrent l'ouverture dans leurs missions, l'identification implique bien souvent de désigner des responsables des données, des agents qui s'occupent de leur ouverture et de leur mise à jour, mais aussi parfois doivent s'assurer de leur qualité et répondre aux questions des usagers. Le travail de l'organisation qui accompagne les projets d'*open data* est une des raisons pour lesquelles ils sont souvent promus et portés par des services dédiés à la « modernisation » de l'État ou à la transformation des pratiques de l'administration comme le SGMAP auquel Etalab est rattaché.

Au fil de ce que j'ai décrit jusqu'ici, on comprend qu'au fur et à mesure des discussions avec les agents et de l'avancement de l'enquête, les équipes en charge de projets d'*open data* ne se contentent pas d'inventorier ou de découvrir les données qui pourront être ouvertes. Elles délimitent progressivement le périmètre des données qui sont concernées par l'ouverture. On a vu, dans le second chapitre notamment, que les lois d'accès à l'information

publique comme la loi CADA en France spécifiaient déjà une définition précise d'un document administratif ou d'une information publique. Mais les équipes en charge de l'*open data* formulent une demande bien précise : des données et si possible sous leur forme brute. Les responsables de projet d'*open data* ne rencontrent pas des données brutes disponibles et reconnues en tant que telles. Ils doivent régulièrement spécifier leur demande et préciser ce qui différencie les données brutes.

> Je leur ai dit au départ que la donnée statistique, c'est une donnée retravaillée. Je leur ai dit « tout fichier statistique qui est produit dans votre département, il est sous-tendu à la base par une base de données, un système d'information. » Je leur disais « pensez systèmes d'information, pensez pas fichier stats. Ce qui vient des systèmes d'information bruts de décoffrage, c'est ça que je veux »
> (T.Y., un agent de la mission Etalab)

Au-delà de leur caractère brut, les fichiers et les bases de données auxquels ils s'intéressent ne sont pas, de manière évidente, considérés comme des données par celles et ceux qui les produisent et les manipulent. Les agents en charge de l'identification inspectent les outils quotidiens du travail de l'administration et caractérisent de données ce qui est considéré dans les services comme des fichiers, des documents, des systèmes d'information, voire des chiffres dans une brochure.

> Nous, on explique aux services qu'on s'intéresse à des données, c'est-à-dire soit des fichiers métiers, soit des extractions de bases de données. On part vraiment de la compréhension de ce qu'ils produisent et on leur dit « voilà, la donnée pour nous c'est ça. » […] Typiquement, des gens vont nous présenter des trucs, moi je bosse avec ça, bah y'a des chiffres. Ils vont te présenter une brochure papier ou un rapport d'activité annuel. Il y a des chiffres, quelques tableaux qui se baladent, un ou deux camemberts. Nous on y va, on leur dit « si ça peut être de la donnée. » Si on reprend les chiffres, au final, on va avoir un tableau de cinquante lignes et ça fera sens.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

Dans une approche constructiviste, on pourrait ainsi dire que les données sont construites socialement, qu'elles n'existent que parce qu'elles sont catégorisées en tant que telles. Mais, comme le rappelle Lemieux (2012), une telle approche risque de rendre le réel artificiel en répandant l'idée selon laquelle la réalité pourrait ne pas être socialement construite. Elle risque aussi d'opérer, ce que Dorothy Pawluch et Steve Woolgar (1985) appellent « un charcutage ontologique » (*ontological gerrymandering*) en séparant une réalité objective et

réaliste, celle des chercheurs, et en renvoyant les représentations des acteurs dans le domaine de la construction sociale. Surtout, elle coupe les représentations de leurs fondements pratiques et matériels. Il n'y aurait pas de données matériellement, seulement une catégorie que les acteurs attribuent à des objets. La sociologie pragmatique a dépassé ces critiques en s'intéressant aux épreuves, aux moments lors desquelles les acteurs font l'expérience de l'instabilité du monde social et où les représentations sont confrontées à la matérialité du monde. En demandant aux acteurs de se prononcer sur ce qui différencie les données des autres objets, l'identification constitue une de ces épreuves. Pour échapper aux impasses théoriques du constructivisme, Latour (2015) a emprunté à Émile Souriau (Souriau, 2009) la notion d'instauration. Esthéticien, Souriau explique que le potier instaure plutôt qu'il construit la sculpture pour rendre compte de la manière dont la matière résiste au potier. Par rapport au constructivisme, la notion d'instauration offre l'avantage de prendre en compte la dimension matérielle et concrète d'un réel multiple qui participe, en résistant, en surprenant, en esquivant, à sa propre émergence.

> L'instauration possède l'insigne avantage de ne pas réutiliser tout le bagage métaphorique du constructivisme — qui serait pourtant d'un emploi facile et presque automatique dans le cas de l'œuvre si évidemment « construite » par l'artiste. Parler d'« instauration » c'est préparer l'esprit à engager la question de la modalité à l'envers exact du constructivisme. Dire, par exemple, qu'un fait est « construit » c'est inévitablement (et je suis bien payé pour le savoir) désigner à l'origine du vecteur le savant, selon le modèle du Dieu potier. Mais à l'inverse, dire d'une œuvre d'art qu'elle est « instaurée », c'est se préparer à faire du potier celui qui accueille, recueille, prépare, explore, invente — comme on invente un trésor — la forme de l'œuvre. […] si les faits sont construits, alors le savant les construit de rien ; ils ne sont eux-mêmes que de la boue saisie par le souffle divin. Mais s'ils sont instaurés par le savant ou par l'artiste, alors les faits comme les œuvres tiennent, résistent, obligent — et les humains, leurs auteurs, doivent se dévouer pour eux, ce qui ne veut pourtant pas dire qu'ils leurs servent de simple conduit. (Latour, 2015)

L'identification des données constitue donc un geste d'instauration à part entière ; instauration aussi bien technique qu'organisationnelle et politique. Les données ne sont pas révélées comme ouvrables ou découvertes parmi une masse d'autres données déjà disponibles. Elles sont instaurées en données à ouvrir au fil de l'enquête qui transforme le statut des fichiers, bases de données, documents, systèmes d'information qui équipent le travail des agents. Même lorsque l'exploration elle-même ne s'avère pas particulièrement complexe, cette instauration demeure essentielle dans le processus d'ouverture. Au sein

d'une organisation internationale où j'ai pu enquêter, seules certaines données très spécifiques avaient été identifiées dans l'inventaire réalisé pour l'*open data*. Ces données n'avaient pas été difficiles à désigner : elles émanaient du département qui était chargé du programme *open data* et faisaient déjà l'objet d'une publication en partie payante. Même si elles apparaissaient comme évidentes aux yeux des personnes responsables de ce projet d'*open data* très particulier, ce choix opérait lui aussi une instauration de ces données et ces données seulement comme données ouvertes. Ainsi, dans les documents qui ont circulé lors de la mise en place du projet, les informations qui ne figuraient pas dans l'inventaire n'étaient pas écartées du processus d'ouverture à l'issue de négociations particulières ou de choix politiques argumentés, mais simplement qualifiés de non-data sans autre forme de procès. Qu'elle résulte d'une exploration au long cours, ou d'une désignation fluide et quasi « naturelle », l'opération d'identification est donc générative. Elle engendre une certaine réalité (Law, 2009), un périmètre de données qui sont instaurées non seulement comme « ouvertes » (ouvrables, dans un premier temps) ou « brutes », mais aussi comme « données » tout court.

Jusqu'ici, j'ai essayé de montrer comment les responsables de projet d'*open data* mettaient en œuvre l'identification des données, à travers des méthodes d'inventaires, des explorations progressives dans les services de l'administration, en ciblant des usages ou en organisant un réseau. Mais, pour l'instant, j'ai mis de côté les difficultés et les oppositions qui peuvent émerger lors de l'identification des données. C'est l'objet du chapitre suivant dans lequel nous allons voir que la circulation des données demandée par les politiques d'*open data* engendre des frictions et des tensions.

Chapitre 4

Les frictions de l'identification : quelques « bonnes raisons organisationnelles » de ne pas ouvrir des données

Au début du chapitre précédent, j'évoquais un « jeu de rôle » d'identification de données qui s'est déroulé lors de l'Open Data Bootcamp de la région Ile-de-France, mais je ne me suis pas arrêté sur le contenu de la fiche d'identification qui était distribué aux participants. Quand on l'observe (figure 28), on y retrouve les principaux champs demandés par les portails *open data* : l'organisme concerné, le nom du jeu de données, la description, les mots-clefs, la désignation d'un producteur une personne dédiée à sa maintenance. Dans ce qui est conçu comme une simulation, la fiche semble inverser le processus habituel d'ouverture des données en demandant, dès la première identification, le remplissage des métadonnées (un aspect que nous aborderons en détail dans le dernier chapitre). Un champ en bas à droite de la fiche mérite notre attention : il permet aux participants de signaler les difficultés qu'il ou elle pourrait rencontrer lors de l'ouverture. Les organisateurs de l'évènement ont choisi de réduire les difficultés à quatre possibilités prenant la forme de cases à cocher. Selon ce document, les difficultés posées par l'ouverture peuvent être de quatre ordres : financière, juridique, stratégique et technique.

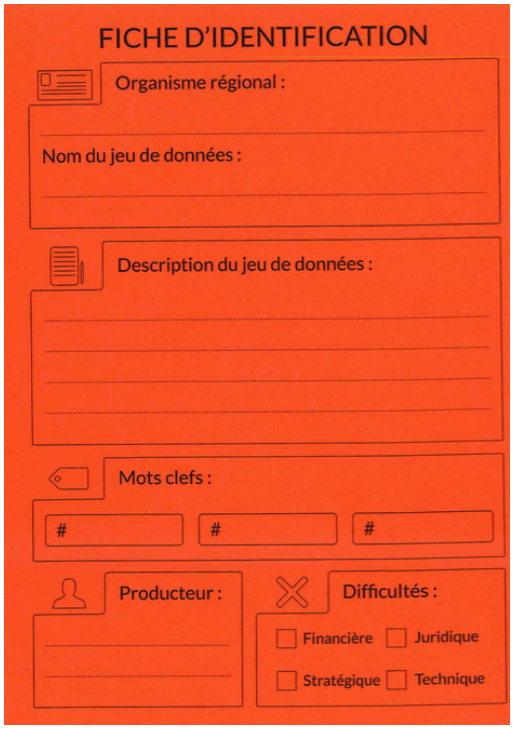


Figure 28. Fiche d'identification d'une base de données dynamique distribuée lors de l'Open Data Bootcamp de la région Ile-de-France.

Lorsque les fiches sont restituées aux animateurs, ces derniers lisent la fiche et annoncent les cases cochées. Ils n'entrent pas dans le détail des difficultés et des contraintes auxquels les gestionnaires auraient à faire face s'ils devaient ouvrir les données. Après avoir lu la description et convenu de la catégorie, ils collent la fiche et l'ajoutent dans la catégorie correspondante sur le tableau (figure 29).

Figure 29. Les fiches sont collées sur le tableau au fond de la salle. Image communiquée par un des organisateurs.

Cet évènement est un cas très singulier dans mon enquête, la simulation permet aux organisateurs d'écarter temporairement les difficultés qui peuvent survenir lors de l'ouverture des données. Pour certains acteurs évoqués dans le premier chapitre, l'ouverture immédiate et complète des données brutes qu'ils réclament doit, comme dans cette simulation, s'abstraire des réticences et des difficultés que peuvent exprimer les gestionnaires de données. C'est le cas de Hans Rosling, le médecin suédois que j'ai évoqué dans le troisième épisode, car il était mentionné par Tim Berners-Lee dans sa conférence TED. Selon lui, les gestionnaires de données souffrent d'une grave pathologie qui les accroche à leurs données. Lors d'une conférence à la Banque Mondiale en 2010, il lui a donné un nom qu'a repris Tim Berners-Lee : le DbHD, *Database Hugging Disorder* (figure 30). Il a félicité l'organisation internationale basée à Washington d'avoir éradiqué ce « syndrome » par son initiative d'*open data*.

> I come from the medical profession and we were sort of very concerned with H1N1, it was a severe disease we thought but we realized in fact that swine flu wasn't really that bad. I was albeit very concerned by DbHD. DbHD has been a more chronic disorder, it's Database Hugging Disorder. [rires et applaudissements dans la salle] And I congratulate deeply the World Bank from having cleared itself completely from this disorder.[93]
> (Hans Rosling lors d'une conférence à la Banque Mondiale en 2010)

Figure 30. Hans Rosling présente le *Data base Hugging disorder* lors d'une conférence à Washington à la Banque Mondiale en 2010.

[93] Youtube, « Mindset upgrade for a multipolar world, Washington: World Bank », https://www.youtube.com/watch?v=5OWhcrjxP-E, consulté le 1 février 2015.

La notion de DBHD, si elle peut être amusante, n'est pas très utile dans le cadre de ce travail. Plutôt que de considérer comme pathologique le comportement des agents, de balayer d'un revers de main les raisons qu'ils invoquent, essayons plutôt de comprendre ce qui les empêche d'ouvrir leurs données en prenant au sérieux leurs arguments. Ma démarche ici s'inspire de Garfinkel et Bittner (1967) lorsqu'ils ont essayé de comprendre les difficultés auxquelles ils faisaient face en exploitant des dossiers médicaux comme données d'une étude. Alors qu'il leur manquait des informations essentielles dans de nombreux dossiers pour mener à bien leur enquête telles que le lieu de naissance, la profession ou le suivi des échanges entre les patients et le personnel, il aurait été tentant pour eux de dénoncer la « mauvaise qualité » du remplissage de ces dossiers. Mais ils ont montré que les dossiers médicaux ont « de bonnes raisons organisationnelles » d'être « mal » remplis. Le personnel de l'hôpital collecte des informations pour ses propres missions dans un contexte organisationnel orienté vers un certain type d'actions, le soin des patients, et non la recherche en sciences sociales. Ils montrent que les dossiers médicaux sont des écrits organisationnellement situés. Faire passer les agents qui les produisent pour des malades à soigner revient à occulter le contexte organisationnel dans lequel ces données ont été produites.

Dans ce chapitre, je vais m'intéresser plus spécifiquement aux négociations qu'engagent les responsables de projet d'*open data* avec les agents administratifs lors de l'identification des données. Plutôt que de parler de « freins » ou d'« obstacles » à l'ouverture des données pour décrire les principales raisons de ne pas ouvrir des données, je reprends ici la notion de « frictions » proposée par Edwards (2010) pour décrire les difficultés qu'engendrent la circulation et le partage de données dans différentes disciplines scientifiques. En partant notamment de l'histoire de la climatologie, il souligne par ce terme le coût qu'implique la réutilisation de données qui ont été produites dans des configurations techniques et disciplinaires hétérogènes. Pour alimenter les modèles qui visent à mesurer le réchauffement climatique, il faut par exemple rassembler des enregistrements qui ont été effectués dans des lieux et des temps très différents, mais aussi à des fins et avec des moyens extrêmement variés. Edwards explique que ces données ne portent pas en elles les qualités suffisantes pour être utilisées par les scientifiques en question. La récolte des données nécessaires à la mise en calcul d'un climat global, à l'échelle de la terre entière, n'est donc

pas un processus transparent qui se résumerait à un échange de flux d'information brute. Au-delà de la climatologie, c'est la circulation des données qui génère irrémédiablement des frictions (Edwards et al, 2011).

> *Friction resists and impedes. At every interface between two surfaces, friction consumes energy, produces heat, and wears down moving parts. Edwards' metaphor of data friction describes what happens at the interfaces between data "surfaces": the points where data move between people, substrates, organizations, or machines—from one lab to another, from one discipline to another, from a sensor to a computer, or from one data format (such as Excel spreadsheets) to another (such as a custom-designed scientific database) (Edwards, 2010). Every movement of data across an interface comes at some cost in time, energy, and human attention. Every interface between groups and organizations, as well as between machines, represents a point of resistance where data can be garbled, misinterpreted, or lost. In social systems, data friction consumes energy and produces turbulence and heat—that is, conflicts, disagreements, and inexact, unruly processes.*

Les frictions ne sont donc pas un trouble à soigner, pour reprendre les termes médicaux du DbHD, mais une constante de la circulation des données. Pour mieux comprendre les contraintes organisationnelles et techniques auxquelles les agents font face lors de l'ouverture des données, je vais m'intéresser ici à quatre grandes sources de frictions que j'ai retrouvées dans la plupart des terrains et des services que j'ai explorés. La première porte sur les difficultés d'extraction. Quand les données sont gérées à travers un système d'information, il faut parvenir à les collecter à même leur espace de stockage et à les extirper de la nasse sociotechnique qui les entoure. Les gestionnaires de bases de données doivent alors entreprendre de nouvelles explorations dans les serveurs et les disques durs et concevoir les « moulinettes » qui permettent de défaire les données des systèmes d'information dans lesquelles elles sont produites. La qualité des données constitue une autre source de friction. En effet, dans une situation proche de Garfinkel et Bittner, les projets d'*open data* s'intéressent très souvent à des données qui n'ont pas été conçues au départ pour sortir des réseaux sociotechniques de l'organisation. Si elles étaient publiées telles quelles, ces données pourraient paraitre de mauvaise qualité alors même que leurs usagers en interne n'y voyaient rien à redire jusque là. Dans un troisième temps, nous verrons que des données peuvent être exclues du périmètre de l'ouverture lorsque les agents anticipent des risques liés à la sécurité qui pourraient survenir avec leur réutilisation. Enfin,

la question de la transparence peut prévenir l'ouverture des données, les agents ne disposant généralement pas du mandat pour libérer des données qui pourraient servir à l'opposition politique. Dans les cas où les données sont jugées « sensibles », de mauvaise qualité ou que leur ouverture pourrait faire courir un risque sur la carrière des agents, les données doivent passer à travers des circuits de validation plus ou moins formalisés et obtenir l'approbation de la hiérarchie pour être ouvertes.

L'extraction : des assemblages de données à défaire

Dans les négociations qui surviennent entre les responsables de projet *open data* et les gestionnaires de données, une première source de frictions porte sur l'accessibilité même des données. Là où certains voient les données comme une ressource disponible et prête à circuler, il faut en fait souvent désarticuler les assemblages dans lesquels elles sont prises. En effet, quand les données sont stockées dans des systèmes d'information qui sont les outils de travail quotidien des agents, elles ne prennent pas la forme de fichiers qui pourraient être transférés facilement comme c'est le cas avec les tableurs. D'un point de vue matériel, les informations y sont stockées, organisées et traitées dans des bases de données dont l'export est rarement prévu dans les fonctionnalités. Pour parvenir à exporter les données et éventuellement automatiser leur ouverture, les responsables de projet *open data* doivent généralement faire appel aux gestionnaires de bases de données. Mais pour mieux comprendre ces opérations, il nous faut d'abord faire un rapide retour en arrière historique et introduire le fonctionnement élémentaire d'une base de données.

Dans la majorité des situations étudiées ici, les bases de données sont fondées sur le modèle relationnel (Campbell-Kelly, 2007 ; Haigh, 2013 ; Dagiral & Peerbaye, 2013 ; Driscoll, 2012). Développé dans les années 1970, celui-ci a été pensé pour faciliter le traitement des informations en proposant des « vues utilisateurs » qui ne donnent pas à voir l'organisation physique des données (Castelle, 2013). Selon leur concepteur, les bases de données relationnelles doivent même « protéger » les usagers d'avoir à connaitre l'organisation physique des données (Codd 1970), la multiplication des vues devant permettre de faciliter la variété des usages (Dagiral & Peerbaye, 2013). Après avoir identifié des données et obtenu leur ouverture, les responsables de projets d'*open data* formulent leurs demandes d'ouverture en s'appuyant sur ces vues. Or, les informations y sont présentées dans un langage courant, celui employé au quotidien par les usagers dans leur travail. Dans la vue physique, les données prennent la forme de tables nommées dans un langage bien différent de celui de la vue utilisateurs. Dans la figure 31, le responsable de l'ouverture des données

d'une ville consignait les noms des tables dans lesquelles les informations étaient contenues afin d'assurer la continuité du projet en son absence. Dans les deux premières colonnes du document, il mettait en correspondance les données avec le nom des tables dans lesquelles elles sont enregistrées dans la base pour relier l'éclairage public à IMALUX.V_LANTERNE ou les caméras de surveillance à REF.cameras.

Données	Nom dans la base	Date de MEL	Date de la dernière MAJ	Fréquence de MAJ
Arbres Alignement	IMAVOI.ARBRE_ALIGN	19/01/11	24/01/2013	Annuelle
Arbres remarquables	IMAVOI.ARBRE_REMARQUABLE	13/04/11	24/01/2013	Annuelle
Caméra de surveillance	REF.cameras REF.Camera_vision	13/07/11	13/07/11	Annuelle
Eclairage public	IMALUX.V_LANTERNE	25/07/11	12/06/12	Mensuelle
Espaces boisé/vert	REF.parcjard_p	5/05/11 31/08/11	31/08/11	Annuelle
ERP	REF.Equipement	10/05/11	23/08/11	Hebdomadaire
Filaire	IMAVOI.V_Filaire	25/07/11	25/01/2013	Hebdomadaire
Jardin et parc	REF.parcjard_p	22/01/2011	22/01/2011	Annuelle
Monument Hist	PLU.AC1_monuments_historiques	26/02/11	12/09/11	Annuelle
Ouvrage d'art	IMAVOI.VOIRIE_OA	22/07/11	22/07/11	Mensuelle
PAE	PLU.pae	06/04/11	12/06/12	Mensuelle
Passages piétons	IMAVOI.VOIE_MTP_PP	04/07/11	12/06/12	Hebdomadaire
ZAC	PLU.zac_apres_2008	06/04/11	12/09/11	Mensuelle
ZAD		06/04/11		

Figure 31. Tableau « Mise à jour des données » issu du service *open data* d'une ville.

Pour faire le lien entre les tables et se repérer dans les méandres des bases de données, les responsables de projet *open data* font appel aux gestionnaires de bases de données. Rattachés aux services informatiques, les *database managers* ont pour mission de déployer et de maintenir les systèmes d'information de leur institution. Les gestionnaires de données traduisent les demandes d'ouverture des agents en une commande d'extraction qui vise précisément les tables concernées.

> Les gens du génie urbain travaillent sur la base directement, ils voient toutes les informations qu'ils ont. […] Mais, au génie urbain, les informations qu'ils voient, c'est habillé dans un logiciel. Au lieu d'avoir par exemple, « Idscore_voie. » dans ton logiciel, il va y avoir marqué « identifiant ». Et souvent c'est masqué, ces choses-là parce que c'est pas évident de travailler avec des termes barbares donc les logiciels habillent souvent cette base de données. C'est pour ça qu'on a besoin du support pour nous dire « mais en réalité, derrière ce que vous voyez, il y a ça, il y a ça, il y a ça. » […] En plus, des fois, tu peux avoir des informations sur une voie construites à

> partir de plusieurs tables différentes. Donc les données peuvent être réparties à des tas d'endroits.
> (Y.N., gestionnaire de bases de données, Montpellier)

Ici, le gestionnaire connait l'organisation de la base et parvient à recréer le lien entre la vue et la table. Mais, dans bien des situations, les vues utilisateurs construisent des liens entre des tables hétérogènes ou vont appeler des informations dont le nom de la table ne correspond pas à l'interface. Pour les gestionnaires, l'extraction demande de disposer du schéma de la base qui permet de reconstituer l'organisation physique des données puisque l'usager n'y a pas accès selon le modèle relationnel.

> On a besoin du schéma de la base […] Souvent c'est représenté sous forme de carrés qui représentent les différentes tables. Et tu traces des liens sur chaque table, tu mets l'identifiant de la table en fait ce qui permet de te dire que l'information à cet endroit-là est unique. Par exemple, pour l'état civil, tu as la liste des communes. Chaque commune est unique. L'information de la commune, tu ne vas pas la réécrire partout où tu l'utilises sinon tu dupliquerais à chaque fois l'information. Donc, tu vas avoir ta liste des communes avec ce qui t'intéresse… le nom de ta commune, la taille de la commune, le nombre d'habitants… Toi, quand tu donnes ta date de naissance et ta commune de naissance, c'est complètement inutile que toutes les informations sur la commune se retrouvent associées à ton nom. […] Donc, dans la table où il y a la liste des gens, au lieu de mettre le nom de la commune, on va mettre l'identifiant de la commune. L'identifiant est unique et lié à la table des communes, ça, c'est le fameux lien.
> (Y.N., gestionnaire de bases de données, Montpellier)

Lorsque le système d'information a été conçu par un prestataire externe, le schéma est rarement connu des gestionnaires de bases de données. Pour les éditeurs de logiciel, le schéma de la base peut être considéré comme un élément soumis à la propriété intellectuelle qu'ils refusent de communiquer. Ils peuvent aussi considérer que ce support doit faire l'objet d'une prestation facturée. Les gestionnaires de données se retrouvent donc à négocier avec les prestataires pour obtenir le schéma de la base.

> La plupart des sociétés sont extrêmement réfractaires à nous donner le schéma. Exemple, les actes qui reviennent de la préfecture […] pour savoir où est ce fameux acte, sous quel format il est stocké, à quel endroit. Et bien, la société n'a pas voulu nous donner ces informations-là. […] eux, ça leur permet de facturer en fait. Et donc ils vont te faire payer un bras.

(Y.N., gestionnaire de bases de données, Montpellier)

La bibliothèque, c'est un logiciel qu'ils utilisent depuis longtemps. Appels d'offres, genre il y a dix ans. C'est un logiciel produit par une boîte américaine qui a genre trois clients en France et qui ne s'en occupe pas beaucoup. C'est un logiciel complètement opaque, ils ne maitrisent pas du tout ce qu'il y a dedans, ce que fait le logiciel et ce qu'il peut sortir à la fin, ils ne peuvent pas trop y toucher. Ils n'ont pas d'accès direct à la base de données, ils sont obligés de passer par le formulaire que leur a gentiment fourni le prestataire. […] Comme on n'a pas accès à la base de données, on ne peut pas aller piocher dedans comme on veut. Donc on essaie de détourner ce système pour pouvoir quand même pouvoir aller interroger la base de données, mais c'est pas évident.
(L.K., responsable projet ***open data***, Rennes)

Certaines entreprises qui travaillent essentiellement avec des organisations publiques coopèrent plus facilement en cas de demandes d'ouvertures de données. C'est le cas d'une société évoquée par YN dont les logiciels prévoient des possibilités d'extraction et qui lui a fourni une assistance directe. Pour cette société, l'ouverture des données constitue un avantage compétitif, j'ai eu plusieurs occasions de la voir sponsoriser des conférences ou des salons dédiés à l'*open data*. Mais en l'absence du schéma de la base de données, les gestionnaires doivent enquêter pour comprendre le schéma qui lie les tables entre elles. Dans ce travail d'exploration, il leur faut fouiller au-delà des interfaces de visualisation, dans les entrailles des serveurs, à la racine même de la base de données. Leur enquête peut s'appuyer sur les producteurs de la base de données qui vont leur donner des indices sur son organisation physique. Dans un cas évoqué par YN, le service indiquait les dates des conseils municipaux ce qui lui a permis de retrouver la table et les fichiers concernés dans le serveur qui faisait la liaison avec la préfecture.

L'ouverture d'une base de données passe donc par cette première étape qui consiste à reconstituer son schéma, parfois au terme de négociations avec les prestataires ou d'une longue enquête pour comprendre l'organisation de la « vue physique. » Mais ce n'est pas souvent suffisant pour parvenir à effectivement ouvrir les données. Une fois le schéma reconstituer, il faut souvent développer un script informatique, une « moulinette », qui accomplit l'extraction de la base de données en rassemblant les informations éparpillées dans différentes tables et parfois plusieurs serveurs.

> On a été chercher dans toute leur configuration comment ils organisaient leurs données ? Comment ils stockaient ça ? Est-ce qu'ils stockaient ça dans une base de données ? Est-ce qu'ils stockaient ça dans des fichiers tout ça ? Il y a eu un gros gros travail de recherche là-dessus.
> (Y.N., gestionnaire de bases de données, Montpellier)

> En fait, ce qui est complexe, il faut bien comprendre c'est qu'au départ pour la plupart des systèmes des applications qu'on a chez nous qu'on a acheté, elles ne sont pas du tout conçues pour faire de l'*open data*. Donc, c'est compliqué. On est obligé, nous, de développer des moulinettes, des tas de choses pour pouvoir sortir des données proprement.
> (N.N, Technicien informatique, Keolis Rennes)

> C'est une requête SQL […] qui s'attaque à la base de données de surf parce que SPORT [le système qui fournit les vues utilisateurs], qui étrangement n'a pas été conçu pour sortir directement ces tableaux, ces tableaux de données agrégées. Il est conçu pour sortir des tableaux, le tableau mensuel d'un capteur, mais pas tous les capteurs en table dans un dans un seul tableau Excel. Bon, il a été conçu comme ça, mais en même temps, c'est pas grave parce qu'on sait le faire malgré tout via une requête SQL
> (V.C., analyste de données, ville de Paris)

Le développement de ces moulinettes dépend de la capacité des gestionnaires à se repérer dans l'organisation physique des données. Même si de grands principes se retrouvent dans la manière dont les données sont effectivement stockées sur les disques durs, la « vue physique » est toujours spécifique, comme le sont les manières d'organiser les placards personnels.

> Il faut te dire que rien n'est universel là-dedans. La manière dont tu ranges tes données, c'est comme la manière dont tu ranges tes chaussettes à la maison. Chacun peut les ranger de manière différente. On a tous le même placard, mais on les range tous de manières différentes.
> (Y.N., gestionnaire de bases de données, Montpellier)

Chaque outil d'extraction est donc toujours fait sur mesure, et le travail est d'autant plus complexe que les bases de données et les logiciels qui y donnent accès, voire les différentes versions d'un même logiciel, se sont accumulés au sein des institutions. C'est parfois une véritable foule d'instruments à laquelle les informaticiens ont affaire, dont l'exploration représente un coût très important.

> Ce qui peut aussi poser problème (...) c'est que chaque logiciel étant unique, les formats de données sont tous différents, et les schémas de répartitions des données sont tous différents, donc une procédure que tu as utilisée pour un logiciel ça ne sera pas la même pour un autre, même si tu reprends un peu les bases. Le corps est à peu près le même, mais les informations, elles, ne seront pas stockées de la même manière, donc il faudra refaire ce processus d'analyse pour chaque base de données différente. Et des bases de données, on doit en avoir peut-être au moins cinquante différentes. Donc, c'est extrêmement long d'extraire ces données-là. À la mairie, on a des données depuis plus de trente ans, qui en plus sont arrivés à l'époque sur les grands systèmes IBM qui sont différents des systèmes Windows, qui sont différents des systèmes Linux. On a à peu près de tout à la mairie. Du coup, c'est très compliqué d'extraire quelque chose de précis. (H.B., Chef de projet ***open data***, Montpellier)

Pour mettre en œuvre une politique d'*open data*, il faut donc être capable de récolter les données à même leur espace de stockage. Quand ce dernier prend la forme d'une base de donnée ou plus généralement d'un système d'information, les données doivent être extirpées d'une véritable nasse sociotechnique dont l'épaisseur se mesure à la complexité des explorations, des bricolages et des « moulinettes » abordés précédemment. L'ouverture des données amène ainsi les responsables de système d'information, les gestionnaires de données et les chefs de projet *open data* à repenser les conditions de leur « souveraineté » par rapport à leurs prestataires[94]. En l'absence de dispositions juridiques précises, les bricolages peuvent être assimilés à des détournements, voire à des ruptures contractuelles. Ce point est essentiel pour remettre en question l'idée selon laquelle les données publiques seraient des ressources dormantes qui ne demanderaient qu'à être libérées pour être exploitées. Cette vision de la donnée comme une « commodity » (Ribes & Jackson, 2013), une marchandise qui circulerait de manière fluide, est mise à mal à la fois par le coût et les ajustements, voire les bricolages, qui constituent le travail d'extraction. D'un certain point de vue, le travail d'extraction consiste donc pour les institutions à reprendre la main sur les données, en désarticulant de manière très concrète les assemblages sociotechniques qui les lient à certaines entreprises privées. Plus généralement, cette désarticulation donne à voir sous un angle très pratique, le feuilletage des infrastructures informationnelles. Comme

[94] A Montpellier, les prestataires sont désormais contraints de fournir le schéma de la base de données avant la signature du contrat. La ville de Paris a entrepris une démarche similaire et un groupe de travail de l'association OpenDataFrance élabore des dispositions contractuelles pour contraindre les prestataires à garantir l'extraction des données.

l'ont montré Star et Ruhleder (Star & Ruhleder, 1996), toute infrastructure repose sur une autre et est prise dans des jeux d'interdépendance complexes qui rendent délicate toute opération qui viserait à la singulariser. Ce n'est généralement qu'au terme d'un couteux travail que les données sont extraites, isolées des bases qui les ordonnaient et assuraient leur accessibilité ordinaire, afin d'être déplacées et inscrites dans un nouvel assemblage, dédié à leur ouverture.

La qualité : des données qui n'ont pas été conçues pour leur ouverture

Au-delà des cas où l'ouverture est rendue délicate par les systèmes d'information et les bases de données, les « bonnes raisons » de ne pas ouvrir des données portent aussi sur leur contenu. Nous l'avons aperçu dans le chapitre précédent : l'ouverture « travaille l'organisation » et attribue à certains agents la responsabilité inédite de gérer les données et d'assurer leur ouverture. Du point de vue de l'équipe en charge du projet, attribuer cette responsabilité peut distribuer l'ouverture et l'inscrire de manière plus pérenne dans le fonctionnement de l'organisation. Mais, du point de vue des agents, cette responsabilité interroge les conséquences possibles de la réutilisation et les risques qu'ils encourent en ouvrant volontairement les données dont ils ont la charge. Le premier risque que je vais aborder porte sur la qualité des données. Dans de très nombreux cas, l'ouverture concerne des données qui étaient rarement, voire jamais, sorties des réseaux sociotechniques de l'organisation et dont la qualité était généralement jugée suffisante pour les usages internes pour lesquelles elles sont produites. Pour une responsable de projet *open data*, les interrogations sur la qualité constituent la raison la plus récurrente pour laquelle les agents s'opposent à l'ouverture des données.

> La tarte à la crème de l'***open data***, c'est de partir du principe que les services ne veulent pas diffuser leurs données, que les services sont très possessifs de leur savoir, qu'il y aurait beaucoup de réticences… Je ne sais pas si Rennes est un cas à part, mais, de mémoire, je n'en ai eu quasiment aucune. Personne ne m'a dit « non » juste par principe. Par contre, c'est vrai qu'au début, ils disaient « oui, mais si on a des erreurs ? Est-ce que nos données sont vraiment bonnes ? » C'était vraiment toujours lié à la qualité de la donnée.
> (N.L., Responsable de la communication, Rennes Metropole)

L'idée, formulée notamment par Hans Rosling, selon laquelle les agents s'accrocheraient à leurs données et refusent par principe leur ouverture occulte le contexte sociotechnique dans lesquelles les données sont inscrites. Les projets d'*open data* s'intéressent à des

données qui n'ont pas été conçues au départ pour sortir des réseaux sociotechniques de l'organisation. Ils mettent donc à l'épreuve des données qui, si elles étaient publiées telles quelles, pourraient passer pour des données de mauvaise qualité, alors même que leurs usagers en interne n'y voyaient rien à redire jusque là. Pourtant, pour certains des acteurs évoqués dans le premier chapitre, cette question de la qualité des données doit tout simplement être évacuée du processus d'ouverture. Par exemple, Rufus Pollock suggérait dans son billet de blog repris par Tim Berners-Lee d'urger l'ouverture des données brutes lorsque les gestionnaires veulent les nettoyer et en réduire la complexité. Pour la Sunlight Foundation dans ses dix principes de l'*open data*, il faut d'abord ouvrir les données sans les modifier et attendre que le retour des usagers permette d'améliorer leur qualité. Ces préconisations se retrouvent dans certains des documents officiels qui cadrent les politiques publiques d'*open data*. En particulier, le vademecum d'Etalab considère que, puisque les données publiques sont déjà utilisées par l'administration pour ses missions de service public, elles peuvent être ouvertes sans que les gestionnaires se préoccupent de leur qualité[95]. Mais, en pratique, lorsqu'ils réclament l'ouverture d'un jeu de données, les responsables de projet *open data* ne peuvent pas simplement balayer de la main la question de la qualité et suggérer d'attendre leur amélioration par la discussion avec les usagers.

> Leur réflexe naturel c'est de dire « dans ce cas-là, il ne vaut mieux pas les ouvrir parce que je ne veux pas fournir des données dont je sais qu'elles ne sont pas bonnes. » C'est aussi peut-être un peu la peur qu'on voie que leur travail entre guillemets n'est pas bien fait, que les gens pointent du doigt « là ça ne va pas, etc. » Tandis que la solution, encore une fois très utopique, serait de se dire « non il vaut mieux plutôt les ouvrir et les gens vont pointer les erreurs qui permettront d'améliorer les données. »
> (L.K., responsable projet ***open data***, Rennes)

Selon cette cheffe de projet *open data*, l'ouverture immédiate des données publiques sous leur forme primaire se révèle « utopique », car les agents ne cessent de s'interroger sur la présence possible d'erreurs et d'inexactitudes dans des données qui n'ont jamais été mises

[95] « Les données publiques sont produites ou reçues dans le cadre d'une mission de service public. Elles sont donc généralement d'une qualité permettant le travail quotidien de l'administration et, en fonction de leur destination initiale, une utilisation statistique pertinente. […] Toutefois, les grands systèmes d'information de l'État et des collectivités territoriales, tout comme ceux des entreprises, peuvent parfois comporter des erreurs. L'existence de ces erreurs ne doit pas ralentir la démarche d'ouverture et de partage des données publiques. L'ouverture et le dialogue avec les réutilisateurs favorisent le signalement d'erreurs éventuelles. »

à l'épreuve du public. Dans un autre cas, celui de la ville de Paris, le service en charge du projet, le secrétariat général, a fixé comme priorité au réseau des correspondants d'ouvrir un grand nombre de données sans se préoccuper de la qualité des fichiers publiés. Mais, pour le correspondant d'une direction, cette « doctrine » de l'*open data* n'est pas envisageable, car d'éventuelles critiques du public sur la qualité des données pourraient rejaillir sur l'image du service et remettre en cause le professionnalisme de son travail.

> Moi, je suis vraiment pour l'ouverture des données à partir du moment où les données sont de qualité. C'est un avis qui est complètement personnel, on avait aussi des directives du secrétariat général qui étaient de ne pas hésiter à ouvrir des données même si elles étaient incomplètes, quitte à corriger dans le temps. […] À travers le jeu de données, on voit ce qu'on était capable de produire nous. Je n'étais pas avec la doctrine, d'ouvrir un maximum de jeux de données, quitte à ce qu'ils ne soient pas de très bonne qualité. Parfois on préfère attendre et prendre le temps de bien se structurer.
> (D.L., correspondant ***open data***, service informatique d'une direction, ville de Paris)

Pour les agents, l'ouverture des données peut constituer une prise de risque. Comme les données ouvertes peuvent servir au contrôle et à l'évaluation des politiques publiques, les critiques à l'égard de leur faible qualité ou de la présence d'erreurs peuvent avoir de lourdes conséquences pour la carrière des agents, d'autant plus que le projet d'*open data* ne fait généralement pas partie des missions qui leur sont assignées.

> [L'accidentologie] Ce sont des données sensibles, ça parle de morts, de blessés. On n'est pas encore assez sûr de ces données-là pour les diffuser. […] Ce sont des données qui sont relativement précises, mais il y a des problèmes de géocodage, on va peut-être oublier des accidents importants. Pour l'instant, on ne se sent pas prêt.
> (V.N., Responsable informatique d'une direction, Montpellier)

> Les agents flippent de te filer de la data parce qu'ils imaginent que toi, automatiquement, sans relecture, tu vas la sortir sur le site. Et du coup, ils se disent « oui, mais s'il y a un problème, je serai responsable du fait que cette donnée est sortie sur le portail. » […] Les agents ont peur de se mettre dans la merde. Puis déjà, il y a un truc de base, l'***open data*** ne figure dans aucune fiche de poste à part la mienne et celle de mon chef.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

De ce point de vue, les projets d'*open data* constituent des épreuves. En faisant migrer les données dans un cadre nouveau, ils rendent potentiellement centrales certaines de leurs dimensions qui étaient peu pertinentes dans leurs cadres d'usages initiaux. Des absences jamais remarquées deviennent des manquements, des approximations des erreurs.

> En ouvrant les données, on s'est rendu compte que, entre les écrans et ce qu'on publiait, ce n'était pas toujours pareil. Et ça a permis de faire remonter un bug à l'industriel. Ils étaient tout étonnés qu'on leur demande de corriger un bug qui existe depuis quinze ans, mais, comme personne n'avait ouvert des données en temps réel, personne n'avait pris la peine de vérifier que dans cette base était la même que sur les écrans.
> (N.N, Technicien informatique, Keolis Rennes)

> À partir du moment où vous diffusez des données, vous vous exposez à ce qu'on les analyse. L'exemple qu'on prend souvent, c'est les données sur les stations de vélos. Il y a un certain nombre de gens qui nous ont dit « ah, mais vous avez mis une station de vélo à tel endroit, mais en fait, elle est trois ou dix mètres plus loin. » Il y a eu des prévisions d'implantations des stations et puis, dans la réalité, elles ont peut-être été déplacées parce qu'il y a eu des travaux ou, dans la configuration du quartier, c'était mieux de le mettre à tel endroit et non pas à tel autre. Donc, ça fait que… ça nous a obligés à réinterroger l'emplacement réel des stations.
> (N.L., Responsable de la communication, Rennes Métropole)

On retrouve ici une question largement discutée en STS et au-delà, à propos des « *bad records* » (Garfinkel & Bittner, 1967) et des « *false numbers* » (Lampland, 2010). Au sein des organisations, les données dites « métiers » ne sont pas justes ou vraies en elles-mêmes. Leur faible degré de précision ou leur manque d'harmonisation n'ont aucun impact sur leur efficacité, au contraire. Il existe de nombreuses bonnes raisons organisationnelles, pour reprendre les termes de Garfinkel, pour que ces données persistent, tout simplement parce que leur justesse et même leur « vérité » sont ancrées dans les pratiques de ceux qui les manipulent et les mobilisent. En sciences (Zimmerman, 2008 ; Almklov, 2008) et dans les institutions statistiques (Desrosières, 2015), la question de la qualité est essentiellement liée au niveau de standardisation et d'harmonisation de données produites dans des contextes différents. Pour les données publiques, c'est d'abord par la confrontation entre des domaines de pratiques aux enjeux différents qu'une donnée est stigmatisée comme fausse ou mauvaise. En amont de l'ouverture, cela place les gestionnaires de données dans une situation délicate. D'une part, ils ne connaissent pas a priori les publics qui vont se saisir de

données qui souvent ne sont jamais sorties des réseaux sociotechniques de l'organisation. D'autre part, la présence d'erreurs ou les éventuelles critiques à l'égard de la qualité des données peuvent constituer un risque pour leur carrière ou leur service alors même que les responsables de projet *open data* leur demandent souvent de ne pas en tenir compte. À ce titre, il est particulièrement intéressant de revenir sur la formulation du deuxième principe « *Quality and Quantity* » de la charte du G8 qui demande l'ouverture de données de qualité tout en réclamant, si possible, leur publication sous leur forme brute et non modifiée : « *We will release high-quality open data that are timely, comprehensive, and accurate. To the extent possible, data will be in their original, unmodified form and at the finest level of granularity available.*" Au premier abord, on pourrait lire la demande de publier à la fois des données de qualité et, dans la mesure du possible, de les maintenir intactes de toute modification, comme contradictoire voire comme un prolongement de l'oxymore des données brutes. À la lumière des différents cas exposés précédemment, on peut plutôt y voir une tentative de mise à jour de la position de principe qui demande aux agents d'ignorer la qualité des données dans le processus d'ouverture. Cette formulation, qui a été retenue dans des termes très proches à Ottowa en 2015 lors de l'élaboration d'une charte internationale de l'*open data*, assume que la qualité, loin d'être une variable à écarter, constitue une source majeure de frictions dans l'ouverture des données. L'*accountability*, promue comme une vertu des politiques d'*open data*, ne consiste pas uniquement pour les agents à rendre des comptes sur la conduite des politiques publiques, mais aussi sur la qualité des données qu'ils produisent.

La sécurité : anticiper les dangers de la réutilisation

Au-delà des problèmes d'extraction et de la question de la qualité, de nouvelles frictions apparaissent lorsque les agents prévoient les risques qui pourraient survenir de l'utilisation des données. Dans le chapitre précédent, nous avons vu que l'anticipation des usages des données peut être une ressource pour orienter l'identification, mais elle peut parfois être une contrainte. Comme ils se voient souvent attribuer la responsabilité des données, certains agents prévoient les conséquences possibles de l'ouverture. Dans plusieurs cas en particulier, les gestionnaires de données ont exercé une grande vigilance quant aux risques possibles de leur réutilisation pour la sécurité et l'intégrité des habitants. Lorsqu'ils envisagent l'utilisation des données pour des usages malveillants, des données peuvent être exclues du périmètre concerné par l'ouverture.

> Il reste des données qu'on ne peut pas publier pour différentes raisons. Par exemple, il y a des données sensibles du type sur le réseau d'eau. Bon là, c'est vraiment pour

> des questions purement de sécurité, on ne peut pas publier les emplacements parce que ça pourrait permettre des actes malveillants qui peuvent avoir des conséquences assez graves.
> (G.H., Directeur des systèmes d'information, Montpellier)

En principe, ces données sur le réseau d'eau de la ville étaient ouvrables. Légalement, ce sont des données publiques, car elles ne comportent pas d'informations nominatives et ont été produites dans le cadre d'une mission de service public. Techniquement, elles pouvaient être extraites sans grandes difficultés, car le logiciel prévoit une fonction d'export. Enfin, la conduite d'un inventaire, antérieur au projet d'*open data*, a permis de les localiser dans les bases de données du service. Mais, en mettant les données au banc d'essais des risques, le gestionnaire de ces données sur le réseau d'eau a inclus des acteurs malveillants parmi les usagers possibles et a considéré que ces informations pouvaient servir à cibler des actes criminels. Dans d'autres cas, certains invoquent le fait que les usagers pourraient se servir des données pour commettre des actes illégaux. Par exemple, ils pourraient exploiter les données pour repérer les points sensibles des infrastructures urbaines tels que l'alimentation en électricité de l'éclairage public.

> L'éclairage public, on sait qu'il y a des types aux métiers pas très recommandables qui vont éteindre l'éclairage pour certains lampadaires en trifouillant dans les câbles, histoire de faire du trafic tranquille sans qu'il y ait trop de lumière pour les déranger. Alors si on libérait le câblage, ils pourraient savoir donc d'où vient le jus de telle armoire pour éclairer telle rue, péter telle armoire pour avoir tout un quartier dans le noir pour faire du trafic tranquillement. Bon, c'est un peu tiré par les cheveux, mais voilà le risque existe et puis, bon, il n'y a pas vraiment d'intérêt à libérer cette donnée. Alors, on libère uniquement le positionnement des mats.
> (V.N., Responsable informatique d'une direction, Montpellier)

Les oppositions à l'ouverture liées à la question de la sécurité n'émanent pas nécessairement des gestionnaires des données. Dans deux cas situés au sein des services de la ville de Paris, c'est la hiérarchie administrative qui a refusé l'ouverture de données proposées par les agents, au motif qu'elles pourraient servir à dégrader des équipements publics ou dérober des éléments du patrimoine de l'institution.

> J'avais également proposé les listes de pavoisement, mais ça a été refusé. Le pavoisement c'est là où nous on met des drapeaux. Donc, c'était un petit data set qui faisait environ 100 lignes avec drapeaux européens et drapeau français. Donc,

> après c'est sûr que quand on réfléchit bien, ça peut donner le challenge pour aller les piquer, mais bon.
> (Q.M., chargé de mission informatique d'une direction, ville de Paris)

> Après, il y a des données sur les équipements de la ville. Alors là, ce n'est même pas envisageable. Encore moins sur les terrains vagues, on ne va pas les mettre en ligne pour qu'ils soient dégradés. Là ce n'est même pas la peine de demander, c'est non d'entrée de jeu.
> (R.W., chargé de mission informatique d'une direction, ville de Paris)

Ces résistances liées à la question de la sécurité restent toutefois relativement rares et se situent uniquement dans deux terrains de mon enquête. Quelle que soit la récurrence de ce motif d'opposition à l'ouverture, ces cas font ressortir un point important. Distingués lors de l'identification comme responsables des données, les agents anticipent souvent la manière dont elles pourraient être utilisées. Pour prévoir les usages, ils imaginent comment de nouveaux réseaux sociotechniques pourraient s'attacher aux données ouvertes, y compris en incluant des criminels parmi les publics potentiels des projets d'*open data*. Ces cas remettent donc en question tous les principes évoqués dans le premier chapitre qui ne donnaient jamais à voir un usage potentiellement « mauvais » de la donnée, l'ouverture étant toujours pensée comme un bien au service d'un autre, ou un bien en soi (Dodier 2005 ; Dodier 2003).

La transparence : un mandat à obtenir

Une des raisons couramment évoquées pour expliquer les résistances à l'ouverture des données serait une aversion des agents à la transparence qui préféreraient « travailler dans l'opacité » et empêcher les citoyens de remettre en cause la mise en œuvre des politiques publiques. Dans les cas étudiés, les agents n'expriment pas une opposition systématique et définitive à la divulgation d'informations sur le fonctionnement des institutions, mais ils ne disposent en fait pas du mandat pour libérer des données qui pourraient servir à l'opposition politique. Les données servant l'objectif de transparence de l'action publique doivent parfois passer des circuits de validation qui sont mis en place pour décider de leur ouverture, une des transformations organisationnelles qui accompagnent les projets d'*open data*. Ces procédures plus ou moins formalisées conditionnent l'ouverture de données jugées « sensibles » à la validation de la hiérarchie.

> On voulait publier des jeux de données qui étaient assez sensibles, entre autres sur les collections vivantes du Jardin botanique, certaines sont sur les serres d'Auteuil

> où il y a un déménagement en ce moment. Et il vaut mieux éviter de publier des jeux de données quand il y a une actualité politique assez importante. Donc c'est pour ça qu'on a créé ce processus de validation au niveau de la direction.
> (D.L., correspondant ***open data***, service informatique d'une direction, ville de Paris)

> RW : C'est au chef du service de dire « oui on peut le publier » ou « non on ne peut pas le publier. » […] Moi, je n'ai pas vraiment mon mot à dire.
> SG : Et il décide en fonction de quoi le chef du service ?
> RW : En règle générale, si c'est des données sensibles… Par exemple, les demandes de logements c'est ultrasensible. Avec la pénurie de logements à Paris, il y a un peu plus de 130 000 demandeurs et on va pouvoir attribuer environ 6 000 logements par an. C'est un sujet extrêmement sensible.
> (R.W., chargé de mission informatique d'une direction, ville de Paris)

Comme pour les enjeux de sécurité abordés précédemment, la sensibilité est évaluée en anticipant sur la manière dont les données pourraient être utilisées. Nous avons vu auparavant que, lorsque les agents anticipent une utilisation des données qui pourrait nuire à la sécurité de la population ou à l'intégrité du patrimoine public, le processus d'ouverture pouvait être interrompu. Ici, si les données peuvent être utilisées « contre » l'administration et remettre en cause la conduite des politiques publiques, la simple demande d'ouverture qui émane du processus d'identification ne suffit pas à couvrir les agents pour les risques liés à la transparence. Pour être ouvertes, ces données doivent, en quelque sorte, obtenir un « visa » par lequel la hiérarchie protège les agents en cas d'utilisation contestataire des données.

Les procédures de validation, plus ou moins élaborées selon les contextes organisationnels, peuvent aboutir à ce que certains services soient écartés de l'identification des données. Dans plusieurs cas, cela concerne des services dits « supports » dédiés aux ressources humaines, à la gestion technique ou logistique. Ces composantes de l'organisation ont pour mission d'assister les équipes en charge de la mise en œuvre des politiques publiques. Dans l'identification, les données de gestion que produisent ces services peuvent répondre à l'objectif de transparence qui fonde souvent les projets d'*open data*. Mais l'ouverture volontaire de ces données peut poser problème dans des services qui n'ont pas pour mission ni pour habitude d'interagir avec les habitants.

> Nous, on est assez pauvre en données ouvertes parce qu'on est une direction support des services. Il faut comprendre que comme on est une direction support,

> on n'est pas à destination des parisiens en direct. Enfin nous, les parisiens on ne les connait pas et donc on ne travaille pas pour eux. […] L'*open data*, on ne sera pas dedans tant que ce ne sera pas conçu pour servir à une certaine transparence de fonctionnement de l'administration et non pas que des services aux habitants. Nous, on ne peut rien publier.
> (Q.M., chargé de mission informatique d'une direction, ville de Paris)

Cet extrait soulève un point important : l'identification et la sélection des données sont des épreuves qui invitent les acteurs à se positionner quant à l'utilité des projets d'*open data* et à leurs objectifs, là où les descriptifs des projets affichent des « biens en soi » (Dodier 2005 ; Dodier 2003) tels que la transparence, l'innovation ou la modernisation de l'État. En sélectionnant les données ouvertes, les chefs de service doivent parfois trancher en faveur d'un objectif ou d'un autre. Dans le cas de ce service, seul un jeu de données est accepté, car il pourrait permettre de créer des services pratiques pour les habitants, l'objectif retenu par la direction. Toutes les données proposées pouvant révéler le fonctionnement de l'administration ont été écartées par les dirigeants du service.

> J'avais proposé sept ou huit jeux de données. Le seul qui reste, c'est celui des cabines téléphoniques qui parle assez bien en termes d'utilité du jeu dans une présentation géographique. Tout ce qui pourrait montrer un petit peu le fonctionnement de l'administration, il y a un veto de la direction qui ne veut pas que l'habitant lambda s'immisce dans le fonctionnement de la collectivité.
> (Q.M., chargé de mission informatique d'une direction, ville de Paris)

Dans d'autres cas, certains services sont écartés directement par l'équipe en charge du projet d'*open data* avant même de formuler une demande d'ouverture auprès des agents et de leur hiérarchie. Dans un cas, le responsable du projet *open data* souhaiterait obtenir des données relatives à la rémunération et à l'absentéisme des agents de l'administration municipale. Sa direction s'y oppose invoquant une « culture » française de la transparence qui exclut systématiquement ces données des pratiques de divulgation d'informations publiques.

> Il y a des services que je pense que je n'irai jamais voir comme la Direction des Ressources Humaines (DRH). Les salaires, les taux d'absentéisme, ce genre de choses, ce n'est pas du tout dans la culture française et latine. Pour les faire rire, quand il y a des réunions avec les chefs, je mets grille salariale et absentéisme, on sait pertinemment que c'est une boutade. Je trouve que c'est une information importante, mais je ne trouve pas ça scandaleux qu'on me la refuse. C'est plus une

> réflexion…limite de société, mais je ne pense pas que ce soit à l'***open data*** de Montpellier de faire ça.
> (H.B., Chef de projet ***open data***, Montpellier)

On note dans son discours qu'il serait favorable à la publication d'informations détaillées sur l'activité et la rémunération des agents de l'institution, mais qu'une telle publication demande une transformation profonde des politiques de transparence qui va bien au-delà des limites du projet. Les pratiques de transparence des pays anglo-saxons, où les premiers projets d'*open data* ont été fondés en partie sur la publication de telles données, constituent une ressource pour renforcer sa revendication. Dans les témoignages de la plupart des responsables de projet *open data* que j'ai interrogés, je retrouve un militantisme affirmé et assumé en faveur de la transparence qui guide leur action très proche du cas bien étudié des professionnels de la participation dont la carrière peut dépendre de l'affirmation d'un éthos militant (Mazeaud, 2012). C'est cet engagement qui incite ce chef de projet *open data* à remettre sur la table, sur le ton de l'humour certes, l'ouverture de données telles que les salaires ou l'absentéisme.

Ces cas soulignent l'épaisseur des procédures de validation qui s'imposent aux agents et aux responsables de projet d'*open data* dès lors qu'ils veulent obtenir l'ouverture de données jugées « sensibles » d'un point de vue politique. Mais l'objectif de transparence souvent assigné aux politiques d'*open data* demande tout de même l'ouverture de certaines de ces données malgré les oppositions qui peuvent s'exprimer à travers les circuits de validation. Après avoir tenté de convaincre les gestionnaires d'ouvrir leurs données, les responsables de projet *open data* peuvent alors se tourner vers la hiérarchie politique pour contraindre l'ouverture des données. Dans le cas d'Etalab, ces requêtes ont été renforcées par le rattachement de la mission et de son directeur au Premier ministre. Comme nous l'avons vu précédemment dans le deuxième chapitre, l'affichage de la transparence du candidat sortant à l'élection présidentielle constituait une des raisons derrière la création de data.gouv.fr. Pour obtenir l'ouverture de certaines données réclamées par Etalab, les équipes des cabinets ministériels ont appuyé les demandes d'Etalab et ont parfois contraint les agents à ouvrir les données malgré les éventuelles oppositions de leur hiérarchie.

> On avait des réunions régulières, on faisait le point des données qui pouvaient être mises à disposition assez rapidement et puis les points d'achoppement. Et puis il y

> avait aussi les exigences d'Etalab, ils tapaient du poing sur la table et disaient je veux les données des [subventions]. Il y avait une exigence assez forte de la part d'Etalab. Nous on faisait intervenir le cabinet, on leur dit « Etalab met la pression sur le [bureau des subventions] et ils ne veulent pas donc voilà on est passé.
> (Q.H., Correspondant du réseau Etalab, ministère)

Le travail d'identification des données est ainsi ponctué de moments de négociation lors desquels les responsables de projet *open data* peuvent s'appuyer sur leur hiérarchie politique pour contourner les procédures de validation et obtenir des données portant sur la transparence de l'action publique. L'appui des élus et des membres de leurs cabinets peut, dans certains cas, justifier la prise de risque politique qui peut résulter d'une action de transparence et donner aux agents le mandat d'ouvrir des données « sensibles. »

> Pour qu'un projet comme ça arrive à fonctionner, il faut qu'il y ait un portage politique, un élu qui le porte politiquement sinon au premier arbitrage, le projet saute.
> (K.B., Responsable innovation numérique, Montpellier)

> On a beau avoir des contraintes, des textes, s'il n'y a pas une volonté du politique, quand il n'y a pas l'échelon décisionnel qui pousse, rien ne se fait dans les administrations parce que les gens ils ne font pas du zèle inutilement. Si ce n'est pas une contrainte pour eux, ils vont faire le service minimum. […] Si on n'a pas de référents dans les cabinets, on va travailler dans le vide parce qu'on n'est pas dans l'***open data*** par défaut. Ce n'est pas dans l'ADN des directions. S'il n'y a pas une impulsion politique, on est mort parce que les directeurs ne vont pas s'exposer sur des trucs qu'on ne leur demande pas.
> (Q.H., Correspondant du réseau Etalab, ministère)

Les responsables de projet d'*open data* doivent généralement s'appuyer sur leurs soutiens politiques pour obtenir l'ouverture de certaines données qui se révèlent trop « sensibles » pour que leurs demandes mènent à la publication des données. Les requêtes du service en charge de l'*open data* sont parfois traitées avec indifférence par des agents tandis que, relayées par un responsable politique, elles peuvent être traitées prioritairement.

> Quand tu as les chefs de service et les cabinets politiques avec toi, tu peux débloquer des situations. […] Tout à l'heure, j'ai vu des membres du cabinet de Jean-Paul Huchon [le président de la région]. Ils viennent taper à la porte d'un DGS en disant « il faudrait sortir les données, c'est Jean-Paul qui les demande. » En face, ils font « OK, très bien je te les envoie dans la journée. » Toi, tu es là depuis trois

> mois à bosser dans l'***open data***, ils vont te dire « on n'a pas » ou « on l'a, mais je ne te les donne pas », soit ils vont te faire patienter.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

Promue comme une des vertus de l'*open data* par certains des acteurs évoqués précédemment tels que la Sunlight Foundation, le renouveau de la transparence de l'action publique, ne peut, en pratique, se réaliser sans l'accord des autorités politiques. Lorsque les agents anticipent des risques politiques de contestation, les responsables de projet *open data* engagent souvent une négociation dans laquelle la hiérarchie politique a le dernier mot pour attribuer un mandat d'ouverture. Sans cet appui, les données publiques jugées « sensibles » peuvent rarement traverser les circuits de validation. Ces résultats sont aussi particulièrement intéressants au moment où le modèle de transparence par les données (*data-driven transparency*) discrédite les formes de révélation fondées sur la narration et l'interprétation (Birchall, 2014) et devient un produit d'export pour certains gouvernements tels que les États-Unis, la France ou le Royaume-Uni (Birchall, 2015). Au sein d'instances internationales telles que l'Open Government Partnership, la Banque Mondiale ou l'OCDE, l'ouverture des données est promue comme un outil essentiel pour améliorer la transparence de l'action publique, favoriser la participation des citoyens et accroitre leur confiance[96]. Les gouvernements des pays en développement sont parfois incités financièrement à l'ouverture de leurs données par le biais de subventions et la mise à disposition de méthodologies « clés en main » permettant de lancer un portail *open data*. Or, les cas évoqués précédemment rappellent que, même en France, le troisième pays dans le monde à avoir adopté une législation de droit d'accès à l'information publique (Boustany, 2013), l'ouverture de données sensibles ne peut pas reposer uniquement sur la volonté des agents, mais s'inscrit dans un contexte juridique et politique qui encadre la circulation des données.

Conclusion

Au terme de ce chapitre, nous avons pu identifier quelques-unes des sources de frictions qui empêchent les agents d'ouvrir des données localisées lors de la phase d'identification. La première porte sur les conditions techniques dans lesquelles les données sont produites. Quand les données sont gérées dans des systèmes d'information, elles ne prennent pas la

[96] Voir par exemple les rapports suivants : United Nations - Department of Economic and Social Affairs. (2013). Open Government Data for Citizen Engagement in Managing Development (pp. 1–104) ou Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing.

forme de fichiers qui pourraient s'échanger simplement. Pour les ouvrir, il faut souvent parvenir à désarticuler des systèmes d'information qui comprennent rarement une fonctionnalité d'export. Pour accéder aux données brutes, les *database managers* doivent souvent reconstituer le schéma de la base, un document rarement fourni par les prestataires, qui donne les clés de l'organisation des bases de données. Une fois qu'ils ont décodé les liens entre les éléments de la base, ils doivent aussi parvenir à mettre la main sur les données en concevant des « moulinettes », des outils sur mesure qui désarticulent les assemblages dans lesquels les données étaient ordonnées et rendues intelligibles à leurs usagers quotidiens. En dehors de ces cas, nous avons vu que le contenu même des données pouvait provoquer des frictions qui empêchent la circulation des données. En s'intéressant à des données qui n'ont pas été conçues au départ pour sortir des réseaux sociotechniques de l'organisation, les projets d'*open data* mettent à l'épreuve la qualité de données dont les usagers internes ne voyaient rien à redire jusque là. Les responsables de projet d'*open data* reprennent parfois une position de principe, formulée par certains acteurs évoqués dans le premier chapitre, selon laquelle les données doivent être ouvertes telles quelles, sous leur forme brute puisque le retour des usagers permettra d'en améliorer la qualité. Mais une telle position de principe se révèle bien souvent irréaliste. En effet, la question de la qualité ne peut pas être négligée tant ses répercussions peuvent être importantes pour les agents et leur hiérarchie. Des critiques à l'égard de la qualité peuvent rejaillir sur l'image du service, des erreurs ou des approximations pouvant être perçues comme le signe d'un travail de mauvaise qualité. Les projets d'*open data* opèrent ainsi une véritable transmutation des données que les agents ne peuvent pas ignorer. Avant d'ouvrir leurs données, les agents s'interrogent sur les conséquences de leur éventuelle ouverture en tentant d'imaginer les nouveaux réseaux sociotechniques dans lesquelles elles pourraient s'insérer. Ils peuvent alors s'opposer à leur ouverture lorsqu'ils anticipent de potentiels usages criminels ou illégaux. La préfiguration des usages intervient aussi dans le cas de l'ouverture de données « sensibles » politiquement. Contrairement à l'idée reçue selon laquelle les agents s'opposeraient à la transparence, nous avons vu que c'est précisément une « bonne raison organisationnelle » qui empêche l'ouverture de ces données. En effet, des procédures de validation des données sont souvent mises en place par la hiérarchie administrative, de manière plus ou moins formelle selon les services ou les organisations. Elles soumettent l'ouverture des données à la validation des responsables de service qui doivent attribuer un « visa » par lequel la hiérarchie protège les agents en cas d'utilisation contestataire des

données. Pour contourner ces procédures, les responsables de projet d'*open data* peuvent parfois s'appuyer sur le soutien des élus qui assument les risques politiques de la réutilisation et donnent le mandat d'ouvrir ces données « sensibles. »

Dans ce chapitre, je me suis concentré essentiellement sur des données dont l'ouverture posait des difficultés. Mais, dans le travail d'identification, il arrive aussi que les responsables de projet d'*open data* trouvent des données dont l'ouverture ne provoque pas de frictions. C'est particulièrement le cas dans les services où les informations que les agents manipulent sont déjà désignées comme des données. Pour ces services, les données ne sont pas un sous-produit des activités de gestion, elles constituent la mission qui oriente les activités des agents. La production de données est, par exemple, au cœur des missions des Services d'Information Géographique (SIG) dans les collectivités locales et des Services Statistiques Ministériels (SSM) dans les ministères. Lors de l'identification, les responsables de projet d'*open data* sont souvent renvoyés vers ces services lorsqu'ils réclament des données.

> Après, on connaissait les services qui gèrent le plus de données. Par exemple, le service du SIG. On sait qu'ils vont avoir énormément de données géographiques donc c'est un des premiers services qu'on est allé voir aussi.
> (H.B., Chef de projet ***open data***, Montpellier)

> Pour eux au début, l'***open data*** c'était de la donnée statistique. Une des idées reçues dans la tête de beaucoup d'administrations, c'était « on va vous mettre en relation avec le département statistique, vous verrez, ça se passera bien. »
> (T.Y., un agent de la mission Etalab)

Dans ces services, les agents sont déjà dédiés à la production de données, c'est même la mission première qui leur est assignée. Le travail d'instauration des données ne constitue pas une épreuve dans ces services ; les agents manipulent déjà au quotidien des données. Ce qui distingue les produits de ces services en tant que données, c'est qu'elles sont déjà prises dans des réseaux sociotechniques qui assurent leur circulation et leur exploitation. Cela ressort aussi très clairement dans le cas du projet d'*open data* de Rennes qui a débuté par l'ouverture de données de transport, jugées « faciles à ouvrir. »

> Donc il y a eu d'abord l'ouverture des données du vélo. Pourquoi le vélo ? Parce que c'était facile, il faut être honnête. On avait déjà un système qui donnait en temps réel sur le site Internet la disponibilité du nombre de places et du nombre de vélos. […]

> Finalement, il suffisait juste d'ouvrir ce flux de données en ***open data***, il était déjà presque codé parce qu'il était déjà sur le site Internet. […] On a aussi mis la disponibilité des parcs relais. En fait, les parcs relais, on avait déjà un flux entre guillemets « en ***open data*** privé ». Tous les quarts d'heure, on envoyait à la Dir ouest, l'organisme qui gère les grandes routes sur l'agglomération et notamment la rocade, un fichier qui leur donnait en temps réel le nombre de places disponibles dans le parc relais, s'il était fermé, ouvert. Et ça s'affiche sur la rocade.
> (N.N, Technicien informatique, Keolis Rennes)

Que ce soit pour les données de disponibilité des vélos en libre-service ou pour celles sur les places dans les parcs relais, on aperçoit dans l'extrait précédent l'épaisseur des réseaux sociotechniques qui assurent leur circulation. Cet extrait sous-entend l'existence d'une chaine de traitement et de mise en forme de ces données. Ces circuits ressemblent à ceux qu'a identifiés Desrosières (2005) à propos de la transformation des fichiers de gestion de l'administration, locaux et situés, en un savoir à portée générale. Comme pour les statistiques publiques, ils assurent un début de transmutation, c'est-à-dire la transformation d'une substance en une autre et permettent le passage d'un monde de signification, celui de la gestion d'un système de transport, à un autre, celui de l'information voyageur. L'ouverture de ces données s'est révélée « simple » puisqu'elle a consisté à orienter ces réseaux vers un nouvel assemblage dédié à leur ouverture. À l'inverse, pour les données pour lesquelles les frictions sont les plus fortes, les réseaux sociotechniques qui assurent leur circulation restent à construire et le travail d'instauration doit encore être effectué : il faut désigner des rôles et des responsabilités, reconstituer les schémas d'organisation des bases de données, développer les « moulinettes » qui permettent l'extraction, négocier leur ouverture, inspecter la qualité des données, s'assurer qu'il n'y a pas d'informations « sensibles » qui demanderaient une validation hiérarchique… Le contraste entre ces cas de données « faciles à ouvrir » et celles pour lesquelles l'ouverture provoque d'importantes frictions est particulièrement intéressant pour répondre à la question « que sont les données ? » Dans les cas précédents, c'est à partir du moment que ces objets sont pris dans des réseaux sociotechniques et des chaines de traitement dédiés à leur circulation, leur exploitation et leur mise en visibilité qu'ils sont considérés comme des données.

Cela nous amène à nous intéresser au traitement des données lors de leur ouverture. Jusque là, on pourrait prétendre que les données « elles-mêmes » dans leur « contenu » ne sont pas vraiment affectées par l'ouverture. Or, nous allons voir dans le chapitre suivant que le

processus de l'ouverture réclame souvent des transformations qui assurent concrètement l'intelligibilité des données et leur instauration progressive en une donnée ouverte à de nouveaux traitements.

Chapitre 5

Transformations et transmutations : la fabrique des données brutes

Revenons-en à l'événement Open Data Bootcamp de la région Ile-de-France. Au début de la réunion, les organisateurs ont distribué un vadémécum, un document prévu pour répondre aux questions des producteurs de données. J'ai déjà eu l'occasion de m'y référer, il prend la forme de questions et de réponses. Une d'entre elles est formulée ainsi « comment publie-t-on concrètement des données sur data.iledefrance.fr ? » Le document explique qu'il y a deux méthodes pour publier des données sur le portail *open data* (figure 32).

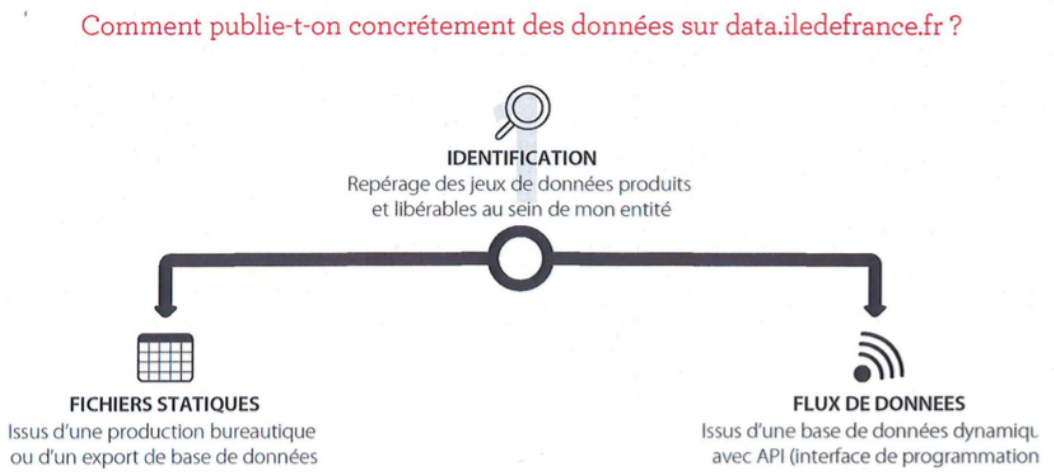


Figure 32. Schéma représentant les méthodes de publication de données, extrait du vadémécum de l'*open data* de la région Ile-de-France.

La première est qualifiée de méthode de versement manuel : « le référent transfère les fichiers correspondants sur un espace de stockage en ligne. Les jeux transmis font ensuite l'objet d'une vérification de qualité et d'éventuelles opérations de traitement (conversion de format, enrichissement, *geocoding*, etc.) par l'équipe de coordination de la démarche *open data*. » La deuxième méthode, le versement automatisé « concerne les jeux de données disponibles sous la forme de flux à partir d'une base de données dynamique. » Le document indique que « la mise en œuvre de cette méthode nécessite de vérifier la faisabilité technique, voire de développer un connecteur spécifique. » Les organisateurs indiquent que le versement manuel est le cas plus fréquent pour les jeux de données de la région. Une

autre question du vadémécum demande « pourquoi les données doivent-elles être publiées dans un format brut et quels sont les différents formats proposés ? » En réponse, le document explique que « pour permettre une réutilisation simple par le plus grand nombre, il est recommandé de présenter ces données dans des formats ouverts (exemples : CSV, JSON, XML, RDF) qui permettent la réutilisation sans restriction d'accès ni de mise en œuvre, par opposition à un format fermé ou propriétaire. Dans la mesure du possible, l'ouverture des données publiques requiert la diffusion des données brutes dans des formats normalisés qui permettent une réutilisation simplifiée dans des applications. »

Après avoir abordé les principes de l'*open data*, leur mise en politique publique, les explorations et les négociations qui ont conduit à la décision d'ouvrir certaines données, je vais m'intéresser ici aux conditions concrètes dans lesquelles les données sont ouvertes. À travers ces extraits, on comprend que le choix des formats semble déterminer en partie la capacité des données à être facilement réutilisées, mais le vadémécum explique brièvement les raisons de l'utilisation des formats ouverts. Il semblerait aussi que certains formats normalisent les données afin d'accroitre leur potentiel de réutilisation. Enfin, le document évoque pour la méthode manuelle des vérifications de qualité et des opérations de traitement des données. Les données ne restent donc pas toujours brutes, elles sont parfois modifiées et retraitées avant leur ouverture.

Dans le chapitre précédent, je me suis intéressé aux contraintes d'extraction que pose l'ouverture de données produites dans des systèmes d'information. Dans la première partie, je vais m'intéresser aux transformations effectuées sur les données produites dans un autre environnement, celui des tableurs, l'outil bureautique le plus commun pour stocker, traiter et utiliser des données. Les données n'y sont pas nécessairement disposées sous la forme de tables, car le tableur propose un environnement dans lequel les données sont traitées, visualisées et mises en forme. Les formats fréquemment utilisés pour transmettre des données tabulaires tels que le PDF ou Excel sont critiqués pour « freiner » la réutilisation des données. Certains militants de l'*open data* vantent l'usage du format CSV, un standard flexible qui présente les données sous forme de valeurs lisibles dans tout éditeur de texte. Or, les opérations de transformation des fichiers (d'Excel à CSV par exemple) ne se résument pas à l'usage du menu « enregistrer sous » ou de convertisseurs automatisés, elles exigent des transformations importantes pour conserver l'intégralité et l'intégrité des

informations contenues dans les fichiers produits par les tableurs. Par ailleurs, les standards de données sont présentés par certains comme des solutions qui permettent une utilisation automatisée des données et leur interopérabilité, quels que soient les contextes de leur production. Je m'intéresserai, dans la deuxième partie de ce chapitre, au format GTFS, un standard développé au départ par Google, qui s'est rapidement imposé comme la norme dans l'ouverture de données de transports. Son utilisation implique de repenser les catégories employées dans les bases de données de transport et demande une transformation profonde des infrastructures de production de données. Enfin, j'aborderai les transformations qu'exercent manuellement les agents sur les données en amont de l'ouverture. Regroupées sous le terme d'édition, ces opérations consistent à modifier le contenu des données avant leur publication. Pour les gestionnaires de données, l'édition permet de garantir l'intelligibilité des données et aussi de protéger les agents des risques juridiques, politiques ou communicationnels qui pourraient découler de l'ouverture. Ces opérations de conversion, de structuration et d'édition visent à désencastrer les données de réseaux sociotechniques dans lesquelles elles ont été produites et à les transformer en données « ouvertes », disponibles pour de nombreux types de traitements. Les étudier de près va nous permettre de poursuivre notre investigation sur la « nature » des données.

Convertir

Dans un grand nombre de cas, les données sont produites et traitées dans l'environnement du tableur. D'un point de vue historique, Visicalc a été le premier logiciel de tableur. Lancé en 1979, il a simplifié la production et le traitement des données par l'invention de la feuille de calcul, un espace où les valeurs numériques peuvent être directement manipulées sans avoir à connaitre de langage de programmation (Campbell-Kelly 2007). Les fondements de la feuille de calcul restent globalement inchangés depuis 1979[97], mais les logiciels qui ont succédé à Visicalc (Lotus 123, Microsoft Excel, OpenOffice) ont apporté de nouvelles fonctionnalités telles que l'ajout de graphiques, le formatage du texte des cellules ou encore la création de tableaux croisés dynamiques. Mais, pour certains responsables de projets d'*open data*, ces fichiers ne sont pas considérés comme pleinement ouverts du fait de leur

[97] Une feuille de calcul consiste en une matrice de cellules, indexées par la combinaison d'une lettre désignant les colonnes et d'un nombre pour les lignes. Ce système permet de désigner une cellule (ex. : A5), mais aussi une colonne (A), une ligne (5) ou encore un ensemble de cellules (A5:B12). Chaque feuille de calcul peut comporter plusieurs tables, accessibles par des onglets en bas de la fenêtre. L'utilisateur ne voit qu'une seule table à la fois, mais peut appeler les cellules d'une autre table pour ses calculs. Chaque cellule peut contenir soit du texte soit une valeur numérique soit une formule de calcul (Campbell-Kelly, 2003, p.329).

format. Par exemple, ce DSI, un des responsables du projet *open data* de Montpellier, considère que les formats employés ne permettent pas d'exploiter les données : « souvent, l'information était diffusée d'une manière assez inexploitable soit dans des documents en PDF soit dans des fichiers Excel. Mais il faut changer le format si on veut les publier pour que ce soit réellement exploitable derrière. » Comment expliquer que les formats PDF et Excel ne soient pas considérés comme exploitables ? Pour qui et pour quelles raisons ? Comment les données sont-elles transformées par et pour les standards ?

Une des revendications essentielles formulées par certains des acteurs évoqués dans le premier chapitre a consisté à réclamer l'abandon du format PDF pour la publication des données gouvernementales. Ce format, très critiqué par certains, est bâti sur des fondements et un modèle d'intelligibilité qui ont fait son succès. Conçu par Adobe à partir de 1991 et adopté par l'ISO depuis 2008 comme un standard ouvert, le format PDF tente de reproduire l'impression du papier, « *the look of printedness* », lors de l'échange de documents numériques dans des contextes professionnels. Comme le format mp3 qui anticipe que le corps humain ne perçoit qu'une faible partie du spectre audio (Sterne, 2006 ; Sterne, 2012), le PDF a été conçu pour un certain type d'usagers. Il ancre une division du travail entre le producteur du document et le destinataire dont le rôle sera essentiellement la lecture du document (Gitelman, 2014). En créant une séparation entre l'auteur et le lecteur et en isolant les activités d'écriture et de lecture, le format PDF assure que le document aura une apparence identique quel que soit le support, mais cette fixité frustre certains usages. À ce titre, Gitelman, dans le chapitre de *Paper Knowledge* sur le PDF, évoque brièvement quelques critiques à l'égard du format PDF. Elles portent en particulier sur le fait que certains documents sont composés d'images qui doivent être traitées par des outils de reconnaissance de caractères dits d'OCR (*Optical Character Recognition*) pour que le texte soit indexé (Gitelman, 2014). En effet, l'extraction d'un tableau publié en PDF est loin d'être automatique, elle demande souvent une interprétation manuelle. Dans les projets d'*open data*, les limites du format PDF se sont révélées en juillet 2014 avec la publication des déclarations de patrimoine et d'intérêts des parlementaires suite à l'adoption de la loi sur la transparence de la vie publique du 11 octobre 2013. La plupart des déclarations étaient manuscrites, scannées par la Haute Autorité pour la Transparence de la Vie Publique (HATVP) créée par la loi de 2013. La HATVP s'était engagée à publier les déclarations en *open data* sur son site web. En juillet 2014, elle a diffusé un fichier au format CSV qui

renvoyait vers les liens les déclarations publiées dans une centaine de fichiers PDF[98]. À la suite, Regards Citoyens a publié un communiqué dans lequel l'association conteste le fait que les données soient effectivement ouvertes[99].

> Si la Haute Autorité pour la Transparence met à disposition un jeu de données recensant les élus et les déclarations qu'elle contrôle, les informations contenues dans les déclarations d'intérêts ne sont en revanche pas à proprement parler en Open Data : elles n'ont pu être publiées par la HATVP scannées sous la forme de PDF images rendant l'exploitation de ces informations malaisée au vu du grand nombre d'informations mises en ligne.

Regards Citoyens a publié une plateforme participative dans lequel plus de 8000 citoyens ont numérisé le texte des déclarations. Sans revenir sur les détails de cet épisode, notons que Regards Citoyens nie qu'un tel fichier PDF composé d'images puisse être une donnée ouverte. Il est ainsi fréquent que les partisans de l'*open data* contestent la qualité de donnée ouverte lorsque le format PDF est employé. Lors de l'Open Data Bootcamp de la région Ile-de-France, un des organisateurs répondait à la question posée par un participant « pouvez-vous donner quelques exemples de formats fermés ? » Il répond « PDF c'est bien documenté, c'est ouvert, mais paradoxalement, mais XLS c'est un format fermé [...] c'est un format de document, mais c'est pas de la donnée en fait. » Du fait qu'il faille en extraire les données contenues dans un fichier PDF soit par reconnaissance de caractères soit par des outils qui vont identifier les tableaux, la qualité même de données est remise en cause par ces acteurs. Pourtant, pour certains, ces fichiers sont des données, des bits qui entrent et sortent d'un système informatique (Blanchette, 2011). Au-delà de la qualité même de donnée, Regards Citoyens considère que les informations numérisées par les internautes sont des « données brutes. » On touche là à une des ambiguïtés de la notion, car c'est précisément le travail des internautes qui a permis de transformer les centaines de documents PDF en des données « brutes ». Ce cas montre qu'une donnée peut devenir brute même lorsqu'elle a été façonnée par des milliers de mains. Nous aurons l'occasion d'y revenir en détail, mais ce cas révèle de

[98] Haute Autorité pour la transparence de la vie publique, « Open data », http://www.hatvp.fr/open-data.html, consulté 25 juillet 2014.

[99] Regards Citoyens, « 8000 personnes libèrent en une semaine les données manuscrites des déclarations d'intérêts des parlementaires ! », http://www.regardscitoyens.org/8000-personnes-liberent-en-une-semaine-les-donnees-manuscrites-des-declarations-dinterets-des-parlementaires/, consulté le 4 août 2014.

manière saillante « l'oxymore des données brutes » (Bowker, 2005 ; Gitelman, 2013). Ce n'est qu'après avoir été manipulées et transformées que les données sont devenues brutes.

Quant au format Excel, les groupes d'intérêt se revendiquant de l'*open data* en nient régulièrement le caractère ouvert. La Sunlight Foundation considère dans ses dix principes pour l'ouverture des données que l'utilisation de formats propriétaires empêche des usagers d'exploiter les données[100].

> *Sometimes that program is unavailable to the public at any cost, or is available, but for a fee. For example, Microsoft Excel is a fairly commonly-used spreadsheet program which costs money to use. Freely available alternative formats often exist by which stored data can be accessed without the need for a software license. Removing this cost makes the data available to a wider pool of potential users.*

Microsoft Excel représente le tableur le plus utilisé, il propose par défaut l'utilisation de formats élaborés par Microsoft. Par défaut, Microsoft employait le XLS, un format dont les spécifications sont la propriété de l'entreprise. Pour exploiter les données publiées dans ce format, les logiciels doivent interpréter ses spécifications par retro-engineering. LibreOffice, l'alternative libre à la suite Office, qui a même été pendant un temps poursuivi par Microsoft pour l'utilisation du format XLS. En 2006, Microsoft s'est engagée à rendre le format ouvert suite notamment à des demandes de la commission européenne. Elle a développé des formats déclinés progressivement dans tous les logiciels de la suite bureautique Office, dont le XLSX pour Excel. Bien que le format a été standardisé par l'ISO, Regards Citoyens comme d'autres acteurs du logiciel libre tels que l'APRIL (l'Association pour la Promotion de l'Informatique Libre) considèrent que ce format n'est pas ouvert. Ils se réfèrent au cadre général d'interopérabilité de l'Union Européenne qui définit un standard ouvert, non seulement par ses spécifications publiques et gratuites, mais aussi par la gouvernance ouverte de son processus d'élaboration : « le standard est adopté et sera maintenu par une organisation sans but lucratif et ses évolutions se font sur base d'un processus de décision ouvert accessible à toutes les parties intéressées (consensus ou vote à la majorité, etc.) »[101]

[100] Sunlight Foundation, « Open Data Policy Guidelines », http://sunlightfoundation.com/opendataguidelines/, consulté le 20 décembre 2014.

[101] Regards Citoyens, « Non, Excel et Word ne sont pas des formats ouverts ! », http://www.regardscitoyens.org/non-excel-et-word-ne-sont-pas-des-formats-ouverts/, consulté 25 novembre 2015.

S'ils décident de suivre les préconisations de Regards Citoyens ou de la Sunlight Foundation, les agents ne pourront pas publier directement leurs fichiers Excel. Ils devront donc se tourner vers un standard ouvert, Regards Citoyens en préconise trois en particulier pour faciliter l'interopérabilité des données : le RDF, le format OpenDocument utilisé dans Libre Office et le CSV. Comme je l'ai expliqué dans le chapitre 1, je n'ai pas rencontré dans mon enquête de cas d'usages du format RDF. Le format Open Document, en revanche, est fréquemment employé dans les portails *open data*. Mais, il apparait dans mon enquête que certains développeurs considèrent que ce format est difficile à interpréter. Un des responsables de data.gov.uk m'a ainsi expliqué qu'il le déconseille aux producteurs de données : « *With ODS [le format OpenDocument pour les feuilles de calcul], we have a lot of problem with that. A lot of hardcore users told us 'please, don't use it!' We push the community not to use ODS until the format is better understood*." Certains développeurs peuvent déconseiller ce format ouvert, car les données dans les feuilles de calcul ne sont pas uniquement sous forme tabulaire. Les feuilles peuvent comprendre des graphiques, du formatage ou des tableaux dynamiques qui servent à l'élaboration de documents, mais peuvent entraver la réutilisation automatique des données. Enfin, le troisième format, le CSV, est recommandé à la fois par les militants de l'ouverture des formats et les politiques publiques d'*open data*. CSV signifie *Comma Separated Values*, valeurs séparées par des virgules. Chaque fichier CSV contient du texte codé selon des standards internationaux tels que ASCII ou Unicode. Les éditeurs de texte ou les navigateurs web peuvent ouvrir les fichiers CSV, mais les données n'y sont pas représentées comme une feuille de calcul, mais comme un texte. Chaque ligne d'un fichier CSV contient le même nombre de valeurs, des séquences de texte sont séparées par un caractère. Comme le suggère le nom du format, une virgule fait normalement office de séparateur, mais une espace, un point virgule ou une barre de tabulation sont aussi acceptés par les logiciels. Le format CSV précède l'entrée sur le marché des ordinateurs personnels, il est utilisé depuis 1967 par le langage de programmation d'IBM, le Fortran et il est interprété par la grande majorité des tableurs et de nombreux systèmes de gestion de données. Les fichiers CSV sont facilement utilisables dans la plupart des langages de programmation du fait que le texte soit codé dans les standards informatiques les plus répandus. Le CSV est particulièrement utilisé pour échanger des données tabulaires entre des programmes ou des systèmes informatiques.

Lorsque les militants de l'ouverture des données prônent l'usage du CSV, ils le décrivent comme un standard robuste, établi et stabilisé. Par exemple, Rufus Pollock en 2007 implorait les administrations à publier des fichiers CSV plutôt que de belles interfaces web qui vieillissent vite et dont il est difficile d'extraire les données : « *please, please just give me a plain old csv file and a plain old url […] ascii text, csv files and plain old sql dumps (at least if done with some respect for the ascii standard) don't date — they remain forever in style*. »[102] On pourrait croire en lisant Pollock que le CSV est standardisé de longue date et que ses spécifications ne font plus débat. Pourtant, les premières tentatives de standardisation du CSV datent de 2005 quand un ingénieur, Yakov Shafranovich, a publié une *Request for Comments* à l'*Internet Engineering Task Force* (IETF), une organisation qui promeut l'utilisation de standards ouverts sur Internet[103]. La RFC 4180, est catégorisée comme « *informational* » par l'IETF, cela signifie que l'organisation ne recommande pas officiellement de la suivre. La RFC est néanmoins citée comme le standard de facto pour un fichier CSV. Elle spécifie notamment que la première ligne du fichier doit inclure un en-tête qui définit les colonnes et que des guillemets doubles doivent délimiter chaque champ. Le codage des caractères n'est pas défini, mais la RFC suggère l'utilisation de l'ASCII. Pourtant, ce standard exclut les caractères non latins et les accents alors que des alternatives plus compréhensives comme l'Unicode (UTF-8) existent. Pour Palme (2009), le maintien de l'ASCII entretient un « impérialisme » de la langue anglaise en ignorant les spécificités des langues étrangères. Si les concepteurs de logiciels suivent les recommandations de la RFC, les fichiers produits par des usagers non anglophones seront mal interprétés. Les efforts de standardisation du CSV se poursuivent. En particulier, le W3C (*World Wide Consortium*) en charge des standards du web a lancé un groupe de travail sur le sujet dans la lignée des travaux de son fondateur, Tim Berners-Lee sur l'ouverture des données (voir chapitre 1, épisodes 3 et 4). Le W3C espère établir un standard de métadonnées qui servirait à décrire les paramètres de chaque fichier sans exiger l'application de spécifications strictes.

Les politiques publiques d'*open data* placent comme objectif l'utilisation de standards ouverts. Par exemple, le vadémécum de l'*open data* du gouvernement français recommande l'utilisation de formats ouverts, en particulier du CSV et déconseille fortement l'usage du

[102] OKFN blog, « Give Us the Data Raw, and Give it to Us Now », http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/, consulté le 12 avril 2015.
[103] IETF, « RFC 4180 - Common Format and MIME Type for Comma-Separated Values (CSV) Files », http://tools.ietf.org/html/rfc4180, consulté 25 mars 2014.

PDF pour ouvrir les données. L'utilisation de formats ouverts fait aussi l'objet d'évaluations que ce soit par des acteurs publics ou des groupes d'intérêt. Le modèle en cinq étoiles de Tim Berners-Lee, conçu au départ pour indiquer la « marche à suivre » est devenu un véritable outil de *benchmarking*. Par exemple, en décembre 2011, au lancement de data.gouv.fr, Regards Citoyens a publié un billet de blog qui propose une première évaluation du nouveau portail. L'association y reprend le modèle de Tim Berners-Lee pour en faire un outil d'évaluation : « en examinant le catalogue, nous n'avons pu recenser que quelques dizaines de données en CSV et XML contre plusieurs centaines sous des formats propriétaires de Microsoft, loin de respecter les objectifs fixés par le gouvernement. Un sérieux effort reste donc encore à faire par Etalab sur ce point. Suivant la norme édictée par l'inventeur du web, ce n'est qu'une petite moyenne que l'on peut simplement accorder à data.gouv.fr pour sa sortie. » On le voit à travers cet extrait, l'emploi de formats ouverts est considéré comme un indicateur essentiel d'une « bonne » politique d'*open data*. Les responsables de projet *open data* y ont aussi recours pour encourager les agents à utiliser des standards ouverts. Au Royaume-Uni, l'équipe en charge de data.gov.uk a créé un outil pour attribuer un score à chaque fichier publié sur data.gouv.fr pour son niveau de conformité avec l'échelle en cinq étoiles de Tim Berners-Lee. Le gouvernement britannique s'est servi du modèle pour benchmarker les services de l'État. L'outil compare les notes moyennes des jeux de données publiés par chaque département. Dans un communiqué de presse publié en décembre 2012, le cabinet annonce avoir établi un score d'ouverture pour chaque département gouvernemental: "*The average openness score for all departments is 52%, based on the percentage of the datasets published by each department and its arms-length bodies that achieve three stars and above against the Five Star Rating for Open Data*." Tim Davies, doctorant à l'université de Southampton, avec qui j'ai coécrit un article sur les standards de données[104], a vivement critiqué ce score d'ouverture. Dans un billet de blog, il explique que le score incite les départements à supprimer les fichiers non lisibles par les machines pour obtenir une meilleure note. Pourtant, Tim Berners-Lee voulait encourager les administrations à ouvrir leurs données, quel que soit le format. Dès qu'un fichier est publié avec une licence ouverte, il y obtient une étoile. Le score d'ouverture figure toujours sur data.gov.uk en tant qu'un des critères pour explorer les jeux de données publiés sur le portail (figure 33).

[104] Notre article « The Daily Shaping of State Transparency: Standards, machine-readability and the configuration of Open Government Data policies » paraitra en 2016 dans la revue Science and Technology Studies.

Figure 33. Score d'ouverture des jeux de données. Capture d'écran de la page *datasets* de data.gov.uk (juillet 2015)[105].

Ces cas montrent que l'utilisation de standards ouverts dont le CSV en particulier, n'est pas seulement guidée par un souci d'interopérabilité et de lisibilité par les machines. L'utilisation du CSV devient donc un standard de performance (Busch, 2011), un signe d'une « bonne politique d'ouverture des données. »

En l'absence de spécifications contraignantes, le CSV reste un format relativement flexible. Les francophones peuvent ainsi toujours utiliser un point virgule comme séparateur (et donc garder la virgule comme séparateur décimal), coder la date selon leur norme ou utiliser Unicode pour que les accents (ou un tréma dans mon cas personnel) contenus dans leurs noms soient pris en compte. Mais cette flexibilité peut devenir une contrainte pour les usagers qui doivent spécifier les règles utilisées par chaque fichier ou adapter leur code à la variété de fichiers CSV acceptables. Dans un tableur, il faut souvent spécifier l'encodage du texte, le caractère de séparation et le séparateur décimal pour chaque fichier (figure 34).[106]

[105] Remarquons que l'algorithme évalue la majorité des jeux de données au niveau zéro. Cela signifie soit qu'il considère qu'ils ne sont pas ouverts, soit qu'il ne parvient pas à les évaluer.

[106] Les dernières versions d'Excel et de LibreOffice, les deux tableurs les plus utilisés, sont configurées pour détecter automatiquement les paramètres des fichiers CSV.

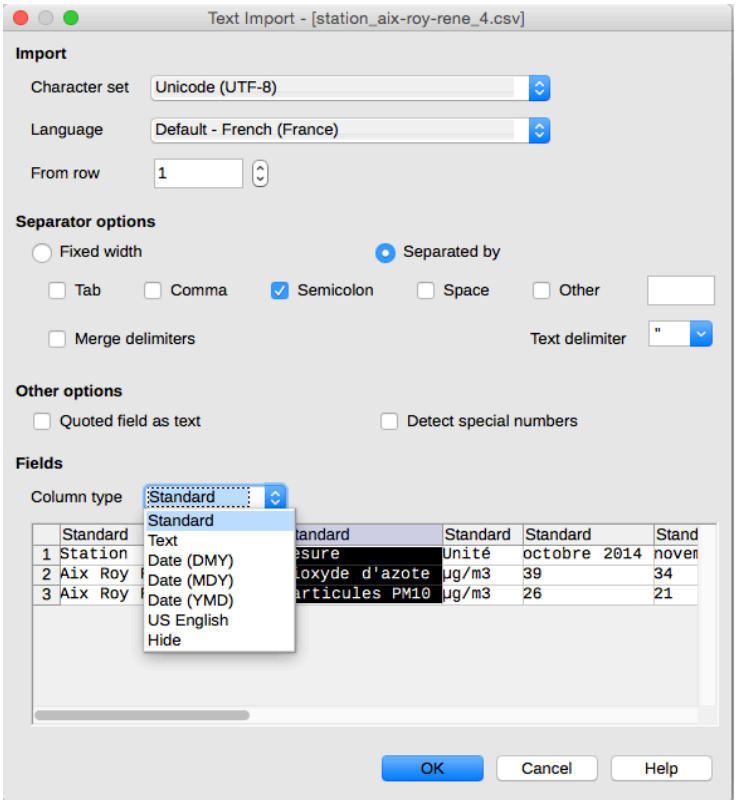


Figure 34. Fenêtre de dialogue dans le logiciel Libre Office 4.2 à l'ouverture d'un fichier CSV.

Pour faciliter l'utilisation de fichiers CSV, des métadonnées peuvent spécifier les paramètres utilisés. Par exemple, la Haute Autorité de Transparence de la Vie Politique (HATVP) a publié une notice descriptive d'un fichier CSV qui contient les liens vers les déclarations d'intérêt des parlementaires. Elle prend la forme d'un document PDF de trois pages qui décrit le contenu de chaque colonne et commence par une description des paramètres du fichier CSV.

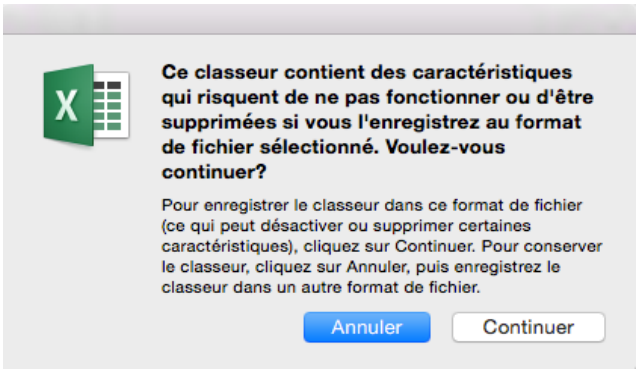
> Le fichier liste.csv est un fichier texte permettant de décrire un tableau. Chaque ligne de texte correspond à une ligne d'un tableau. Sur une même ligne, un séparateur est placé entre chaque colonne du tableau.
> – Encodage des caractères : UTF-8
> – Retour à la ligne : CR+LF
> – Séparateur : Point-virgule (« ; »)
> – Identificateur de chaine : Guillemet droit double (« " »)
> – Première ligne : En-tête du tableau

> Il est à noter que pour ouvrir ce document sur les tableurs, l'encodage UTF-8 nécessite généralement une manipulation spécifique. Ex : Excel 2010/Données/A partir du texte/Origine du fichier=65001 : Unicode (UTF-8[107])

La HATVP a dû spécifier comment ouvrir le fichier CSV, car certaines versions de Word sont alignées par défaut sur les paramètres suggérés par la RFC. Or, les cellules comprennent du texte avec des accents, l'encodage ASCII recommandé par la RFC 4180. À travers cet exemple, on voit bien les contraintes que peut imposer le format CSV pour les usagers non anglophones.

Mais, si on quitte le point de vue des usagers et qu'on porte notre attention sur le travail des administrations, comment concrètement un fichier Excel passe-t-il au format CSV ? Est-ce aussi simple que d'utiliser le menu « enregistrer sous » du tableur et changer le format ? Quel est le coût de l'utilisation de ce format ouvert pour les travailleurs des données ? La plupart des fichiers Excel demandent des transformations avant d'être convertis en CSV. Je vais l'illustrer par un document qui explique en détail les modifications et les « bonnes pratiques » à suivre pour que les fichiers Excel soient correctement interprétés en CSV. Il a été conçu par le service numérique de la région Ile-de-France pour former les producteurs de données à l'usage de ce format[108]. Ce document montre que les fichiers doivent être profondément transformés avant d'être convertis en CSV.

Lors du passage en CSV, Excel publie par défaut un message d'avertissement lors de l'enregistrement d'un fichier en CSV (figure 35).



[107] Haute Autorité pour la transparence de la vie publique, « Open data », http://www.hatvp.fr/open-data.html, consulté 25 juillet 2014.

[108] Christophe Libert sur Slideshare, « OpenData : quelques bonnes pratiques sur Excel », http://fr.slideshare.net/christophelibertidf/bonnes-pratiquesexcel-cc27juin2013, consulté le 25 juillet 2015.

Figure 35. Message d'avertissement lors de l'enregistrement d'un fichier en CSV dans Microsoft Excel (version 15.11 pour Mac)

Ce message prévient l'usager que des caractéristiques ne seront pas prises en compte par le format, mais le logiciel ne spécifie pas lesquelles. L'usager doit donc tester par lui même le fichier converti afin de comprendre le fonctionnement du format. Avant de convertir un fichier Excel ou LibreOffice en CSV, l'usager doit procéder à une série de transformations pour que les données ne soient pas altérées. Le document de la région Ile-de-France suggère une série de transformations à effectuer en vue de la conversion des fichiers Excel en CSV. La première consiste à fusionner les onglets. En effet, un fichier CSV ne peut contenir qu'une seule feuille de calcul par fichier alors que les tableurs permettent de travailler sur plusieurs feuilles de calcul délimitées par des onglets en bas de la fenêtre. Cette fonctionnalité peut rendre visible un découpage territorial ou temporel comme dans la figure 36 où les onglets séparent les départements. Les onglets peuvent aussi servir à délimiter les données de leur formes agrégées ou visualisées sous forme de graphiques.

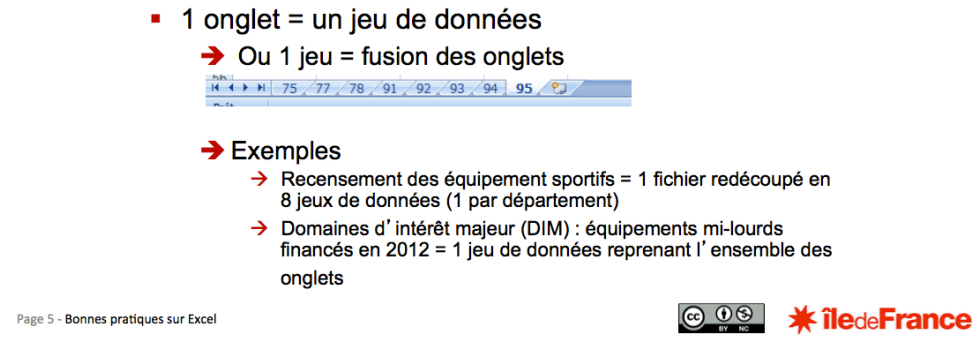


Figure 36. Recommandations relatives sur les onglets. Extrait du document « bonnes pratiques sur Excel » de la région Ile-de-France.

Selon le document de la région, deux options s'offrent aux agents pour contourner cette limite du format CSV. La première consiste à créer un fichier par feuille comme dans l'exemple du recensement des équipements sportifs. Cette opération peut être longue et laborieuse. Sauf à utiliser des scripts automatisés qui peuvent se trouver sur le web, elle nécessite d'enregistrer manuellement en CSV chaque onglet du fichier. L'autre option consiste à fusionner les tableaux, mais cela suppose que les données qu'ils contiennent soient compatibles entre elles.

Deuxièmement, la conversion au format CSV va aussi changer le contenu de certaines cellules. Comme on le voit dans la figure 37, les agents peuvent fusionner des cellules pour regrouper des titres de colonnes ou des valeurs identiques. Or, le format CSV ne comprend qu'un nombre défini de cellules et de colonnes par lignes il n'est pas possible de fusionner plusieurs cellules.

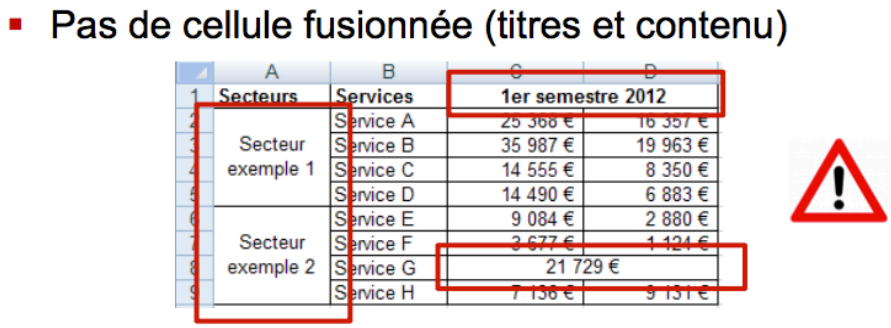


Figure 37. Recommandations sur les cellules fusionnées. Extrait du document « bonnes pratiques sur Excel » de la région Ile-de-France.

Par défaut, les convertisseurs des tableurs déplacent la valeur vers la cellule en haut à gauche d'un ensemble fusionné. Ils remplacent les autres cellules par une valeur vide. Dans l'exemple précédent converti en CSV, seules les lignes 2 et 6 auront un secteur renseigné, la colonne D n'aura plus de titre et la cellule D8 sera vide. Il faut donc corriger manuellement le fichier converti en CSV pour que le sens du tableau ne soit pas altéré. Par ailleurs, les auteurs du document de la région Ile-de-France avertissent que le passage en CSV va afficher les lignes masquées, des informations pouvant être divulguées si les producteurs de données n'y prêtent pas attention.

Enfin, le format CSV ne transmet pas d'informations par le formatage. Un fichier CSV ne comprend que du texte encodé de manière standardisée, le format ne définit pas de mise en forme. Les informations transmises par la couleur du texte, le gras, l'italique ou encore la taille des caractères doivent être converties en valeur textuelle pour subsister au format CSV. Par exemple, dans le cas de la figure 38, le code couleur transmet le type de musée et des précisions sur les périodes de fermeture. En CSV, ces informations devront figurer soit dans le tableau dans une colonne supplémentaire soit dans les métadonnées qui accompagnent le fichier.

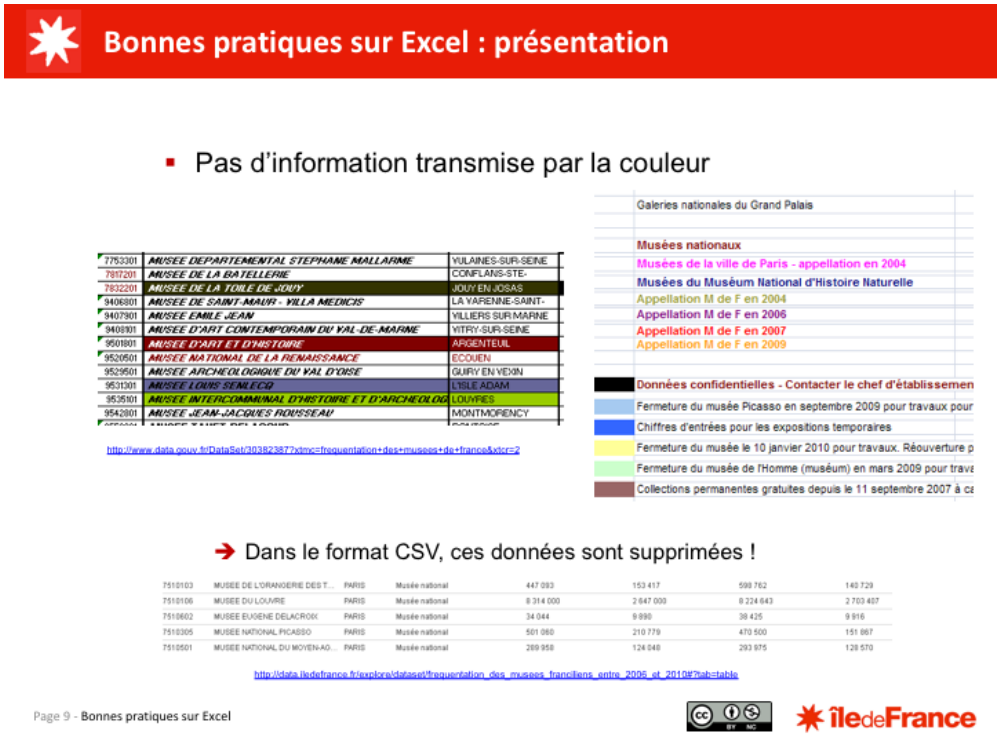


Figure 38. Recommandations sur la présentation du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Ile-de-France.

Le changement de format demande donc une transformation importante des fichiers. Convertir réclame un travail bien plus conséquent qu'utiliser le menu « enregistrer sous » des tableurs et changer le format. La transformation des fichiers par et pour le format CSV se révèle coûteuse en temps et en énergie pour les producteurs de données. Le changement de standard crée des frictions dans l'ouverture, choisir le CSV au lieu de conserver le format utilisé par le tableur par défaut consiste en un « investissement » (Thévenot, 1986) pour les gestionnaires de données. Ces transformations impliquent pour les agents de renoncer à une partie de leur temps de travail (pour un projet qui rentre rarement dans leurs missions). Il est espéré que cet investissement dans la lisibilité des données par les machines sera rentabilisé par une réutilisation facilitée des données et la création de services, de visualisations et d'applications qui sont une des retombées attendues de l'ouverture. Même si l'usage du CSV est recommandé et fait parfois l'objet d'évaluation, les responsables de projet *open data* doivent persuader les agents de réaliser cet investissement, en convertissant leurs données en CSV.

> Sur le CSV, j'ai fait un gros travail de pédagogie avec les mecs avec qui je bosse. Je leur ai montré des petits scripts que [un développeur d'Etalab] avait développé et je

> leur ai dit « mettez vous à la place d'un informaticien. » Ou je leur montrais Umap [un outil de cartographie en ligne développé par OpenStreetMap] que je trouve assez génial. Quand tu veux charger dans Umap, ce n'est pas du XLS, c'est du CSV ou du vectoriel donc je leur dis « voilà, regardez, vous ouvrez des menus déroulants, le fichier XLS n'est pas possible. Donc si le mec veut faire une cartographie de votre fichier, il peut pas. Déjà c'est un blocage. »
> (T.Y., un agent de la mission Etalab)

On le voit, le travail de conviction ne se fonde pas uniquement sur des arguments purement « techniques » portant sur les avantages du format. Il s'appuie aussi sur la convocation de figures de l'usager. Par exemple, un développeur sélectionné dans l'équipe peut jouer le rôle de porte-parole des « informaticiens » dont les scripts ne sont pas compatibles avec le format Excel. Lorsqu'ils tentent de convaincre les producteurs de données, les chefs de projet *open data* représentent l'usage de formats propriétaires et non lisibles par les machines comme un obstacle qui va réduire la capacité de réutilisation des données. Ils vont aussi présenter des outils qui imposent l'usage du format CSV. Ces outils pourront servir directement aux agents eux-mêmes à condition qu'ils fassent l'effort de convertir leurs données. Mais la conviction ne suffit pas toujours à justifier le coût du changement de format et les responsables de projet *open data* préfèrent parfois l'ouverture dans un premier temps dans le format Excel.

> On ne pouvait imposer le CSV. Il fallait passer par le XLS parce que c'est le format de l'administration quoi. Et je suis désolé je peux pas leur demander l'impossible. Ils ne sont pas tous des geeks, ils n'ont pas tous des formations sur les outils bureautiques sinon juste une formation très standard pour faire du copier-coller… Parfois cette culture des ayatollahs du libre, moi ça me fatigue un peu aussi. On fait du mieux qu'on peut. Les mecs ne passeront pas au CSV du jour au lendemain.
> (T.Y., un agent de la mission Etalab)

La demande d'utilisation du format CSV rompt avec les pratiques d'usage et d'échange de fichiers dans l'administration où le format Excel est souvent la norme. J'ai pu directement observer les réticences des producteurs de données à l'usage du CSV lors de la réunion Open Data Bootcamp de la région Ile-de-France évoquée en introduction du chapitre. Lors de la présentation du portail, un des organisateurs évoque les possibilités d'export des données. Il évoque brièvement les formats « on a csv, json, et enfin Excel, mais ça c'est pas bien. » On entend des rires dans la salle, que j'interprète alors comme de la moquerie ou de la désapprobation. Une personne dans la salle réagit « Excel est un format que plein de

logiciels lisent, c'est un standard de fait, il n'y a aucune raison de mettre à l'écart Excel parce que c'est un format commercial. » L., un des organisateurs le coupe et le corrige « c'est un format propriétaire et fermé. » Cette personne reprend « moi je préfère avoir un truc en Excel qu'en JSON. » Les organisateurs préfèrent couper ce débat, car nous étions en fin de réunion et ils avaient déjà largement plaidé pour l'utilisation de ces formats. À travers ce bref récit, on voit bien que les réticences des agents ne s'expliquent pas uniquement par le coût de la conversion des fichiers, bien souvent les agents n'ont pas l'habitude de les utiliser et ils imposent de changer des pratiques bien ancrées.

Du fait de l'investissement que requiert l'usage du CSV, les producteurs de données refusent souvent d'utiliser ce format. Une division du travail s'opère alors entre les gestionnaires de données qui transmettent les fichiers et les chefs de projet *open data* qui les transforment. Ces derniers opèrent les transformations sur le fichier et s'assurent de la bonne traduction de l'ensemble des informations qui figurent dans le fichier.

> CD : Quand on reçoit un fichier, on l'ouvre et il y a des trucs genre du fusionné, du gras, de la couleur. De toute façon, quand tu le passes en CSV, tout saute. Dans certains fichiers, les mecs ont mis de la couleur qui a une signification. Alors, que dans le CSV il n'y a pas de couleur donc, tu es obligé de créer d'autres colonnes.
> SG : Tu fais comment dans ce cas-là ?
> CD : Eh ben… à la mano. […] Si on ne comprend pas, on s'en réfère au producteur de données qui nous a envoyé le fichier. […] On essaie d'éduquer tous nos référents à la bonne formalisation à la base de leur fichier. Comme ça ça nous évite effectivement de retravailler à chaque fois. […]
> SG : vous dites quoi aux agents quand il y un code couleur dans leur fichier ?
> CD : Nous, on lui dit que c'est mal et que ce n'est pas comme ça qu'il faudra faire les prochaines fois. Après, je pense que ça se fera par le travail dans le temps sur la data. Dans deux ans, nos référents, ils sauront très bien qu'un fichier Excel, il ne faudra plus le traiter avec de la couleur, du gras, des cellules fusionnées ou des titres incompréhensibles.
> (C.D., chargé de projet ***open data***, région Ile-de-France)

> Ça m'arrive, encore aujourd'hui, de repasser derrière des fichiers le soir ou le weekend et de les enregistrer en CSV puis je les mets sur la plateforme. C'est ma petite contribution pour aider les ministères. […] J'informe évidemment mon correspondant « cette fois-ci, je l'ai fait pour ma pomme, mais la prochaine fois si tu peux m'éviter de faire ça le weekend… » J'essaie de les faire culpabiliser un peu, ça marche.
> (T.Y., un agent de la mission Etalab)

Certaines informations ne sont plus transmises lors du passage au format CSV. Les responsables de projet *open data* doivent alors les faire figurer dans le fichier ou dans les métadonnées. On le voit dans les extraits précédents, les responsables de projet *open data* conçoivent cette division de travail comme temporaire. Ils prévoient que les gestionnaires de données vont prendre en compte ces contraintes dans l'élaboration de leur fichier. Ce chargé de projet *open data* estime que la « culture de la donnée » va mettre un terme au formatage des cellules et va réduire les fichiers à une forme tabulaire. Or, l'usage du gras, d'un code couleur, d'onglets ou de tableaux croisés, s'ils peuvent être perçues comme une nuisance pour la réutilisation automatisée des données ont une valeur d'usage considérable qui favorisent la spatialisation de l'information (Kirsch, 1995) et facilitent aussi bien le repérage que la recomposition des informations (Beltrame & Jungen, 2013). Contraindre les agents à se conformer au format tabulaire et textuel du CSV risque donc de réduire la flexibilité cognitive du tableur au profit de l'intelligibilité des données pour les machines et d'aligner les pratiques de production et d'échange des données dans les administrations aux exigences techniques de leur ouverture.

Structurer

Jusqu'ici, les standards auxquels nous nous sommes intéressés concernent l'encodage des caractères et proposent des grands principes dans l'organisation des données. Or, certains standards concernent aussi la structure des fichiers ainsi que les termes, les catégories et les nomenclatures qu'ils contiennent. En science, la standardisation des données a déjà été bien étudiée. Latour (1993) a montré que c'est par la standardisation que les inscriptions acquièrent de nouvelles propriétés matérielles (elles deviennent immuables) qui permettent leur circulation et leur combinaison dans des « centres de calcul » (Latour 2006). Cette réduction des particularités locales permet une amplification des phénomènes observés par les instruments, ils deviennent alors plus généraux et plus mobiles. Dans le cas des projets d'*open data*, l'« harmonisation » des données entre les producteurs permettrait de réutiliser les données sans avoir à adapter des fichiers produits dans des contextes locaux. L'idéal d'une réutilisation « sans friction » des données qui a animé le développement des grandes infrastructures informationnelles en science se retrouve aujourd'hui dans les projets d'*open data*. L'ouverture des données publiques a ainsi été accompagnée de la création de plusieurs standards internationaux dans des domaines variés tels que les aides au développement (IATI), la loi (Legal XML) ou encore le tourisme (projet européen Citadel in the Move). Au niveau national, en France, l'association Open Data France tente d'« harmoniser » la

structure de plusieurs jeux de données produits par les collectivités locales. Tous ces projets exigent que la structure et le format des données soient uniformisés, seules les valeurs requises par le standard doivent varier entre les institutions. On voit avec cet enjeu de structuration un nouvel aspect des transformations que peuvent subir les données dans le processus de leur ouverture. Je vais m'intéresser ici à un standard international dont j'ai pu observer l'implémentation à Rennes. Il s'agit du GTFS (General Transit Feed Specification), un standard dédié à l'ouverture des données des horaires de transports. J'ai pu enquêter sur l'adoption de ce standard au sein de Keolis Rennes et comprendre comment les spécifications du standard ont reconfiguré les pratiques locales de production de données.

Le standard GTFS a été développé à partir de 2005 par un ingénieur de Google, Chris Harrelson, après une demande de Trimet, la régie des transports urbains de la ville de Portland dans l'Oregon[109]. Harrelson développait le projet Google Transit qui visait à inclure les horaires des transports publics dans Google Maps. Sa collaboration avec Trimet a permis de définir les spécifications du standard qui s'alignait largement sur les pratiques de l'entreprise de Portland. Google, qui apparaissait au départ dans le nom du standard, a ouvert ses spécifications et publié des outils qui valident la bonne implémentation de la norme dans les données ouvertes. Google inclut régulièrement de nouvelles données ouvertes au format GTFS dans les calculs d'itinéraire de Google Maps. Aujourd'hui, le GTFS est devenu le standard de facto pour les données de transport. Les développeurs le recommandent aux agences de transport en vue de l'ouverture de leurs données. Dans le cas de Rennes, l'ouverture des données de Keolis était guidée par le développement de l'information voyageur sur application mobile. Le standard a été conçu pour l'utilisation automatisée des données de transport par les développeurs. En choisissant le GTFS, les données sont configurées pour leur utilisation par les scripts informatiques conçus par des développeurs.

> À un moment donné, on s'est dit « bon, on a ouvert les données vélo, super. Demain on ouvre les données bus, bus métro, enfin les autres données dont on dispose. » Le souci c'est qu'on s'est dit « mais dans quel format ? » On ne savait pas trop. Donc, toujours pareil, on était sur un terrain en friche, on est allé demander aux

[109] Streetsblog San Francisco, « How Google and Portland's TriMet Set the Standard for Open Transit Data », http://sf.streetsblog.org/2010/01/05/how-google-and-portlands-trimet-set-the-standard-for-open-transit-data/, consulté le 25 mars 2014.

> développeurs et ils nous ont dit « A notre avis, le GTFS est un bon format, populaire, documenté, facile d'accès, commençons avec ça. »
> (J.B., Responsable marketing, Keolis Rennes)

> On publie nos données sous forme de fichier plat sous le format GTFS qui est très bien foutu, qui est une norme de Google, mais qui peut être complexe à comprendre. On s'était dit « le mec qui arrivera à nous sortir une appli avec ça, il va être costaud. » Et il y a eu le concours lancé par Rennes métropole sur l'*open data* en octobre 2010 et dès janvier, on a eu des applications qui utilisaient le GTFS, on était épatés.
> (N.N, Technicien informatique, Keolis Rennes)

Lors d'une réunion d'une entreprise de transport, j'ai pu aussi entendre que « le grand public ne peut pas utiliser le GTFS. » Sans un outil qui va combiner les différents fichiers qui composent un flux GTFS ou sans maitrise de la programmation informatique, l'affichage d'un horaire de transport est en effet très complexe. Chaque « flux » GTFS est composé d'une série de fichiers textuels en CSV et compressés dans une archive ZIP. Chaque fichier texte standardisé en CSV qui le compose détaille un aspect des horaires des transports publics : les entreprises prestataires, les arrêts, les lignes, les trajets, le calendrier, les jours spéciaux ainsi que les informations sur les tarifs et les transferts possibles. La construction du « flux » ne demande pas la production de tous ces fichiers, certains sont optionnels, mais les spécifications imposent des champs spécifiques et détaillés qui ne doivent pas varier entre les fournisseurs de données. Pour un usager qui souhaiterait juste consulter l'horaire d'un bus, il faudrait connaitre l'identifiant de l'arrêt dans stops.txt, retrouver la ligne dans routes.txt, identifier le prochain trajet dans trips.txt et finalement connaitre l'heure du passage du bus à l'arrêt dans le fichier stop_times.txt. Clairement, le format GTFS n'a pas été conçu pour la consultation des données. Elles doivent être combinées par des scripts automatisés pour que l'usager parvienne à afficher les horaires dans une interface.

En plus de configurer les données pour les développeurs, le standard GTFS réclame que les fournisseurs de données décrivent leurs horaires à travers des normes partagées. Le GTFS ne se contente pas de définir l'encodage des fichiers ou la mise en forme des données comme le fait le format CSV. Il requiert que les fournisseurs de données standardisent la structure des fichiers ainsi que les termes, les catégories et les nomenclatures qu'ils contiennent. Or, les bases de données comprennent toutes sortes de valeurs qui peuvent être considérées comme des erreurs ou des anomalies, mais qui ont une utilité dans les

activités de gestion (Garfinkel & Bittner, 1967). Bien que l'horaire soit un des domaines les plus investis par la standardisation afin d'assurer la coordination des activités à distance (Busch, 2011), cette donnée peut prendre des formes tout à fait singulières dans les bases de données d'une organisation. Dans le cas suivant, les gestionnaires inscrivent dans les horaires des passages au-delà de 24h. Ce que les développeurs pourraient décrire comme une aberration facilite le travail de gestion. Indiquer le passage d'un bus à 25h30 assure le suivi du travail des chauffeurs ou le traitement des journées avec une circulation extraordinaire. Or, le standard GTFS réclame une certaine description des horaires, cette information pourrait être traitée comme une aberration ou une erreur par les outils qui sont fondés sur ses spécifications.

> En gros, on avait des journées qui, au lieu de se terminer à minuit pile, terminaient à 26h20. […] On gère une journée de travail et des journées salariées donc ce n'est pas le lendemain.
> (J.B. et V.M, Responsables marketing, Keolis Rennes)

> Et il y a notamment le cas conceptuellement difficile à imaginer du 25h30, mais ça permet d'être sûr que 25h30 c'est bien rattaché à la journée de la veille. C'est vrai que quand on voit 25h30 la première fois, on se demande si ça n'est pas une erreur. […] Donc on est obligé de préciser 30 h parce que c'est bien lié à la journée de la veille. Et, s'il n'y a pas de journée la veille parce que c'est le premier mai, il n'y aura pas de départ à 7h le matin. Des fois, c'est un peu compliqué, il y a des journées qui se chevauchent : le premier bus le matin part à 4h, mais le dernier bus de la veille arrive à 5h. C'est-à-dire qu'il y a deux journées qui se chevauchent. […] C'était compliqué à expliquer aux développeurs que, quand ils font une recherche à 4h du matin, il faut aussi chercher les départs qui sont à plus de 25 h.
> (N.N, Technicien informatique, Keolis Rennes)

En incorporant une définition d'un arrêt, d'une ligne ou d'une agence, le standard demande d'adopter des définitions communes des objets qui sont décrits dans les données. Au sein des organisations qui participent à une démarche d'ouverture des données, ce processus de « commensuration » (Espeland & Stevens, 1998) exige parfois de repenser les données dans leur conception et d'exclure ces « erreurs » qui ont pourtant une grande utilité dans le travail quotidien des gestionnaires. Ici, Keolis Rennes a préféré ne pas adopter une mesure commune, l'entreprise a conservé ses horaires au-delà des 24h. En partageant ses référentiels avec le public, l'entreprise garde les particularités de ses bases de données pour que l'ouverture n'entrave pas le bon déroulement des activités de gestion. Le fichier qui en

résulte ne respecte pas strictement les spécifications du GTFS qui sont vérifiées par un outil de certification développé par Google. Cela peut éventuellement expliquer pourquoi Google n'a toujours pas intégré les données d'horaires de Keolis Rennes dans Maps près de cinq ans après leur ouverture. Ce cas rappelle que les infrastructures de données supportent rarement la diversité des ontologies et excluent les entités qui résistent aux systèmes de classification standardisés (Bowker, 2000).

Dans d'autres cas, l'adoption du standard implique de modifier en profondeur les pratiques de production de données. Lorsque le standard demande une définition unique, il faut aligner tous les acteurs qui interviennent dans la production de la base de données des horaires. Le gestionnaire de données doit alors convaincre les producteurs de modifier leurs routines de travail. Pour produire des données de qualité dans lesquelles les entités sont uniformément nommées, des changements dans les bases de données doivent être pris en compte à tous les niveaux de la production du flux GTFS.

> J'ai dû faire un gros travail pour expliquer aux gens que l'*open data*, ça générait de nouveaux potentiels de clientèle, des gens qui n'auraient peut-être pas pris forcément le bus. […] Pour pouvoir conserver ce potentiel de clientèle, il fallait absolument avoir de la donnée propre. Alors évidemment, ça impliquait d'être plus vigilants à ce qu'on écrivait. Comme toujours, le plus gros travail n'était pas technique, c'était vraiment organisationnel et humain. Il fallait expliquer aux gens tous les impacts d'une erreur dans une base. On l'a bien expliqué que ça arrivait, tout le monde fait des erreurs, mais qu'il y avait juste à être vigilant et si jamais il y avait une erreur, la corriger au plus vite pour que derrière ça redescende bien. […] On s'est rendu compte que lors du changement d'un nom d'arrêt, la personne ne redescendait pas toujours l'info au service qui concevait les horaires. Donc j'ai fait remonter aux différentes personnes qui ont remis en place un process d'information « quand je change un nom d'arrêt, j'envoie un mail à untel. » C'est tout bête, mais comme avant ça ne se voyait pas, le nom d'arrêt à la limite on s'en fout du moment qu'on arrive à concevoir les horaires. Il y a eu un travail qui a été fait pour cartographier le processus et mettre en évidence ce qui n'allait pas pour l'améliorer. (N.N, Technicien informatique, Keolis Rennes)

Pour convaincre les producteurs d'adopter de nouvelles pratiques de travail, il soutient que l'amélioration de la qualité des données encouragera les développeurs à créer des applications qui faciliteront l'usage des transports en commun et vont in fine permettre de créer une nouvelle clientèle au bus. L'ouverture des horaires au format GTFS demande de

transformer en profondeur les pratiques de production de ces données. Elle impose de créer un référentiel, une base unique et partagée des horaires des arrêts. L'adoption du standard a aussi demandé de repenser des catégories qui n'avaient jusqu'alors jamais été contestées. Une information essentielle comme celle de la localisation des arrêts doit être repensée en fonction des usages vers lesquels les données vont être orientées. Pour un usage mobile, la position exacte de l'arrêt devient essentielle pour que les usagers repèrent leur itinéraire. Or, cette position de l'arrêt ne faisait pas l'objet d'une collecte précise, les agents s'en servaient uniquement à des fins de maintenance. Ces données ne servant pas à la communication au public, la précision n'était pas au cœur des préoccupations de leurs producteurs.

> Le positionnement des arrêts était crucial et on se retrouvait finalement à se demander « mais quelle est la définition de l'arrêt ? » On avait un arrêt qui était sur un plan sauf qu'il y a deux côtés de la rue, qu'un arrêt de la rue peut être à vingt mètres ou quarante mètres de l'autre… Le positionnement de l'arrêt, c'est là où il y a le zébra où le bus s'arrête ou c'est le positionnement de l'arrêt où le client attend ? […] Et ce qui fait que parfois l'arrêt est plutôt en amont de la zone d'arrêt du véhicule ou en aval. Et on se retrouvait avec tous les cas de figure et finalement, il a fallu affiner la démarche. […] Ça nous a permis de reconstruire la donnée en croisant avec les données du SIG. Nos données étaient peut-être un peu plus aléatoires parce qu'on les utilisait à des fins de construction ou de plan, on n'a pas un besoin énorme de précision.
> (J.B., Responsables marketing, Keolis Rennes)

Le standard, dans ses spécifications, oriente ainsi les données vers leur usage mobile. Son adoption transforme des bases de données produites à des fins de maintenance en des éléments essentiels de l'information voyageur. Dans ce cas, les gestionnaires ont dû se coordonner avec le service d'information géographique de la collectivité locale qui disposaient d'informations plus précises. Comme dans le cas des échanges de données scientifiques, la production de données standardisées destinées à être ouvertes passe dans ce type de cas par la mise en œuvre de « chorégraphies d'acteurs », de synchroniser des acteurs et des pratiques locales afin de produire des données « de qualité » (Ribes & Jackson, 2013).

Editer

Les transformations que nous avons vu jusque là sont essentiellement guidés par des standards qui déterminent le format, la structure des données. Le GTFS, un cas très

particulier dans mon enquête, va au-delà en imposant des définitions des objets qui sont contenus dans les données. Mais les processus d'ouverture mettent parfois en marche des transformations plus radicales encore, que l'on peut rassembler sous le terme « édition. » Éditer, c'est par ce terme que je regroupe toutes les opérations qui consistent à modifier le contenu des données avant leur publication. En statistique, *data editing* est le terme anglo-saxon qui désigne les opérations par lesquelles les statisticiens traitent et transforment les données issues des sources administratives (Desrosières, 2005). Pour les statisticiens, ces opérations servent à transformer des données produites à des fins de gestion en un savoir agrégé de portée générale. Dans cette transmutation des données, les conventions d'équivalence permettent de qualifier un cas individuel en une catégorie (Desrosières, 2000). Dans l'ouverture des données publiques, il s'agit là aussi de changer d'univers de sens, mais les modifications n'obéissent généralement pas à des règles ou à des conventions précisées d'avance ou partagées entre les institutions. Les données sont façonnées directement par les agents selon deux préoccupations principales : rendre intelligibles les données et réduire les risques de leur ouverture. Là encore, la préfiguration des usages guide les transformations : les données sont rendues compréhensibles pour encourager certains usages ou à l'inverse plus « inoffensives » pour en prévenir d'autres. Les transformations apportées aux données avant leur diffusion sont souvent qualifiées de nettoyage dans les sciences comme dans les administrations. Elles désignent le travail par nature invisible et « transparent » par lequel les idiosyncrasies, le désordre et les traces du travail des données brutes sont effacés pour produire des données « certifiées », prêtes à être traitées dans de nouveaux réseaux sociotechniques (Millerand, 2012 ; Walford, 2013). J'ai préféré ici le terme plus générique d'édition, car le nettoyage laisse entendre que les transformations que nous allons aborder ici consistent à effacer des erreurs ou des « saletés » qui pourraient encrasser les rouages de la réutilisation. Or, lorsque les agents préviennent certains usages jugés risqués, ils ne considèrent pas les entités effacées comme un rebut, mais comme des objets sensibles qui pourraient avoir des conséquences néfastes pour leur carrière, le service ou le projet d'*open data*.

Pourquoi rendre intelligibles les données ? Les bases de données et les tableurs des administrations abondent d'informations formulées dans le vocabulaire de travail des producteurs de données. Les écrits professionnels (Pène, 1995) sont pour une grande part des écrits abrégés (Fraenkel, 1994) : toute organisation repose sur des formes langagières

réduites par lesquelles des entités sont identifiées de manière opératoire. Ces abréviations sont rarement indexées dans un glossaire ou dans des métadonnées, elles forment un savoir tacite (Collins, 1974) propre à un agent ou une branche de l'organisation. Or, ces idiosyncrasies sont traitées comme des brèches à réparer dans le processus d'ouverture des données.

> Au départ, la plupart des systèmes qu'on a achetés chez nous ne sont pas du tout conçus pour faire de l'*open data*. Donc, c'est compliqué, on est obligé de développer des moulinettes pour sortir des données proprement. […] À quatre-vingt-dix pour cent ce sont des données purement techniques avec par exemple, les libellés commerciaux au lieu que ce soit marqué « onze stade rennais » c'est marqué « onze STRE » par exemple. Parce que c'est un code qui suffit largement quand les départements concernés conçoivent les horaires, « onze STRE », ils savent à quoi ça correspond. Le voyageur, ça ne lui parle pas du tout, donc il a fallu pouvoir croiser certaines bases chez nous qui ont les bons libellés. Aujourd'hui, pour construire le lot GTFS, on croise avec six à sept bases.
> (N.N, Technicien informatique, Keolis Rennes)

Le croisement de plusieurs bases de données peut parvenir à rendre intelligibles ces termes difficilement interprétables par les usagers des données et des applications les réutilisant. Ici, il a fallu croiser des données de gestion comprenant des abréviations et acronymes singuliers avec une base dédiée à l'information voyageur pour créer le fichier des arrêts exigés par le standard GTFS. Mais cette méthode s'applique dans un contexte organisationnel où l'information est produite par différents services et selon différentes finalités. Elle exige aussi de mettre en correspondance les deux bases, un croisement qui n'a pu se faire que par la connaissance approfondie des pratiques de production de données de ce technicien dédié à leur maintenance.

Au-delà du contenu de chaque valeur, la mise en intelligibilité des données mène les agents à repenser des catégories qui étaient jusqu'alors solidement ancrées dans les pratiques de travail. Une catégorie comme un espace vert prend un sens totalement différent lorsque les données ne s'adressent plus uniquement aux agents en charge de leur entretien. Les responsables de projet d'*open data* peuvent imaginer les usages pour détecter les incompréhensions possibles lors de la réutilisation des données.

> Le jeu de données qui était extrait était un jeu de données général sur tous les espaces vers de la ville. Par exemple une jardinière, vous savez celles qui sont sur les trottoirs, une jardinière qui fait une certaine taille est dite espace vert. Donc on trouve ça débile de mettre dans un jeu de données un parc comme le parc Montsouris et les jardinières. Donc on a mis plutôt : parcs et jardins de la ville, les squares, les jardinières… Par type de ce que nous on appelle espace vert, on avait fait un jeu de données.
> (D.L., correspondant ***open data***, service informatique d'une direction, ville de Paris)

En changeant la finalité des données, les agents évaluent l'intelligibilité des catégories dans les contextes d'utilisation qu'ils préfigurent. Dans ce cas précis, ce correspondant souhaite le développement d'applications mobiles pour les habitants et les visiteurs afin de localiser l'espace vert le plus proche. Or, dans une telle application, la présence d'une jardinière constituerait une aberration pour l'usage qu'il envisage. Impossible de se détendre ou de se promener dans un espace aussi réduit que celui d'une jardinière. La présence de ces espaces réduits dans la base des espaces verts n'est pas une erreur à corriger, elle rappelle que les ontologies varient selon l'orientation des données : pour les jardiniers, un espace vert désigne tous les lieux de végétation à entretenir ; pour un habitant, c'est un espace d'agrément mis à disposition par la ville. Configurer les données pour certaines utilisations demande souvent de repenser les catégories employées dans les fichiers de gestion des administrations et instaure de nouvelles catégories. Lors de l'édition, les agents peuvent aussi être amenés à faire le tri dans les variables en fonction de leur intelligibilité pour les publics vers lesquels les données sont configurées. Les champs qui comportent des informations « métier », servant à la description ou à la coordination du travail de gestion, sont jugés inutiles et exclus de l'extraction.

> On voulait vraiment partir de notre base de métier et essayer de faire un jeu de données de qualité. La qualité, ça passait par enlever des champs qui ne servent à rien. […] Le jeu de données sur les espaces verts, il y a 50 colonnes et il y a bien 35 qui ne servent à rien. Globalement, les attributs qu'on a voulu communiquer, c'était la surface du jardin, la surface aquatique dans le jardin, les informations d'accessibilité, les ouvertures, la dénomination, le nom de l'espace vert et le numéro d'identifiant. Et encore l'identifiant, je ne crois pas qu'on le publie.
> (D.L., correspondant ***open data***, service informatique d'une direction, ville de Paris)

La mise en intelligibilité peut aussi concerner l'ordre des colonnes. En changeant d'univers de sens, cet ordre peut être bouleversé pour refléter les nouveaux impératifs vers lesquels

les données sont orientées. Lorsqu'elles passent entre les mains d'un agent en charge de la communication, l'ordre des colonnes peut devenir un facteur de la lisibilité du message de transparence que la ville tente de faire passer en ouvrant les données sur les logements sociaux qu'elle finance.

> Je remets les colonnes dans l'ordre qui moi me parait plus pertinent pour les gens. Pour que ce soit un peu plus lisible, je mets d'abord l'adresse, le bailleur et puis aussi l'année d'agrément. D'ailleurs après, je crois que j'ai appelé ça « année de financement » plutôt qu'« année d'agrément », ça revient à peu près au même, mais c'est un terme un peu plus clair.
> (B.N., responsable communication d'une direction, une ville)

Ces données sur les logements sociaux sont d'abord produites afin de gérer les attributions et le financement des logements sociaux. Le responsable de leur ouverture décide de faire figurer d'abord l'adresse, car il considère que ces données vont servir à localiser les logements sociaux financés par la ville à proximité des usagers. En éditant les données, il préfère aussi modifier un titre de colonne écrit dans le langage vernaculaire des politiques publiques. La préfiguration des usages éprouve l'intelligibilité des abréviations, des catégories, de la structure ou du contenu même des fichiers de gestion employés au quotidien des administrations. Dans bien des cas, les agents en charge de leur ouverture considèrent l'intelligibilité de ces informations comme une condition essentielle de la qualité des données. Mais l'édition des données n'est pas uniquement guidée par un souci de lisibilité et de compréhension des données par les publics qui vont les utiliser. Les gestionnaires transforment aussi les données lorsqu'ils évaluent les risques qui peuvent survenir de la diffusion et la réutilisation des informations. Les gestionnaires les effacent lorsqu'elles peuvent compromettre leur carrière, leur service, le bon déroulement des missions de service public ou encore leur hiérarchie. Cette évaluation des risques, qui débute lors de la sélection des données à ouvrir, se poursuit lorsque les agents vérifient à même les données les risques des informations que contiennent les fichiers à ouvrir. Contrairement aux cas précédents, les informations qui sont éditées ne sont pas perçues comme une erreur ou une aberration qu'il faudrait corriger. Elles sont perçues comme une faille dans la préfiguration des risques qui avait été opérée dès leur sélection.

Certaines données sont ainsi passées au crible de l'évaluation des risques avant leur publication. Un gestionnaire de données m'a, par exemple, montré en détail les

transformations qui étaient apportées à des données de leur extraction jusqu'à leur ouverture. Les données, qui portent sur la localisation et le financement des logements sociaux, étaient déjà diffusées sous forme de fichiers PDF suite à un engagement électoral de transparence. Cet engagement a précédé le projet d'*open data* de la ville qui a changé les conditions de leur diffusion. Au format PDF, les possibilités de leur traitement automatique étaient réduites, il fallait prévoir plusieurs heures de travail pour retranscrire ou convertir ces informations avant de pouvoir les exploiter. Dans un format lisible par les machines, les données sont configurées pour être manipulées et transformées par des outils informatiques. Ce changement de format a incité les gestionnaires à reconsidérer la sensibilité de ces données et à éditer encore plus qu'auparavant les informations.

> On fait une requête de la base Access et après, on enlève les colonnes qu'on veut et on donne ce qu'on a bien envie de donner. C'est vrai que quand [VC, responsable de la communication] m'a dit que c'était mis sous format Excel, ça nous a posé des questions. Sous format Excel, les gens peuvent en faire ce qu'ils veulent ce qui n'était pas le cas avec le PDF. Et les infos n'étaient pas toujours à jour. C'est pour ça que j'ai, notamment, sur notre donnée « date de livraison », ça peut être délicat par rapport aux mairies d'arrondissements si la date est fausse. Aussi par rapport aux associations qui cherchent des immeubles à squatter, même si généralement, ils les connaissent.
> (S.N., gestionnaire de données, une ville)

Pour réduire les risques politiques de l'ouverture des données, les gestionnaires de données ont effacé un champ, celui de la date de livraison qui figurait autrefois dans les fichiers PDF. L'édition a servi à réduire les risques que ces données renforcent le contrôle des mairies d'arrondissements ou permettent aux associations de se servir des immeubles en réhabilitation comme logement temporaire. Mais l'exclusion de ce champ ne constitue que la dernière des transformations qui sont appliquées aux données avant leur ouverture. Avant d'être transmises pour leur diffusion, les données sont extraites de la base de données qui sert au suivi du financement et de la construction des logements sociaux. À une occasion, le responsable de la communication a pu recevoir un « export brut » qui comprend l'intégralité des champs de la base de données. Lors de notre entretien, il retrouve le fichier « brut » qu'il compare avec l'extrait qu'il reçoit de la part des gestionnaires de données.

> [Le service du financement], c'est eux qui ont les données brutes. Une fois, j'ai reçu un fichier avec plus d'informations. Désolé je ne le retrouve pas. [Il cherche dans son

> ordinateur] tiens celui-là ! il est tel quel, il n'est pas touché ! C'était au tout début, le fichier brut. La typiquement, numéro unique, je ne sais pas à quoi il fait référence. Ensuite, bailleur bah ça c'est resté. L'arrondissement, l'adresse, mode de réalisation pareil ça reste, l'année de financement. Par contre, après, il y a plus de détails sur la sorte de financement, l'année de livraison, le mois de livraison, la livraison prévisionnelle, le mois de livraison prévisionnel…
> (B.N., responsable communication d'une direction, une ville)

Après l'entretien, j'ai eu accès à ce fichier qui comprend les données et la liste des variables pour les deux versions. L'export brut et la version publiée sur le site de la ville y figurent dans deux onglets (figures 39 et 40). La majorité des colonnes sont exclues du fichier en vue de la publication sur le site de la ville, les données effacées concernent les opérations en cours de réalisation et celles qui ont été abandonnées.

Liste des variables de l'onglet "Livraisons 010109 complet"	
N° unique ordre 2006	Ces deux variables constituent la clé unique de chaque opération Pour les opérations agréées avant 2001, le numéro unique n'a pas été complété
Bailleur	Nom du bailleur
Adresse	Adresse de l'opération
Mode de réalisation selon agrément initial	
Nature du programme	Logements familiaux, Foyers, etc.
Année d'agrément	Année d'agrément
PLA TS ag	Logements agréés
PLA LM ag	
PLA ag	
PLA I ag	
PLUS ag	
dont PLUS CD ag	
PLS ag	
TOTAL ag	
PLA TS liv	Logements livrés
PLA LM liv	
PLA liv	
PLA I liv	
PLUS liv	
PLS liv	
TOTAL liv	
Année liv	Année de livraison
Mois liv	Mois de livraison
Année liv prev	Année de livraison prévisionnelle
Mois liv prev	Mois de livraison prévisionnelle
PLA TS modif	Variation entre le nombre de logements agréés initialement et le nombre de logements effectivement livrés
PLA LM modif	
PLA modif	
PLA I modif	
PLUS modif	
PLS modif	
TOTAL modif	
PLA TS abandon	Nombre de logements abandonnés (négatif)
PLA LM abandon	
PLA abandon	
PLA I abandon	
PLUS abandon	
PLS abandon	
TOTAL abandon	
Année liv - agr	Année de livraison - Année d'agrément
Année liv prev - agr	Année de livraison prévisionnelle - Année d'agrément
Enquete 012009	Opération enquêtée en janvier 2009
Observations	Commentaires divers
Commentaires 1	
Commentaires 2	

Figure 39. Liste des variables de l'« export brut » du fichier des logements sociaux.

Liste des variables de l'onglet "Livraisons 010109 pour site de la ville	
Adresse	Adresse simplifiée
Bailleur	Bailleur
PLA I	Nombre de logements PLA I (Agréments corrigés = Agréments initiaux + modifs)
PLUS	Nombre de logements PLUS (Agréments corrigés = Agréments initiaux + modifs)
PLS	Nombre de logements PLS (Agréments corrigés = Agréments initiaux + modifs)
TOTAL	Nombre de logements TOTAL (Agréments corrigés = Agréments initiaux + modifs)
Année d'agrément	
Année de livraison	
Année de livraison prévisionnelle	
Nature du programme	

Figure 40. Liste des variables sélectionnées pour l'ouverture du fichier des logements sociaux.

À l'inverse de la base de données de gestion qui décrit l'avancement du financement et de la construction, le fichier publié donne une vision statique des logements sociaux. Les données ouvertes ne concernent que des opérations déjà décidées et mises en œuvre, celles sur lesquelles le public ne peut plus intervenir. Tout ce qui figure dans les commentaires, qui sert à coordonner les acteurs en charge des opérations, est effacé du jeu de données. On y trouve par exemple une correction du nombre de logements financés suite à une enquête, le transfert d'opération d'un bailleur à un autre ou le suivi des travaux de certains prestataires. Avant l'édition, les données servent à la coordination et au suivi. D'outil de travail dynamique, elles deviennent une liste statique qui établit l'emplacement et les conditions des logements financés par la ville. On peut interpréter cet exemple d'un point de vue purement politique : l'édition des données a permis d'écarter les informations qui pourraient servir à contrôler la construction ou le financement des logements sociaux avant sa mise en œuvre. L'édition aurait donc servi à euphémiser ces données et à réduire leurs usages possibles par des opposants. Mais on peut aussi avoir une lecture plus organisationnelle de ce cas. Il montre aussi que le « risque » est attaché à des aspects dynamiques de la base de données, qui sont en cours de mise en œuvre. Une fois publiées, ces informations pourraient être interprétées comme des choses stables, définitives et perdre du flou intrinsèque à cet état. Par rapport à la loi CADA, les agents administratifs ont même le droit de refuser la publication d'informations préparatoires à une décision afin de ne pas « paralyser l'action administrative[110]. »

[110] L'article 2 de la loi du 17 juillet 1978 dispose que le droit à communication ne concerne pas les documents préparatoires à une décision administrative « tant qu'elle est en cours d'élaboration ». […] Le site de la CADA précise ce point de la sorte : « La CADA subordonne la communication des documents préparatoires à l'intervention de la décision qu'ils préparent (conseil n° 20073363 du 13 septembre 2007),

Outre la réduction des risques politiques de contestation, les données peuvent être éditées pour répondre à l'obligation légale d'effacement des données personnelles. Leur divulgation constitue un point sensible de l'ouverture des données publiques, car la loi CADA de 1978, sur laquelle les politiques d'*open data* sont bâties, exclut les données personnelles des informations publiques pouvant être diffusées par les agents. Pour qu'elles soient publiées et réutilisables, les données comportant des informations personnelles doivent soit faire l'objet du consentement des personnes soit être anonymisées. Or, l'effacement des informations nominatives ne suffit pas toujours à anonymiser une base de données. Les identifiants uniques de chaque usager peuvent servir à réidentifier les individus dans une base de données ne comportant aucun nom. Pour garantir l'anonymat, il est donc souvent suggéré d'introduire des identifiants aléatoires pour empêcher de tels cas de réidentification[111]. Comme les risques juridiques de publier des informations personnelles, même anonymisées, sont importants, les agents s'imposent la plus grande prudence lorsqu'ils éditent des données publiques avant leur ouverture. Bien souvent, ces données sont même exclues d'emblée du périmètre des données ouvrables. Lorsque les agents obtiennent tout de même l'autorisation d'ouvrir de telles données, ils prennent les plus grandes précautions lors de l'édition des données. C'est le cas d'un des jeux de données les plus réutilisés sur les portails *open data* des villes, les prénoms des nouveau-nés. Pourtant, ce sont des données agrégées qui indiquent le nombre de fois que chaque prénom est attribué dans la commune. Mais les agents en charge de son ouverture craignent qu'on puisse identifier des individus à partir des prénoms attribués à peu d'enfants. Dans la crainte de divulguer des informations personnelles, les agents recherchent des recommandations pour ouvrir, sans risques, ce jeu de données très réclamé par les réutilisateurs. Un acteur

notamment lorsqu'une information précoce risquerait de paralyser l'action administrative en mettant trop tôt sur la place publique des éléments d'information qui, nourrissant l'action du responsable d'une décision, peuvent le faire hésiter entre plusieurs solutions avant de prendre parti. » [Cada, « Le document ne doit plus être préparatoire », http://www.cada.fr/le-document-ne-doit-plus-etre-preparatoire,6135.html, consulté le 30/06/2016.]

[111] En France, en 2014, un rapport du Sénat (rapport d'information n° 469) portait sur les risques pour la vie privée de l'ouverture des données publiques. Il a signalé un seul cas d'ouverture de données pouvant donner lieu à réidentification. Elles concernent l'INSEE qui a publié des données socio- économiques dites carroyées, au niveau de carreaux de 200m de côté. Dans des zones de faible densité, il était possible de déduire les informations socio-économiques des habitants sans que les données les désigne. L'INSEE a depuis changé sa méthodologie.

associatif qui organise des ateliers avec ces données s'exaspère de la grande prudence avec lesquels les agents créent des règles d'anonymisation des données.

> Rennes a fait comme Nantes et apparemment Nantes a fait comme Paris. Mais Paris ne parle pas d'une recommandation de la CNIL, il dit juste « pour des raisons de respect de la vie privée. » Donc, j'ai trouvé celui qui a écrit ça à Paris […] je finis par avoir une interview avec lui. Je lui dis « c'est quoi cette histoire ? », il me dit « mes services d'état civil m'ont dit que l'INSEE recommandait de ne pas publier les prénoms attribués trois fois et moins », je lui ai dit « trois fois, pas six ? », il me dit « Oui, six mais j'ai pris ma petite marge. » Tu vois le truc ? Il entend ça et puis il dit « oui, je ne veux pas d'embrouilles, je prends ma petite marge. » Et les autres le reprennent et ça devient une recommandation de la CNIL.
> (D.V., Responsable associatif, Rennes)

L'anonymisation des données n'obéit pas à des règles précises bien que les agents craignent les répercussions politiques et juridiques de la divulgation d'informations personnelles. En l'absence de « bonnes pratiques » ou de la garantie que les données ne contiennent plus d'informations personnelles, les agents redoublent de précaution lorsqu'ils traient des données personnelles. Quand un doute persiste, les agents décident de les exclure du champ de l'ouverture des données publiques.

> J'ai un oui de principe du service cimetière. Il est entièrement d'accord pour que tous les plans de toutes les divisions et concessions, même avec la position des sépultures, soient publiés. On n'ira pas écrire la sépulture avec le nom, parce que je crois qu'on n'a pas le droit de le faire. […] Nous, il faut qu'on fasse attention à l'information qui est publiée. On avait décidé de s'arrêter là pour éviter d'avoir des problèmes juridiques. Après si ça se trouve, il n'y en a pas, et s'il n'y en a pas, peut-être qu'on le fera.
> (D.L., correspondant ***open data***, service informatique d'une direction, ville de Paris)

Du fait du risque de réidentification, les agents tendent à exclure du périmètre de l'ouverture les données concernées. Comme mon enquête intervient dans les premières années des politiques d'*open data*, les agents préfèrent éviter un tel risque pour ne pas remettre en cause les projets d'*open data* dans leur ensemble[112]. On le voit, les agents

[112] Fin janvier 2013, peu après l'alternance lors de laquelle l'avenir d'Etalab et de l'ouverture de données étaient parfois mis en doute, le sénateur socialiste Gaëtan Gorce publie un billet de blog suivi d'une question au gouvernement adressée à Fleur Pellerin, ministre déléguée à l'Économie numérique. Il demande au gouvernement « de stopper les développements de l'open-Data tant qu'un cadre juridique respectueux de la vie privée n'aura pas été arrêté » craignant le recoupement des données publiées pour

perçoivent des risques très variés lorsqu'ils éditent les données avant leur publication. Sans prétendre recenser ces risques, notons qu'ils répondent jusqu'alors à la préfiguration d'un mésusage des données que leur édition va tenter de corriger. Mais, dans certains cas, le risque ne porte pas sur le contenu des données, mais sur l'évaluation de leur qualité.

L'édition ne sert donc pas uniquement à garantir l'intelligibilité des données, elle peut servir de protection pour les agents lorsqu'ils perçoivent un risque juridique, politique ou communicationnel qui pourrait survenir de l'ouverture et de la réutilisation des données. Comme en science, ce travail manuel de façonnage certifie les données (Walford, 2013). L'édition garantit ainsi l'intelligibilité des données et assure que les risques non prévus lors des négociations avant l'ouverture ont été évalués et prévenus. C'est cette certification qui survient après que les données ont été soigneusement éditées qui les autorise à circuler dans de nouveaux réseaux sociotechniques où elles seront traitées et de nouveau transformées. L'étude du travail d'édition rappelle, là encore, que les données ne circulent pas dans de nouveaux réseaux sans frictions (Edwards, 2010 ; Edwards et al, 2011). Considérer que les données pourraient être ouvertes telles quelles, sans que leurs gestionnaires n'aient à les rendre intelligibles ou n'atténuent les risques de l'ouverture, procède du même raisonnement que lorsque les politiques publiques incitent les statisticiens à réutiliser les fichiers de gestion des administrations, car elles seraient une source « économique » pour les institutions statistiques. Or, la transformation des sources administratives demande un lourd investissement pour les statisticiens : « elle est enfin et surtout coûteuse, en argent, en temps de travail et en matière grise, ce qui relativise l'idée, qui reste répandue, que les sources administratives sont "économiques", en ce qu'elles éviteraient le coût du recueil initial, comme s'"il n'y avait qu'à se baisser pour les cueillir". » (Desrosières, 2005) Même si l'édition des données dans les cas étudiés ici ne répond pas à une exigence d'objectivité comme dans les sciences et les statistiques, le passage d'un monde de significations à un autre demande un lourd investissement pour les agents afin que d'une part leurs données soient effectivement utilisables et que d'autre part l'ouverture des données ne compromette pas leur carrière. L'absence de l'intégration de ces coûts dans les politiques publiques repose sur l'invisibilité du travail que je viens de décrire ici.

identifier les individus dans les données pouvant instaurer un « fichage généralisé. » Henri Verdier et son prédécesseur, Séverin Naudet, lui ont répondu dans Les Echos pour assurer qu'aucune donnée personnelle n'a été publiée sur data.gouv.fr. Ce cas rappelle la sensibilité politique des questions de protection de la vie privée pour les projets d'open data.

Conclusion

Au terme de ce chapitre, nous avons pu voir que le processus d'ouverture des données est ponctué de transformations. Dans les cas où les données sont traitées dans l'environnement d'un tableur, les agents administratifs doivent souvent convertir leurs données avant leur ouverture. En effet, des formats couramment utilisés comme le PDF ou le XLS sont déconseillés par les responsables de projet *open data* car ils réduisent la capacité des données à être utilisées par les machines. Mais le passage des données vers un format ouvert comme le CSV ne se résume pas à l'utilisation d'un convertisseur automatisé, les agents doivent souvent transformer leurs données en profondeur pour éviter la perte d'informations comprises dans la mise en forme (gras, italique, fond de couleur), dans des cellules fusionnées ou dans les onglets de leur tableur. Par ailleurs, des standards, comme le GTFS dans le domaine des transports, imposent des spécifications qui dépassent la structure des données. Dans l'optique d'une réutilisation « sans frictions » de données interopérables, ils réclament l'utilisation de définitions normalisées des objets qu'elles désignent et imposent une structuration très précise du contenu des données. Dans sa conception, ce standard demande un investissement coûteux dans la réutilisation des données par les machines. Il permet de cibler un public en particulier des développeurs dans le but de créer des applications au service de l'information voyageur. Enfin, au-delà des transformations imposées par les standards, les agents peuvent intervenir directement sur le contenu des données. Les opérations que j'ai regroupées sous le terme d'édition visent à assurer l'intelligibilité des données et à protéger les agents des risques juridiques, politiques ou communicationnels qui pourraient découler de l'ouverture. Toutes ces transformations sont tournées vers un usage nouveau des données « métiers » qui progressivement sont instaurées en des données ouvertes à de nombreux usages. C'est une véritable transmutation qui s'opère en coulisses, le changement d'une substance en une autre, des informations administratives qui sont progressivement instaurées en données ouvertes.

Deux points sensibles se dégagent de ces transformations. Premièrement, nous avons vu que la nouvelle orientation des données vers les publics soulève un problème central, celui de l'intelligibilité. Mais, au fil de l'exploration des opérations de transformation, nous avons pu voir se distinguer un double horizon de l'intelligibilité, vers les machines (et les développeurs qui les programment) et vers le grand public, les humains. Dans le premier cas, les standards et les formats sont l'opérateur essentiel de l'intelligibilité. Dans le second, c'est le travail d'édition qui permet de rendre compréhensibles par le plus grand nombre

les catégories et les termes employés dans les fichiers de gestion de l'administration. Ces deux horizons remettent en cause l'idée selon laquelle les données brutes seraient exploitables telles quelles, que leur ouverture ne sera pas coûteuse et ne demanderait pas de transformation. Par ailleurs, cette double orientation de l'intelligibilité rejoint la distinction que faisait Tim Berners-Lee dans sa conférence TED entre des données intelligibles par les machines et des documents compréhensibles par les humains. Il avait notamment formulé l'idée selon laquelle certaines données sur le web ne sont pas des données, car elles ne sont pas intelligibles par les machines : « *we haven't got data on the web as data*. » Pour certains, l'intelligibilité par les machines permettrait de distinguer les données de l'information[113] et pourrait tracer une frontière entre de « bonnes » et de « mauvaises » données. Par exemple, l'OKFN a lancé un projet intitulé « *Bad Data* » qui signale certains jeux de données, en explique les raisons de leur mauvaise qualité et en publie un correctif. Dans le cas d'un jeu de données sur nombre de passagers dans les transports publics londoniens, l'OKFN pointait du doigt un fichier publié un CSV, un format pourtant vanté pour sa lisibilité par les machines[114].

> *This is a CSV provided by data.london.gov.uk about Transport for London (TfL) passenger numbers. The problem is the CSV is so messy only a human could use it! What specifically is wrong?*
> *—The first column is missing a heading (one guesses this should be "date"?)*
> *—Dates are not of a recognizable format instead being of form: "2006/2007 - 1". One assumes this should be a month or similar (but its not entirely clear if these are months since 13 items in a year!)*
> *—Percentage sign written into percentage column*
> *—Large number of trailing blank rows and columns*

On le voit, au-delà des standards et des formats, l'intelligibilité des données par les machines s'affirme aussi par le biais de porte-paroles des non-humains, des « entrepreneurs de cause » (Cobb & Elder, 1972) qui définissent des critères et rendent publiques leurs revendications.

[113] Par exemple, lors de la numérisation des déclarations d'intérêt des parlementaires, Regards Citoyens avait publié un communiqué dans lequel l'association pointait les problèmes des données publiées sous la forme de fichiers PDF et considérait qu'elles ne sont pas « à proprement parler en Open Data. »[1] Par la suite, l'association avait publié un autre communiqué dans lequel elle distinguait des documents, les déclarations d'intérêt des parlementaires, et les données brutes des déclarations d'intérêts des parlementaires numérisées par les citoyens publiées sous la forme d'un fichier CSV

[114] Open Knowledge Foundation, « Bad data - Passenger Numbers for Humans Only », http://okfnlabs.org/bad-data/ex/tfl-passenger-numbers/, consulté le 25 novembre 2015.

À travers les préconisations officielles formulées dans le sens du développement de la « culture des données » ou de la « modernisation de l'administration », on voit se dessiner l'horizon possible d'un alignement des usages initiaux des données « métiers » pour qu'elles se conforment en amont, dans leur utilisation quotidienne par les agents, aux contraintes de la lisibilité des données par les machines. Par exemple, un responsable du projet d'*open data* de la région Ile-de-France évoquait la formation des agents à la « culture des données. » Son but à terme consisterait à intégrer dans la conception des fichiers les grands principes du format CSV pour éviter d'avoir à opérer des transformations à chaque mise à jour des fichiers. À Rennes, les agents de Keolis Rennes m'ont indiqué mettre en œuvre une refonte du système d'information de gestion des horaires de bus pour uniformiser le nom des arrêts ou respecter la mise en forme des horaires imposée par le standard GTFS. Or, nous avons pu voir qu'il y a de « bonnes raisons » d'utiliser le formatage ou un code couleur dans un tableur, des fonctionnalités qui favorisent la spatialisation de l'information et facilitent la manipulation des données. De même, un horaire comme 25h30, s'il parait incompréhensible au premier abord, a un sens dans le cadre de la gestion des ressources humaines d'un réseau de bus. Les standards mettent ainsi à l'épreuve l'orientation des données et interrogent sur les conséquences de l'instauration des informations administratives en données ouvertes. Restent-elles d'abord des outils de gestion ? Ou faut-il considérer que c'est leur qualité de données ouvertes qui prime sur les activités de gestion quotidiennes de l'administration ? Les situations de réorganisation interne, que j'évoque ici comme une suite possible des politiques d'ouverture de données, montrent qu'il existe deux grandes directions possibles pour prendre en considération le travail de fabrication des données brutes. Une fois ce travail éprouvé, puis reconnu, c'est-à-dire une fois que l'on assume que l'ouverture des données a un coût, qu'elle représente même un investissement, on peut l'assumer comme une série d'opérations à mener a posteriori sur les données métier. Il faut alors inventer des postes, comme celui de *data editor* qui a été créé en 2013 dans l'équipe d'Etalab, et redéfinir des rôles au sein de l'organisation. On peut au contraire chercher à intégrer ce travail en amont, en transformant la nature même des données sur les sites de leur production et dans leurs premiers usages. La différence entre les deux directions ne tient pas tant à la part organisationnelle de la fabrique des données brutes [elle est présente à chaque fois], mais à la définition sous-jacente de ce que l'on entend par données. Dans le premier cas, la multiplicité des données et la nécessité d'en faire coexister des versions différentes au sein

de l'institution sont assumées. Dans le second cas, le caractère générique des données leur aspect « brut » est considéré comme un bien en soi, sur lequel il faut aligner les idiosyncrasies professionnelles.

L'inversion du travail de transformation des données, de l'aval vers l'amont de l'ouverture, prend le risque de créer le même type de situations que Garfinkel et Bittner (1967) décrivent dans leur article, des situations marquées par le décalage entre des registres de pertinence peu compatibles, celui de la gestion des soins versus celui de la recherche en sciences sociales dans leur cas. Pour l'ouverture des données, le décalage entre le registre de pertinence des activités de gestion de l'administration et celui de l'ouverture de données interroge en profondeur les publics de ces données ouvertes. Que ce soit dans les grands principes de l'ouverture ou dans les politiques publics d'*open data*, ces publics sont définis dans des termes très larges comme des « communautés » de réutilisateurs, de développeurs, d'innovateurs, de porteurs de projet ou encore de *civic hackers*. Dans le processus d'ouverture, nous avons pu voir que les usagers sont rarement présents, tout au plus sont-ils imaginés comme on a pu le voir dans les cas où l'identification et l'édition des données sont guidées par la préfiguration des usages. La production de données intelligibles peut s'avérer extrêmement délicate sans la présence d'usagers avec lesquels négocier localement tel ou tel aspect, discuter de la lisibilité d'une catégorie ou s'accorder sur la structure des données.

Au-delà même de la question de l'intelligibilité, ce sont les politiques d'*open data* qui sont entièrement fondées sur l'existence de publics de données formulant une demande de réutilisation. Tim Berners-Lee a appelé le public de la conférence TED à crier « *we want raw data* », mais, passée l'exhortation, la demande de données brutes peut se révéler beaucoup moins criante pour ceux qui ouvrent des données. Après avoir instauré des données, les agents administratifs et les responsables de projets d'*open data* doivent aussi souvent instaurer des publics pour ces données. C'est l'objet du chapitre suivant dans lequel nous allons explorer trois instruments les interfaces de visualisation, les métadonnées, les concours qui font exister les publics de données à l'origine des politiques d'*open data*.

Chapitre 6

L'instauration des publics de données

Faisons un dernier crochet par la réunion Open Data Bootcamp de la région Île-de-France. Au cours de la présentation de la démarche d'*open data*, à plusieurs reprises, les organisateurs sont interrompus par les remarques de certains participants. Sans revenir sur l'ensemble de ces objections, elles sont particulièrement fortes après qu'un des organisateurs évoque les actions que la région conduit pour encourager la réutilisation de ses données. Laurent, un des organisateurs, présente un évènement dédié aux développeurs et aux porteurs de projet, un « hackathon » organisé par la région qui a joué un rôle déclencheur dans le lancement du projet d'*open data*.

> Il y a des formes incitatives sur la réutilisation qu'on peut déployer, on l'a déjà fait : en fait le hackathon[115] qu'on a organisé en mars à l'IAU [Institut d'Aménagement et d'Urbanisme] qui portait sur le schéma directeur d'aménagement de la région. […] Ça a permis d'accélérer la démarche d'ouverture des données et ça a produit immédiatement un démonstrateur sur la capacité des réutilisateurs à traiter des jeux de données et à produire des choses avec.

À la suite, un participant prend la parole et objecte : « ça marche certainement très bien, mais l'*open data* n'a pas les effets souhaités au niveau de la réutilisation. Les expériences de Rennes ou de Nantes qui devaient découler sur des créations d'applications sur iPhone ont été assez décevantes. » Pierre, un autre organisateur, lui répond en citant plusieurs exemples d'innovations tels que le GPS qui, après plusieurs décennies d'expérimentation soutenues par les pouvoirs publics, ont créé une véritable filière économique. Pour lui, les politiques d'*open data* doivent soutenir financièrement les premières initiatives qui réutilisent les données ouvertes : « Là où ça [l'innovation] se joue, c'est sur notre capacité en tant qu'acteur public à stimuler l'innovation sur notre territoire. Elle se crée parce que nous mettons à disposition des moyens et des données pour nos citoyens. »

[115] Contraction de hacking et de marathon, un hackathon est un évènement compétitif de durée limitée lors duquel les participants, souvent des développeurs, des designers et des entrepreneurs, développent des projets. Un jury évalue généralement la qualité des réalisations puis attribue une récompense symbolique et/ou financière aux meilleurs projets.

Plus tard dans la réunion, de nouvelles objections sont formulées par les participants quand François, un autre intervenant, présente le portail *open data* de la région et montre aux participants une de ses fonctionnalités, la visualisation des données sous la forme de cartes. Il utilise l'exemple d'un jeu de données tiré de data.gouv.fr, les équipements sportifs en France, qui a été republiée par la région en incluant uniquement les installations du territoire. Un participant interrompt l'exposé et objecte : « on sort un peu de l'objectif de l'*open data* parce que vous retraitez de la donnée et vous la mettez en forme, ça devient une application. » Laurent lui répond : « on a des outils qu'on met à disposition pour des gens qui n'ont pas les compétences de produire ses réutilisations par leurs propres moyens […] Mais vous avez raison de dire que c'est un début de réutilisation de la donnée brute. » François complète cet argumentaire : « Sur le portail, on a vraiment tenu à ce qu'on ait de la visualisation vis-à-vis du citoyen, car l'*open data* brut n'est pas lisible par n'importe qui. » Une participante réagit et interrompt les organisateurs : « on en revient à la question du début sur "qu'est ce que c'est qu'une donnée brute ?" et "est-ce qu'on peut diffuser de la donnée brute ?" Vous voyez bien qu'avec ces outils de prévisualisation, ce n'est pas si évident que ça d'utiliser de la donnée brute. » Laurent réagit : « mais on ne dit pas que c'est évident » puis la discussion part sur le contenu du fichier des équipements sportifs.

Ce court récit, qui rassemble des objections exprimées à plusieurs moments de l'évènement, montre que le travail d'ouverture ne s'arrête souvent pas à la publication des données sur les portails d'*open data*. On l'a vu dans le premier chapitre, l'existence d'un public pour les données est un présupposé qui sous-tend les définitions de l'*open data*. C'est le sens de la conférence TED de Tim Berners-Lee qui demande à l'auditoire de constituer un public qui réclame des données brutes. Pareillement, Rufus Pollock, dans son billet « *Give us the data raw, give us the data now* » considère que dès lors que les données brutes sont publiées, il ne fait aucun doute qu'elles seront réutilisées : « *many interfaces can be written to that data (and not just a web one) and it is likely (if not certain) that a better interface will be written by someone else (albeit perhaps with some delay)*. » Mais pour les responsables de projet d'*open data* qui ont aussi souvent pour mission d'encourager et de stimuler la réutilisation des données, l'existence des publics n'a rien d'un donné. Dans les extraits précédents, l'équipe en charge de l'*open data* de la région Ile-de-France a dû « démontrer » l'existence d'un public pour les données à travers un hackathon. Elle organise des évènements, encourage parfois

financièrement la réutilisation des données et propose des outils pour permettre de visualiser les données brutes sans avoir des compétences techniques avancées.

En effet, on a pu voir que les projets d'*open data* comme celui du gouvernement français sont évalués en fonction du nombre de réutilisations. Pour Etalab, ses objectifs en terme de réutilisation sont même inscrits dans le projet de lois de finances. Les agents d'Etalab doivent « faire vivre » leurs données, montrer qu'elles sont utiles et utilisées[116]. La réutilisation est aussi évaluée par des organisations non-gouvernementales comme la Web Foundation créée par Tim Berners-Lee qui publie chaque année l'Open Data Barometer, un classement par pays de l'ouverture des données qui comporte trois critères principaux : *readiness* qui mesure les politiques de liberté de l'information et d'expression en place dans chaque pays ; *implementation* qui évalue l'avancement des politiques d'*open data* ; *impact* qui apprécie l'utilisation des données ouvertes. Sans impact, sans réutilisation des données, un gouvernement ne peut pas être bien classé dans l'Open Data Barometer, un des principaux outils d'évaluation externe des politiques d'ouverture de données. Ce problème ne se pose pas uniquement à l'échelle des gouvernements, mais aussi au sein des collectivités locales qui mettent en place des politiques d'*open data*. Par exemple, à Montpellier et à Rennes, les projets d'*open data* étaient assignés à des objectifs en terme de création de services. Par ailleurs, pour certains agents, si les données ne trouvent pas de public, il n'y a pas de raison d'accomplir le travail conséquent que demande leur ouverture. Par exemple, lors de la réunion Open Data Bootcamp, un correspondant du réseau de la région Ile-de-France demandait qu'on analyse les statistiques de téléchargement pour éviter d'avoir à ouvrir des données qui ne sont pas utilisées. Il a demandé par ailleurs que les usagers s'enregistrent avant de télécharger les données, remettant en cause le sixième principe formulé à Sebastopol intitulé « *non-discriminatory* » qui demande que les données soient accessibles à tous sans inscription préalable.

> Il faut un mécanisme d'analyse des statistiques d'utilisation parce qu'on risque d'avoir à maintenir des jeux de données ce qui est coûteux alors qu'en fait, ça se

[116] L'association Open Knowledge Foundation France a été associée à la candidature de la nouvelle version de data.gouv.fr aux awards de l'Open Governement Partnership. Le dossier de candidature rédigé par les agents d'Etalab, qui nous a été soumis en tant que partenaire issu de la société civile, comporte une phrase qui illustre bien les objectifs assignés à leur travail : « Data is valuable if it's being used, and more than a 1000 reuse examples of data reuse have been posted. » Une donnée n'a donc d'intérêt que si elle est utilisée, j'aurai l'occasion d'analyser cette phrase plus en détail dans la conclusion de cette thèse.

> trouve, ça ne sert à rien. Donc il faut qu'on puisse avoir de l'information sur ce qui est fait et par qui. Je préférerais un modèle d'***open data*** « identifié », où on demande un mail pour s'assurer d'où l'information. Sinon, c'est un miroir aux alouettes, on va produire des tas de données, on va tous se contraindre à faire des choses parce qu'il faut produire des indicateurs pour la région et, au final, ça risque nous couter plus cher avec un faible rendu. Donc il faut pouvoir mesurer l'effet de tout ce qu'on rend disponible via l'***open data***.

Si la preuve de l'utilité des données n'est pas fournie, c'est tout l'édifice de l'ouverture qui semble progressivement s'effondrer. Les responsables de projets d'*open data* ne doivent pas seulement instaurer ces données, mais aussi souvent instaurer leurs publics. Le vocabulaire de l'instauration est aussi particulièrement utile ici, il évite de laisser croire que les publics de données sont une pure création. Comme l'a montré Vinciane Despret (2015) sur un tout autre sujet, instaurer ne consiste pas à tirer un être du néant, mais à le « mener à l'existence » et à l'aider à devenir ce qu'il est. L'instauration permet de rendre compte des différentes manières par lesquels les responsables de projets d'*open data* contribuent à faire exister des publics multiples qui se lient aux données.

Ce sixième et dernier chapitre s'intéresse donc aux instruments qui instaurent les publics des données ouvertes. On aperçoit dans les extraits précédents des instruments[117] qui, chacun à leur manière, configurent les publics des données ouvertes. J'en évoque trois en particulier : les métadonnées, la visualisation des données et les concours de réutilisation. Lorsqu'ils tentent de favoriser la réutilisation des données, les agents peuvent aussi miser sur l'utilisation de métadonnées, mais leur exactitude et leur exhaustivité ne suffisent pas à atténuer les frictions qui accompagnent la réutilisation des données. En présentant les données directement sous la forme de tableaux, graphiques ou cartes, certains portails tentent de réduire les frictions de la réutilisation pour des publics n'ayant pas les compétences techniques d'ouvrir et exploiter les fichiers. Pour ceux qui ont en charge la gestion des données, ces fonctionnalités apportent de nouvelles contraintes dans le processus de l'ouverture en intégrant dans les portails des interprétations spécifiques des standards et en réclamant de nouvelles transformations des fichiers. Enfin, les projets d'*open data* donnent souvent lieu à l'organisation de concours qui incitent, de manière financière ou symbolique, les développeurs et les entrepreneurs à réutiliser les données sous la forme

[117] Par instrument, j'entends un « dispositif sociotechnique qui organise des rapports sociaux spécifiques entre la puissance publique et ses destinataires. » (Lascoumes & Le Gales, 2005)

de services et d'applications. Les assemblages sociotechniques qui en découlent ne parviennent généralement pas à se maintenir. Ils peuvent toutefois servir en interne à justifier l'existence d'un public pour les données ouvertes, un des présupposés qui fondent les politiques d'*open data*.

Les métadonnées : réduire les frictions de l'ouverture et de la réutilisation

Au-delà de l'édition que j'ai abordée dans le chapitre précédent, il existe un moyen d'améliorer l'intelligibilité des données sans les transformer directement : la production de métadonnées. Ce sujet a déjà été grandement étudié par les *Infrastructure Studies* qui proposent des ressources essentielles pour mieux comprendre les enjeux de la production de métadonnées (Baker & Bowker, 2007 ; Millerand et al, 2009 ; Zimmerman, 2008 ; Edwards, 2010 ; Edwards et al, 2011). Ces travaux montrent que la réutilisation des données implique un travail complexe de coordination qui passe souvent par des interactions directes avec leurs producteurs initiaux. Courriers électroniques, coups de téléphone et réunions sont des ressources essentielles pour clarifier le contenu d'un jeu de données. Comment avez-vous mesuré exactement ? Où se trouvaient les sondes ? Pourquoi y a-t-il une valeur étonnante ici ? Et pourquoi manque-t-il des valeurs sur cette colonne ? Pourquoi les unités de mesure changent-elles entre ces deux années ? Ces questions qui peuvent paraitre triviales sont en fait vitales à la réussite d'un projet fondé sur le partage de données. Dans l'idéal d'une science universelle et globalisée, fondée sur la transmission fluide de données, ces échanges ne sont pas considérés comme complètement satisfaisants. Dans de nombreux projets de collaboration scientifique à grande échelle, la solution n'a pas consisté à institutionnaliser ces interactions en face à face observées par Edwards et ses collègues (2011), ni à assumer le coût que représentent ces ajustements collectifs, mais à investir dans de nouvelles données, complémentaires aux données principales, afin de minimiser le recours aux échanges interpersonnels. Ces données, appelées « métadonnées », sont censées apporter toutes les informations nécessaires à la compréhension et l'appropriation des données initiales. Leur efficacité est assurée, encore une fois, par un investissement fort dans leur standardisation (Millerand et al, 2009). En quelques années, des domaines très variés ont vu ainsi naître un nombre considérable de standards de métadonnées, devenus le véritable Graal de la collaboration scientifique internationale.

> *Extensive, highly structured metadata are often seen as a holy grail, a magic chalice both necessary and sufficient to render sharing and reusing data seamless, perhaps even automatic.* (Edwards et al., 2011, p. 672).

Sans surprise, Edwards, Mayernik, Batchelle, Bowker et Borgman montrent qu'aux *data frictions*, les métadonnées envisagées comme seules ressources nécessaires au partage efficace de données scientifiques, ne font qu'ajouter des *metadata frictions*. Le fantasme d'un langage transparent et complet se traduit donc, dans les situations concrètes, par un travail supplémentaire d'ancrage, de contextualisation, un coût que les scientifiques peinent à absorber, d'autant plus qu'il n'est pas calculé dans les budgets alloués à ces projets (Edwards, 2010).

Dans les projets d'*open data*, la production de métadonnées fait partie des préconisations officielles qui fondent les politiques d'ouverture de données. Par exemple, l'annexe technique de la charte du G8, dans son deuxième principe « *Quality and Quantity* » considère la production de métadonnées complètes et fiables comme un des instruments essentiels par lesquels les usagers peuvent parvenir à s'approprier les données : « *We will: use robust and consistent metadata (i.e. the fields or elements that describe the actual data); […] ensure data are fully described, as appropriate, to help users to fully understand the data*. » Au niveau national, les recommandations formulées par Etalab dans son vademecum de l'ouverture des données envisagent les métadonnées comme un outil essentiel pour que les usagers découvrent la « bonne » donnée parmi la masse de fichiers compris sur les portails d'*open data* et parviennent à l'utiliser.

> La qualification des métadonnées et l'indexation sont une étape essentielle pour faciliter la réutilisation des données publiques. Les données sont très difficiles à retrouver si elles ne sont pas indexées et elles sont difficilement réutilisables si elles ne sont pas décrites avec précision.
> Ces informations complémentaires décrivant les données sont appelées « métadonnées ». Etalab propose ainsi des champs de descriptions normalisées à tous les producteurs de données publiques afin de leur permettre de spécifier le contexte et le contenu des données. Il leur est notamment demandé de caractériser leurs données (titre, description, mots clés…) en répondant aux questions suivantes : Qui a produit les données ? Quand les données ont-elles été produites ? Quelle est la période temporelle concernée ? Quelles sont les zones géographiques couvertes ? Quelles sont les thématiques des données ?
> Par ailleurs, pour faciliter la réutilisation la plus large possible des données publiques, Etalab recommande que tout jeu de données soit accompagné d'une description du contenu du jeu de données. Ce document annexe peut se révéler très important pour les réutilisateurs.

Au milieu de l'extrait précédent, on trouve une liste de questions qui correspondent aux champs proposés par le portail pour décrire les métadonnées. Ces champs sont liés à un standard de métadonnées, le Dublin Core Metadata Initiative (DCMI), qui réclame le remplissage d'un certain nombre de champs standardisés tels que le nom du jeu de données, sa description textuelle, sa couverture géographique, la période couverte par les données, le contact de la personne responsable, des mots clés ou encore la date de mise à jour. Cet investissement dans la standardisation des métadonnées ambitionne de constituer des catalogues de données interopérables. Des métadonnées uniformes permettent ainsi leur « moissonnage », leur exploitation automatique dans des « catalogues de catalogues », des portails qui donnent accès à de multiples sources de données. C'est le choix qu'a pris Etalab dans la deuxième version de data.gouv.fr qui, en plus des données publiées par l'État, « moissonne » les portails de collectivités locales, d'associations ou d'organisations internationales pour permettre un accès direct à ces données depuis le portail. Si elles facilitent l'indexation dans les moteurs de recherche et permettent le « moissonnage », ces spécifications ne suffisent pas à documenter les conditions de production des données et à décrire les catégories qui y figurent. Pour cela, Etalab recommande de fournir une description du contenu du jeu de données. Les producteurs de données peuvent soit les indiquer dans la partie description des métadonnées, un champ de texte de libre qui généralement précède l'accès au fichier, soit dans un document annexe, une notice par laquelle les producteurs de données peuvent accompagner les usagers dans la réutilisation des données. Mais ces recommandations sont généralement peu contraignantes et les gestionnaires de données avec lesquels je me suis entretenu n'évoquaient pas le remplissage des métadonnées comme une épreuve.

Néanmoins, pour certains services dédiés à la production de données tels que les SIG, le remplissage des métadonnées peut constituer un travail à part entière, pour lesquels certains agents disposent d'une véritable expertise acquise après plusieurs années de partage de données et soumises à la réglementation qui impose une normalisation des métadonnées. En particulier, la directive INSPIRE de 2007 leur demande de remplir les métadonnées selon une norme européenne et de décrire le contenu des données, les objets qui y figurent,

selon des modèles standardisés[118]. Lors d'une réunion interne de présentation de la nouvelle version du portail de la ville de Paris, une gestionnaire de données géographiques a ainsi souligné le décalage entre les pratiques de production de métadonnées en vigueur dans son service et celles mises en place pour le projet *open data*. Son service maintient déjà un catalogue de données dans lequel l'accès et la réutilisation sont conditionnés à la lecture de métadonnées extensives qui contiennent des restrictions d'usages. Selon elle, la plupart des données publiées sur le portail *open data* ne comportent pas une description suffisante des données et ne préviennent pas assez les réutilisateurs des limites de leur réutilisation.

> Les métadonnées, ça me parle beaucoup, c'est mon sujet, mon domaine. […] Au niveau de Paris, j'ai du mal à me retrouver. Nous quand on nous a posé la question de la publication de nos données, tout le monde était extrêmement frileux. La donnée, c'est une donnée métier. Une fois qu'elle sera sortie, comment cette donnée va être interprétée si elle n'est pas suffisamment documentée ? […] La limite que je trouve à la diffusion des données telle que vous le pratiquez, c'est qu'il faut accompagner les gens dans les usages pour leur donner un warning « c'est pas fait pour faire ça ». C'est un gros travail de constituer un catalogue de données géographiques, mais c'est indispensable.
> (Extrait d'une réunion interne de présentation du portail ***open data***, ville de Paris)

Dans les services où le partage des données constitue une pratique nouvelle, la production de métadonnées fait rarement l'objet d'une telle expertise. Leur production constitue donc un travail inédit pour les agents qui doivent qualifier les données dont ils assument la gestion quotidienne. Dans ce cas, les métadonnées ne servent pas uniquement à faciliter la réutilisation, elles sont conçues comme un cadrage qui spécifie leurs orientations initiales et tente de prévenir leur exploitation hors des usages pour lesquels elles ont été prévues. Ce cas donne à voir une autre manière de garantir l'intelligibilité des données sans mettre en œuvre de nouvelles transformations.

Pour les responsables de projet d'*open data*, les métadonnées constituent un instrument qui permet, entre autres, de réduire les frictions. En effet, comme en sciences, les métadonnées sont considérées comme un des moyens par lesquels les usagers peuvent parvenir à réutiliser les données sans avoir à interagir avec leurs producteurs. Par exemple, lors de la

[118] Merrien, F., & Leobet, M. (2011). La directive Inspire pour les néophytes. Mission de l'information géographique (MIG) du ministère de l'Écologie, du Développement durable, des Transports et du Logement.

présentation du portail pendant la réunion Open Data Bootcamp de la région IDF, un des présentateurs, affirme que les métadonnées sont généralement suffisantes pour permettre la réutilisation des données : « à partir d'un moment où les données sont suffisamment intelligibles, c'est aussi au réutilisateur de se débrouiller, d'interpréter ce qui est a été dit dans les métadonnées. » Mais, pour les responsables de projet d'*open data*, les métadonnées ont une autre vertu : elles permettraient aussi de réduire les frictions qui peuvent survenir en amont, lors de l'ouverture des données. Comme on l'a vu dans le quatrième chapitre, la question de la qualité de données brutes des administrations constitue un des motifs récurrents d'opposition à l'ouverture exprimés par les agents. Or, lors des négociations en vue d'obtenir l'ouverture des données, les responsables de projet d'*open data* brandissent souvent les métadonnées comme un moyen par lequel les agents peuvent expliquer ce qui pourrait être interprété comme des erreurs ou des incohérences.

> Plutôt que de publier des données complètement à jour, au mètre près, que tout soit tout bon, on a choisi de publier les données en état parce que même si elles sont justes à 80 ou 90 %, ça permet quand même que les gens s'en emparent et ça peut permettre qu'elles soient complétées par les retours du public. Donc, il ne faut pas s'en priver. Le tout c'est d'être au clair dès le départ, au moment où on publie la donnée d'expliquer ce qu'elle contient.
> (G.H., Directeur des systèmes d'information, Montpellier)

> Ça peut paraitre long et fastidieux, mais c'est extrêmement important parce que ça correspond en fait à toute la légende de votre jeu de données. Ça permet de préciser les conditions de production de votre jeu de données et ça permet en fait derrière de restreindre les mauvaises interprétations qu'on pourrait faire de votre jeu de données, de façon à ne pas créer des attentes qui seraient déceptives pour les réutilisateurs.
> (Extrait de la réunion Open Data Bootcamp, région Ile-de-France)

Certains responsables de projet d'*open data* considèrent donc les métadonnées comme un « lubrifiant » permettant de réduire les frictions qui surviennent lors de l'ouverture. En quelque sorte, elles seraient un moyen par lequel les agents pourraient assumer les « bonnes raisons » de produire des données que certains usagers pourraient juger mauvaises (Garfinkel & Bittner, 1967). Or, comme nous l'avons vu, l'ouverture attribue de nouvelles responsabilités : en ouvrant des données, les agents doivent non seulement rendre des comptes sur la conduite des politiques publiques, mais aussi sur la qualité de leurs données. Ouvrir des données qui pourraient être stigmatisées comme « de mauvaise qualité » expose

les agents à des critiques publiques, à des commentaires qui pourraient rejaillir sur l'image de leur service ou même sur la perception de la qualité de leur travail en tant qu'agent. Par exemple, les métadonnées n'ont pas suffi à convaincre les responsables d'un service de l'ouverture de données dont les agents ne sont pas certains de la qualité et dont ils craignent la sensibilité.

> Pour l'accidentologie, ce n'est pas consolidé, les données sont encore brutes, inexploitables. Et puis voilà, ce sont des données sensibles, ça parle de morts, de tués, de blessés. […] On n'est pas encore assez sûr de ces données pour les diffuser. Voilà c'est l'avis du service. Moi, je serais pour qu'on les diffuse avec des métadonnées qui vont bien.
> (V.N., Responsable informatique d'une direction, Montpellier)

Les métadonnées sont donc un des outils par lesquels les agents se prémunissent d'éventuelles critiques à l'égard de la qualité de leurs données. Mais elles ne sont en aucun cas la « baguette magique » qui pourrait permettre d'obtenir systématiquement l'ouverture des données.

Dans le processus de l'ouverture, en plus d'être un instrument par lequel les responsables de projet d'*open data* tentent de réduire les frictions, les métadonnées peuvent aussi contribuer à l'instauration des données. Dans l'introduction des troisièmes et quatrièmes chapitres, j'ai évoqué un « jeu de rôle » lors duquel les agents ont rempli une fiche d'identification (figure 28) ensuite disposée sur un tableau représentant le portail. Sur chacune de ces fiches, les agents ont dû remplir un nom du jeu de données, la désignation d'un responsable, une description, des mots clés. Sur ces fiches, ils remplissent en fait certains des champs standardisés requis par les portails pour les métadonnées. Cet exercice, qui était conçu comme fictif au départ, permet aux responsables de projet d'*open data* d'instaurer les fichiers, les documents, les outils utilisés au quotidien par les agents comme des données. En remplissant des métadonnées, ils contribuent à les faire exister comme des données, et pas seulement comme des outils de travail quotidien, une première étape dans leur identification et leur instauration en tant que données. Cette idée se retrouve chez Birchall (2014) pour qui les métadonnées contribuent à désigner les données parmi l'ensemble des objets informationnels.

> *Although the term "data" comes from the Latin word datum, meaning "something given," data is not simply objectively out there in the world already provided for us. The specific ways in which metadata is created, organized and presented helps to produce (rather than merely passively reflect) what is classified as data and information—and what is not.*

Le simple acte de remplir des métadonnées constitue donc une étape essentielle dans l'instauration des données. Elles sont un des instruments par lesquels les fichiers des administrations deviennent des données et par lesquels des publics parviennent à se lier à ces données.

La visualisation : transformer les données pour les rendre intelligibles à un plus large public

Pour comprendre l'intérêt des fonctionnalités de visualisation de données proposées par les portails *open data*, il me faut d'abord revenir en détail sur la question des formats de données. Le modèle en cinq étoiles proposé par Tim Berners-Lee postule que l'utilisation de formats ouverts et lisibles par les machines facilite la réutilisation des données par le public : « *you get more stars as you make it progressively more powerful, easier for people to use*. » Or, comme nous l'avons dans le chapitre précédent, l'utilisation du CSV, un format ouvert et lisible par les machines qui correspond à la troisième étoile du modèle de Berners-Lee, peut constituer une contrainte pour des usagers qui ne possèdent pas des compétences techniques avancées. En effet, pour ouvrir un fichier au format CSV, l'usager doit souvent saisir manuellement des paramètres pour que le fichier s'affiche correctement sous la forme d'un tableau (figure 34). C'est particulièrement le cas en France où la virgule sert généralement de délimitateur décimal alors que c'est le point dans les pays anglo-saxons. Or, les logiciels et les scripts informatiques sont souvent paramétrés par défaut pour que la virgule délimite les cellules, d'où le nom *Comma Separated Values* même si le format autorise l'usage d'autres caractères tel que le point-virgule utilisé généralement en France pour séparer les cellules. De manière très concrète, les contraintes du format ont une conséquence importante pour les publics des données. Le choix du format, pour lequel les politiques publiques et les militants de l'ouverture préconisent généralement le CSV, constitue une épreuve lors de laquelle les responsables de projet d'*open data* doivent se prononcer sur le public attendu des données.

> J'ai mis les données dans plusieurs formats pour essayer de trouver un juste milieu entre le format très brut donc, par exemple le format CSV qui est un truc très carré,

> très utilisable pour les développeurs même si ça l'est un peu moins pour les gens qui veulent juste voir à quoi ça ressemble.
> (L.K., responsable projet ***open data***, Rennes)

> Après, moi, je ne suis pas forcément un grand technicien, il y a beaucoup de fichiers sur lesquels je préfère les ouvrir en XLS plutôt qu'en CSV. Je te le dis de manière très transparente. Avec la nouvelle plateforme, ils peuvent me mettre tous les formats qu'ils veulent. Donc je leur ai dit « CSV OK, mais tu mets aussi le XLS avec. »
> (T.Y., un agent de la mission Etalab)

Dans ces deux extraits, il ressort que les grands principes du format CSV intègrent, directement dans la conception du standard, un « script » (Akrich, Bijker & Law, 1992), des compétences et des formes d'action présupposées. Par sa conception, le format CSV attend en effet que l'usager comprenne et s'adapte aux paramètres du fichier pour l'utiliser correctement dans un tableur ou un autre outil informatique. Pour les responsables de projet d'*open data*, le choix du format CSV peut ainsi configurer les données pour des usagers avec des compétences techniques avancées. Ainsi, l'impératif de lisibilité des données par les machines peut donc en partie réduire leur lisibilité par les humains. Pour éviter que les données ne s'adressent uniquement à des publics disposant de connaissances techniques avancées, les portails d'*open data* proposent donc souvent des fonctionnalités de visualisation qui permettent aux usagers d'afficher les données sous la forme de tableaux, de graphiques ou de cartes sans même avoir à ouvrir le fichier dans un autre logiciel que le navigateur. Ces fonctionnalités, comme nous l'avons vu en introduction, sont au cœur du projet de la région Ile-de-France sur lequel je vais me concentrer ici pour comprendre le travail que demande l'utilisation de ces fonctionnalités de visualisation. Lors de la réunion OpenDataBootcamp, les animateurs ont insisté à plusieurs reprises sur l'intérêt de ces fonctionnalités pour permettre à un public plus large de se saisir des données.

> Pierre présente les fonctionnalités de visualisations : « Si je passe sur le tableau, là j'ai la donnée brute à partir d'un fichier CSV, c'est pour ça qu'il n'y a pas de couleur. Donc grâce à l'excellent Open Data Soft [la société qui a développé le portail de la région], nous pouvons visualiser ces données quand elles sont géocodées sur une carte. […] Sur les colonnes de mes données tabulaires, je peux aussi créer des filtres, ce qu'on appelle des facettes, que je trouve ici à gauche. […] », François complète « on est plus sur une prévisualisation de la donnée pour pouvoir la faire parler, la visualiser autrement que sous une forme tabulaire et avoir une première vision de ce qui a dans nos données. Alors il faut faire un minimum de traitement sur le jeu de données pour avoir des visualisations. »

Les fonctionnalités de visualisation demandent donc d'apporter de nouvelles transformations aux données. Le travail d'instauration des publics crée de nouvelles frictions dans le processus d'ouverture des données. Mais comment les données sont-elles traitées pour que le public puisse les « faire parler » directement depuis le portail ? Comment les données brutes sont-elles transformées pour être exploitées par les portails d'*open data* ?

Pour saisir l'épaisseur de ce travail, revenons au document « bonnes pratiques sur Excel » de la région Île-de-France évoqué dans le chapitre précédent à propos de la conversion des données au format CSV. Lorsqu'on consulte ce document lisible en ligne, plusieurs transformations des fichiers Excel sont préconisées alors même que la conversion des données ne modifie pas leur présentation dans un tableur. Par exemple, dans la figure 41), il est recommandé aux agents de ne disposer qu'un tableau par feuille. Or, j'ai fait l'expérience de convertir ce fichier en CSV dans LibreOffice en utilisant les paramètres par défaut. Dans LibreOffice, après conversion au format CSV, les cellules sont disposées de la même manière dans le fichier CSV que dans le format OpenDocument utilisé par défaut dans le tableur. Seul le formatage (texte en gras et contour des cellules) a disparu (figure 42).

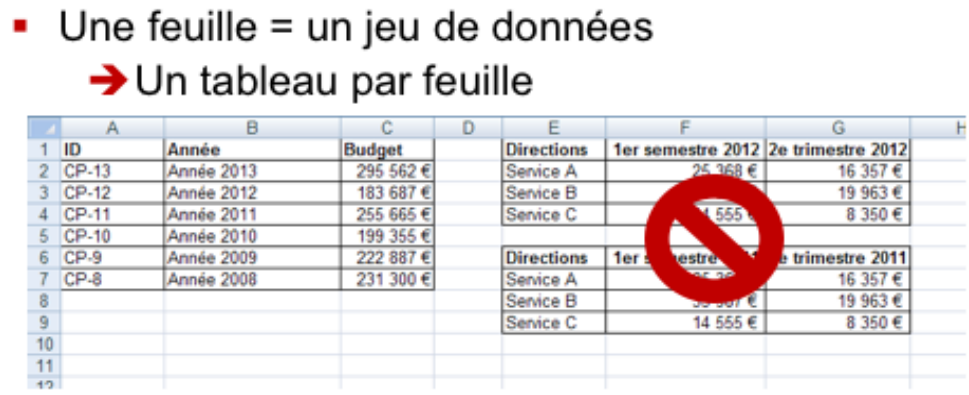


Figure 41. Recommandations sur la structure du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Île-de-France.

	A	B	C	D	E	F	G
1	ID	Année	Budget		Directions	1er semestre 2012	2e trimestre 2012
2	CP-13	Année 2013	295 562 €		Service A	25 368 €	16 357 €
3	CP-12	Année 2012	183 687 €		Service B	Valeur illisible	19 963 €
4	CP-11	Année 2011	255 665 €		Service C	4 555 €	8 350 €
5	CP-10	Année 2010	199 355 €				
6	CP-9	Année 2009	222 887 €		Directions	1er semestre 2011	1e trimestre 2011
7	CP-8	Année 2008	231 300 €		Service A	Valeur illisible	16 357 €
8					Service B	Valeur illisible	19 963 €
9					Service C	14 555 €	8 350 €
10							

Figure 42. Les tableaux de la première illustration de la figure 41 convertis en CSV et affichés dans Libre Office (paramètres par défaut du convertisseur).

À première vue, ces transformations ne sont donc pas justifiées par une perte d'information qui pourrait survenir lors de la conversion des données. J'ai donc interrogé l'auteur du document qui m'a expliqué qu'elles sont liées aux spécifications du format CSV inscrites dans le portail développé par la société française OpenDataSoft. Les fonctionnalités qui permettent de visualiser les fichiers CSV sous la forme de tableaux, de cartes ou de graphiques réclament de se conformer à des spécifications bien précises du standard.

> Les transformations en CSV sont ici optimisées pour la plateforme OpenDataSoft qui, dans ses premières versions, était terriblement restrictive (notamment pour les cellules vides). Par ailleurs, pour que la visualisation « automatique » (histogrammes, ou cartographie) fonctionne, une feuille ne peut regrouper plusieurs tableaux de données parfois hétérogènes.

Le portail, dans la version utilisée par la région Île-de-France en 2013, s'aligne vraisemblablement sur la définition du CSV proposée dans la RFC 4180 de 2005 qui est devenu son standard de facto. Selon ses spécifications qui précisent certains aspects du format CSV, chaque ligne désigne un enregistrement qui correspond à une série de valeurs qui s'appliquent à un objet souvent désigné par un identifiant unique. Chaque enregistrement doit comporter le même nombre de valeurs et la RFC réclame que chaque valeur soit définie par un titre dans la première ligne du fichier. En plus de ne faire figurer qu'un seul tableau par feuille de calcul, l'intégration de ces spécifications dans les portails d'*open data* demande d'autres formations. Dans le document « bonnes pratiques sur Excel », l'équipe en charge de l'ouverture des données de la région demande ainsi de ne pas faire figurer de valeurs en dehors du tableau de données. Ces valeurs sont qualifiées

d'« orphelines » : pour exister selon ces spécifications du format CSV, elles doivent figurer dans un autre tableau, dans une autre série d'enregistrements (figure 43).

Figure 43. Recommandations sur la structure du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Île-de-France.

Les spécifications de la RFC4180 demandent aussi de faire figurer les titres des colonnes sur la première ligne du fichier. Pour que le fichier soit correctement interprété, les informations qui figurent sur la première ligne doivent être effacées ou déplacées si elles ne font pas partie du tableau de données comprenant les enregistrements (figure 44).

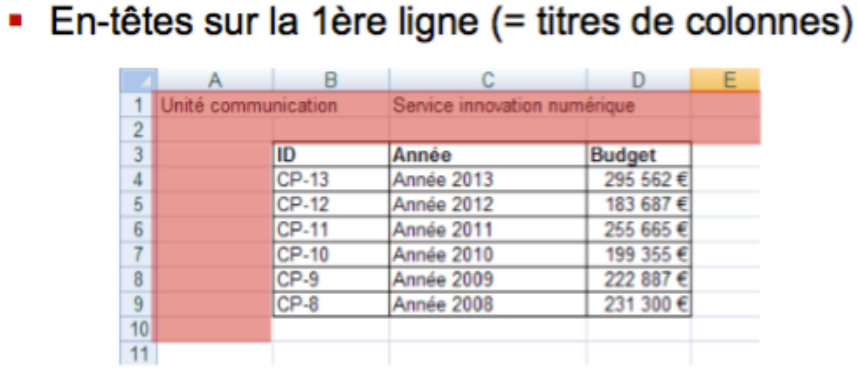


Figure 44. Recommandations sur la structure du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Île-de-France.

Les gestionnaires de données doivent aussi ne pas faire figurer de cellule vide dans la première ligne pour que chaque colonne comporte un titre comme le réclament les fonctionnalités de visualisation du portail (figure 45).

Figure 45. Recommandations sur la structure du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Île-de-France.

Les opérations évoquées précédemment permettent uniquement l'affichage des données sous la forme d'un tableau comportant des fonctionnalités de tri ou sous forme de graphiques. Pour que les données soient visualisées sous la forme de cartes, les agents doivent effectuer de nouvelles transformations afin que les adresses contenues dans les fichiers puissent correspondre à des coordonnées géographiques qui puissent être affichées sous la forme de points sur une carte. Cette opération dite de « géocodage » impose une nouvelle couche dans le travail de transformation des données. Bien qu'ils s'appuient sur les interfaces de programmation (API) des services de cartographie en ligne comme Google Maps ou Open Street Map pour automatiser la mise en correspondance des adresses avec les coordonnées géographiques, les agents doivent intervenir à plusieurs reprises pour uniformiser le format des adresses, corriger les erreurs et contrôler les résultats. Ce processus crée généralement des erreurs et des incohérences qui sont prises en compte par les API de géocodage[119] et par les responsables de projet *open data* qui peuvent mettre en place des procédures de nettoyage et de validation des données (figure 46).

[119] La plupart des API de géocodage renvoient pour chaque adresse un score de fiabilité qui évalue la précision de la mise en correspondance d'une adresse avec des coordonnées géographiques.

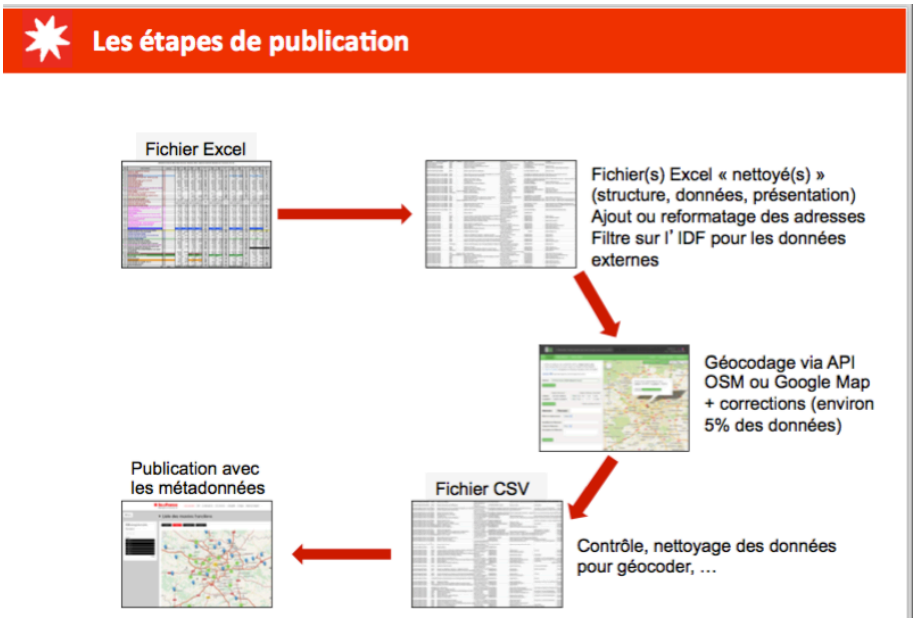


Figure 46. Recommandations sur la structure du fichier. Extrait du document « bonnes pratiques sur Excel » de la région Île-de-France.

La visualisation crée donc de nouvelles frictions dans le processus de l'ouverture des données. Dans le chapitre précédent, la conversion en CSV demandait déjà une importante transformation des données afin de garantir leur intégrité d'un format à l'autre dans l'environnement du tableur. Ces logiciels, qui constituent pour une majorité d'usagers le principal outil de traitement des données, sont ouverts à la multitude d'interprétations du format CSV qui régissent ce standard[120]. Mais tous les outils qui exploitent le format CSV n'assurent pas l'interopérabilité des données, c'est le cas ici des fonctionnalités de visualisation des tableurs qui intègrent certaines spécifications dans leur code. Les gestionnaires de données et les responsables de projets d'*open data* doivent donc souvent effectuer de nouvelles transformations pour se conformer au standard adopté par le portail. Dans le cas évoqué précédemment, l'utilisation de ces fonctionnalités demande de s'aligner avec les spécifications de la RFC 4180 qui réduit la feuille de calcul à une série d'enregistrements comportant pour chaque ligne le même nombre de valeurs. Le travail de transformation des données se poursuit lorsque les agents veulent obtenir un rendu

[120] C'est l'approche promue au sein de l'IETF, l'organisme à l'origine des standards d'Internet, qui consiste à accepter la diversité des spécifications tout en produisant des fichiers qui appliquent strictement le standard. La RFC 4180, le standard de facto pour le format CSV, préconise ainsi que les développeurs acceptent les différentes versions du format : « Due to lack of a single specification, there are considerable differences among implementations. Implementors should "be conservative in what you do, be liberal in what you accept from others" (RFC 793 [8]) when processing CSV files. »

cartographique de leurs données ce qui demande de faire correspondre les adresses en coordonnées géographiques et de corriger les erreurs qui peuvent survenir.

Dans le domaine des sciences, Latour (1993) a montré, en suivant la trajectoire des inscriptions produites par les instruments d'une équipe de scientifiques en mission à Boa Vista en Amazonie, que les « données » scientifiques connaissent un double mouvement de réduction et d'amplification. Réduction, car les inscriptions perdent progressivement leur matérialité et leurs particularités locales au fur et à mesure de leur standardisation ; amplification, car elles gagnent par la même occasion en généralité et en capacité de combinaison. Les cas précédents montrent une situation similaire où, transformation après transformation, les données deviennent de plus en plus lisibles et exploitables en même temps qu'elles perdent les traces et les particularités des données « métier » : le formatage disparait, les informations en dehors du tableau principal deviennent des données « orphelines », les particularités locales ou nationales des adresses sont converties en coordonnées standardisées et globales… Mais, si en sciences ce travail fait partie intégrale de la production du savoir même s'il est souvent considéré comme un « sale boulot » attribué aux travailleurs en bas de l'échelle hiérarchique (Hugues, 1962), la transformation des données brutes de l'administration est censée généralement être prise en charge par les usagers selon certains des « principes » évoqués dans le premier chapitre. Par exemple, pour Tim Berners-Lee, les agents de l'administration doivent d'abord fournir les données à leur état brut avant de les retravailler pour leur exploitation : « *make a beautiful website, but first give us the unadulterated data*. » Du point de vue des agents, la prise en charge de ce travail dans le processus de l'ouverture constitue un nouvel investissement (Thévenot, 1986) dans la lisibilité des données par les machines, mais aussi par les humains. En effet, ces fonctionnalités de visualisation anticipent que les usagers ne disposent pas nécessairement des compétences techniques ni de la disponibilité pour exploiter et visualiser ces données à travers des outils ad hoc. C'est un certain type de publics de données qui est inscrit à travers l'intégration de ces outils à même les portails. À travers le cas du *Transparency Agenda* (TA) du gouvernement britannique, Ruppert (2012) a montré comment les politiques d'*open data* produisent des publics de données en érigeant la visualisation de données comme une technologie neutre pouvant renouveler les conditions de l'*accountability* de l'État.

> *The production of data publics is mediated by software developers, journalists, think tanks, lobbyists, watchdog organisations, data visualisers, and bloggers who are the*

> *predominant experts rather than auditors, policy analysts, academics and statisticians. […] TA similarly positions numbers, spreadsheets and visualisation software as neutral technologies that enable the production of "objective" accounts of "facts" without the intervention of experts.*

En créant un agencement entre les données de l'État et des *data publics*, les technologies de médiation fondées sur la visualisation et l'interaction avec les données créent un nouvel agencement entre les données de l'État et leurs publics. Ces outils, souvent présentés comme aussi « neutres » que les données, configurent le citoyen comme un acteur de l'évaluation et de l'audit des politiques publiques, disposant des outils et des données brutes, inaltérées permettant d'inspecter le fonctionnement de l'État. Mais les cas précédents montrent que la production des publics ne consiste pas seulement à la mise en place d'un nouvel agencement. Elle s'effectue à travers un travail très concret effectué par les agents qui adaptent les données aux réseaux sociotechniques qui permettent de produire ces publics.

Les assemblages temporaires des concours de réutilisation de données

Jusque là, les instruments d'instauration des publics de données que j'ai étudiés s'articulent autour des portails de données ouvertes et se situent en amont de l'ouverture des données. Ici, je vais m'intéresser à une forme d'action publique qui accompagne souvent les politiques d'*open data* : les concours de développement de services. À travers ces évènements compétitifs, souvent qualifiés de marathon de développement ou « hackathons » quand ils sont limités à quelques jours, la participation par le développement de services est incitée, voire stimulée par l'attribution de prix dont la valeur est soit financière soit symbolique. Ces prix sont généralement attribués par un jury composé d'usagers et/ou d'experts du numérique. Il existe une littérature abondante sur la participation du public, en sciences politiques particulièrement, qui a notamment considéré la participation des citoyens comme une « politique de l'offre » (Gourgues, 2012) et qui a observé que les dispositifs participatifs peinent couramment à attirer leur public (Hibbing & Theiss-More, 2002). Par rapport aux dispositifs participatifs classiquement étudiés comme les budgets participatifs ou les concertations en ligne, les concours divergent sur deux points. Premièrement, le dispositif participatif prend la forme d'un concours où des prix sont attribués avec une rétribution financière ou symbolique : la participation est donc encouragée, voire provoquée. Les prix distinguent certaines contributions dont la pertinence est approuvée par l'institution. Deuxièmement, la technicité ne porte pas tant sur les sujets sur lesquels les participants

sont amenés à travailler (Ferretti, 2007 ; Gourgues, Rui & Topçu, 2013), mais sur les modalités de la participation. Les concours demandent d'attirer des utilisateurs techniquement compétents pour créer les réutilisations des données ouvertes sous la forme de services qui sont attendues par les jurys. Ces dispositifs contrastent aussi avec le discours des acteurs évoqués dans le premier chapitre ou encore avec celui de Daniel Kaplan qui figure en introduction de ce mémoire. Pour ces acteurs, l'existence d'une demande de réutilisation des données constitue une prémisse fondamentale des politiques d'*open data*. L'association Regards Citoyens a donné un bon exemple de ce postulat lorsqu'elle a publié un « petit guide de l'*open data* ». Dans ce document[121], elle affirmait que si les administrations publiaient leurs données en respectant les « principes » de l'*open data* tels que les dix principes de la Sunlight Foundation ou l'Open Definition, les données devraient trouver leurs publics sans même qu'il n'y ait besoin d'avoir recours à de tels dispositifs.

> Si les données ont bien été libérées sous conditions Open Data, les réutilisations arriveront sans doute d'elles-mêmes. Ne perdez pas donc votre temps avant même l'ouverture à préparer des communications, hackathons, sites officiels de réutilisation… […] Les jeux de données que vous avez rendus publics vont certainement intéresser des réutilisateurs.

Comment expliquer alors l'organisation de ces évènements qui stimulent la participation du public ? Comment ces dispositifs tentent-ils de démontrer le postulat de l'existence de publics de données ? Quel est leur rôle dans le processus de l'ouverture des données ? Explorant dans cette section ce qui se déroule après l'ouverture des données, j'étudie ici deux concours, situés à Rennes et à Montpellier, dont je retracerai les origines et la trajectoire afin de comprendre dans quelle mesure ces dispositifs contribuent à instaurer des publics de données.

Historiquement, les concours répliquent un des rites des hackers, les « marathons » de développement qui se déroulent pendant les conférences des communautés issues du logiciel libre. Lors de ces évènements, les participants se regroupent pour développer les logiciels pendant d'intenses sessions d'écriture de code : « les participants s'adonnent frénétiquement au hacking et rendent visibles les liens qui existent entre eux, leur donnant

[121] Regards Citoyens, « Apprenons des échecs de la DILA, épisode 1 : « Comment faire de l'Open Data ? » », http://www.regardscitoyens.org/apprenons-des-echecs-de-la-dila-episode-1-comment-faire-de-lopen-data/#guide, consulté le 12 aout 2015.

une plus grande intensité »[122] (Coleman, 2013). Les concours de réutilisation de données ouvertes sont aussi liés à l'émergence de deux concepts — le *crowdsourcing* et l'innovation ouverte — qui se sont imposés dans les entreprises comme des modèles d'innovation à part entière. Le *crowdsourcing* considère que la participation massive (et généralement bénévole) des internautes peut être une source majeure de création de valeur voire de renouvellement de la démocratie (Surowiecki, 2004 ; Shirky, 2008 ; Colin & Verdier, 2013). De son côté, l'innovation ouverte consiste à diffuser une partie de l'information stratégique à des acteurs extérieurs qui, par la mise en compétition ou leur nombre, parviendront à mieux innover que les ressources internes dont dispose l'organisation (Chesbrough, 2006 ; Tapscott & D. Williams, 2008). Ces analyses, issues principalement d'essais prédictifs sur l'avenir des technologies, ont idéalisé la créativité des hackers et ont inspiré les premières compétitions. Dans le domaine des données ouvertes, *Apps for Democracy* a été un des premiers concours de développement de services. Il s'est tenu à Washington en 2008 (Demeyer, 2012), quelques mois après le lancement de l'App Store d'Apple, une des premières plateformes de services mobiles qui a généré des revenus conséquents pour les développeurs d'applications mobiles. Cet évènement a inspiré le développement de nombreux concours d'application notamment par la publication d'un guide[123] qui a détaillé pas à pas comment faire participer les développeurs à un concours d'applications et a permis l'essaimage de ce modèle dans de nombreuses villes. En particulier, le guide d'*Apps for Democracy* a insisté sur les retombées économiques de l'évènement de 2008, la valeur des 47 applications créées étant estimée à 2 300 000 $ pour un coût de l'opération de 50 000 $[124]. En démontrant la valeur économique des concours d'application, *Apps for Democracy* a constitué des ressources essentielles pour l'essaimage de cette forme de participation.

À Rennes et Montpellier, l'organisation des concours de service réutilisant les données ouvertes était fortement liée à un acteur en particulier, la FING, qui a contribué à diffuser ce modèle de participation. Au départ, à Rennes, le projet d'*open data* est né d'une situation

[122] Traduction personnelle

[123] « How to run your Apps for Democracy Innovation Contest », accessible en ligne : https://www.howto.gov/sites/default/files/documents/createanappsfordemocracy%5B1%5D.pdf

[124] L'estimation repose sur un judicieux calcul qui multiplie le coût de développement évalué à 50 000$ par application pour déterminer la valeur crée lors du concours. Or, la valeur n'équivaut pas nécessairement au coût ; elle dépend principalement du marché et de l'utilité du bien. On peut donc raisonnablement dire que cette valeur a été largement sur-évaluée probablement pour servir à la communication des organisateurs.

inédite : la création d'une application mobile permettant de connaitre la disponibilité des vélos en libre-service. En 2009, un développeur a extrait automatiquement les données du site sur les vélos en libre-service pour proposer une application mobile facilitant la recherche d'une borne de stationnement. Craignant que cette application utilisant des éléments de la marque « le Vélo Star » ne soit considérée comme un service officiel, les agents du département marketing de Keolis Rennes ont demandé à une agence de communication numérique de préparer une application mobile proposant les mêmes fonctionnalités que le service créé par le développeur et d'identifier comment l'extraction automatique des données du site Vélo Star pouvait être bloquée. Plutôt que de répondre à ces deux demandes, le dirigeant de l'agence de communication, se revendiquant des principes de l'*open source*, est parvenu à convaincre les agents de Keolis Rennes d'ouvrir leurs données. Quelques mois après, au début de l'année 2010, un portail donnant accès aux données du vélo en libre-service était prêt à être publié. Lors d'un voyage d'études, un dirigeant de Keolis Rennes a présenté le portail d'accès aux données du vélo au maire dans le but d'obtenir son accord pour son lancement prochain, la ville étant propriétaire des données. En présence des responsables des transports de la ville et de la métropole, le maire a autorisé le lancement du site. Mais à cette occasion, la municipalité s'est engagée à aller plus loin que la seule ouverture des données du vélo. Sur proposition des agents du service de la communication de la ville, il a accepté que l'administration municipale prenne part à une expérimentation proposée par la FING, le programme « réutilisation des données publiques locales[125]. » En tant que territoire « pilote », la ville et la métropole de Rennes ont dû adhérer à certaines conditions imposées par la FING en échange de son accompagnement et de son réseau. En plus d'ouvrir des données publiques pendant une période minimale de six mois, le programme exigeait aussi d'encourager à leur réutilisation par la mise en place d'évènements et d'un concours qui récompensera les meilleurs usages des données.

Lancé en octobre 2010 à la Cantine[126] à Paris, le concours a mobilisé un réseau d'acteurs variés : des agents qui ont participé à des évènements, des partenaires qui le finançaient,

[125] Fondation Internet Nouvelle Génération. « Réutilisation des données publiques locales : présentation du programme ». Consulté le 19 novembre 2014. http://fing.org/?Presentation,448.

[126] La Cantine était un des premiers espaces de coworking ouverts à Paris par l'association Silicon Sentier. C'était un point de ralliement régulier pour certaines communautés de hackers dans lequel se tenaient

des médias associés qui se sont engagés à le couvrir et des institutions nationales qui venaient s'inspirer de ce projet « pionnier » de l'*open data* en France. Ouvert aux entreprises et aux particuliers, il était doté de 50 000 € de prix, une enveloppe partagée en cinq prix thématiques pour inciter les développeurs à exploiter toutes les données mises à disposition. Les participants avaient cinq mois pour développer leurs applications. Pendant cette période, ses organisateurs ont programmé une série de rencontres entre les candidats et les producteurs des données à la Cantine numérique rennaise, un lieu similaire à l'espace parisien qui a ouvert en novembre 2010. Le 30 mars 2011, les résultats du concours ont été annoncés en présence des membres du jury et de représentants des partenaires. Cette cérémonie a donné lieu à une remise de prix (figure 47) aux développeurs des applications récompensées. Une majorité des applications primées étaient dédiées à l'information des usagers des transports et des vélos en libre-service pour proposer un calcul d'itinéraires, des incitations à l'intermodalité ou pour fournir une aide au stationnement. Les autres applications ont tenté de valoriser le patrimoine historique et environnemental et de proposer des guides dans la ville.

Figure 47. Remise des prix du concours le 30 mars 2011. Photo : ville de Rennes Stéphane Priou

À Montpellier, la ville avait, depuis vingt ans, engagé un programme d'investissement nommé « Pégase » sur les infrastructures numériques à travers le déploiement de la fibre

des conférences régulières sur les sujets numériques. La Cantine a déménagé en 2013 dans un espace plus grand et a été renommé pour devenir Numa.

optique. Aux alentours de l'année 2010, un élu en charge des questions numériques a proposé de faire évoluer les politiques publiques vers le développement de contenus et des services pour les habitants. À cette même période, le directeur des systèmes d'information (DSI) de la ville a découvert le concept d'*open data* lors d'une rencontre européenne des DSI. L'ouverture des données était promue comme un outil facilitant la création de services pour les habitants ; cet argument a convaincu le directeur des systèmes d'information de lancer un programme d'innovation numérique comportant un volet dédié à l'ouverture des données.

> À Stockholm, il y a eu un colloque des DSI des villes un peu partout dans le monde et un des sujets de discussion c'était de dire justement qu'il fallait absolument mettre à disposition tout le patrimoine d'information qu'on avait dans les collectivités pour que pour que tous ceux qui voulaient faire quelque chose avec, principalement les entreprises, puissent effectivement s'en emparer et produire des services à partir de ces informations. Sur Montpellier, l'idée a fait son chemin et courant 2010, on est parti sur un programme qui s'appelle Montpellier Territoire Numérique qui avait différents axes dont un est précisément axé sur l'ouverture des données.
> (G.H., Directeur des systèmes d'information, Montpellier)

La DSI a initié le projet Montpellier Territoire Numérique (MTN) avec pour objectif principal, comme à Rennes, la création de services et d'applications pour les habitants. En plus de l'ouverture des données de la ville et d'un appel à projets pour la création de services réutilisant des données ouvertes, le projet Territoire Numérique comportait plusieurs volets parmi lesquels l'installation d'écrans interactifs dans l'espace public urbain et la création d'un lieu de travail collaboratif sur le modèle de la Cantine à Paris. La ville s'est rapprochée de la FING et a recruté un ancien salarié de l'association pour piloter le projet. Au-delà du rôle de passeur de la FING dans la circulation du modèle des concours, il apparait en retraçant les origines de ces deux projets que les politiques publiques d'*open data* ne fixent pas seulement comme objectif aux administrations de diffuser des données auprès du public. On avait pu le voir précédemment avec Etalab à qui des objectifs quantifiés de réutilisation sont assignés (figure 21), les données doivent être réutilisées pour que les politiques d'*open data* remplissent leurs promesses. Ainsi, dans les deux terrains étudiés ici, les chefs de projet d'*open data* n'ont pas seulement pour mission l'ouverture des données, mais leur réutilisation, en particulier pour la création de service à destination des habitants.

En incitant les usagers à participer en réutilisant les données, les concours constituent un instrument particulier pour susciter l'émergence des publics de données. Pour mieux comprendre comment les concours contribuent à instaurer ces publics, la notion d'intéressement développée par Michel Callon (1986) propose un cadre théorique intéressant qui souligne le caractère temporaire de certains assemblages sociotechniques même si l'objet de son analyse peut paraitre bien éloigné des cas précédents. Callon y décrit le travail de trois chercheurs qui tentent d'importer dans la baie de Saint-Brieuc une technique japonaise qui rend possible la culture intensive des coquilles Saint-Jacques. Dans cette même baie, les pêcheurs font face à une chute de la production et participent à une expérience proposée par les chercheurs. L'article a pour objet de montrer comment les chercheurs parviennent à convaincre les marins-pêcheurs et les collègues scientifiques du bien-fondé de l'expérience. L'auteur y définit la notion d'intéressement comme « l'ensemble des actions par lesquelles une entité [...] s'efforce d'imposer et de stabiliser l'identité des autres acteurs qu'elle a définis dans sa problématisation. » Callon lui donne un sens littéral : intéresser, c'est se placer entre (inter-esse), s'interposer entre les acteurs compris dans la problématisation qui doivent passer à travers des points de passage obligé dans lesquels ils peuvent s'allier et changer temporairement leur identité. Lorsque l'intéressement est réussi, il devient enrôlement qui désigne « le mécanisme par lequel un rôle est défini et attribué à un acteur qui l'accepte. »

En gardant ce cadre d'analyse, la problématisation qui guide l'organisation des concours peut être résumée à cette question : existe-t-il un public pour nos données ? À ce titre, l'extrait en introduction de la réunion de la région Ile-de-France donne un exemple particulièrement parlant. Les organisateurs de la réunion Open Data Bootcamp évoquaient un hackathon à l'Institut d'Aménagement et d'Urbanisme dans ces termes : « ça a produit immédiatement un démonstrateur sur la capacité des réutilisateurs à traiter des jeux de données et à produire des choses avec. » Comme à Montpellier ou à Rennes, l'existence de publics pour les données était un postulat auquel le concours peut apporter des réponses. Le concours assigne à chaque acteur une identité propre. A minima, les modalités des concours exigent que les développeurs réutilisent les données ouvertes par la collectivité qui l'organise et créent un service gratuitement même s'il n'est pas primé par le concours. De plus, dans le cadre du concours, des moments et des lieux d'échange sont mis en place pour organiser les discussions entre les développeurs et les producteurs de données. À

Rennes, cela a pris notamment la forme d'un forum en ligne dans lequel les développeurs peuvent échanger avec les producteurs de données.

> On leur avait dit : « ayez la posture d'aller sur le forum, il faut qu'il ne soit pas modéré ». Et répondez quand les gens disent « votre donnée, elle est nulle parce qu'elle est fausse », dites leur « merci, où et comment ? Corrigez-la vite et répondez aussi aux questions qu'il y a eu sur la licence qui, au début, interdisait les usages commerciaux. » […] On les a beaucoup accompagnés sur le concours. C'est une cible qu'ils ne connaissaient pas ; ils découvraient ce monde des geeks, un peu hackers, férus de démocratie à l'anglo-saxonne qui peut heurter. Nous on insistait aussi pour qu'ils signent en leur nom en disant que c'est des gens qui aiment bien savoir à qui ils parlent quoi. Puis on leur conseillait d'aller les voir après autour d'une bière parce qu'il va falloir que vous alliez boire des coups, enfin il faut voir les gens quoi. Et ce qui fera la réussite, c'est la proximité.
> (A.M., Dirigeant, agence de communication web, Rennes)

Dans l'espace du forum, les agents sont incités par les organisateurs du concours à adopter une certaine attitude, un « savoir-être » fondé sur l'échange et la prise en compte des critiques, une attitude proche de celle qui est demandée aux agents spécialisés de la participation (Mazeaud, Sa Vilas Boas & Berthomé, 2012). Les concours organisent aussi des moments de rencontre entre les producteurs des données et les participants, c'était le cas à Rennes et à Montpellier où cela été organisé à plusieurs reprises. Ces rencontres peuvent permettre d'initier le travail complexe de coordination entre les producteurs initiaux des données et leurs réutilisateurs auquel des métadonnées exhaustives ne peuvent pas se substituer (Edwards et al, 2011). Par ailleurs, les concours imposent des règles et encadrent la participation. Tout d'abord, cela peut paraitre évident, mais les porteurs de projet doivent obligatoirement exploiter au moins un jeu de données ouvertes par l'organisation à l'origine du concours. Cette condition est essentielle pour que de nouveaux assemblages se créent entre les données et leurs publics et donc répondent à la problématisation. Par ailleurs, les concours imposent souvent des conditions quant aux types de résultats qui peuvent être primés. À Rennes, l'enveloppe de 50 000€ était distribuée en six prix thématiques correspondant aux critères établis par les partenaires du concours[127]. Par exemple, le prix de l'accessibilité attribué par la région Bretagne récompensait « l'application qui favorisera l'accès aux services ou aux transports pour les personnes à mobilité réduite, déficients

[127] Rennes métropole en acces libre, « Les prix et le jury », http://www.data.rennes-metropole.fr/le-concours/les-prix-et-le-jury/, consulté le 19 novembre 2013.

visuels ou auditifs » ; celui de l'écomobilité soutenait « les applications favorisant les modes de déplacements doux »… Les définitions des prix orientent ainsi les participants vers certaines données, souvent dans le domaine des transports, et réclament parfois un certain type de réutilisation, sous la forme d'applications mobiles ou de visualisations de données.

> Le dispositif d'aide est permanent et, de temps en temps, on fait des focus. Le premier a été fait autour de la réutilisation des données ***open data***, le second autour des données de santé et le troisième sera sûrement fait autour de la visualisation de données ou quelque chose comme ça.
> (H.B., Chef de projet ***open data***, Montpellier)

> On a été embêté à l'époque du jury parce que les catégories avaient été pensées par les partenaires financiers. En fait, il y avait le prix de la mobilité, le prix de machin, le prix de truc et puis il y a eu un vote en ligne : ben le premier prix c'était une appli de transport. Et enfin c'était gênant parce qu'il y avait beaucoup trop de trucs autour des transports. […] L'***open data*** qu'on a vu dans le résultat du concours, ce n'est pas un ***open data*** politique, c'est un ***open data*** purement utilitariste. Ce sont des applis qui rendent des services, des applis qui ne sont pas particulièrement politiquement marquées.
> (D.V., Responsable associatif, Rennes)

Comme on a pu le voir précédemment lors de la sélection des données, la définition des prix du concours constitue une nouvelle épreuve lors de laquelle les responsables de projet d'*open data* doivent se positionner quant à l'utilité des projets d'*open data* et à leurs objectifs. En encadrant la participation dans le dispositif du concours vers la création de service, c'est un certain de type de publics de données qui émerge, non pas pour favoriser la transparence et l'*accountability*, mais pour développer des services qui complètent voire pour certains pourraient se substituer au service public (Bates, 2012 ; Birchall, 2015). Le dispositif du concours encadre donc la participation : les réutilisations qui échappent à ses modalités sont de fait défavorisées et la probabilité de leur création sera plus faible en l'absence d'un dispositif de soutien.

En intéressant les développeurs aux données, le concours ne permet pas uniquement la création de services annoncée officiellement. Il forme un assemblage dans lequel les données se retrouvent liées à un public ce qui répond au problème qui occupe les responsables de projet d'*open data*. Ces derniers peuvent alors prouver à leurs interlocuteurs le postulat essentiel de l'*open data* : l'existence de publics de données.

> Lors du concours, on a vu assez rapidement qu'il y avait énormément de potentiel et de personnes qui étaient aptes à développer des choses que nous on ne développerait pas. Et qui seraient des services complémentaires qui viendraient enrichir du coup, notre donnée.
> (Q.C., référent SIG, direction des jardins)

Au-delà des développeurs et des porteurs de projet, le concours permet aussi d'instaurer un autre public : celui des usagers de service. Dans la communication qui accompagne les projets d'*open data*, les données ne sont pas uniquement présentées comme des ressources mises à disposition de leurs usagers directs, qui sont souvent qualifiés de « réutilisateurs », mais aussi comme une offre de service qui peut toucher un public large, bien au-delà des personnes coutumières de l'exploitation de données.

> On fait des petits fascicules, ça nous permet de communiquer. En fait, les gens, quand tu leur parles d'***open data*** (là je parle de citoyens normaux), ils ne savent pas ce que c'est généralement et puis, quand je leur explique, ils ne comprennent pas forcément. Mais quand tu leur montres que ça sert à faire une application pour calculer l'itinéraire pour les aveugles, que c'est parce que les données sont ouvertes, que les gens peuvent faire ça, là, ça devient de suite beaucoup plus concret.
> (H.B., Chef de projet ***open data***, Montpellier)

Comme dans le cas évoqué précédemment des outils de visualisation intégrés dans les portails, les services et les applications qui sont créés lors des concours contribuent à produire de nouveaux publics de données. Comme l'a bien montré Ruppert (2012), ces technologies de médiation fondées sur l'interactivité et la visualisation des données produisent un certain type de public qui n'interagit pas avec les données, mais avec des « boites noires » qui orientent l'usager vers certaines conclusions, vers certains « faits. » Elles produisent des publics multiples composés d'une part des « réutilisateurs », le public qui produit les interfaces par lesquelles les données sont rendues lisibles et d'autre part des usagers qui interagissent avec des services.

Néanmoins, l'instauration des publics par les concours n'a rien d'acquis. Penser les concours avec les concepts d'intéressement et d'enrôlement souligne le caractère temporaire de ces assemblages. Telles les coquilles Saint-Jacques qui ne se fixent plus après l'expérience des chercheurs, l'intéressement des publics de données se révèle bien souvent

précaire passée la remise des prix. À Rennes, à la suite du concours, de nouvelles données le budget de la commune, les prénoms des enfants, les résultats des élections ou encore la fréquentation de la bibliothèque ont bien été publiées sur le portail à la suite du concours. Mais ces nouvelles données ne semblent pas trouver de publics contrairement à celles qui avaient été publiées lors du concours.

> Après le concours qui avait suscité énormément de réactions, beaucoup de créations d'applications, qui avait très bien marché, il y avait eu en gros creux où il ne se passait plus rien. Il se disait que plus personne ne veut utiliser les données, alors qu'il y en a des nouvelles qui sortent et c'est quand même dommage.
> (L.K., responsable projet ***open data***, Rennes)

Alors que le concours avait fait émerger des services et des publics, la faible réutilisation des nouvelles données publiées amène des questionnements en interne quant à la persistance des publics de données. De leur côté, les agents de la ville ont repris leurs missions habituelles sans forcément prolonger le « savoir-être » fondé sur l'échange continu avec les développeurs et la prise en compte de leurs remarques qui était préconisé par les organisateurs du concours. Un membre du jury a décrit la phase qui suivait le concours comme un « trou noir » dans lequel le travail de production des publics qui animait une partie de l'administration municipale lors du concours a perdu rapidement en priorité.

> Après le concours, c'est le trou noir. Pourquoi c'est le trou noir ? Le concours est sorti donc, après, à cette époque-là, il y a pleins d'interrogations et on va dire qu'autant les bonnes étoiles étaient alignées à une époque autant là d'un seul coup elles commencent à se désaligner chez tous les acteurs. […] L'impression de réactivité qu'on avait donnée jusqu'ici, d'un coup on rentre dans un schéma où c'est beaucoup plus lent tu vois. Tout ce qui avait été pensé dans l'après-concours et qui était pour moi vraiment indispensable, par exemple une vraie communication autour de certaines applis, n'a pas été fait.
> (D.V., Responsable associatif, Rennes)

En l'absence de nouveau soutien technique, financier ou communication, certains développeurs n'ont plus mis à jour leurs services ce qui a remis en cause le public des usagers. En effet, les évolutions régulières des systèmes d'exploitation mobiles peuvent rendre une application rapidement obsolète pour les usagers. Par ailleurs, dans le domaine des transports, si les applications n'ont pas été mises à jour avec la dernière version des données, les services perdent leur utilité pour des usagers en situation de mobilité à la

recherche d'horaires fiables. En l'absence d'intéressement à la réutilisation des données, certains porteurs de projet ont progressivement abandonné le développement de leurs services que ce soit à Rennes ou à Montpellier.

> Avec le concours, il y a eu une vingtaine d'applications d'information voyageur sur les transports en commun qui ont été créées. Deux ans plus tard, il y en a maintenant peut-être une dizaine qui est à jour parce que les horaires changent, etc. Donc, il y a un certain nombre d'applis qui ont fait l'effort de continuer à maintenir leurs systèmes et d'autres qui ont renoncé parce qu'ils ne sont pas financés pour ça.
> (N.L., Responsable de la communication, Rennes Metropole)

> Les applications manquent de pérennité. […] Comme il n'y a pas de financement et bien ce n'est pas réparé. Et comme ce n'est pas réparé et bien ça tombe en panne, ça n'évolue pas. Il n'y a pas de com en plus : es applications c'est bien si les gens savent qu'elles existent ; si elles ne savent pas, elles sont mortes. Pour faire connaitre une application, il faut faire beaucoup de buzz et beaucoup de communication.
> (K.,B. Responsable innovation numérique, Montpellier)

Finalement, les concours de réutilisation de données ouvertes ne servent pas uniquement à la production de services, leur objectif affiché. Ils tentent de répondre au postulat essentiel des politiques d'*open data* : l'existence d'une demande de réutilisation des données, de publics qui disposent des capacités techniques et de la disponibilité de s'en saisir pour inscrire les données ouvertes dans de nouveaux réseaux. Par rapport aux fonctionnalités de visualisation ou aux métadonnées proposées dans les portails, l'instauration des publics n'est pas uniquement permise par l'instrument du concours, elle est stimulée et incitée par un dispositif d'intéressement qui crée un assemblage temporaire entre les données et ses publics.

Conclusion

Dans l'extrait qui figure en introduction de ce mémoire, Daniel Kaplan disait à propos des données publiques qu'« il se trouve qu'il y a des gens qui vous les demandent. » À travers les cas que nous avons pu voir dans ce chapitre, le postulat de l'existence de publics de données n'a pas nécessairement de fondements pratiques lorsqu'on se place du point de vue de celles et ceux qui ont la charge d'un projet d'*open data*. Ruppert (2012) a introduit cette notion de publics de données pour montrer comment une politique d'*open data*, le *Transparency Agenda* du gouvernement britannique, a été conçue pour renouveler la transparence publique. Elle montre que cette politique « imagine » des publics de données

qui pourraient renouveler et remplacer le contrôle de l'action réalisé par les experts et les fonctionnaires dont c'est la mission. L'ouverture des données détaillées sur les dépenses publiques, les salaires de fonctionnaires ou les avantages en nature reçus par les élus, qui découle du *Transparency Agenda*, renouvelle la transparence étatique en présumant de l'existence d'un public doté des compétences techniques pour manipuler ces données.

> *For Prime Minister David Cameron, putting information that was previously held by a few into the hands of all will give rise to "real people power" and a "post bureaucratic age" where government officials no longer are the keepers, arbiters, and interpreters of data. Power over data will ostensibly reside with an imagined public. Yet most people lack the tools or expertise to make sense of the terabytes of data being released*. (Ruppert, 2012)

Il y aurait donc en quelque sorte un script (Akrich, 1987) qui configurerait les politiques d'*open data* pour un public dont il serait attendu un certain type de compétences. Dans le même ordre d'idées, Barry (2001) a montré que la prolifération des technologies supposait l'émergence de « citoyens technologiques », capables de comprendre leur fonctionnement, de connaitre les risques associés à ces technologies et d'opérer des choix dans leur vie quotidienne. Sans faire référence aux travaux de Barry, Ruppert montre en quelque sorte que les politiques d'*open data* attendent une nouvelle sorte de citoyens technologiques avec les compétences techniques, l'intérêt et la disponibilité pour exploiter les données et exercer le contrôle de l'action publique en renforcement des garde-fous prévus par la loi. On peut apporter deux critiques principales à son analyse. Premièrement, le *Transparency Agenda* est un cas très particulier de politique d'*open data* orientée vers la transparence de l'action publique. En effet, il a été adopté en 2010 par le nouveau gouvernement de David Cameron en réaction au scandale des dépenses des parlementaires britanniques et pour renforcer le contrôle des élus jugés défaillant par certains. Or, même s'ils revendiquent la transparence comme un bénéfice de l'ouverture de données, beaucoup de projets d'*open data* tentent de favoriser un tout autre objectif, celui de la création de service pratique. Concrètement, cela peut orienter en partie le processus de l'ouverture comme nous avons pu le voir à propos du travail d'identification, lorsque les responsables de projets d'*open data* sélectionnent les données à ouvrir en fonction de leur utilisation potentielle dans des services ou lorsque les agents choisissent un format comme le GTFS configuré pour la réutilisation dans des applications. Deuxièmement, l'article de Ruppert ne se préoccupe pas de savoir jusqu'à quand et à quel point les publics peuvent rester imaginés. Or, pour les responsables de

projet d'*open data* et les agents, l'absence de publics constitue un problème très concret qui fait l'objet d'évaluation et qui, s'il n'est pas résolu, peut remettre en question tout l'édifice de l'ouverture des données. Au même titre que les données, les publics eux-mêmes font dans certains cas l'objet d'un véritable processus d'instauration.

Face au problème de la réutilisation, et de son absence, on a pu voir dans ce chapitre le travail que mettent en œuvre les responsables de projet d'*open data* pour instaurer des publics. Pour cela, ils mobilisent des instruments qui contribuent à assurer l'impératif d'intelligibilité des données évoqué dans le chapitre précédent. Ces instruments (j'aurais bien sûr pu en évoquer d'autres[128]) prennent des formes très différentes. Les deux premiers cas s'inscrivent dans l'espace du portail et ne provoquent pas directement l'émergence de publics. Nous avons ainsi pu voir que les métadonnées sont souvent présentées comme une « baguette magique » censée permettre une réutilisation sans frictions. Les responsables de projet d'*open data* les envisagent aussi comme un moyen de faciliter l'ouverture des données pour éviter leur édition et leur mise en intelligibilité. Cet instrument projette un public très générique de réutilisateurs contrairement aux fonctionnalités de visualisation qui permettent aux projets d'*open data* de cibler un public plus large qui n'est pas uniquement caractérisé par ses capacités techniques. Ces fonctionnalités de visualisation constituent une forme de réponse à une critique qui a été formulée par Michael Gurstein (2011). Selon ce chercheur canadien, les politiques d'*open data* risquent d'*empower the empowered*, de renforcer le pouvoir de ceux qui en ont déjà si elles ne s'assurent pas que les citoyens, au-delà du seul public des développeurs, disposent de la capacité d'utiliser les données ouvertes. Mais ces outils de visualisation peuvent aussi être critiqués, car ils réclament d'importantes transformations des données. Pour certains, la visualisation constitue une interprétation des données brutes qui va à l'encontre de certains des principes de l'*open data*. Lors de la réunion Open Data Bootcamp, une personne évoquée en introduction objectait par exemple que les agents de la région Ile-de-France « [sortaient] un peu de l'objectif de l'*open data* » parce qu'ils retraitaient des données brutes. Transformer les

[128] Par exemple, dans la réunion Open Data Bootcamp, les organisateurs ont évoqué par exemple les « cartoparties », des évènements courts lors desquels les participants contribuent à OpenStreetMap sur un territoire ou un domaine particulier. Au sein de l'Open Knowledge Foundation, nous avons aussi développé des actions de médiation comme les expéditions de données lors desquels des participants cherchent et exploitent des données ouvertes sur une thématique précise. La FING porte le projet d'Infolab, des lieux souvent créés avec l'aide des collecivités locales qui, à la manière d'un FabLab, proposent des outils et de l'assistance dans l'exploitation de l'information et de données.

données brutes pour favoriser leur usage par un public plus large constituerait une dénaturation de leur état « pur » et non modifié. Par ailleurs, nous avons pu voir que les standards opèrent des formes de réduction des données afin de favoriser l'interopérabilité et la lisibilité des données par les machines. Néanmoins, cela pourrait entraver l'intelligibilité des données pour les humains. À travers une ethnographie de l'exploitation de données géologiques sur une plateforme pétrolière, Almklov (2008) a montré que la standardisation décontextualise les données. Pour exploiter les données, les géologistes doivent les « recontextualiser » et réintroduire certains des éléments locaux qui avaient été évacués lors de leur réduction. Pour permettre leur intelligibilité par les machines, les outils de visualisation et les standards qui y sont inscrits peuvent évacuer certaines informations contextuelles et certaines traces de leur passé de données « métiers » qui pourraient pourtant être essentielles pour leur réutilisation. La mise en place de fonctionnalités de visualisation pourrait donc être aussi critiquée comme une forme de simplification des données.

Le cas des concours, étudié dans la troisième partie de ce chapitre, présente une autre forme d'instauration des publics. Là où les métadonnées et les outils de visualisation agissent sur l'intelligibilité des données pour les usagers, les concours stimulent la création d'applications et de services. Par des incitations financières et symboliques, ils tentent de créer des assemblages entre les données et les machines qui se révèlent souvent temporaires et incertains. Du point de vue des responsables de projet d'*open data*, les concours peuvent permettre de fournir la démonstration de l'existence des publics de données. Rosental (2009 ; 2002) a souligné l'importance grandissante des démonstrations devenues un outil de persuasion qui a débordé au-delà des mathématiques et des sciences et techniques. Les démonstrations permettent de convaincre rapidement et d'éviter de longs débats, elles doivent prouver du fonctionnement de l'objet ou du mécanisme qui est exposé. Pris sous cet angle, les concours fournissent la démonstration de l'existence des publics et de leur capacité à proposer des services pratiques dès lors que les données sont ouvertes. Mais ces démonstrations restent fragiles et l'instauration du public qu'elles permettent est finalement provisoire. Le cas de Rennes a ainsi montré que, passée la remise des prix, dernier temps de la démonstration, les publics de données se sont révélés incertains en dehors de l'assemblage temporaire du concours. Par ailleurs, les concours cadrent la participation : les partenaires financiers fixent le contenu des prix, les organisateurs peuvent imposer une thématique ou certaines données à réutiliser. Généralement, ils favorisent l'émergence d'un

public d'entrepreneurs qui témoigne de l'orientation de certaines politiques d'*open data* vers la création de valeur économique. Cela rejoint une des critiques adressées aux promoteurs de l'*open data* qui n'auraient pas assez pris en compte les conséquences politiques de leurs revendications. Yu et Robinson (2012) ont montré que les principes de Sebastopol étaient guidés essentiellement par des considérations techniques qui ne prennent pas en compte le contenu des données. C'est ce que nous avons pu voir dans le premier chapitre de ce mémoire. Un gouvernement dictatorial comme celui de la Corée du Nord pourrait ainsi ouvrir des données en respectant ces principes sans que cela contribue à renforcer la transparence et l'*accountability* souvent promus comme une des vertus de l'ouverture de données.

> *It is easy to imagine that a closed regime might disclose large amounts of data conforming to these eight requirements without in any way advancing its actual accountability as a government [...] An electronic release of the propaganda statements made by North Korea's political leadership, for example, might satisfy all eight of these requirements and might not tend to promote any additional transparency or accountability on the part of the notoriously closed and unaccountable regime.* (Yu & Robinson, 2012)

Des critiques plus fortes encore ont été formulées quant à l'agenda des promoteurs de l'*open data*. Certains ont par exemple reproché aux principaux acteurs qui ont défini les « grands principes » de l'ouverture des données d'être guidés par des considérations économiques et d'avoir des intentions politiques libérales non avouées. Pour Bates (2012), par exemple, les acteurs qui ont promu les politiques d'*open data* ont stratégiquement écarté les questions politiques en présentant leurs revendications comme politiquement neutres. Or, selon lui, l'ouverture des données sert notamment à la privatisation des services publics. Puisque les données sont ouvertes et que des publics sont censés les réutiliser, les usagers n'auraient qu'à choisir parmi les applications proposées par les acteurs privés plutôt que de se reposer sur le service public. Cette version critique de l'*open data* est loin d'être infondée. Elle est par exemple explicite dans les propos de Tim O'Reilly, un des organisateurs et l'hôte de la réunion de Sebastapol. Dans un article intitulé « *Governement as a Platform* » (O'Reilly, 2010), celui explique ainsi que l'État doit se contenter de fournir les données et les interfaces de programmation (API) au lieu de développer des services complets. Dans ce modèle, c'est aux développeurs de proposer des services, de manière lucrative ou non, pour remplacer les sites gouvernementaux, en utilisant les informations publiques dans de nouveaux contextes.

En encourageant les développeurs à exploiter les données publiques pour créer des services, les concours organisés par les institutions publiques peuvent prêter à la critique d'une libéralisation des services publics masquée sous une couche technologique supposément neutre[129].

Au-delà des critiques qui pourraient être formulées à l'égard de ces dispositifs, ce chapitre soulève le problème important de l'absence de publics. Comment expliquer que certaines données ouvertes ne trouvent pas de publics ? Pourrait-il en être autrement ? Dans Le Public et ses problèmes, Dewey définit un public comme ceux qui sont affectés par un problème et qui se mobilisent collectivement pour le résoudre. Les publics se constituent dans une démarche d'enquête collective où le problème est discuté et où, au fil des épreuves, des pistes de résolution se dessinent. Joëlle Zask (2008) a souligné l'importance de l'accès à l'information dans la conduite de l'enquête et de la constitution des publics : « un public est l'ensemble des gens ayant un plein accès aux données concernant les affaires qui les concernent, formant des jugements communs quant à la conduite à tenir sur la base de ces données et jouissant de la possibilité de manifester ouvertement ses jugements. » Par ailleurs, la thèse de Noortje Marres (2005) sur le débat entre John Dewey et Walter Lippmann a montré qu'en l'absence de problème à résoudre, il n'y a pas de publics. Elle le résume sous la formule « *no issue, no public*. » Dans une perspective pragmatiste, on peut ainsi dire que si le public n'utilise pas les données, c'est que celles-ci ne répondent pas, a priori, à des problèmes qui rassemblent suffisamment de personnes. On a pu voir que le processus d'ouverture des données inclut rarement des usagers potentiels « en chair et en os » qui

[129] La critique du libéralisme déguisé de la notion de « governement as a platform » est apparue en France dans un article de Sabine Blanc dans la Gazette des Communes du 6 juillet 2015 à l'occasion de la publication de la stratégie d' « État plateforme » élaborée par la Direction interministérielle des systèmes d'information et de communication (DISIC) en 2014 et 2015. Cette dernière est bâtie en grande partie sur l'ouvrage *L'âge de la multitude* (2013) co-écrit par Henri Verdier, l'actuel AGD et directeur de la DINSIC (direction interministérielle du numérique et du système d'information et de communication de l'État), et Nicolas Colin, auteur d'un rapport sur la fiscalité des données et fondateur de The Family, un accélérateur de start-ups. Dans leur ouvrage, ils se réfèrent directement à Tim O'Reilly pour imaginer des services publics « auto-organisés par des communautés de citoyens. » Sabine Blanc s'est interrogée sur la vision politique qui guide ce projet : « l'État plate-forme sera-t-il une façon douce de réduire la voilure sur les services publics, en en déléguant d'emblée une partie de la création de la version numérique à des acteurs extérieurs, et en n'assurant que le strict minimum au niveau de l'État ? » in La Gazette des Communes, « L'État plate-forme, vraie source de services publics innovants ou cache-misère ? », http://www.lagazettedescommunes.com/323547/letat-plate-forme-vraie-source-de-services-publics-innovants-ou-cache-misere/, consulté le 7 février 2015.

pourraient guider le travail d'identification et de transformation des données. Des usagers sont certes pris en compte lors de l'identification, certaines données étant sélectionnées en fonction de leur usage potentiel (cf. chapitre 3). Néanmoins, ces données ne ciblent pas un public au sens de Dewey, des personnes qui se mobiliseraient pour mener une enquête et résoudre un problème commun. Les données ainsi identifiées, tout comme les concours, configurent en fait un certain type d'usagers (Woolgar, 1991), généralement autour de la figure du développeur, pour sa capacité technique à produire les services et applications pour lesquels les projets d'*open data* sont en partie évalués.

On pourrait pourtant imaginer une autre manière d'instaurer des données qui partirait des problèmes des publics pour guider l'identification et la transformation des données. Le modèle de la statistique publique pourrait par exemple servir d'inspiration pour créer des espaces de concertation dans lesquels les publics pourraient faire de leurs demandes d'information publique. Une institution, le Conseil National de la Statistique (CNIS), assure en effet depuis 1984 la concertation entre les usagers des informations statistiques et ses producteurs. Il est composé de représentants des institutions, des syndicats, des organisations patronales, de la recherche et de la société civile. Sujobert a montré comment des syndicalistes, des chercheurs et des associations sont parvenus à intervenir au sein de cette instance pour faire modifier les méthodes de calcul des inégalités employées par l'INSEE (Sujobert, 2014). Cette méthode de concertation pourrait servir d'inspiration pour inclure dans le travail d'instauration des données des publics qui ont besoin d'information pour porter leurs revendications.

Conclusion

Pour conclure ce mémoire, mesurons le chemin parcouru avant d'ouvrir de nouvelles pistes. En retraçant six moments de définition présentés sous la forme d'épisodes, le premier chapitre a montré comment, depuis près de dix ans, des acteurs ont réclamé une transformation des politiques en matière d'information publique. Ces revendications ont pris de multiples formes (des manifestes, des principes, des outils de classement ou encore des traités) et trouvent des ressources dans des mouvements qui ont réclamé l'ouverture des standards de l'informatique, des télécommunications ou des sciences. Ces acteurs ont inventé une nouvelle catégorie dans les informations publiques, les données ouvertes, des ressources librement réutilisables d'un point de vue juridique et technique dans de nouveaux contextes, pour d'autres usages. Dans la lignée des mouvements du logiciel libre, ces revendications se sont appuyées sur deux grands leviers d'action, les licences et les standards de données. J'y ai repéré une tension entre deux modèles de l'ouverture : l'un formulant des critères très techniques pour l'ensemble des données, l'autre s'intéressant à leur contenu en distinguant des données essentielles et prioritaires. Au-delà de ces divergences, ces acteurs ont mis en lumière les données brutes qui constitueraient une matière première non altérée capable de circuler de manière fluide et à faible coût. Ces données brutes seraient le matériau informationnel avant son traitement, capable de réduire les asymétries d'information et de décentraliser les centres de calcul. Surtout, ces acteurs ont porté l'attention sur les données, là où la législation considérait essentiellement des documents administratifs ou des informations publiques.

En suivant la trajectoire de 2011 à 2016 d'une institution, Etalab, le deuxième chapitre a réduit la focale pour montrer comment ces grands principes ont été traduits en politiques publiques. La politique mise en œuvre par Etalab s'est revendiquée de l'*open data* afin de valoriser la transparence du président sortant et de combler « le retard » de la France. La traduction des grands principes de l'*open data* a transformé les politiques de réutilisation des informations publiques dans le sens de la gratuité et de la lisibilité des données par les machines. La trajectoire de cette institution souligne l'instabilité de la politique d'*open data* ; ses missions, son cadre légal et ses objectifs sont mouvants et ses actions restent liées à ses attaches politiques. La nomination d'un administrateur général en charge d'organiser la

circulation et l'exploitation des données a montré que les politiques d'*open data* ont aussi eu pour effet de porter l'attention sur les données considérées comme une « ressource inexploitée », le « nouveau pétrole » gisant sous les organisations. Les projets d'*open data* ont fait entrer les données brutes en politique : ils ont transformé l'État en organisant des lieux dédiés à leur circulation et en instituant leur exploitation dans de nouveaux réseaux comme un objet de gouvernance.

Lorsqu'on quitte le terrain de l'élaboration de ces principes et de ces politiques et que l'on s'intéresse à l'*open data* en train de se faire, le contraste entre les discours qui considèrent les données comme une ressource disponible et la réalité du travail des agents administratifs est saisissant. Le processus d'ouverture des données débute par une première épreuve, l'identification. Les responsables de projets d'*open data* ne rencontrent pas des entités localisables et reconnues comme données, ils doivent les identifier. Ils ne disposent pas non plus de catalogues dans lesquels ils pourraient sélectionner les données à ouvrir en priorité au commencement d'un projet d'*open data*. L'identification prend plutôt la forme d'explorations au cours desquelles les agents qui en ont la responsabilité arpentent les services à la recherche de données à ouvrir. Au cours de ces explorations, ils suivent la multiplicité des pistes qui se présentent à eux lors de nombreuses rencontres avec des agents administratifs, en s'inspirant des données publiées dans d'autres organisations ou de cas d'usage identifiés ailleurs. Pour répondre aux objectifs quantitatifs de publication de données qui leur sont souvent assignés, les projets d'*open data* « travaillent l'organisation » par la mise en place de réseaux de correspondants qui distribuent l'identification et attribuent des responsabilités inédites à des agents. L'identification engendre une nouvelle réalité : elle instaure comme donnée des informations, des documents ou des fichiers servant au travail de gestion administrative.

L'intérêt de la notion d'instauration, proposée par l'esthéticien Souriau et reprise par Latour, c'est qu'elle insiste sur la matérialité de l'objet instauré. Les données ne sont pas créées ou inventées, elles le deviennent par le travail et les efforts de celles et ceux qui les instaurent. Cette matérialité qui résiste est aussi au cœur de la notion de friction proposée par Edwards (2010) pour caractériser les désordres et les résistances qui surviennent immanquablement lors de la circulation des données. Le quatrième chapitre prolonge le précédent en s'intéressant aux sources de friction qui empêchent l'ouverture de données

localisées lors de l'identification. Plutôt que d'évacuer les arguments avancés par les agents lorsqu'ils s'opposent ou résistent à l'ouverture de leurs données, j'ai pris au sérieux les arguments qu'ils avancent et les contraintes auxquelles ils font face pour comprendre les « bonnes raisons organisationnelles » pour lesquelles ils n'ouvrent pas leurs données. Les systèmes d'information constituent la première source de friction que j'ai identifiée. Lorsqu'ils ouvrent des données, les responsables de projet d'*open data* font face à l'épaisseur pratique des systèmes d'information avec lesquels les agents assurent les activités de gestion dont ils ont la charge. Pour obtenir des données jamais sorties des réseaux sociotechniques de l'organisation, les gestionnaires de bases de données doivent conduire une véritable enquête pour reconstituer leur structure et mettre en place des « moulinettes », des outils qui désarticulent les assemblages pour permettre l'extraction. Les frictions portent aussi sur le contenu des données. En s'intéressant à des objets qui n'étaient jamais sortis des réseaux de l'organisation, les projets d'*open data* mettent à l'épreuve la qualité des informations produites par les agents. Leur ouverture au public peut mettre en lumière des manquements ou des erreurs jamais remarquées ou qui pouvaient s'accoutumer d'approximations pour les usages pour lesquels elles étaient conçues. L'ouverture sans modifications de ces données constitue un risque pour les agents, car les critiques à l'égard de leur qualité ou de leur « saleté » peuvent être perçues par la hiérarchie comme le signe d'un travail bâclé. Ces critiques pourraient rejaillir sur la carrière des agents ou la réputation du service alors même que le projet d'*open data* ne fait pas partie des missions officiellement assignées. Dans certains cas, les données peuvent comporter des informations qui, entre les mains d'usages malveillants, peuvent servir à dégrader le patrimoine public ou à mettre en danger les populations. Enfin, l'exposition d'informations « sensibles » politiquement peut constituer un risque pour les agents qui ne peuvent pas « jouer la transparence » sans avoir l'aval de la hiérarchie administrative et politique. Ces cas révèlent l'épaisseur de procédures plus ou moins formelles, de circuits de validation plus ou moins établis qui conditionnent l'ouverture à l'obtention d'un mandat délivré par la hiérarchie politique et administrative.

Lorsqu'on s'intéresse aux transformations apportées aux données avant leur ouverture, comme c'est le cas dans le cinquième chapitre, on constate l'« oxymore des données brutes » que Bowker puis Gitelman ont souligné. Dans les administrations, les données sont souvent produites dans l'environnement du tableur, l'outil le plus courant pour stocker, traiter, visualiser et échanger des données. Mais l'impératif de lisibilité des données par les

machines, affirmé par les responsables de projet d'*open data* de manière plus ou moins contraignante, réclame de convertir ces données. Même si le CSV est un format vanté pour sa simplicité, sa flexibilité et sa lisibilité par les machines, il demande des transformations en profondeur des fichiers. En effet, certaines informations inscrites sous la forme d'un texte en gras, de la couleur, ou de cellules fusionnées, disparaissent lors de la conversion et doivent être traduites sous une autre forme. En choisissant le CSV, les agents consentent à un investissement dans l'intelligibilité des données par les machines. Guidés par le même objectif, certains standards imposent une structure de fichiers et des définitions précises des objets qu'ils contiennent. C'est le cas du GTFS, un format conçu pour rendre interopérables et exploitables les horaires de transport dans des applications mobiles. L'implémentation de ce standard nécessite des transformations organisationnelles et infrastructurelles conséquentes, un investissement supplémentaire dans l'intelligibilité des données par les machines. On entrevoit avec ce cas un horizon des politiques d'*open data* vers l'« harmonisation » des données entre les institutions, une standardisation dont on mesure le coût considérable avec le cas du GTFS. Enfin, les transformations peuvent intervenir du fait même des agents afin de réduire les risques pour leur carrière que comporte la publication de données qui pourraient être jugées comme « de mauvaise qualité », « sales » ou « incompréhensibles ». Dans certains cas, l'édition des données peut aussi servir à rendre intelligibles les catégories administratives et anonymiser certaines informations pour respecter les obligations légales.

Le travail largement invisible de l'ouverture de données repose sur l'existence d'une demande de réutilisation par des publics. Ces publics doivent disposer des compétences et de la disponibilité pour produire les visualisations, les services et les applications qui sont attendus à la suite des projets d'*open data*. En plus d'ouvrir des données, les agents en charge de ces projets doivent donc aussi s'assurer de leur réutilisation. Or, en pratique, les usages des données n'arrivent souvent pas d'eux-mêmes. C'est un autre pan du travail invisible de l'ouverture des données qu'a révélé l'enquête. Dans le sixième chapitre, je me suis intéressé à trois instruments qui contribuent à instaurer des publics de données. Pour permettre au « grand public » d'exploiter les fichiers mis à disposition, les portails proposent des fonctionnalités de visualisation qui évitent les frictions du téléchargement et de l'ouverture. Mais ces fonctionnalités imposent leurs spécifications de standard comme le CSV et déplacent plus en amont les contraintes de l'exploitation de données. Il faut alors bien

souvent appliquer de nouvelles transformations aux données pour qu'elles deviennent interprétables par les scripts informatiques. Dans l'environnement du portail, les métadonnées contribuent aussi à l'intelligibilité des données brutes. Dans les sciences comme dans les projets d'*open data*, ces dernières sont souvent présentées comme une sorte de « baguette magique » qui permettrait de réutiliser des données « sans friction », sans que les usagers aient à échanger avec les producteurs. Les responsables de projets d'*open data* s'en servent pour convaincre les agents d'ouvrir les données brutes, en leur permettant de signaler les erreurs ou les manquements que pourraient percevoir les usagers. Enfin, les concours montrent un autre aspect du travail des équipes en charge des projets d'*open data*. Les instruments par lesquels elles instaurent des publics peuvent aller jusqu'à inciter symboliquement et financièrement les réutilisations. En se plaçant entre les développeurs et les données, dans un dispositif qui incite et encadre la participation, les concours créent de nouveaux agencements sociotechniques autour des données. Mais leur solidité est mise à l'épreuve, une fois les prix remis, lorsque l'intéressement s'arrête et révèle le côté temporaire et incertain des agencements créés.

Revenons sur ce qui est fait aux données. Que faire de l'idée même de données brutes au regard de ce que nous avons observé ? Comment interpréter la demande récurrente de données brutes, primaires ou non modifiées par rapport au travail d'instauration et de transformation que j'ai mis en lumière ? J'ai évoqué, dans la section sur l'édition, l'enquête ethnographique sur la production de données scientifiques en Amazonie conduite par Antonia Walford. Un chapitre entier de sa thèse porte sur la notion de données brutes en sciences. Elle y montre que celles-ci apparaissent dans les mains de ceux qui les collectent, puis de ceux qui les transforment, comme des entités ambiguës qui doivent encore être consolidées pour produire des résultats scientifiques. En quelque sorte, les données brutes « attendent » leur inscription dans un réseau sociotechnique stabilisé (Walford, 2013). Cette inscription consiste notamment à écarter le « bruit » et les « artefacts » qui parasitent l'accès à la réalité observée par les instruments. Le nettoyage permet de passer de données brutes à des données certifiées qui peuvent circuler d'un monde à l'autre, d'une discipline à l'autre. Le lien avec les projets d'*open data* dans les administrations publiques est évident puisqu'une partie du travail effectué sur les données vise également à leur mise en intelligibilité. Cependant, dans les programmes d'*open data*, le processus est en quelque sorte inversé. Les objets instaurés en données ouvertes ont déjà eu une longue vie sociale et

sont ancrés dans des usages parfois anciens. Ces données sont déjà inscrites dans des réseaux sociotechniques qui les stabilisent et les orientent vers des pratiques spécifiques. L'enjeu des tâches mises en œuvre pour assurer leur ouverture ne consiste pas à les débarrasser des traces de leur fabrication, mais au contraire d'en élargir l'usage possible. Les activités opérées en coulisses tentent de désencastrer des données situées dans les réseaux sociotechniques de leur production. En éliminant les traces de leur vie passée de données « métiers », les agents administratifs les transforment pour devenir des données intelligibles et ouvertes à de nouveaux traitements. Bien entendu, ce désencastrement des données « métier » ne veut pas dire que les données puissent exister par elles-mêmes, une fois « libérées. » L'ouverture opère un réencastrement des données dans un nouveau réseau sociotechnique avec ses propres formes de réduction et de clôture, notamment par le biais des formats et des standards.

Plus que de remettre en cause la définition des « données brutes », explorer les coulisses des projets d'*open data* permet d'insister sur le travail des données. On l'aura compris, les données ne flottent pas dans les nuages et ne tombent pas du ciel ; elles sont manipulées, fabriquées, transformées, ajustées dès lors qu'elles doivent circuler et être exploitées. En d'autres termes, la circulation fluide des données, qui fonde les politiques d'*open data*, ne s'opère pas sans coûts et sans frictions. Néanmoins, si les opérations mises en œuvre pour ouvrir concrètement certaines données étaient décrites comme difficiles, et leur invisibilité parfois douloureuse, leur mise en lumière ne débouchait jamais sur une remise en cause frontale de l'idée même de données brutes. Comment comprendre cela ? Comment ne pas voir une opposition très claire entre l'injonction à l'ouverture de données brutes, « à l'état pur » et des opérations qui modifient en profondeur les données pour assurer leur ouverture ? C'est une responsable de projet d'*open data* qui nous a permis, lors de l'écriture avec Jérôme Denis d'un article à paraître, d'apporter la réponse en expliquant, dans ses propres termes, comment le travail opéré sur les données conditionnait de l'existence des données brutes.

> Pour les statistiques de fréquentation, par exemple, c'était leur fichier de travail [du département]. C'était un fichier Excel qu'ils avaient mis en forme selon ce dont ils avaient besoin. Ils avaient fait un tableau avec leur propre titre de colonnes, des couleurs... […] Donc, c'était vraiment leur fichier de travail. Or, nous, on ne voulait pas ça. Nous, on voulait des données plus brutes, c'est-à-dire pas de

> commentaires, pas de tableaux, pas de mise en forme, juste vraiment les données au jour le jour, statistiques. Moi, je me suis occupée de ce travail-là, rebrutifier les données en fait, pour qu'elles soient vraiment le plus simple possible à utiliser ensuite pour les développeurs.
> (L.K., responsable du projet d'***open data***, Rennes)

Plutôt que d'opposer d'un côté les demandes de données brutes et les transformations que nous avons pu observer de l'autre, il faut plutôt voir comment les deux s'articulent en comprenant le processus de l'ouverture des données comme un travail de rebrutification. L'identification, l'extraction, la conversion, les transformations et la production de l'intelligibilité humaine et technique ne luttent pas contre la demande de données brutes, mais y répondent. À partir de ces résultats de cette enquête dans les coulisses de l'*open data*, nous avons tout intérêt à prendre au sérieux l'idée de rebrutification et à comprendre les données brutes, pas seulement comme un oxymore pour les sciences humaines (Bowker, 2000 ; Gitelman, 2013), mais comme un oxymore avec lequel les travailleurs des données doivent composer au quotidien. De ce point de vue, les données brutes ne sont pas une illusion, mais une chose complexe et fragile à fabriquer.

Par ailleurs, au-delà de la question des données brutes, une telle posture pragmatiste, qui prend au sérieux le vocabulaire et les pratiques des acteurs, peut aider à reconsidérer la notion même de données. Comme Borgman (2015) l'affirme et semble le regretter, la nature des données, ce qu'elles sont, reste extrêmement vague.

> The inability to anchor the concept in ways that clarify what are data and are not data in a given situation contributes mightily to the confusion about matters such as data management plans, open data policies and data curation. (Borgman, 2015, p. 28–29)

En effet, il est particulièrement frappant de ne trouver nulle part une définition claire et stable de ce qu'est (ou n'est pas) une donnée dans les textes et dans les discours qui fondent les politiques d'*open data*. Néanmoins, une telle absence de définition stable est-elle tant un problème ? Peut-être pas. Ou peut-être elle nous invite à faire un pas de côté et à reconsidérer comment nous comprenons les données. En observant les personnes qui assument concrètement l'ouverture des données, ce flou n'a rien de surprenant. Nous avons vu que les données sont rarement considérées comme telles dans les administrations, à l'exception de certains services dédiés à leur production, et ne prennent pas la forme

d'informations numériques prêtes à circuler automatiquement. Une fois que nous avons conscience du travail qui est accompli pour instaurer progressivement les données, nous pouvons nous rappeler du jeu de mots de Latour (1993) sur les données : « décidément, on ne devrait jamais parler de "données", mais toujours d'"obtenues". » En présupposant l'existence de données brutes dans les administrations et en prônant la circulation fluide des données non modifiées qui doivent être non seulement de bonne qualité, mais aussi intelligibles par les humains et les machines, les militants de l'*open data* ont rendu le travail d'obtention des données invisible. L'absence de considération pour le travail des données réalisé par les agents administratifs peut être vue comme un moyen pour les militants de l'*open data*, porte-paroles des futurs usagers des données, de ne pas avoir à l'accomplir.

Il ressort donc très clairement de cette enquête que les données sont le produit d'un travail méticuleux et invisible de façonnage, mais ce n'est pas qu'une première étape pour comprendre le processus de leur ouverture. Car, au terme de cette enquête, il faut aussi respécifier ce que nous entendons par « données. » Dans un chapitre de *Raw Data is an Oxymoron*, Rosenberg (2013) est revenu en détail sur les origines du mot data en s'appuyant sur une analyse lexicographique des livres numérisés dans Google Books et ECCO, une base de données sur les livres anglophones au XVIIIe siècle. Il montre que le terme est entré dans le dictionnaire en 1646, sous sa forme plurielle, pour désigner le matériau de base de l'analyse et du calcul, sans que son rapport à la réalité soit pris en considération. Au milieu du XVIIIe siècle, le mot « *data* » signifiait aussi les faits et les preuves produits par l'expérimentation et la collection avant de désigner, au XXe siècle, aux informations sous forme numérique. La donnée scientifique, explique-t-il, a pendant longtemps désigné le matériau de base de l'analyse et du calcul, sans que son rapport à la réalité soit pris en considération. La donnée était un point de départ du travail scientifique, ce avec quoi il fallait faire, et pas forcément un « bon » représentant du réel. Étudier les pratiques d'*open data* de l'intérieur invite à ne pas négliger cet aspect au seul motif que nous avons insisté sur les difficultés à obtenir des données prêtes à être ouvertes. Les données qui ont été scrupuleusement façonnées peuvent être considérées comme un produit. Mais ce produit doit aussi être pensé comme un *input*, un intrant destiné à être transformé par ses futurs utilisateurs, un point de départ qui peut être utilisé sans avoir à questionner ses conditions de véracité. De ce point de vue, les données s'apparentent à des « dons », des objets soigneusement façonnés par les agents administratifs offerts au public. La notion d'oxymore

proposé par Bowker est toujours utile dans cette perspective puisqu'elle permet de révéler le travail invisible effectué par celles et ceux qui consentent ce « don ». Néanmoins, ce « don » se fait dans des conditions très particulières dans lesquels les agents doivent souvent répondre aux injonctions de leur hiérarchie et aux revendications de ceux qui vont le recevoir. Par ailleurs, la multiplicité des sens du mot data que Rosenberg a identifié nous permet aussi de comprendre que les données ne peuvent pas être désignées comme des objets fixes, identifiables par une série de caractéristiques définies une fois pour toutes. On peut alors, comme Leonelli (2015), considérer les données comme une catégorie relationnelle, mais à condition de répondre à deux critères, leur utilisation comme preuve potentielle et la capacité à circuler. Or, ces critères ne peuvent pas s'appliquer pour les données publiques. En effet, hors des lieux de production du savoir scientifique, le critère d'administration de la preuve ne suffit pas à clarifier ce qui distingue les données des autres objets informationnels. Au terme de cette enquête, je propose de reformuler la question. Suivant le questionnement proposé par Engeström à propos des outils (1990), plutôt que de demander « *what is data* ? », nous devrions demander « *when is data* ? ». En effet, j'ai montré que, à travers le processus de l'ouverture, les fichiers, les documents, les nombres, les textes ou les images des administrations ne deviennent des données qu'à partir du moment où elles peuvent servir de point de départ à l'assemblage de nouveaux réseaux sociotechniques par leurs usagers.

En considérant les données comme un point de départ de nouveaux réseaux sociotechniques, on voit se dessiner une des premières lignes de fuite de ce travail. Dans le dernier chapitre, j'ai montré l'épaisseur du travail que mettaient en œuvre les agents pour faire exister les publics de données, mais, dans certains cas, certaines données ne trouvent tout bonnement pas de publics. Quel peut être le statut des données qui n'ont pas trouvé de publics ? Sont-elles encore considérées par les agents comme des données même si elles ne constituent pas un point de départ ? Le travail d'instauration pourrait alors s'estomper et ces données pourraient ne plus être considérées comme telles. Si des réseaux sociotechniques ne se tissent pas autour des données, les agents pourraient ne plus consentir à l'investissement considérable qui permet de façonner des données ouvertes, d'autant plus que ce travail est largement invisibilisé et n'entre pas dans les missions des administrations. Au-delà de la question des données, on pourrait prolonger l'enquête pour étudier les évolutions des projets d'*open data*. En effet, la majorité des cas étudiés ici

concerne des données ouvertes récemment qui souvent n'étaient jamais sorties des réseaux sociotechniques de l'organisation. Mais que se passe-t-il quelques années après l'ouverture ? Dans quels cas l'ouverture des données entre-t-elle dans les routines des agents et dans les missions des services ? Comment évoluent les missions des responsables de projets d'*open data* ? Comment l'apparition de nouveaux rôles et de nouvelles compétences dans l'administration, comme celles de *data editor* ou de *data scientist*, change-t-elle les conditions concrètes de l'ouverture des données ? Comment se déroule l'ouverture des données quand elle devient une obligation légale ? Et comment les agents parviennent-ils à distinguer les données des autres objets informationnels ? Enfin, une troisième piste de poursuite de cette enquête concerne les publics des données. On a pu voir, tout au cours du processus de l'ouverture, que les données étaient souvent configurées pour des développeurs. Mais on pourrait imaginer que les politiques d'*open data* soient conçues pour un public plus large, que les données ouvertes ne ciblent pas uniquement les développeurs. Au sein de l'Open Knowledge Foundation, j'ai participé depuis plus de deux ans au lancement du projet Ecole des Données (ecoledesdonnees.org) qui, à travers des cours et des activités de médiation réutilisables gratuitement et librement, essaie de permettre à quiconque d'utiliser des données ouvertes sans compétences préalables. C'est aussi le sens de mon prochain projet qui donnera une suite pratique à cette thèse. Avec Joël Gombin, politologue spécialiste de l'utilisation de données électorales, nous montons datactivi.st, une société coopérative qui proposera des activités de conseil, de formation et de médiation sur l'ouverture des données et qui mettra en pratique ce qu'elle prêche en publiant sous licence libre le contenu des formations qu'elle produit. À travers datactivi.st, nous allons essayer de nous adresser en particulier à des associations et des militants pour voir quelles formes pourraient prendre le « datactivisme », à travers par exemple la critique des données publiées, la production de nouvelles en réponse aux données ouvertes de l'administration ou le militantisme pour l'ouverture de données spécifiques. Par rapport aux pratiques « statactivistes » (Bruno, Didier & Prévieux, 2014), nous voulons essayer de comprendre ce qui change lorsque des militants sont équipés de données. Nous nous lançons officiellement le 23 septembre 2016 à Aix-en-Provence en marge de la Data Litteracy Conference organisée par la FING, une conférence qui part du principe que « les données sont désormais une affaire trop importante pour être laissées entre les mains des spécialistes et les considère comme un élément de la "littératie" qui, au même titre que la lecture et l'écriture, peut être accessible

à tous.[130] Sans entrer dans les implications d'une telle notion, on peut noter que cet évènement organisé par la FING reboucle avec le discours de son directeur évoqué en introduction, Daniel Kaplan. Il rappelle que les publics de données ne peuvent pas rester un postulat et qu'au même titre que les données, ils doivent souvent être instaurés.

[130] « Data Literacy Conference », http://dataliteracyconference.net/, consulté le 12 juillet 2016.

Bibliographie

Akrich, M. (1987), « Comment décrire les objets techniques ? », *Techniques et Culture* (1), pp. 49-64.

Akrich, M. (2012), « Les listes de discussion comme communautés en ligne : outils de description et méthodes d'analyse », *Papiers de recherche du CSI* (025).

Akrich, M., Bijker, W. & Law, J. (1992), "The De-scription of Technical Objects", in *Shaping Technology—Building Society: Studies in Sociotechnical Change,* Cambridge, MA, .

Almklov, P.G. (2008), "Standardized Data and Singular Situations", *Social Studies of Science* 38(6), pp. 873–897.

Baker, K.S. & Bowker, G.C. (2007), "Information ecology: open system environment for data, memories, and knowing", *Journal of Intelligent Information Systems* 29(1), pp. 127–144.

Barry, A. (2001), *Political Machines. Governing a Technological Society. ,* London, The Athlone Press.

Bates, J. (2012), "'This is what modern deregulation looks like' : co-optation and contestation in the shaping of the UK's Open Government Data Initiative", *Journal of community informatics* 8(2).

Beltrame, T.N. & Jungen, C. (2013), « Cataloguer, indexer, encoder », *Revue d'anthropologie des connaissances* 7, 4(4), p. 747.

Birchall, C. (2014), "Radical Transparency?", *Cultural Studies ↔Critical Methodologies* 14(1), pp. 77–88.

Birchall, C. (2015), "'Data.gov-in-a-box': Delimiting transparency", *European Journal of Social Theory* 18(2), pp. 185–202.

Blanchette, J. (2011), "A material history of bits", *Journal of the American Society for Information Science and Technology* 62(6), pp. 1042–1057.

Borgman, C. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World,* Cambridge, MA, The MIT Press.

Boustany, J. (2013), « Accès et réutilisation des données publiques. Etat des lieux en France », *Les cahiers du numérique* 9(1), pp. 21-37.

Bowker, G.C. (2000), "Biodiversity datadiversity", *Social Studies of Science* 30/5(643).

Bowker, G.C. (2005), *Memory Practices in the Sciences,* The MIT Press.

Breton, P. (2004), *L'utopie de la communication : le mythe du « village planétaire »,* La Découverte.

Broca (2013), *Utopie du logiciel libre. Du bricolage informatique à la réinvention sociale,* Neuvy-en-Champagne, Le Passager Clandestin.

Bruno, I. & Didier, E. (2013), *Benchmarking : l'État sous pression statistique,* Paris, Zones.

Bruno, I., Didier, E., & Prévieux, J. (dir.), (2014), *Statactivisme. Comment lutter avec des nombres,* Paris, Zones.

Busch, L. (2011), *Standards: recipes for realities,* Cambridge (MA), The MIT Press.

Callon, M. (1986), « Elements pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins pêcheurs dans la baie de Saint Brieuc », *L'Année Sociologique* 36, pp. 169-208.

Campbell-Kelly, M. (2007), "The rise and the rise of the spreadsheet", in *The History of Mathematical Tables. From Sumer to Spreadsheets,* sous la direction de Campbell-Kelly et al, Oxford University Press.

Castelle, M. (2013), "Relational and Non-Relational Models in the Entextualization of Bureaucracy", *Computational Culture*.

Cefaï, D. (1996), « La construction des problèmes publics. Définitions de situations dans des arènes publiques », *Réseaux* 14(75), pp. 43-66.

Chesbrough, H. (2006), *Open Business Models: How to Thrive in the New Innovation Landscape,* Cambridge (MA), Harvard Business School Publishing.

Chignard, S. (2012), *Open data. Comprendre l'ouverture des données publiques,* Limoges, Fyp Editions.

Cobb, R. & Elder, C. (1972), *Participation in American Politics: the Dynamics of Agenda Building,* John Hopkins University Press.

Cochoy, F., Garel, J.-P. & de Terssac, G. (1998), « Comment l'écrit travaille l'organisation : le cas des normes ISO 9000 », *Revue française de sociologie* XXXIX(4), pp. 673-699.

Coleman, G. (2013), *Coding Freedom. The Ethics and Aesthetics of Hacking,* Princeton University Press.

Colin, N. & Verdier, H. (2012), *L'Age de la Multitude. Entreprendre et gouverner après la révolution numérique.* Paris, Armand Colin.

Collins, H. (1974), "The TEA Set: Tacit Knowledge and Scientific Networks", *Science Studies* 4(2), pp. 165–186.

Dagiral & Peerbaye, A. (2013), « Voir pour savoir. Concevoir et partager des "vues" à travers une base de données médicales », *Réseaux*, pp. 163-196.

Chrzanowski. (2011). *Data.Gov.Uk, l'ouverture des données publiques au Royaume-Uni,* Londres, Ambassade de France au Royaume Uni.

Demeyer, T. (2012), "Apps For Amsterdam", *Journal of community informatics* 8(2).

Denis, J. & Pontille, D. (2012), « Travailleurs de l'écrit, matières de l'information », *Revue d'anthropologie des connaissances* 6, 1(1), p. 1.

Despret, V. (2015), *Au Bonheur des morts,* Paris, La Découverte.

Desrosières, A. (2000), *La politique des grands nombres : histoire de la raison statistique,* Paris, La Découverte.

Desrosières, A. (2005), « Décrire l'Etat ou explorer la société : les deux sources de la statistique publique », *Genèse* (58), pp. 4-27.

Desrosières, A. (2015), « La qualité est-elle la condition de l'harmonisation européenne ? », in *Prouver et Gouverner. Une analyse politique des statistiques publiques. ,* Paris, La Découverte, .

Dewey, J. (2010), *Le public et ses problèmes,* Paris, Gallimard.

Dodier, N. (2003), *Leçons politiques de l'épidémie de sida,* Paris, Editions de l'EHESS.

Dodier, N. (2005), « L'espace et le mouvement du sens critique », *Annales. Histoire, sciences sociales* (1), pp. 7-31.

Dodier, N. & Bazanger, I. (1997), « Totalisation et altérité dans l'enquête ehnographique », *Revue Française de Sociologie* 38(1), pp. 37-66.

Donovan, K.P. (2012), "Seeing Like a Slum: Towards Open , Deliberative Development", *Georgetown Journal of Informational Affairs* 13(1), pp. 97–104.

Driscoll, K. (2012), "From Punched Cards to ' Big Data ': A Social History of Database Populism", *communication +1* 1.

Edwards, P. (2010), *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming,* Cambridge, The MIT Press.

Edwards, P. et al (2011), "Science friction: Data, metadata, and collaboration", *Social Studies of Science* 41(5), pp. 667–690.

Engenström, Y. (1990), "When is a tool? Multiple meanings of artifacts in human activity ", in *Learning, working and imagining,* Helsinki, Orienta-Konsutit, pp. 171–195.

Espeland, W.N. & Sauder, M. (2007), "Rankings and Reactivity: How Public Measures Recreate Social Worlds", *American Journal of Sociology* 113(1), pp. 1–40.

Espeland, W.N. & Stevens, M.L. (1998), "Commensuration as a social process", *Annual review of sociology* 24(1998), pp. 313–343.

Fecher, B. & Friesike, S. (2014), « Open Science : One Term, Five Schools of Thought », in *Opening Science,* sous la direction de S. Friesike & S. Bartling, New York, Springer, .

Ferretti (2007), "Why Public Participation in Risk Regulation? The Case of Authorising GMO Products in the European Union ", *Science as Culture* 4(16).

Flichy, P. (2001), « La place de l'imaginaire dans l'action technique », *Réseaux* 109(5), p. 52.

Fraenkel, B. (1994), « Le style abrégé des écrits de travail », *Cahiers du français contemporain* (1), pp. 177-194.

Garfinkel, H. & Bittner, E. (1967), « 'Good' organizational reasons for 'bad' clinic records », in *Studies in Ethnnomethodology,* sous la direction de H. Garfinkel, Englewood-cliffs, Prentice-Hall, pp. 186-207.

Gitelman (dir.), (2013), *"Raw Data" Is an Oxymoron,* Cambridge (MA), The MIT Press.

Gitelman, L. (2014), *Paper Knowledge. Toward a Media History of Documents,* Durham (NC), Duke University Press.

Glaser, B. & Strauss, A. (1967), *The discovery of grounded theory: strategies for qualitative research,* New York, .

Goffman, E. (1973), *La mise en scène de la vie quotidienne 1 : La présentation de soi,* Minuit.

Gourgues, G. (2012), « Penser la participation publique comme une politique de l'offre : une hypothèse heuristique », *Quaderni* (79).

Gourgues, G., Rui, S. & Topçu, S. (2013), « Critique de la participation et gouvernementalité : Lectures Critiques », *Participations* 2(6).

Grosjean, M. & Lacoste (1999), « L'analyse des communications de travail : perspectives et méthodes », in *Communication et intelligence collective. Le travail à l'hôpital. ,* Paris, PUF, pp. 45-75.

Gurstein, M. (2011), "Open data: Empowering the empowered or effective data use for everyone?", *First Monday* 16(2).

Haigh, T. (2013), "How Data Got its Base : Information Storage Software in the 1950s and 1960s", *Annals of the History of Computing* 31(4), pp. 6–25.

Hansen, H.K. & Flyverbom, M. (2014), "The politics of transparency and the calibration of knowledge in the digital age", *Organization* 1(18).

Hassenteufel (2011), *Sociologie Politique : l'Action Publique,* Paris, Armand Colin.

Hibbing, J. & Theiss-More, E. (2002), *Stealth Democracy,* Cambridge, Cambridge University Press.

Hugues, E.C. (1962), "Good People and Dirty Work", *Social Problems* 10(1), pp. 3–11.

Johnson, J.A. (2013), Annual conference of the Midwest Political Science Association, *From Open Data to Information Justice.* Chicago, Illinois, pp. 0–20.

Jounin, N. (2014), *Voyage de classes,* Paris, La Découverte.

Kafka, B. (2012), *The Demon of Writing. Powers and Failures of Paperwork,* New York, Zone Books.

Kelty, C. (2008), *Two Bits,* Duke University Press, Durham.

Kirsch, D. (1995), "The intelligent use of space", *Artificial intelligence* (73), pp. 31–68.

Knorr-Cetina, K. (1981), *The manufacture of knowledge. An essay on the constructivist and contextual nature of science ,* Oxford, Pergamon.

Lafontaine, C. (2004), *L'empire cybernétique : des machines à penser à la pensée machine,* Paris, Seuil.

Lampland, M. (2010), "False numbers as formalizing practices", *Social Studies of Science* 40(3), pp. 377–404.

Lascoumes, P. & Le Gales, P.L. (dir.), (2005), *Gouverner par les instruments,* Paris, Presses de Sciences Po.

Lascoumes, P. & Le Galès, P. (2007), *Sociologie de l'action publique,* Armand Colin.

Latour, B. (1993), « Le pédofil de Boavista, montage photo-philosophique », *Petites leçons de sociologie des sciences* , pp. 171-225.

Latour, B. (2006), « "Les 'vues' de l'esprit" Une introduction à l'anthropologie des sciences et des techniques », in *Sociologie de la traduction. Textes fondateurs,* sous la direction de M. Akrich, B. Latour, & M. Callon, Paris, Presse des mines, .

Latour, B. (2015), « Sur un livre d'Etienne Souriau : Les Différents modes d'existence », in *Etienne Souriau. Une ontologie de l'instauration,* sous la direction de F. Courtois-L'Heureux & A. Wiame, Paris, Vrin, pp. 17-53.

Latour, B. & Woolgar, S. (2000), *La vie de laboratoire. La production des faits scientifiques,* La Découverte.

Law, J. (2009), "Seeing Like a Survey", *Cultural Sociology* 3(2), pp. 239–256.

Lemieux, C. (2012), « Peut-on ne pas être constructiviste ? », *Politix* 100(4), p. 169.

Leonelli, S. (2015), “What Counts as Scientific Data? A Relational Framework ”, *Philosophy of Science* .

Lessig (2009), “Against Transparency. The perils of openness in government ”, *The New Republic,*.

Loveluck, B. (2012). *La liberté par l’information. Généalogie politique du libéralisme informationnel et des formes de l’auto-organisation sur internet.* Thèse pour l’obtention du doctorat de l’EHESS, sous la direction de M. Marcel Gauchet.

Lynch, M. (1985), *Art and Artifact in Laboratory Science. A Study of Shop Work and Shop Talk in a Research Laboratory,* London, Routledge.

Pontille, D, & Torny, D. (à paraître), « L’Open Access : du militantisme aux marchés régulés. », article de travail disponible auprès des auteurs.

Marcus, G. (1995), “Ethnography in/of the world system: the emergence of multi-sited ethnography”, *Annual review of anthropology* (24), pp. 95–117.

Marres, N. (2005), “No issue, no public: Democratic deficits after the displacement of politics”, PhD Dissertation in philosophy, University of Amsterdam, Amsterdam.

Mazeaud, A. (2012), « Administrer la participation : l’invention d’un métier entre valorisation du militantisme et professionnalisation de la démocratie locale », *Quaderni* 3(79), pp. 45-58.

Mazeaud, A., Sa Vilas Boas, M. & Berthomé, G. (2012), « Penser les effets de la participation sur l’action publique à partir de ses impensés », *Participations* (1).

Mazzarella, W. (2006), “Internet X-Ray: E-Governance, Transparency, and the Politics of Immediation in India”, *Public Culture* 18(3), pp. 473–505.

Millerand, F. (2012), « La science en réseau », *Revue d’anthropologie des connaissances* 6, 1(1), p. 163.

Millerand, F. et al (2009), “Metadata standards: Trajectories and enactment in the life of an ontology”, in *Standards and their stories,* Itahaca, pp. 149–165.

O’Reilly, T. (2010), “Government as a Platform”, *innovations* 6(1), pp. 13–40.

Palme, J. et al (2009), “ASCII Imperialism”, in *Standards and their stories,* .

Parasie, S. (2013), « Des machines à scandale. Éléments pour une sociologie morale des bases de données », *Réseaux* 2, pp. 178-179.

Pawluch, D. & Woolgar, S. (1985), “Ontological Gerrymandering: The Anatomy of Social Problems Explanations”, *Social Problems* 32(3).

Peixoto, T. (2013), “The Uncertain Relationship between Open Data and Accountability: A Response to Yu and Robinson’s ‘The New Ambiguity of Open Government’”, *UCLA Law Review* 60(200), pp. 200–213.

Peugeot, V. (2010), « Les enjeux publics, économiques et citoyens de l’ouverture des données : l’expérience britannique », *Actes de la conférence DocSoc 2010* , pp. 1-17

Pène, S. (1995), « Les écrits et les acteurs. Circulation des discours et empreinte des objets », *Etudes de communication* (16).

Raman, B. (2012), “The Rhetoric and Reality of Transparency : Transparent Information , Opaque City Spaces and the Empowerment Question”, *Journal of community informatics* 8(2).

Ribes, D. & Jackson, S.J. (2013), “Data Bite Man: The Work of Sustaining a Long-Term Study”, in *Raw Data is an Oxymoron,* sous la direction de L. Gitelman, Cambridge, The MIT Press, pp. 147–166.

Ronai, M. (1994), « L’Etat comme machine informationnelle », *Revue Française d’Administration Publique* (72), pp. 571-580.

Ronai, M. (1996), « De l’activisme informationnel à la régulation », in *Le Communicateur,* pp. 24-32.

Ronai, M. (1997), « Données publiques : accès, diffusion, commercialisation », *Problèmes politiques et sociaux* (773-774), p. 68.

Rosenberg, D. (2013), “Data before the fact”, in *Raw Data is an Oxymoron,* sous la direction de L. Gitelman, The MIT Press, .

Rosental, C. (2002), « De la démo-cratie en Amérique. Formes actuelles de la démonstration en Intelligence Artificielle », *Actes de la Recherche en Sciences Sociales,* (141-142), pp. 110-120.

Rosental, C. (2009), « Anthropologie de la démonstration », *Revue d’anthropologie des connaissances* 3, 2(2), p. 233.

Ruppert, E. (2012), “Doing the Transparent State : open government data as performance indicators”, in *A World of Indicators: The production of knowledge and justice in an interconnected world,* sous la direction de S.J. Park & J. Mugler, Cambridge, Cambridge University Press, pp. 51–78

Russell, A. (2014), *Open Standards and the Digital Age. History, Ideology, and Networks,* Cambridge, Cambridge University Press.

Schafer, V. & Thierry, B. (2015), « L’ogre et la toile. Le rendez-vous de l’histoire et des archives du web », *Socio* (4), pp. 75-95.

Sfez, L. (1976), *Critique de la décision,* Presses de Sciences Po, Paris.

Shirky, C. (2008), *Here Comes Everybody Power of Organizing Without Organizations,* Pinguin Books.

Souriau, E. (2009), *Les différents modes d’existence. Suivi de « l’Œuvre à faire »,* Paris, PUF.

Star, S.L. (1999), “The Ethnography of Infrastructure”, *American Behavioral Scientist* 43(3), pp. 377–391.

Star, S.L. & Ruhleder, K. (1996), “Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces”, *Information Systems Research* 7(1).

Sterne, J. (2006), “The mp3 as cultural artifact”, *New Media & Society* 8(5), pp. 825–842.

Sterne, J. (2012), *MP3: The Meaning of a Format,* Durham, NC, Duke University Press.

Strasser, B. (2012), Séminaire « Penser l’écosystème des données », *Data-driven science : entre ruptures et continuités.* Paris, Institut des Sciences de la Communication du CNRS, .

Strasser, B.J. (2011), “The Experimenter’s Museum: GenBank. Natural History and the Moral Economies of Biomedicine ”, *Isis* 102(1), pp. 60–96.

Sujobert, B. (2014), « Comment intervenir sur le programme de la statistique publique ? L’exemple des inégalités sociales », in *Statactivisme,* sous la direction de I. Bruno, E. Didier, & J. Prévieux, Paris, Zones, pp. 213-233.

Surowiecki, J. (2004), *The Wisdom of The Crowds,* Doubleday.

Tapscott, D. & D. Williams (2008), *Wikinomics. How Mass Collaboration Changes Everything,* Pinguin Books.

Thévenot, L. (1986), « Les investissements de forme », in *Conventions Economiques,* sous la direction de L. Thévenot, Paris, Presses Universitaires de France, pp. 21-71.

Tkacz, N. (2012), “From open source to open government: a critique of open politics. ”, *Ephemera: Theory and Politics in Organization* 12(4), pp. 386–405.

Triclot, M. (2008), *Le Moment cybernétique. La constitution de la notion d’information,* Seyssel, Editions Champ Vallon.

Trojette (2013), *Ouverture des données publiques. Les exceptions au principe de gratuité sont-elles toutes légitimes ?* Rapport au Premier Ministre.

Turner, F. (2008), *From Counterculture to Cyberculture—Stewart Brand, The Whole Earth Network, and the Rise of Digital Utopianism,* Chicago, .

Vickery, G. (2011), *Review of recent studies on PSI re-use and related market,* Paris, .

Vismann, C. (2008), *Files. Law and Media Technology,* Stanford, Stanford University Press.

Walford, A. (2013), “Transforming Data: An Ethnography of Scientific Data from the Brazilian Amazon. ”, Thesis submitted to the IT University of Copenhagen in compliance with the requirements for the degree of Doctor of Philosophy (PhD), Copenhagen.

Weller, J.-M. (2012), « Comment ranger son bureau ? Le fonctionnaire, l’agriculteur, le droit et l’argent », *Réseaux* (171), pp. 68-101.

Woolgar, S. (1991), “Configuring the User”, in *A Sociology of Monsters: Essays on Power, Technology and Domination,* London, pp. 57–103.

Yu, H. & Robinson, D.G. (2012), “The New Ambiguity of ‘Open Government’”, *UCLA Law Review* 178, pp. 178–208.

Zask, J. (2008), « Le public chez Dewey : une union sociale plurielle », *Tracés* (15), pp. 169-189.

Zimmerman, A.S. (2008), “New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data”, *Science, Technology & Human Values* 33(5), pp. 631–652.

Annexes

L'intégralité des documents publics disponibles sur le web qui ont servi à la rédaction de cette thèse sont archivées sur mon compte Zotero à l'adresse suivante : https://www.zotero.org/samgoeta/items.

Ce sont les « données brutes » de mon travail. N'ayant jusqu'alors comme public que moi-même, ces documents ne font pas l'objet de vérification et les métadonnées ont généralement été produites de manière automatique. Je continue d'utiliser cet outil après la soutenance pour archiver certains documents et pages web importants dans la suite de mes recherches sur l'*open data*.

1. Open Definition (version 1, 2005) [131]

Open Knowledge Definition 1.0
Terminology
The term knowledge is taken to include:

Content such as music, films, books
Data be it scientific, historical, geographic or otherwise
Government and other administrative information
Software is excluded despite its obvious centrality because it is already adequately addressed by previous work.

The term work will be used to denote the item of knowledge at issue.

The term package may also be used to denote a collection of works. Of course such a package may be considered a work in itself.

The term license refers to the legal license under which the work is made available. Where no license has been made this should be interpreted as referring to the resulting default legal conditions under which the work is available.

The Definition
A work is open if its manner of distribution satisfies the following conditions:

1. Access
The work shall be available as a whole and at no more than a reasonable reproduction cost, preferably downloading via the Internet without charge. The work must also be available in a convenient and modifiable form.

2. Redistribution
The license shall not restrict any party from selling or giving away the work either on its own or as part of a package made from works from many different sources. The license shall not require a royalty or other fee for such sale or distribution.

3. Reuse
The license must allow for modifications and derivative works and must allow them to be distributed under the terms of the original work. The license may impose some form of attribution and integrity requirements: see principle 5 (Attribution) and principle 6 (Integrity) below.

4. Absence of Technological Restriction
The work must be provided in such a form that there are no technological obstacles to the performance of the above activities. This can be achieved by the provision of the work in an open data format, i.e. one whose specification is publicly and freely available and which places no restrictions monetary or otherwise upon its use.

5. Attribution
The license may require as a condition for redistribution and re-use the attribution of the contributors and creators to the work. If this condition is imposed it must not be onerous. For example if attribution is required a list of those requiring attribution should accompany the work.

6. Integrity
The license may require as a condition for the work being distributed in modified form that the resulting work carry a different name or version number from the original work.

7. No Discrimination Against Persons or Groups
The license must not discriminate against any person or group of persons.

[131] Open Definition, « Open Knowledge Definition 1.0 », http://opendefinition.org/od/1.0/en/, consulté le 13 janvier 2017.

8. No Discrimination Against Fields of Endeavor
The license must not restrict anyone from making use of the work in a specific field of endeavor. For example, it may not restrict the work from being used in a business, or from being used for military research.

9. Distribution of License
The rights attached to the work must apply to all to whom the work is redistributed without the need for execution of an additional license by those parties.

10. License Must Not Be Specific to a Package
The rights attached to the work must not depend on the work being part of a particular package. If the work is extracted from that package and used or distributed within the terms of the work's license, all parties to whom the work is redistributed should have the same rights as those that are granted in conjunction with the original package.

11. License Must Not Restrict the Distribution of Other Works
The license must not place restrictions on other works that are distributed along with the licensed work. For example, the license must not insist that all other works distributed on the same medium are open.

2. Open Definition (version 2.1, 2015)[132]

Open Definition 2.1
Version 2.1

The Open Definition makes precise the meaning of "open" with respect to knowledge, promoting a robust commons in which anyone may participate, and interoperability is maximized.

Summary: Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.

This essential meaning matches that of "open" with respect to software as in the Open Source Definition and is synonymous with "free" or "libre" as in the Free Software Definition and Definition of Free Cultural Works.

The term work will be used to denote the item or piece of knowledge being transferred.

The term license refers to the legal conditions under which the work is provided.

The term public domain denotes the absence of copyright and similar restrictions, whether by default or waiver of all such conditions.

The key words "must", "must not", "should", and "may" in this document are to be interpreted as described in RFC2119.

1. Open Works
An open work must satisfy the following requirements in its distribution:

1.1 Open License or Status
The work must be in the public domain or provided under an open license (as defined in Section 2). Any additional terms accompanying the work (such as a terms of use, or patents held by the licensor) must not contradict the work's public domain status or terms of the license.

1.2 Access
The work must be provided as a whole and at no more than a reasonable one-time reproduction cost, and should be downloadable via the Internet without charge. Any additional information necessary for license compliance (such as names of contributors required for compliance with attribution requirements) must also accompany the work.

1.3 Machine Readability
The work must be provided in a form readily processable by a computer and where the individual elements of the work can be easily accessed and modified.

1.4 Open Format

[132] Open Definition, « Open Definition 2.1» http://opendefinition.org/od/2.1/en/, consulté le 13 janvier 2017.

The work must be provided in an open format. An open format is one which places no restrictions, monetary or otherwise, upon its use and can be fully processed with at least one free/libre/open-source software tool.

2. Open Licenses
A license should be compatible with other open licenses.

A license is open if its terms satisfy the following conditions:

2.1 Required Permissions
The license must irrevocably permit (or allow) the following:

2.1.1 Use

The license must allow free use of the licensed work.

2.1.2 Redistribution

The license must allow redistribution of the licensed work, including sale, whether on its own or as part of a collection made from works from different sources.

2.1.3 Modification

The license must allow the creation of derivatives of the licensed work and allow the distribution of such derivatives under the same terms of the original licensed work.

2.1.4 Separation

The license must allow any part of the work to be freely used, distributed, or modified separately from any other part of the work or from any collection of works in which it was originally distributed. All parties who receive any distribution of any part of a work within the terms of the original license should have the same rights as those that are granted in conjunction with the original work.

2.1.5 Compilation

The license must allow the licensed work to be distributed along with other distinct works without placing restrictions on these other works.

2.1.6 Non-discrimination

The license must not discriminate against any person or group.

2.1.7 Propagation

The rights attached to the work must apply to all to whom it is redistributed without the need to agree to any additional legal terms.

2.1.8 Application to Any Purpose

The license must allow use, redistribution, modification, and compilation for any purpose. The license must not restrict anyone from making use of the work in a specific field of endeavor.

2.1.9 No Charge

The license must not impose any fee arrangement, royalty, or other compensation or monetary remuneration as part of its conditions.

2.2 Acceptable Conditions
The license must not limit, make uncertain, or otherwise diminish the permissions required in Section 2.1 except by the following allowable conditions:

2.2.1 Attribution

The license may require distributions of the work to include attribution of contributors, rights holders, sponsors, and creators as long as any such prescriptions are not onerous.

2.2.2 Integrity

The license may require that modified versions of a licensed work carry a different name or version number from the original work or otherwise indicate what changes have been made.

2.2.3 Share-alike

The license may require distributions of the work to remain under the same license or a similar license.

2.2.4 Notice

The license may require retention of copyright notices and identification of the license.

2.2.5 Source

The license may require that anyone distributing the work provide recipients with access to the preferred form for making modifications.

2.2.6 Technical Restriction Prohibition

The license may require that distributions of the work remain free of any technical measures that would restrict the exercise of otherwise allowed rights.

2.2.7 Non-aggression

The license may require modifiers to grant the public additional permissions (for example, patent licenses) as required for exercise of the rights allowed by the license. The license may also condition permissions on not aggressing against licensees with respect to exercising any allowed right (again, for example, patent litigation).

The Open Definition was initially derived from the Open Source Definition, which in turn was derived from the original Debian Free Software Guidelines, and the Debian Social Contract of which they are a part, which were created by Bruce Perens and the Debian Developers. Bruce later used the same text in creating the Open Source Definition. This definition is substantially derivative of those documents and retains their essential principles. Richard Stallman was the first to push the ideals of software freedom which we continue.

3. Open Government Data Principles (2007)[133]

Request for Comments

December 7-8, 2007—This weekend, 30 open government advocates gathered to develop a set of principles of open government data. The meeting, held in Sebastopol, California, was designed to develop a more robust understanding of why open government data is essential to democracy.

The Internet is the public space of the modern world, and through it governments now have the opportunity to better understand the needs of their citizens and citizens may participate more fully in their government. Information becomes more valuable as it is shared, less valuable as it is hoarded. Open data promotes increased civil discourse, improved public welfare, and a more efficient use of public resources.

The group is offering a set of fundamental principles for open government data. By embracing the eight principles, governments of the world can become more effective, transparent, and relevant to our lives.

Your comments are welcome!

Open Government Data Principles

Government data shall be considered open if it is made public in a way that complies with the principles below:

1. Complete
All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. Primary

[133] Public Ressource, « Open Government Data Principles », https://public.resource.org/8_principles.html, consulté le 13 janvier 2017.

Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. Timely
Data is made available as quickly as necessary to preserve the value of the data.
4. Accessible
Data is available to the widest range of users for the widest range of purposes.
5. Machine processable
Data is reasonably structured to allow automated processing.
6. Non-discriminatory
Data is available to anyone, with no requirement of registration.
7. Non-proprietary
Data is available in a format over which no entity has exclusive control.
8. License-free
Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.
Compliance must be reviewable.

Definitions

1. "public" means:
The Open Government Data principles do not address what data should be public and open. Privacy, security, and other concerns may legally (and rightly) prevent data sets from being shared with the public. Rather, these principles specify the conditions public data should meet to be considered "open."

2. "data" means:
Electronically stored information or recordings. Examples include documents, databases of contracts, transcripts of hearings, and audio/visual recordings of events.

While non-electronic information resources, such as physical artifacts, are not subject to the Open Government Data principles, it is always encouraged that such resources be made available electronically to the extent feasible.

3. "reviewable" means:
A contact person must be designated to respond to people trying to use the data.

A contact person must be designated to respond to complaints about violations of the principles.

An administrative or judicial court must have the jurisdiction to review whether the agency has applied these principles appropriately.

4. « Raw Data Now », conférence TED de Tim Berners-Lee (transcript officiel, 2008)[134]

Time flies. It's actually almost 20 years ago when I wanted to reframe the way we use information, the way we work together: I invented the World Wide Web. Now, 20 years on, at TED, I want to ask your help in a new reframing.
0:30
So going back to 1989, I wrote a memo suggesting the global hypertext system. Nobody really did anything with it, pretty much. But 18 months later -- this is how innovation happens -- 18 months later, my boss said I could do it on the side, as a sort of a play project, kick the tires of a new computer we'd got. And so he gave me the time to code it up. So I basically roughed out what HTML should look like: hypertext protocol, HTTP; the idea of URLs, these names for things which started with HTTP. I wrote the code and put it out there.
1:10
Why did I do it? Well, it was basically frustration. I was frustrated -- I was working as a software engineer in this huge, very exciting lab, lots of people coming from all over the world. They brought all sorts of different computers with them. They had all sorts of different data formats, all sorts, all kinds of documentation systems. So that, in all that diversity, if I wanted to figure out how to build something out of a bit of this and a bit of this, everything I looked into, I had to connect to some new machine, I had to learn to run some new program, I

[134] TED, « The next web », https://www.ted.com/talks/tim_berners_lee_on_the_next_web/transcript?language=en, consulté le 13 janvier 2017.

would find the information I wanted in some new data format. And these were all incompatible. It was just very frustrating. The frustration was all this unlocked potential.
1:54
In fact, on all these discs there were documents. So if you just imagined them all being part of some big, virtual documentation system in the sky, say on the Internet, then life would be so much easier. Well, once you've had an idea like that it kind of gets under your skin and even if people don't read your memo -- actually he did, it was found after he died, his copy. He had written, "Vague, but exciting," in pencil, in the corner.
2:21
(Laughter)
2:23
But in general it was difficult -- it was really difficult to explain what the web was like. It's difficult to explain to people now that it was difficult then. But then -- OK, when TED started, there was no web so things like "click" didn't have the same meaning. I can show somebody a piece of hypertext, a page which has got links, and we click on the link and bing -- there'll be another hypertext page. Not impressive. You know, we've seen that -- we've got things on hypertext on CD-ROMs. What was difficult was to get them to imagine: so, imagine that that link could have gone to virtually any document you could imagine. Alright, that is the leap that was very difficult for people to make. Well, some people did. So yeah, it was difficult to explain, but there was a grassroots movement. And that is what has made it most fun. That has been the most exciting thing, not the technology, not the things people have done with it, but actually the community, the spirit of all these people getting together, sending the emails. That's what it was like then.
3:24
Do you know what? It's funny, but right now it's kind of like that again. I asked everybody, more or less, to put their documents -- I said, "Could you put your documents on this web thing?" And you did. Thanks. It's been a blast, hasn't it? I mean, it has been quite interesting because we've found out that the things that happen with the web really sort of blow us away. They're much more than we'd originally imagined when we put together the little, initial website that we started off with. Now, I want you to put your data on the web. Turns out that there is still huge unlocked potential. There is still a huge frustration that people have because we haven't got data on the web as data.
4:03
What do you mean, "data"? What's the difference -- documents, data? Well, documents you read, OK? More or less, you read them, you can follow links from them, and that's it. Data -- you can do all kinds of stuff with a computer. Who was here or has otherwise seen Hans Rosling's talk? One of the great -- yes a lot of people have seen it -- one of the great TED Talks. Hans put up this presentation in which he showed, for various different countries, in various different colors -- he showed income levels on one axis and he showed infant mortality, and he shot this thing animated through time. So, he'd taken this data and made a presentation which just shattered a lot of myths that people had about the economics in the developing world.
4:49
He put up a slide a little bit like this. It had underground all the data OK, data is brown and boxy and boring, and that's how we think of it, isn't it? Because data you can't naturally use by itself But in fact, data drives a huge amount of what happens in our lives and it happens because somebody takes that data and does something with it. In this case, Hans had put the data together he had found from all kinds of United Nations websites and things. He had put it together, combined it into something more interesting than the original pieces and then he'd put it into this software, which I think his son developed, originally, and produces this wonderful presentation. And Hans made a point of saying, "Look, it's really important to have a lot of data." And I was happy to see that at the party last night that he was still saying, very forcibly, "It's really important to have a lot of data."
5:43
So I want us now to think about not just two pieces of data being connected, or six like he did, but I want to think about a world where everybody has put data on the web and so virtually everything you can imagine is on the web and then calling that linked data. The technology is linked data, and it's extremely simple. If you want to put something on the web there are three rules: first thing is that those HTTP names -- those things that start with "http:" -- we're using them not just for documents now, we're using them for things that the documents are about. We're using them for people, we're using them for places, we're using them for your products, we're using them for events. All kinds of conceptual things, they have names now that start with HTTP.
6:25
Second rule, if I take one of these HTTP names and I look it up and I do the web thing with it and I fetch the data using the HTTP protocol from the web, I will get back some data in a standard format which is kind of useful data that somebody might like to know about that thing, about that event. Who's at the event? Whatever it is about that person, where they were born, things like that. So the second rule is I get important information back.
6:50
Third rule is that when I get back that information it's not just got somebody's height and weight and when they were born, it's got relationships. Data is relationships. Interestingly, data is relationships. This person was born in Berlin; Berlin is in Germany. And when it has relationships, whenever it expresses a relationship then the other thing that it's related to is given one of those names that starts HTTP. So, I can go ahead and look that thing up.

So I look up a person -- I can look up then the city where they were born; then I can look up the region it's in, and the town it's in, and the population of it, and so on. So I can browse this stuff.
7:30
So that's it, really. That is linked data. I wrote an article entitled "Linked Data" a couple of years ago and soon after that, things started to happen. The idea of linked data is that we get lots and lots and lots of these boxes that Hans had, and we get lots and lots and lots of things sprouting. It's not just a whole lot of other plants. It's not just a root supplying a plant, but for each of those plants, whatever it is -- a presentation, an analysis, somebody's looking for patterns in the data -- they get to look at all the data and they get it connected together, and the really important thing about data is the more things you have to connect together, the more powerful it is.
8:09
So, linked data. The meme went out there. And, pretty soon Chris Bizer at the Freie Universitat in Berlin who was one of the first people to put interesting things up, he noticed that Wikipedia -- you know Wikipedia, the online encyclopedia with lots and lots of interesting documents in it. Well, in those documents, there are little squares, little boxes. And in most information boxes, there's data. So he wrote a program to take the data, extract it from Wikipedia, and put it into a blob of linked data on the web, which he called dbpedia. Dbpedia is represented by the blue blob in the middle of this slide and if you actually go and look up Berlin, you'll find that there are other blobs of data which also have stuff about Berlin, and they're linked together. So if you pull the data from dbpedia about Berlin, you'll end up pulling up these other things as well. And the exciting thing is it's starting to grow. This is just the grassroots stuff again, OK?
9:03
Let's think about data for a bit. Data comes in fact in lots and lots of different forms. Think of the diversity of the web. It's a really important thing that the web allows you to put all kinds of data up there. So it is with data. I could talk about all kinds of data. We could talk about government data, enterprise data is really important, there's scientific data, there's personal data, there's weather data, there's data about events, there's data about talks, and there's news and there's all kinds of stuff. I'm just going to mention a few of them so that you get the idea of the diversity of it, so that you also see how much unlocked potential.
9:40
Let's start with government data. Barack Obama said in a speech, that he -- American government data would be available on the Internet in accessible formats. And I hope that they will put it up as linked data. That's important. Why is it important? Not just for transparency, yeah transparency in government is important, but that data -- this is the data from all the government departments Think about how much of that data is about how life is lived in America. It's actual useful. It's got value. I can use it in my company. I could use it as a kid to do my homework. So we're talking about making the place, making the world run better by making this data available.
10:17
In fact if you're responsible -- if you know about some data in a government department, often you find that these people, they're very tempted to keep it -- Hans calls it database hugging. You hug your database, you don't want to let it go until you've made a beautiful website for it. Well, I'd like to suggest that rather -- yes, make a beautiful website, who am I to say don't make a beautiful website? Make a beautiful website, but first give us the unadulterated data, we want the data. We want unadulterated data. OK, we have to ask for raw data now. And I'm going to ask you to practice that, OK? Can you say "raw"?
10:55
Audience: Raw.
10:56
Tim Berners-Lee: Can you say "data"?
10:57
Audience: Data.
10:58
TBL: Can you say "now"?
10:59
Audience: Now!
11:00
TBL: Alright, "raw data now"!
11:02
Audience: Raw data now!
11:04
Practice that. It's important because you have no idea the number of excuses people come up with to hang onto their data and not give it to you, even though you've paid for it as a taxpayer. And it's not just America. It's all over the world. And it's not just governments, of course -- it's enterprises as well.
11:19
So I'm just going to mention a few other thoughts on data. Here we are at TED, and all the time we are very conscious of the huge challenges that human society has right now -- curing cancer, understanding the brain for Alzheimer's, understanding the economy to make it a little bit more stable, understanding how the world works. The people who are going to solve those -- the scientists -- they have half-formed ideas in their head,

they try to communicate those over the web. But a lot of the state of knowledge of the human race at the moment is on databases, often sitting in their computers, and actually, currently not shared.
11:56
In fact, I'll just go into one area -- if you're looking at Alzheimer's, for example, drug discovery -- there is a whole lot of linked data which is just coming out because scientists in that field realize this is a great way of getting out of those silos, because they had their genomics data in one database in one building, and they had their protein data in another. Now, they are sticking it onto -- linked data -- and now they can ask the sort of question, that you probably wouldn't ask, I wouldn't ask -- they would. What proteins are involved in signal transduction and also related to pyramidal neurons? Well, you take that mouthful and you put it into Google. Of course, there's no page on the web which has answered that question because nobody has asked that question before. You get 223,000 hits -- no results you can use. You ask the linked data -- which they've now put together -- 32 hits, each of which is a protein which has those properties and you can look at. The power of being able to ask those questions, as a scientist -- questions which actually bridge across different disciplines -- is really a complete sea change. It's very very important. Scientists are totally stymied at the moment -- the power of the data that other scientists have collected is locked up and we need to get it unlocked so we can tackle those huge problems.
13:09
Now if I go on like this, you'll think that all the data comes from huge institutions and has nothing to do with you. But, that's not true. In fact, data is about our lives. You just -- you log on to your social networking site, your favorite one, you say, "This is my friend." Bing! Relationship. Data. You say, "This photograph, it's about -- it depicts this person. " Bing! That's data. Data, data, data. Every time you do things on the social networking site, the social networking site is taking data and using it -- re-purposing it -- and using it to make other people's lives more interesting on the site. But, when you go to another linked data site -- and let's say this is one about travel, and you say, "I want to send this photo to all the people in that group," you can't get over the walls. The Economist wrote an article about it, and lots of people have blogged about it -- tremendous frustration. The way to break down the silos is to get inter-operability between social networking sites. We need to do that with linked data.
14:03
One last type of data I'll talk about, maybe it's the most exciting. Before I came down here, I looked it up on OpenStreetMap The OpenStreetMap's a map, but it's also a Wiki. Zoom in and that square thing is a theater -- which we're in right now -- The Terrace Theater. It didn't have a name on it. So I could go into edit mode, I could select the theater, I could add down at the bottom the name, and I could save it back. And now if you go back to the OpenStreetMap. org, and you find this place, you will find that The Terrace Theater has got a name. I did that. Me! I did that to the map. I just did that! I put that up on there. Hey, you know what? If I -- that street map is all about everybody doing their bit and it creates an incredible resource because everybody else does theirs. And that is what linked data is all about. It's about people doing their bit to produce a little bit, and it all connecting. That's how linked data works. You do your bit. Everybody else does theirs. You may not have lots of data which you have yourself to put on there but you know to demand it. And we've practiced that.
15:09
So, linked data -- it's huge. I've only told you a very small number of things There are data in every aspect of our lives, every aspect of work and pleasure, and it's not just about the number of places where data comes, it's about connecting it together. And when you connect data together, you get power in a way that doesn't happen just with the web, with documents. You get this really huge power out of it. So, we're at the stage now where we have to do this -- the people who think it's a great idea. And all the people -- and I think there's a lot of people at TED who do things because -- even though there's not an immediate return on the investment because it will only really pay off when everybody else has done it -- they'll do it because they're the sort of person who just does things which would be good if everybody else did them. OK, so it's called linked data. I want you to make it. I want you to demand it. And I think it's an idea worth spreading.
16:07
Thanks.
16:08
(Applause)

5. « 5-star model », extrait de la page « *Design Issues* » du site de Tim Berners-Lee (2010)[135]

Is your Linked Open Data 5 Star?
(Added 2010). This year, in order to encourage people -- especially government data owners -- along the road to good linked data, I have developped this star rating system.

[135] W3C, « Linked Data - Design Issues », http://www.w3.org/DesignIssues/LinkedData.html, consulté le 28 juillet 2015.

Linked Data is defined above. Linked Open Data (LOD) is Linked Data which is released under an open licence, which does not impede its reuse for free. Creative Commons CC-BY is an example open licence, as is the UK's Open Government Licence. Linked Data does not of course in general have to be open -- there is a lot of important use of lnked data internally, and for personal and group-wide data. You can have 5-star Linked Data without it being open. However, if it claims to be Linked Open Data then it does have to be open, to get any star at all.

Under the star scheme, you get one (big!) star if the information has been made public at all, even if it is a photo of a scan of a fax of a table -- if it has an open licence. The you get more stars as you make it progressively more powerful, easier for people to use.
★ Available on the web (whatever format) but with an open licence, to be Open Data
★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★ All the above, plus: Link your data to other people's data to provide context
How well does your data do? You can buy 5 star data mugs, T-shirts and bumper stickers from the W3C shop at cafepress: use them to get your colleages and fellows conference-goers thinking 5 star linked data. (Profits also help W3C :-).

Now in 2010, people have been pressing me, for governmet data, to add a new requirement, and that is there should be metadata about the data itself, and that that metadata should be availble from a major catalog. Any open dataset (or even datasets which are not but should be open) can be regisetreed at ckan.net. Government datasets from the UK and US hsould be regisetred at data.gov.uk or data.gov respectively. Other copuntries I expect to develop their own registries. Yes, there should be metadata about your dataset. That may be the subject of a new note in this series.

6. Article « *Launching the Open Data Census 2012 !* » publié sur le blog de l'Open Knowledge Foundation (2012)[136]

As government officials, civil society leaders and open data experts gather in Brazil this week for the Open Government Partnership, it is clear that Open Government Data has become a major topic on a global scale. In September last year, 8 governments founded the Open Government Partnership. Little more than six months on, and a further 43 have signed-up and endorsed the Open Government Declaration already. The movement is big and it's growing.

At the close of last year's Open Government Data Camp, the Open Knowledge Foundation announced plans to launch an Open Data Census in 2012. Since then, preparations have been underway. And this week, to coincide with the Open Government Partnership meeting, the Open Data Census is going live!

What is the Open Data Census?
The Open Data Census 2012 is an attempt to monitor the current status of open data across the globe.

The primary focus of the Census is data. Policies are crucial, but as Chris Taggart's analysis of corporate data demonstrates, actual practice can be very different. Focussing on data will also allow us to keep the census very concrete. Analysing policy or even law is a complex process; whether a dataset is 'open' or not is usually a clear yes or no answer.

In this Census, we are interested in the current status of data: is it open, is it accessible, can I use it now?

We hope to gather responses from every country in the world. To find out how to contribute on behalf of your country, read on below!

What will the Open Data Census look at?
In the first incarnation of the 2012 Census, we have decided to look only at ten specific datasets. We hope to expand this in future, and we welcome suggestions for new datasets to include (see below).

For each dataset, we will explore whether it is:

a) available in a digital form

[136] OKFN Blog, « Launching the Open Data Census 2012! », http://blog.okfn.org/2012/04/17/launching-the-open-data-census-2012/, consulté le 30 juillet 2015.

b) machine-readable
c) publicly available, free of charge
d) openly licensed
A yes to all of these questions imply that the dataset is open data.

We are primarily seeking binary yes / no responses – but we have allowed a space for comments in case the situation is not clear cut.

The datasets have been chosen for their breadth and relevance. We have attempted to select data which most governments could reasonably be expected to collect. The ten datasets are:

Election Results (national)
Company Register
National Map (Low resolution: 1:250,000 or better)
Government Budget (high level – spending by sector)
Government Budget (detailed – transactional level data)
Legislation (laws and statutes)
National Statistical Office Data (economic and demographic information)
National Postcode/ZIP database
Public Transport Timetables
Environmental Data on major sources of pollutants (e.g. location, emissions)
It may be that some people have already begun collecting information in some of these areas. We're keen not to duplicate efforts, so please do get in touch if you have information which is relevant.

So what next?
Our biggest challenge is to start gathering responses!

Take part in the Census
To take part in the Open Data Census 2012, please visit: http://opengovernmentdata.org/census/submit/
You should submit one census form per dataset per country.
You can see which countries and datasets have already been submitted at .
If you notice an error in a submitted form or are able to add more information, please submit a new census form for that country and dataset. Please highlight the correction in your comments.
Give us feedback for the future
We welcome all feedback on the Census. We also welcome suggestions for new datasets to include in future Censuses. Please email info[@]okfn.org with your comments
If you would like to become more involved with the Open Data Census 2012, please sign-up to the Open Government mailing list
The Open Government working group welcomes everyone with an interest in Open Government. See our website to find out more.
Watch this Space!
We hope to make the results of the Open Data Census 2012 available later this year. Keep an eye on the blog for more details!

7. G8 Open Data Charter and Technical Annex (2013)[137]

Policy paper
G8 Open Data Charter and Technical Annex

Published 18 June 2013

Contents

[137] Gov.uk, « G8 Open Data Charter and Technical Annex », https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex, consulté le 13 janvier 2017.

1) The world is witnessing the growth of a global movement facilitated by technology and social media and fuelled by information – one that contains enormous potential to create more accountable, efficient, responsive, and effective governments and businesses, and to spur economic growth.

Open data sit at the heart of this global movement.

2) Access to data allows individuals and organisations to develop new insights and innovations that can improve the lives of others and help to improve the flow of information within and between countries. While governments and businesses collect a wide range of data, they do not always share these data in ways that are easily discoverable, useable, or understandable by the public.

This is a missed opportunity.

3) Today, people expect to be able to access information and services electronically when and how they want. Increasingly, this is true of government data as well. We have arrived at a tipping point, heralding a new era in which people can use open data to generate insights, ideas, and services to create a better world for all.

4) Open data can increase transparency about what government and business are doing. Open data also increase awareness about how countries' natural resources are used, how extractives revenues are spent, and how land is transacted and managed. All of which promotes accountability and good governance, enhances public debate, and helps to combat corruption. Transparent data on G8 development assistance are also essential for accountability.

5). Providing access to government data can empower individuals, the media, civil society, and business to fuel better outcomes in public services such as health, education, public safety, environmental protection, and governance. Open data can do this by:

showing how and where public money is spent, providing strong incentives for that money to be used most effectively;
enabling people to make better informed choices about the services they receive and the standards they should expect.
6) Freely-available government data can be used in innovative ways to create useful tools and products that help people navigate modern life more easily. Used in this way, open data are a catalyst for innovation in the private sector, supporting the creation of new markets, businesses, and jobs. Beyond government, these benefits can multiply as more businesses adopt open data practices modelled by government and share their own data with the public.

7) We, the G8, agree that open data are an untapped resource with huge potential to encourage the building of stronger, more interconnected societies that better meet the needs of our citizens and allow innovation and prosperity to flourish.

8) We therefore agree to follow a set of principles that will be the foundation for access to, and the release and re-use of, data made available by G8 governments. They are:

Open Data by Default
Quality and Quantity
Useable by All
Releasing Data for Improved Governance
Releasing Data for Innovation
9) While working within our national political and legal frameworks, we will implement these principles in accordance with the technical best practises and timeframes set out in our national action plans. G8 members will, by the end of this year, develop action plans, with a view to implementation of the Charter and technical annex by the end of 2015 at the latest. We will review progress at our next meeting in 2014.

10) We also recognise the benefits of open data can and should be enjoyed by citizens of all nations. In the spirit of openness we offer this Open Data Charter for consideration by other countries, multinational organisations and initiatives.

1. Principle 1: Open Data by Default
11) We recognise that free access to, and subsequent re-use of, open data are of significant value to society and the economy.

12) We agree to orient our governments towards open data by default.

13) We recognise that the term government data is meant in the widest sense possible. This could apply to data owned by national, federal, local, or international government bodies, or by the wider public sector.

14) We recognise that there is national and international legislation, in particular pertaining to intellectual property, personally-identifiable and sensitive information, which must be observed.

15) We will:

establish an expectation that all government data be published openly by default, as outlined in this Charter, while recognising that there are legitimate reasons why some data cannot be released.
2. Principle 2: Quality and Quantity
16) We recognise that governments and the public sector hold vast amounts of information that may be of interest to citizens.

17) We also recognise that it may take time to prepare high-quality data, and the importance of consulting with each other and with national, and wider, open data users to identify which data to prioritise for release or improvement.

18) We will:

release high-quality open data that are timely, comprehensive, and accurate. To the extent possible, data will be in their original, unmodified form and at the finest level of granularity available;
ensure that information in the data is written in plain, clear language, so that it can be understood by all, though this Charter does not require translation into other languages;
make sure that data are fully described, so that consumers have sufficient information to understand their strengths, weaknesses, analytical limitations, and security requirements, as well as how to process the data; and
release data as early as possible, allow users to provide feedback, and then continue to make revisions to ensure the highest standards of open data quality are met.
3. Principle 3: Usable by All
19) We agree to release data in a way that helps all people to obtain and re-use it.

20) We recognise that open data should be available free of charge in order to encourage their most widespread use.

21) We agree that when open data are released, it should be done without bureaucratic or administrative barriers, such as registration requirements, which can deter people from accessing the data.

22) We will:

release data in open formats wherever possible, ensuring that the data are available to the widest range of users for the widest range of purposes; and
release as much data as possible, and where it is not possible to offer free access at present, promote the benefits and encourage the allowance of free access to data. In many cases this will include providing data in multiple formats, so that they can be processed by computers and understood by people.
4. Principle 4: Releasing Data for Improved Governance
23) We recognise that the release of open data strengthens our democratic institutions and encourages better policy-making to meets the needs of our citizens. This is true not only in our own countries but across the world.

24) We also recognise that interest in open data is growing in other multilateral organisations and initiatives.

25) We will:

share technical expertise and experience with each other and with other countries across the world so that everyone can reap the benefits of open data; and
be transparent about our own data collection, standards, and publishing processes, by documenting all of these related processes online.
5. Principle 5: Releasing Data for Innovation
26) Recognising the importance of diversity in stimulating creativity and innovation, we agree that the more people and organisations that use our data, the greater the social and economic benefits that will be generated. This is true for both commercial and non-commercial uses.

27) We will:

work to increase open data literacy and encourage people, such as developers of applications and civil society organisations that work in the field of open data promotion, to unlock the value of open data;
empower a future generation of data innovators by providing data in machine-readable formats.
6. Technical annex
1) We, the G8, have consulted with technical experts to identify some best practices (part one) and collective actions (part two) that we will use to meet the principles set out in the G8 Open Data Charter.

2) While working within our national political and legal frameworks, we agree to implement these practices as quickly as possible and aim to complete our activities by 2015 at the latest. This will be done in accordance with the timeframes in our national action plans.

3) The Annex constitutes a 'living' set of guidelines that may be subject to amendments after consideration of emerging technology solutions or practical experience gained during the course of implementation of the G8 Open Data Charter.

6.1 Part One - Best Practices
Principle 1: Open Data by Default
4) We recognise the importance of open data and we will establish an expectation that all government data be published openly by default.

5) We will:

define our open data position in a public statement of intent, such as an announcement, strategy or policy, so that our plans for progressing the open data agenda in our jurisdictions are clear;
publish a national action plan to provide more specific details on our plans to release data according to the principles in the G8 Open Data Charter; and
publish data on a national portal so that all government data that has been released can be found easily in one place. A portal may be a central website from which data can be downloaded, or a website which lists all open government data stored at a different location. Each portal will include a registry file that lists all the data and metadata used on the portal, as well as providing APIs for developers. Where it is yet not possible to publish all data on a portal, the location of data will be communicated clearly and not moved without notice.
Principle 2: Quality and Quantity
6) We commit to releasing data that are both high in quality as well as high in quantity. When releasing data, we aim to do so in a way that helps people to use and understand them. This will help to increase the interoperability of data from different policy areas, businesses or countries.

7) We will:

use robust and consistent metadata (i.e. the fields or elements that describe the actual data);
publish and maintain an up-to-date mapping of the core descriptive metadata fields across G8 members to enable easier use and comprehension by people from around the world. This will allow countries, in the G8 and beyond, who do not currently have a data portal to consider adopting the metadata fields included in this mapping;
ensure data are fully described, as appropriate, to help users to fully understand the data. This may include:
Documentation that provides explanations about the data fields used;
Data dictionaries to link different data; and
A user's guide that describes the purpose of the collection, the target audience, the characteristics of the sample, and the method of data collection.
listen to feedback from data users to improve the breadth, quality and accessibility of data we offer. This could be in the form of a public consultation on the national data strategy or policy, discussions with civil society, creation of a feedback mechanism on the data portal, or through other appropriate mechanisms.
Principle 3: Usable by All
8) We agree to release data in a way that helps all people find and re-use them.

9) We will:

make data available in convenient open formats to ensure files can be easily retrieved, downloaded, indexed, and searched by all commonly used Web search applications. Open formats, for example non-proprietary CSV files, are ones where the specification for the format is available to anyone for free, thereby allowing the data contained in a file to be opened by different software programmes.
Principle 4: Releasing Data for Improved Governance
10) We recognise that data are a powerful tool to help drive government effectiveness, efficiency and responsiveness to citizen needs while fuelling further demand for open data.

11) We will:

develop links with civil society organisations and individuals to allow the public to provide feedback on the most important data they would like released;
be open about our own data standards, so that we take into account:
Data that are released by other national and international organisations
The standards emerging from other international transparency initiatives; and
document our own experiences of working with open data by, for example, publishing technical information about our open data policies, practices, and portals so that the benefits of open data can be enjoyed in other countries.
Principle 5: Releasing Data for Innovation
12) We agree that our citizens can use our data to fuel innovation in our own countries and around the world. We recognise that free access to, and reuse of, open government data are an essential part of this.

13) We will:

support the release of data using open licences or other relevant instruments - while respecting intellectual property rights - so that no restrictions or charges are placed on the re-use of the information for non-commercial or commercial purposes, save for exceptional circumstances;
ensure data are machine readable in bulk by providing data that are well structured to allow automated processing and access with the minimum number of file downloads;
release data using application programming interfaces (APIs), where appropriate, to ensure easy access to the most regularly updated and accessed data; and
encourage innovative uses of our data through the organisation of challenges, prizes or mentoring for data users in our individual jurisdictions.
6.2 Part Two - Collective Actions
Action 1: G8 National Action Plans
We will publish individual action plans detailing how we will implement the Open Data Charter according to our national frameworks (October 2013)
We will report progress on an annual basis (via the G8 Accountability Working Group) (2014 and 2015)
Action 2: Release of high value data
We recognise the following as areas of high value, both for improving our democracies and encouraging innovative re-use of data.
Data Category* (alphabetical order) Example datasets
Companies Company/business register
Crime and Justice Crime statistics, safety
Earth observation Meteorological/weather, agriculture, forestry, fishing, and hunting
Education List of schools; performance of schools, digital skills
Energy and Environment Pollution levels, energy consumption
Finance and contracts Transaction spend, contracts let, call for tender, future tenders, local budget, national budget (planned and spent)
Geospatial Topography, postcodes, national maps, local maps
Global Development Aid, food security, extractives, land
Government Accountability and Democracy Government contact points, election results, legislation and statutes, salaries (pay scales), hospitality/gifts
Health Prescription data, performance data
Science and Research Genome data, research and educational activity, experiment results
Statistics National Statistics, Census, infrastructure, wealth, skills
Social mobility and welfare Housing, health insurance and unemployment benefits
Transport and Infrastructure Public transport timetables, access points broadband penetration
In accordance with the principles of “open by default” and “quality and quantity” we will work towards the progressive publication of these data.
As a first step, we will collectively make key datasets on National Statistics, National Maps, National Elections and National Budgets available and discoverable (from June 2013), and we will work towards improving their granularity and accessibility (by December 2013)
We recognise that collective action by all G8 members has the potential to unlock barriers and foster innovative solutions to some of the challenges we are facing. We therefore agree on a mutual effort to increase the supply of open government data available on key functions of our States, such as democracy and environment. We will work on identifying datasets in these areas by December 2013, with an aim to release them by December 2014.
We will set out in our national action plans how and when we will release data under the remaining categories according to our national frameworks (October 2013).
Action 3: Metadata mapping
We have contributed to and commit to maintaining the G8 metadata mapping exercise (June 2013)
This mapping can be viewed on Github and comprises a collective mapping ‘index’ across G8 member’s metadata, and a detailed page on each G8 member use of metadata within their national portal.
*categories and datasets to be finalised by December 2015

8. Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'Etat « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques[138]

Le Premier ministre à Monsieur le ministre d'Etat, Mesdames et Messieurs les ministres, Mesdames et Messieurs les secrétaires d'Etat, Mesdames et Messieurs les préfets

Faciliter l'accès en ligne aux informations publiques dans un souci de transparence de l'action de l'Etat et leur réutilisation afin de favoriser l'innovation constitue une priorité dans la politique gouvernementale de modernisation de l'Etat et de développement de l'économie numérique.

Le conseil de modernisation des politiques publiques a décidé le 30 juin 2010 de la création d'un portail unique « data.gouv.fr ». Le conseil des ministres du 24 novembre 2010 a annoncé la mise en ligne de ce portail avant la fin de l'année 2011. J'ai créé par décret du 21 février 2011 la mission « Etalab » qui est chargée de concevoir ce portail unique interministériel « data.gouv.fr » et de coordonner l'action des administrations de l'Etat en matière de réutilisation des informations publiques. Le portail «data.gouv.fr» s'inscrit dans la continuité du travail de modernisation de l'Etat et de simplification des relations que les usagers entretiennent avec leurs services publics.

Le développement de l'économie numérique et de l'innovation technologique constitue un enjeu majeur tant en termes de croissance et d'emplois, que de compétitivité et d'accès à l'information. En accédant librement aux informations publiques dont disposent les administrations, la communauté des développeurs et des entrepreneurs peut dès lors être en mesure de créer de nouveaux usages et des services applicatifs innovants. En matière d'innovation technologique, l'offre crée souvent la demande. En mettant à disposition ses informations publiques, l'Etat participe à la construction de la société numérique. Cette stratégie d'ouverture des données publiques («Open Data») illustre l'ambition de la politique industrielle et d'innovation du Gouvernement.

La réutilisation libre, facile et gratuite des informations publiques est un levier essentiel pour favoriser la dynamique d'innovation qui sera portée par la communauté des développeurs et des entrepreneurs à partir des données mises en ligne sur « data.gouv.fr ». La créativité des développeurs et des entrepreneurs ne saurait se heurter à des cloisons artificielles qui ont trop souvent constitué des freins au développement de l'innovation dans notre pays. Le portail « data.gouv.fr » illustre ainsi l'engagement de l'Etat en faveur du renforcement de la compétitivité des entreprises françaises, qu'il s'agisse d'entrepreneurs individuels ou de petites, moyennes ou grandes sociétés.

« Data.gouv.fr » proposera des services en ligne afin de renforcer la transparence de la vie publique et la confiance des citoyens dans les institutions de la République. Ces services mettront en valeur le travail des administrations, contribueront à la transparence de l'action de l'Etat et éclaireront le débat public. Ils enrichiront ainsi la vie de notre démocratie.

Il convient donc que le portail « data.gouv.fr » mette à disposition librement, facilement et gratuitement le plus grand nombre d'informations publiques. La politique gouvernementale d'ouverture de ces informations doit être lisible et offrir à tous les réutilisateurs la sécurité juridique nécessaire au plein exercice de leur droit.

La décision de subordonner la réutilisation de certaines de ces informations au versement d'une redevance devra être dûment justifiée par des circonstances particulières. Ces informations devront être au préalable inscrites sur une liste établie par décret.

Les annexes à la présente circulaire précisent le cadre dans lequel les administrations de l'Etat mettent à disposition de « data.gouv.fr » leurs informations publiques.

Je vous demande de veiller personnellement au respect des prescriptions de la présente circulaire par vos services, en particulier celui des délais fixés dans les annexes, et d'inviter les personnes publiques dont vous assurez la tutelle à s'y conformer.

ANNEXES

ANNEXE I

DÉFINITION DU PÉRIMÈTRE ET DES OBJECTIFS DE LA MISSION « ETALAB »

Le décret no 2011-194 du 21 février 2011 crée la mission « Etalab », placée sous l'autorité du Premier ministre et rattachée au secrétaire général du Gouvernement.

[138] Legifrance, « Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'Etat « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques »,
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000024072788, consulté le 13 janvier 2017.

« Etalab » a pour mission de créer le portail unique interministériel « data.gouv.fr » destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'Etat, de ses établissements publics administratifs et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public.
Elle coordonne en outre les actions des administrations de l'Etat et apporte son appui aux établissements publics administratifs afin de faciliter les réutilisations de leurs informations publiques.
Le portail « data.gouv.fr » créé par la mission « Etalab » poursuit les trois objectifs suivants :
– permettre la réutilisation des informations publiques la plus facile et la plus large possible ;
– encourager l'innovation par toute la communauté des développeurs et des entrepreneurs pour soutenir le développement de l'économie numérique ;
– contribuer à renforcer la transparence de l'action de l'Etat, mettre en valeur le travail des administrations et éclairer le débat public.

ANNEXE II
CADRE JURIDIQUE DU DROIT À LA RÉUTILISATION
DES INFORMATIONS PUBLIQUES PAR LES PERSONNES PHYSIQUES ET MORALES
En transposant la directive 2003/98/CE du 17 novembre 2003, l'ordonnance no 2005-650 du 6 juin 2005, qui a modifié la loi no 78-753 du 17 juillet 1978, a instauré le droit pour toute personne physique ou morale de réutiliser les informations publiques des administrations. Le régime est prévu au chapitre II du titre Ier de ladite loi, et les conditions d'application sont précisées dans le titre III du décret no 2005-1755 du 30 décembre 2005.
Les informations publiques correspondent aux informations contenues dans les documents produits ou reçus dans le cadre de la mission de service public des administrations de l'Etat, des collectivités territoriales et des personnes publiques ou privées chargées d'une mission de service public (art. 10 de la loi).
Ne sont toutefois pas des informations publiques selon la loi :
a) Les informations contenues dans des documents dont la communication ne constitue pas un droit en application des articles 1er à 9 de la loi du 17 juillet 1978 ou d'autres dispositions législatives, sauf si ces informations font l'objet d'une diffusion publique ;
b) Les informations produites ou reçues dans le cadre de l'exercice d'une mission de service public à caractère industriel et commercial. Cela concerne non seulement les établissements publics à caractère industriel et commercial mais également les administrations pour la part de leur activité effectuée selon les règles du commerce ;
c) Les informations sur lesquelles des tiers détiendraient des droits de propriété intellectuelle.
L'article 10 de la loi prévoit que les administrations sont tenues de mettre les informations publiques à la disposition de toute personne (physique ou morale). Il précise également que les informations publiques peuvent être réutilisées à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été produits ou reçus. L'échange d'informations publiques entre les autorités mentionnées à l'article 1er de la loi du 17 juillet 1978, aux fins de l'exercice de leur mission de service public, ne constitue pas une réutilisation.
Le législateur a fixé une obligation générale concernant tous les types de réutilisation : sauf accord de l'administration, les informations publiques ne doivent pas être altérées, leur sens ne doit pas être dénaturé et leurs sources et la date de leurs dernières mises à jour doivent être mentionnées (art. 12 de la loi).
L'article 11 de la loi prévoit un régime dérogatoire pour les établissements et les institutions d'enseignement et de recherche ainsi que pour les établissements, organismes ou services culturels qui fixent, le cas échéant, leurs conditions de réutilisation de leurs informations publiques. Ces établissements ainsi que les collectivités territoriales et les personnes de droit public ou de droit privé chargées d'une mission de service public peuvent, s'ils le souhaitent, mettre à disposition leurs informations publiques sur le portail « data.gouv.fr ». Dans ce cas, une convention fixe les conditions de réutilisation de ces informations.
Les informations publiques qui comportent des données à caractère personnel peuvent faire l'objet d'une réutilisation si la personne intéressée y a consenti, ou si l'autorité détentrice est en mesure de les rendre anonymes ou, à défaut d'anonymisation, si une disposition législative ou réglementaire le permet (art. 13 de la loi). Ces réutilisations sont en outre subordonnées au respect des dispositions de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.
Par ailleurs, une personne tierce ne peut bénéficier d'un droit d'exclusivité sur la réutilisation d'informations publiques, sauf si ce droit est nécessaire à l'exercice d'une mission de service public (art. 14 de la loi). La réutilisation d'informations publiques peut donner lieu au versement de redevances (art. 15 de la loi). Le décret d'application de la loi prévoit désormais que le principe de la perception de la redevance au titre de la réutilisation d'informations publiques est arrêté par décret pour chaque base de données ou ensemble d'informations publiques (cf. annexe III).
Lorsqu'elle est soumise au paiement d'une redevance, la réutilisation d'informations publiques donne lieu à la délivrance d'une licence. Celle-ci fixe les conditions de la réutilisation des informations publiques. Ces conditions ne peuvent apporter de restrictions à la réutilisation que pour des motifs d'intérêt général et de façon proportionnée. Elles ne peuvent avoir pour objet ou pour effet de restreindre la concurrence (art. 16 de la loi).
La commission d'accès aux documents administratifs (CADA) est chargée de veiller au respect des dispositions légales relatives à la réutilisation des informations publiques. Elle peut être saisie pour avis de toute décision

défavorable en matière de réutilisation d'informations publiques. Cette saisine est un préalable obligatoire à l'exercice d'un recours contentieux (art. 20 de la loi).

ANNEXE III
L'ÉLABORATION D'UNE LICENCE GRATUITE APPLICABLE AUX INFORMATIONS PUBLIQUES MISES EN LIGNE SUR DATA.GOUV.FR ET L'ÉLABORATION DE LICENCES SPÉCIFIQUES GRATUITES OU PAYANTES POUR DES CAS PARTICULIERS
1. «Data.gouv.fr» met à disposition, librement, facilement et gratuitement, le plus grand nombre d'informations publiques des administrations de l'Etat et de ses établissements publics administratifs. Les réutilisations de ces informations se font dans le cadre d'une licence gratuite.
Cette licence gratuite bénéficie aux administrations et aux réutilisateurs qui disposent ainsi d'un outil juridique adapté à la réutilisation gratuite, à la volonté de renforcer la transparence de l'action de l'Etat et au souhait de favoriser l'innovation et de développer l'économie numérique.
Afin d'élaborer cette licence gratuite, «Etalab» conduit un groupe de travail composé de l'Agence du patrimoine immatériel de l'Etat (APIE), du Conseil d'orientation de l'édition publique et de l'information administrative (COEPIA) et des administrations concernées. La conception de la licence gratuite s'appuie sur les dernières versions des licences libres et gratuites élaborées par les administrations membres de ce groupe de travail.
« Etalab » publiera cette licence gratuite dans un délai de quatre-vingt-dix jours à compter de la publication de la présente circulaire.
2. Des licences gratuites spécifiques peuvent être toutefois adoptées dans les cas où la réutilisation d'un jeu de données déterminé ferait l'objet de conditions particulières.
Les administrations concernées les élaborent et les soumettent à « Etalab », qui les valide et les publie sur « data.gouv.fr ». Elles peuvent solliciter l'appui de l'APIE en cas de besoin.
3. Dans certains cas particuliers, la réutilisation peut faire l'objet d'une redevance, comme le prévoit l'article 15 de la loi du 17 juillet 1978. Il revenait jusqu'ici aux administrations concernées de déterminer les informations publiques dont la réutilisation était soumise à redevance. Le décret no 2011-577 du 26 mai 2011 a complété l'article 38 du décret no 2005-1755 du 30 décembre 2005 pour prévoir que lorsqu'il est envisagé de soumettre à redevance la réutilisation d'informations publiques de l'Etat ou d'un de ses établissements publics administratifs, ces informations ou catégories d'informations doivent être au préalable inscrites sur une liste fixée par décret après avis du COEPIA. Cette liste est rendue publique sur un site internet créé sous l'autorité du Premier ministre. Cela ne concerne que les redevances instituées postérieurement au 1er juillet 2011. La décision de soumettre à redevance une base de données ou un ensemble d'informations publiques est prise au vu d'éléments dûment motivés. Le COEPIA est consulté sur cette décision. Il est saisi par le ministère rapporteur du projet de décret. Il rend son avis dans les conditions prévues par le décret no 2006-672 du 8 juin 2006 relatif à la création, à la composition et au fonctionnement de commissions administratives à caractère consultatif.
Les redevances instituées avant le 1er juillet 2011 ne sont pas remises en cause à la seule condition que l'autorité compétente pour délivrer les licences de réutilisation demande leur inscription sur une liste annexée à celle mentionnée au paragraphe précédent. Cette demande doit avoir lieu au plus tard le 1er juillet 2012 sans quoi les redevances deviennent caduques et les titulaires de licences peuvent réutiliser les informations en cause gratuitement.
Il revient à l'administration d'établir avec le concours de l'APIE le montant et les modalités de la redevance ainsi qu'un projet de licence payante qui peut prévoir des mesures de nature à favoriser l'innovation.

ANNEXE IV
LA DÉSIGNATION D'UN INTERLOCUTEUR UNIQUE POUR « ETALAB » DANS CHAQUE MINISTÈRE
1. Dans un délai de dix jours à compter de la présente circulaire, chaque ministère désigne un interlocuteur unique pour « Etalab » afin de faciliter le recensement et la transmission des informations publiques de son administration.
Cette personne est placée sous l'autorité directe et immédiate du secrétaire général du ministère.
Elle est chargée de coordonner la transmission à « Etalab » des informations publiques de son administration.
En s'appuyant sur le guide technique fourni par « Etalab », dans les conditions précisées à l'annexe V, elle est en particulier responsable, pour son ministère, des points suivants :
– identifier les informations publiques produites ou reçues dans le cadre des missions de service public ;
– coordonner le recensement et la qualification des informations publiques ;
– mettre en place une méthode, avec la direction des systèmes d'information du ministère, pour transmettre régulièrement les informations publiques dans des formats exploitables et accompagnées de leurs informations descriptives (métadonnées) ;
– gérer l'attribution et le contrôle des droits d'accès à « data.gouv.fr », et les réponses de son administration aux questions et aux demandes adressées par les réutilisateurs ;
– coordonner les correspondants des établissements publics administratifs relevant de la tutelle de son ministère.
2. Chaque établissement public administratif de l'Etat est invité à procéder à la désignation d'un correspondant de cet interlocuteur unique.

ANNEXE V
LE RECENSEMENT, LA MISE À DISPOSITION ET LA TRANSMISSION DES INFORMATIONS PUBLIQUES EXISTANTES
1. Les typologies de formats, les volumes de données ainsi que leurs dates de livraison à « Etalab » sont déterminés selon un plan d'actions et un calendrier préparé par «Etalab» en concertation avec chaque ministère.
Etant donné les délais et les impératifs liés à la mise en ligne du portail « data.gouv.fr », chaque ministère rencontre « Etalab » dans un délai d'un mois à compter de la publication de la présente circulaire.
Ces rencontres bilatérales fixent des objectifs quantitatifs et qualitatifs sur le nombre de jeux de données transmis à «Etalab» qui s'accompagnent de leurs dates de livraison. Ces objectifs sont revus tous les trimestres.
2. Outre les informations actuellement recensées dans les répertoires d'informations publiques, chaque ministère répertorie les informations publiques en sa possession. « Etalab » leur fournit un guide technique, élaboré en concertation avec l'APIE, qui aide à identifier, recenser, qualifier et transmettre leurs informations publiques.
Chaque ministère s'assure de la diffusion de ce guide technique dans les administrations centrales et déconcentrées relevant de son périmètre. Il est également adressé aux collectivités territoriales, aux autres personnes de droit public et aux personnes de droit privé chargées d'une mission de service public, si elles souhaitent mettre à disposition leurs informations publiques sur le portail « data.gouv.fr ».
3. Pour faciliter la réutilisation de ses informations publiques, chaque ministère les qualifie en renseignant a minima pour chaque jeu de données le titre, la source, la date de la mise à jour, la licence de réutilisation éventuellement appliquée jusqu'à présent, et les mots clés (métadonnées). Chaque ministère indique également les coordonnées de la personne qui est le correspondant des réutilisateurs pour toute question relative à la réutilisation d'informations publiques ou de jeux de données.
Les informations publiques des administrations, et les métadonnées qui s'y rattachent, sont publiées et mises à jour régulièrement. Le guide technique précise les conditions de publication et de mise à jour.
4. « Data.gouv.fr » peut héberger ou référencer les informations publiques. Le guide technique précise les modalités de l'hébergement et du référencement.
Si l'administration publie des informations publiques sur son site internet, elles sont simultanément accessibles sur « data.gouv.fr ».
Chaque ministère privilégie notamment les formats exploitables comme ceux tabulaires (CSV, ODS, XLS, etc.), textes (ODS, RTF, TXT, etc.), structurés (XML, etc.), géographiques (GML, KML, etc.), sémantiques (RDF, etc.), ou calendaires (iCalendar, etc.).
Les formats exploitables sont notamment recommandés dans le référentiel général d'interopérabilité (RGI), créé par arrêté du 11 novembre 2009. Celui-ci préconise certains formats favorisant l'interopérabilité des systèmes d'information et facilitant la transmission et la réutilisation des informations publiques.
Les solutions logicielles développées par les ministères produisent a minima des formats permettant une réutilisation facile de leurs informations publiques.

9. Décret no 2011-577 du 26 mai 2011 relatif à la réutilisation des informations publiques détenues par l'Etat et ses établissements publics administratifs [139]

Le Premier ministre,
Vu la loi no 78-753 du 17 juillet 1978 modifiée portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal, notamment le chapitre II de son titre Ier ;
Vu la loi no 2009-431 du 20 avril 2009 modifiée de finances rectificative pour 2009, notamment son article 4 ;
Vu le décret no 2005-1755 du 30 décembre 2005 relatif à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques, pris pour l'application de la loi no 78-753 du 17 juillet 1978, notamment ses titres III et VI ;
Vu le décret no 2009-151 du 10 février 2009 relatif à la rémunération de certains services rendus par l'Etat consistant en une valorisation de son patrimoine immatériel ;
Le Conseil d'Etat (section de l'administration) entendu,
Décrète :
Art. 1er. – Le décret du 30 décembre 2005 susvisé est ainsi modifié : 1o A l'article 38, il est ajouté deux alinéas ainsi rédigés :
« Lorsqu'il est envisagé, notamment dans les conditions prévues par l'article 3 du décret no 2009-151 du 10 février 2009 relatif à la rémunération de certains services rendus par l'Etat consistant en une valorisation de son patrimoine immatériel, de soumettre au paiement d'une redevance la réutilisation d'informations publiques

[139] Legifrance, « Décret n° 2011-577 du 26 mai 2011 relatif à la réutilisation des informations publiques détenues par l'Etat et ses établissements publics administratifs », https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000024072772&categorieLien=id, consulté le 13 janvier 2017.

contenues dans des documents produits ou reçus par l'Etat, la liste de ces informations ou catégories d'informations est préalablement fixée par décret après avis du conseil d'orientation de l'édition publique et de l'information administrative. La même procédure est applicable aux établissements publics de l'Etat à caractère administratif.
« Sans préjudice de la publication du répertoire mentionné à l'article 36, la liste mentionnée à l'alinéa précédent est rendue publique sur un site internet créé sous l'autorité du Premier ministre, avec l'indication, soit de la personne responsable des questions relatives à la réutilisation des informations publiques mentionnée au titre IV, soit, pour les établissements publics qui ne sont pas tenus de désigner un tel responsable, du service compétent pour recevoir les demandes de licence » ;
2o Après l'article 48 du même décret, il est inséré au titre VI un article 48-1 ainsi rédigé :
« *Art. 48-1.* – Les redevances instituées au bénéfice de l'Etat ou de l'un de ses établissements publics à caractère administratif avant le 1er juillet 2011 demeurent soumises au régime en vigueur avant cette date sous réserve que les informations ou catégories d'informations concernées soient inscrites, dans un délai maximal d'un an à compter de cette date, sur une liste publiée sur le site internet prévu au quatrième alinéa de l'article 38.
«Le responsable du site internet procède à l'inscription des informations ou catégories d'informations mentionnées à l'alinéa précédent sur simple demande de l'autorité compétente pour délivrer les licences de réutilisation.
« A défaut d'inscription des informations concernées sur la liste mentionnée au premier alinéa ou à défaut de publication de cette liste, avant le 1er juillet 2012, les redevances instituées deviennent caduques et les titulaires de licences peuvent réutiliser les informations en cause gratuitement. »
Art. 2. – Le présent décret sera publié au *Journal officiel* de la République française et entrera en vigueur le 1er juillet 2011.
Fait le 26 mai 2011.

10. Vade-mecum sur l'ouverture et le partage des données publiques (2013) [140]

1. Pourquoi ouvrir et partager les données publiques ?

Une priorité de l'action gouvernementale

Le Gouvernement attache une grande importance à l'ouverture et au partage des données publiques (ou « Open Data »). Cette politique est un axe essentiel de la construction d'un gouvernement plus ouvert et plus efficace. C'est donc une dimension importante de la vie démocratique et de la modernisation de l'action publique. C'est aussi un important levier de stimulation du dynamisme économique et de l'innovation.

Cette priorité est inscrite dans la Charte de déontologie du 17 mai 2012 signée par tous les membres du gouvernement dès le premier Conseil des ministres de la mandature. Elle se traduit par onze décisions prises lors des trois premiers Comités interministériels pour la modernisation de l'action publique (CIMAP), présidés par le Premier ministre le 18 décembre 2012, le 2 avril et le 17 juillet 2013. Une ambitieuse feuille de route stratégique a été adoptée lors du séminaire gouvernemental sur le numérique du 28 février 2013. C'est également un engagement réclamé et souscrit par la France avec l'adoption, le 18 juin 2013, par les chefs d'Etat et de gouvernement du G8, de la Charte du G8 pour l'ouverture des données publiques.

Une démarche pour un gouvernement plus ouvert, plus exemplaire et plus efficace (« Open Government »)

L'ouverture et le partage des données, c'est la manière, pour un Etat moderne, de s'organiser afin de rendre des comptes, d'ouvrir le dialogue, et de faire confiance à l'intelligence collective des citoyens.
C'est aussi - souvent - le moyen de simplifier le fonctionnement interne de l'Etat : les administrations sont les premières bénéficiaires de l'ouverture de ces données qui ont été créées pour les besoins du service public. L'ouverture permet souvent d'améliorer la qualité des données, le partage des données entre administrations permettant de créer des systèmes plus complets et les agents publics gagnant à adosser leur travail sur les don- nées produites par d'autres agents pour des missions proches.
C'est aussi un levier pour construire des relations de travail avec des acteurs passionnés par l'intérêt général, qui vont pouvoir prolonger l'action de l'Etat en concevant de nouveaux services utiles à tous les citoyens. Différents exemples d'ouverture de données publiques montrent combien cette politique permet de fonder de nouvelles relations entre l'Etat et les citoyens : en favorisant la simple consultation et en répondant ainsi aux questions que se posent les usagers du service public, en autorisant la construction de points de vues qui ne

[140] Etalab, « Vade-mecum sur l'ouverture et le partage des données », http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/vademecum-ouverture.pdf, consulté le 13 janvier 2017.

sont pas ceux de l'Etat, en enrichissant les débats de la démocratie locale, en facilitant le développement de services d'aide aux handicapés, en favorisant la naissance de services facilitant l'accessibilité des services publics, cartographies interactives, etc.
C'est enfin un levier pour construire la confiance à travers une action ouverte et transparente, sur le plan national

comme sur le plan des relations internationales.

1. http://www.gouvernement.fr/sites/default/files/fichiers_ joints/donnees-publiques.pdf
2. http://www.etalab.gouv.fr/article-les-chefs-d-etat-reunis-a-loughe-erne-signent-une-charte-du-g8-pour-l-ouverture-des-donnees-publique-118576420.html

Une stratégie d'innovation et de stimulation de l'économie

Avec la révolution numérique, les données prennent par ailleurs une place centrale dans l'économie. Ouvrir et partager les données publiques, c'est organiser la mise en ligne de données essentielles, qui vont enrichir les analyses de nombreux décideurs, permettre de nombreuses économies de temps de travail ou permettre, dans de nombreux secteurs, des prises de décisions mieux informées. C'est créer de grands référentiels partagés par tous les acteurs et encourager le développement de nombreux services à forte valeur ajoutée, par exemple dans le tourisme, le transport, la santé ou la maîtrise de la consommation d'énergie.
C'est donc à la fois une stratégie de souveraineté (organiser soi-même la représentation numérique de notre pays) et, dans bien des cas, un fort levier de développement économique.

Quelles sont les données concernées par l'ouverture des données publiques ?

Toutes les données produites ou détenues par l'administration qui entrent dans le champ des données publiques
(voir définition infra) doivent être partagées, gratuitement, et librement réutilisables.
Prioritairement, il importe d'ouvrir et de partager des données susceptibles de présenter un enjeu démocratique ou un intérêt pour les réutilisateurs. De ce fait, les séries complètes, les données permettant de construire des référentiels, les données fréquemment actualisées, les données géolocalisées ou encore les données portant sur la transparence de l'action publique sont particulièrement utiles.
En annexe 2 sont cités quelques exemples de données fréquemment réutilisées.

2. Le cadre juridique de l'ouverture des données publiques

Qu'est-ce qu'une donnée publique ?

Le langage courant confond parfois les « données publiques » avec « l'ensemble des données accessibles en ligne ». Ce n'est pas le sens de la politique d'ouverture et de partage des données publiques, qui est initialement fondée sur la loi sur l'accès aux documents administratifs et sur la directive européenne sur les informations du secteur public. Cette politique concerne les informations ou données produites ou reçues par une autorité administrative dans le cadre de sa mission de service public, publiées par une autorité administrative ou communicables à toute personne en faisant la demande. Ces informations doivent être présentées sous un format permettant leur traitement automatisé et leur réutilisation.

La loi n° 78-753 du 17 juillet 1978 relative au droit d'accès aux documents administratifs, les définit ainsi dans son article 1er : *« (…) quels que soient leur date, leur lieu de conservation, leur forme et leur support, les documents produits ou reçus, dans le cadre de leur mission de service public, par l'Etat, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission. Constituent de tels documents notamment les dossiers, rapports, études, comptes-rendus, procès-verbaux, statistiques, directives, instructions, circulaires, notes et réponses ministérielles, correspondances, avis, prévisions et décisions. (…) »*.

Le droit d'accès et de réutilisation des données publiques concerne donc les textes, mémorandums, documents, tableaux ou statistiques produits par l'administration dans le cadre d'une mission de service public. Il ne concerne pas les documents préparatoires et non définitifs de l'administration en vue de ses délibérations.

Les informations nominatives, les informations personnelles et les informations protégées par des secrets prévus par la loi (secret de la défense nationale par exemple) sont exclues du champ des données susceptibles d'être rendues publiques, sauf disposition légale ou réglementaire contraire.

Les informations statistiques doivent être publiées dans le respect de la loi de 1951, ainsi que de l'article 285 du Traité instituant la communauté européenne, qui définit le secret statistique.

Qu'est-ce que l'ouverture des données publiques ?

L'ouverture et le partage des données publiques consistent à mettre à disposition de tous les citoyens, sur Internet, toutes les données publiques brutes qui ont vocation à être librement accessibles et gratuitement réutilisables. Le droit d'accès à ces données s'impose à l'Etat, aux collectivités territoriales et à toutes les autres personnes de droit public ou de droit privé chargées d'une mission de service public.
Le droit d'accès aux documents administratifs a été reconnu comme une *« liberté publique »* par le Conseil d'État (CE, 29 avril 2002, X., n° 228830). La Déclaration des Droits de l'Homme et du Citoyen de 1789 prévoyait déjà, dans son article 15, que *« La société a le droit de demander compte à tout agent public de son administration »*.

En 1997, le Gouvernement en a élargi le principe en décidant la mise en ligne gratuite des *« données publiques essentielles »*.

En 2003, la directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public, transposée par l'ordonnance du 6 juin 2005 et le décret du 30 décembre 2005, a permis la réutilisation des documents et des informations publiques des organismes du secteur public. La circulaire du Premier ministre et le décret du 26 mai 2011 ont fixé le principe de la réutilisation libre, facile et gratuite pour tous les citoyens.

Enfin, le décret du 21 février 2011 a créé la mission *Etalab*, qui a été rattachée au SGMAP le 30 octobre 2012. *Etalab* est chargée d'accompagner les administrations dans l'ouverture de leurs données publiques, de piloter le portail national data.gouv.fr et d'animer la communauté des réutilisateurs.

Le Gouvernement a réaffirmé son attachement à la gratuité de la réutilisation des données publiques à l'occasion du Comité interministériel pour la modernisation de l'action publique (CIMAP) du 18 décembre 2012 ainsi que dans la *« Stratégie gouvernementale en matière d'ouverture et de partage des données publiques »* publiée le 28 février 2013.

Les données mises à disposition sur la plateforme data.gouv.fr sont sous « Licence Ouverte/ Open Licence » qui garantit la plus grande liberté de réutilisation tout en apportant la plus forte sécurité juridique aux producteurs et aux réutilisateurs des données publiques :

• en promouvant la réutilisation la plus large et en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ;
• en s'inscrivant dans un contexte international compatible avec les standards des licences Open Data 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).

Pourquoi les données doivent-elles être publiées dans un format brut et quels sont les différents formats proposés ?

L'objectif de l'ouverture des données publiques est de favoriser et de faciliter les réutilisations et les réinterprétations, de la manière la plus automatisée et la plus standardisée possible. Les données brutes – telles qu'elles sont produites ou utilisées par les administrations à des fins de service public – sont en ce sens extrêmement intéressantes. Il est préférable de diffuser ces données dans des formats structurés, sans avoir recours à des options de présentation (couleurs, cellules fusionnées, fichiers à plusieurs onglets..), ni à des fonctions de présentations (macros, liens croisés dynamiques…).

Pour en permettre une réutilisation simple par le plus grand nombre, il est recommandé de présenter ces données dans des formats ouverts (Exemple : CSV, JSON, XML, RDF…) qui permettent la réutilisation sans restriction d'accès ni de mise en œuvre, par opposition à un format fermé ou propriétaire. La circulaire du Premier ministre du 19 septembre 2012, sur l'usage du logiciel libre dans l'administration, encourage l'usage de ces formats réutilisables et ouverts.

Dans la mesure du possible, l'ouverture des données publiques requiert la diffusion des données brutes dans des formats normalisés qui permettent une réutilisation simplifiée dans des applications. Les données peuvent égale- ment être diffusées sous forme de flux accessibles à travers des interfaces de programmation (API).

Il est également recommandé que les données diffusées soient les plus exhaustives et les plus précises possible, diffusées à une granularité fine dans le respect de la loi sur le secret statistique, et qu'elles s'appuient sur des référentiels partagés et des nomenclatures décrites et publiées.
Lorsque de tels formats ouverts n'existent pas, on recommande pour autant de partager ces données dans leurs formats d'origine plutôt que de renoncer à leur diffusion. *Etalab* recommande de rechercher autant que possible le véritable format d'origine, et pas, par exemple, le PDF, développé pour le confort de lecture, qui circule usuellement.

Faut-il indexer ces données avant de les transmettre ?

La qualification des métadonnées et l'indexation sont une étape essentielle pour faciliter la réutilisation des données publiques. Les données sont très difficiles à retrouver si elles ne sont pas indexées et elles sont difficilement réutilisables si elles ne sont pas décrites avec précision.

Ces informations complémentaires décrivant les données sont appelées « métadonnées ». *Etalab* propose ainsi des champs de descriptions normalisées à tous les producteurs de données publiques afin de leur permettre de spécifier le contexte et le contenu des données. Il leur est notamment demandé de caractériser leurs données (titre, description, mots-clés…) en répondant aux questions suivantes :

• Qui a produit les données ?
• Quand les données ont-elles été produites ?
• Quelle est la période temporelle concernée ?
• Quelles sont les zones géographiques couvertes ?
• Quelles sont les thématiques des données ?

Par ailleurs, pour faciliter la réutilisation la plus large possible des données publiques, *Etalab* recommande que tout jeu de données soit accompagné d'une description du contenu du jeu de données. Ce document annexe peut se révéler très important pour les réutilisateurs.

Comment s'assurer de la qualité des données mises en ligne ?

Les données publiques sont produites ou reçues dans le cadre d'une mission de service public. Elles sont donc généralement d'une qualité permettant le travail quotidien de l'administration et, en fonction de leur destination initiale, une utilisation statistique pertinente. Le document annexe présentant les jeux de données pourra, si nécessaire, préciser les méthodes de production et les limites intrinsèques des données proposées.
Toutefois, les grands systèmes d'information de l'Etat et des collectivités territoriales, tout comme ceux des entreprises, peuvent parfois comporter des erreurs. L'existence de ces erreurs ne doit pas ralentir la démarche d'ouverture et de partage des données publiques. L'ouverture et le dialogue avec les réutilisateurs favorisent le signalement d'erreurs éventuelles.
C'est pourquoi, il est recommandé d'intégrer la perspective de l'ouverture des données et le besoin de qualification des jeux de données dans la conception et la rénovation des systèmes d'information.

Peut-on vendre des données publiques ?

Le cadre juridique et réglementaire, rappelé par le Premier ministre au cours du CIMAP du 18 décembre 2012 puis du Séminaire gouvernemental sur le numérique du 28 février 2013, prévoit la gratuité des données publiques comme principe par défaut.
Pour certaines données, liées à l'obligation de rendre des comptes au citoyen, cette gratuité est un pré-requis.
Pour d'autres données, l'expérience a montré que la mise à disposition de ces données gratuites favorisait la création de services à valeur ajoutée économique ou sociale, et donc l'émergence de nouveaux services au public et un surcroît de revenus pour l'Etat.
Cependant, le droit n'interdit pas systématiquement la facturation du coût de mise à disposition des données publiques : il autorise en effet la facturation du coût de la mise à disposition de la donnée, ainsi que celle de services à valeur ajoutée. Cette autorisation est souvent importante pour les opérateurs dont la mission est de produire de l'information, et dont l'équilibre budgétaire peut dépendre de ces revenus complémentaires.

En tout état de cause, il importe que d'éventuelles redevances sur les données ne créent pas de monopoles de fait ou de barrières à l'entrée susceptibles de freiner l'innovation et notamment celles des jeunes entreprises.

Le décret du 26 mai 2011 a prévu qu'à compter du 1er juillet 2011, les informations ou catégories d'informations dont la réutilisation peut être soumise au paiement d'une redevance doivent figurer sur une liste fixée par décret, donc après décision expresse du Premier ministre.
Pour les redevances instituées avant l'entrée en vigueur du décret, les administrations de l'Etat et ses établissements publics à caractère administratif avaient jusqu'au 1er juillet 2012 pour faire inscrire sur une seconde liste les informations ou catégories d'informations concernées. Ces deux listes ont été publiées sur data.gouv.fr.
Cette procédure ne s'applique qu'aux informations publiques faisant l'objet d'une redevance de réutilisation au sens du chapitre II du titre 1er de la loi n° 78-753 du 17 juillet 1978.

A l'occasion du CIMAP du 18 décembre 2012, le Premier ministre a décidé de la création d'une mission d'évaluation des modèles économiques de ces redevances. Cette mission a remis ses conclusions au Premier ministre à l'été 2013, notamment en dressant un « bilan coût-avantage » et en réunissant les « éléments permettant de justifier la pertinence » de ces redevances ainsi que les conditions de leur pérennité. Le Gouvernement annoncera au cours de l'automne 2013 ses décisions concernant la gratuité de nouveaux jeux de données, et les éventuelles évolutions des modèles économiques de certains opérateurs.

Y a-t-il un risque pour la protection de la vie privée ?

Dans la pratique, la démarche d'ouverture et de partage des données publiques par l'Etat ne concerne pas les données à caractère personnel.

Il peut cependant arriver que des informations publiques personnelles soient publiées par l'Etat, après disposition expresse (exemple : lauréats du baccalauréat). Dans ce cas, la loi du 17 juillet 1978 dispose que : « Les informations publiques comportant des données à caractère personnel peuvent faire l'objet d'une réutilisation soit lorsque la personne intéressée y a consenti, soit si l'autorité détentrice est en mesure de les rendre anonymes ou, à défaut d'anonymisation, si une disposition législative ou réglementaire le permet.
La réutilisation d'informations publiques comportant des données à caractère personnel est subordonnée au respect des dispositions de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. »

On rappelle par ailleurs que la loi du 7 juin 1951 organise le secret statistique, qui permet d'assurer :

• aux personnes physiques que la confidentialité sur leur vie personnelle et familiale sera garantie ;
• aux entreprises que le secret commercial sera respecté.

3. Comment se lancer dans une démarche d'ouverture et de partage des données publiques ?

Sur quel support peut-on diffuser les données publiques ?

La plateforme data.gouv.fr peut héberger toutes les données publiques produites notamment par les administrations, les établissements publics ou les collectivités locales.
Par ailleurs, certaines administrations, collectivités locales ou opérateurs ont développé des portails permettant l'ouverture et le partage de données publiques spécifiques, répondant aux contraintes particulières de leur système d'information ou de leur communauté de réutilisateurs. Dans ce cas, il n'est pas nécessaire de dupliquer ces données sur data.gouv.fr, mais il est fondamental d'y placer une fiche de description des données, conte- nant les métadonnées concernées, afin de faciliter les recherches des internautes. Ce recours à la plateforme nationale améliore le référencement des acteurs publics et intensifie le dialogue avec leur communauté de réutilisateurs.

Qui contacter pour engager une démarche d'ouverture de données publiques ?

La mission *Etalab* est chargée de créer et de développer la plateforme data.gouv.fr. Elle anime un réseau de 12 coordonnateurs ministériels « Open Data » placés sous l'autorité directe des secrétaires généraux des ministères. Ce réseau de coordinateurs se réunit tous les mois au sein d'un comité de pilotage de l'Open Data animé par *Etalab*. Ils s'appuient sur des correspondants au sein des directions, bureaux et services de leurs administrations. Il existe donc un coordinateur Open data auprès de chaque secrétaire général d'un ministère.
La feuille de route du gouvernement en matière d'ouverture et de partage des données publiques demande à *Etalab* de veiller à faciliter de plus en plus les conditions techniques de transfert des données vers la plateforme data.gouv.fr. Ces modalités vont donc évoluer rapidement dans le sens d'une grande simplification. En tout état

de cause, si votre entité de service public souhaite s'engager dans une politique d'ouverture et de partage des données publiques, la mission *Etalab* est chargée de vous y aider et de vous en faciliter la démarche.

Comment publie-t-on concrètement les données sur data.gouv.fr ?

Deux méthodes sont possibles pour publier des données publiques sur data.gouv.fr :

• **le versement manuel** : le producteur s'identifie sur l'espace d'administration de data.gouv.fr, décrit les données en renseignant les « métadonnées » associées au jeu de données et transmet ou référence le fichier de données à mettre en ligne. Un jeu de données est chargé en quelques minutes dans l'espace d'administration et ne mobilise qu'une seule personne. Les entités qui le souhaitent peuvent déléguer la validation et/ou la publication des données à un tiers, autre que le producteur.
• **le versement automatisé** : cette démarche concerne les administrations disposant d'importants volumes de données issues de systèmes d'informations ou de données fréquemment mises à jour. *Etalab* propose une interface standardisée, documentée et gratuite, permettant le déversement automatisé de données, et rencontre à la demande les équipes techniques du producteur pour soutenir la mise en place de l'interface.

Quelles sont les retombées d'une démarche d'ouverture des données publiques ?

Ouvrir les données publiques n'est pas seulement un moyen de respecter le principe démocratique de transparence et de motivation de la décision. Cette démarche peut également se révéler très utile :

• pour simplifier les processus internes à l'administration elle-même (notamment en favorisant la circulation du savoir entre les services, et en facilitant le travail quotidien des agents publics) ;
• pour simplifier les démarches des usagers et renforcer les relations de confiance avec les citoyens ;
• pour prolonger et amplifier l'effort de l'administration grâce à des services complémentaires développés par les innovateurs extérieurs ;
• pour attirer à soi des cultures innovantes issues d'horizon divers.

L'ensemble du SGMAP est à la disposition des administrations qui souhaiteraient travailler ces objectifs dans le cadre d'un projet d'ouverture des données publiques.

4. Quelles réutilisations seront faites ?

Qu'est-ce que la réutilisation des données publiques ?

La réutilisation des données publiques peut susciter le développement de nouveaux services comme les applications mobiles, des sites Internet, des visualisations données ou « datavisualisation » notamment par la presse, etc. Elle doit être autorisée sans restrictions autres que celles prévues par la loi CADA (qui demande que ces informations ne soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources et la date de leur dernière mise à jour soient mentionnées).

Les données publiques peuvent être aussi réutilisées par les chercheurs, les enseignants, les étudiants, les responsables associatifs, les citoyens, pour construire de nouveaux points de vue sur la société ou sur l'action publique.

Quelles réutilisations seront faites des données mises en ligne ?

L'objectif d'une politique d'Open Data est d'encourager la créativité, stimuler l'innovation et de favoriser la réutilisation la plus large possible des données publiques en se reposant sur l'intelligence collective et la volonté des citoyens de créer de nouveaux services innovants utiles à tous.

La « Licence Ouverte / Open Licence », sous laquelle les données sont publiées sur data.gouv.fr, rappelle aussi une règle simple : la réutilisation reste de la responsabilité du réutilisateur. Tout usage illégal reste illégal même lorsqu'il est fondé sur des données publiques.

Comment suivre les différentes réutilisations de données ?

Afin d'encourager la réutilisation des données publiques, qu'elles proviennent de l'Etat, des collectivités territoriales ou d'autres entités de service public, *Etalab* a engagé en 2012 et en 2013 une série de quatre concours de création de projets et de services innovants. Il s'agit de l'initiative « Dataconnexions ». Les différents producteurs de données publiques sont particulièrement associés aux projets lauréats. En participant à l'animation de la communauté de l'Open Data, *Etalab* contribue également à mettre en lumière les meilleures réutilisations de données, notamment en assurant leur promotion au sein de l'Etat.

Par ailleurs, les évolutions prochaines du portail data.gouv.fr accorderont une place croissante à l'appropriation par le plus grand nombre des données partagées, à l'enrichissement des données par les utilisateurs, notamment les citoyens, et à la mise en valeur des réutilisations.