ARTICLE DU DOSSIER

De la « donnée » à la « donnée ouverte » : les épreuves de l'ouverture des données

Antoine COURMONT

Chercheur post-doctorant, Centre d'études européennes et de politique comparée, Sciences Po[1]

L'expression « ouverture des données publiques » peut suggérer, à tort, qu'une simple décision administrative suffirait à rendre accessibles des « gisements d'information » préexistants. En réalité, c'est d'un processus d'ouverture qu'il faut parler. On peut le décomposer en trois phases : l'identification, la publicisation, et l'extraction proprement dite. Ce n'est qu'au terme des « épreuves de diffusibilité » que les données peuvent être dites « ouvertes ».

La loi pour une République numérique, dite loi Lemaire, votée en 2016, instaure un principe d'open data « par défaut » pour les administrations publiques[2]. Cela renverse le paradigme en matière de diffusion des informations publiques : l'ouverture devient la règle, la fermeture l'exception. Cette loi marque ainsi le passage d'une logique de demande à une logique d'offre. Sauf exception, les données sont considérées comme ouvertes par défaut. L'analyse ethnographique du processus de mise à disposition des données publiques au sein d'une collectivité territoriale française révèle toutefois l'impensé de ce principe d'open data par défaut : les données ne peuvent être ouvertes par défaut, puisqu'elles acquièrent précisément ces caractéristiques de diffusibilité au cours du processus d'ouverture.

Le travail de catégorisation est en effet l'enjeu central du processus d'ouverture de données. Ouvrir une donnée nécessite en premier lieu d'identifier ce qu'est une donnée, puis de déterminer sa diffusibilité. Pour comprendre comment une donnée devient une donnée ouverte, il faut observer, en situation, ce que les acteurs entendent par le terme de « données », de « données publiques », de « données ouvertes », de « données candidates », de « données personnelles », de « données sensibles », etc. Ce travail de catégorisation est au cœur de la politique d'ouverture de données : plus que d'éventuelles résistances d'acteurs accrochés à leur pouvoir, il détermine pourquoi certaines données sont, ou non, mises à disposition d'un nouveau public.

Les « barrières » qui restreindraient l'ouverture des données occupent une place centrale dans les écrits, militants ou académiques, sur l'*open data*[4]. Janssen et al. pointent le fait que l'ouverture des données implique une transformation institutionnelle d'un système fermé à un système ouvert[5]. Gray et Davies appellent à passer d'une vision de la « libération » des données

1. antoine.courmont@sciencespo.fr

2. Article 6 de la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
3. Les résultats présentés dans cet article sont issus d'une thèse de doctorat en science politique. Courmont A. (2016), Politiques des données urbaines. Ce que l'open data fait au gouvernement urbain, Sciences Po, Paris, 423 p.
4. Martin C. (2014), Barriers to the Open Government Data Agenda: Taking a Multi-Level Perspective, *Policy & Internet*, vol. 6, no 3.
5. Janssen M., Charalabidis Y. et Zuiderwijk A. (2012), Benefits, Adoption Barriers and Myths of Open Data and Open Government, *Information Systems Management* (*ISM*), vol. 29, no 4, 258-268.

à une politique de recomposition de l'infrastructure informationnelle[6]. Les barrières à lever pour permettre ce changement sont nombreuses : héritage institutionnel, emprise politique, aversion au risque, complexité de la démarche, contraintes techniques, incertitude juridique, culture « fermée » des administrations, etc. A partir de l'étude de politiques locales d'open data, Peter Conradie et Sunil Choenni pointent trois principaux facteurs limitant la mise à disposition des données : le stockage décentralisé des données, les sources externes de données, le non-usage de la donnée dans le cœur du service public[7].

En soulignant les nécessaires recompositions institutionnelles, cette littérature pointe le fait que les données ne sont pas autonomes d'un environnement social. Néanmoins, en se focalisant sur les « barrières » de l'open data, ces auteurs s'abstiennent de s'intéresser aux données elles-mêmes. Dans leur perspective, les données préexistent à l'ouverture et elles ne jouent aucun rôle dans ce processus dans la mesure où il suffit de convaincre leurs propriétaires de lever des « barrières » pour les diffuser. Or, l'ouverture des données redéfinit tout autant les données que l'environnement dans lequel elles sont insérées. Pour comprendre comment une « donnée fermée » devient une « donnée ouverte », il est nécessaire d'adopter une perspective relationnelle qui étudie symétriquement les données et l'environnement dans lequel elles sont insérées.

Au cours du processus de diffusion, la donnée subit en une série d'épreuves qui détermine son avenir en définissant sa « diffusibilité[8] ». En catégorisant différemment la donnée, l'épreuve participe au détachement de la donnée de l'infrastructure informationnelle dans laquelle elle est insérée. En effet, la donnée n'est jamais brute, mais elle est toujours étroitement associée à un ensemble de personnes, de pratiques, de technologies, d'institutions qui les produisent, les maintiennent et les utilisent. Pour permettre sa diffusion et son utilisation dans un environnement autre, il est nécessaire de délier l'ensemble de ces attachements afin de rendre la donnée autonome de ce cadre initial. Cela oblige à reprendre une à une les composantes de l'attachement, ce qui exige de prendre en compte la donnée dans toutes ses dimensions : juridiques, techniques, économiques, politiques, etc.

Le travail de détachement est toutefois indissociable d'un travail d'attachement. Dissocier, c'est créer de nouveaux liens autant que d'en défaire d'autres. Pour susciter l'intérêt à l'*open data*, la donnée doit être associée à de nouveaux enjeux (développement économique, émergence de nouveaux services, simplification du travail des agents, image innovante de la collectivité, etc.) et à de nouveaux utilisateurs (entreprises, développeurs, citoyens, agents de collectivités territoriales, etc.). Tout autant que la donnée, le public des données ouvertes est défini au travers de ces épreuves. Des usages potentiels des données sont préfigurés tout au long du processus, un travail de cadrage, bien connu des sociologues des techniques, que Madeleine Akrich a défini comme des scripts[9]. Etudier le processus de diffusion des données c'est analyser symétriquement les médiations qui détachent et les médiations qui attachent.

Ainsi, le travail préliminaire à la mise à disposition des données éprouve à la fois le détachement des données de leur infrastructure informationnelle et leur attachement à de nouveaux utilisateurs. Multiples, ces épreuves de *diffusibilité* peuvent être regroupées en trois catégories : l'identification, la publicisation et l'extraction.

---

6. Gray J. et Davies T. (2015), Fighting Phantom Firms in the UK: From Opening Up Datasets to Reshaping Data Infrastructures ? Paper presented at the Open Data Research Symposium, Ottawa.
7. Conradie P. et Choenni S. (2014), On the barriers for local government releasing open data, *Government Information Quarterly*, vol. 31, p. S10-S17.
8. La diffusibilité est un terme scientifique caractérisant l'aptitude d'une substance fluide à se diffuser (gaz, lumière). Ce terme est préféré à celui, plus courant, de diffusabilité, utilisé pour désigner quelque chose que l'on peut diffuser. En effet, il est plus adapté à l'argument de ce chapitre puisqu'il souligne le fait que la donnée est transformée au cours du processus d'ouverture pour acquérir des propriétés de diffusibilité.
9. Akrich M. (1987), Comment décrire les objets techniques ?, *Techniques et culture*, no 9, p. 49-64.

Identifier

Identifier les données candidates à l'ouverture n'est pas une tâche aisée. La donnée ne préexiste pas à son ouverture : elle n'est pas déjà-là, prête à être mise à disposition. Comme le soulignent Jérôme Denis et Samuel Goëta, les données publiques « *ne sont pas disponibles en l'état, prêtes à être libérées. Leur existence même est loin d'être une évidence*[10] ». Une institution telle que la communauté urbaine de Lyon ne connaît pas son patrimoine de données de manière exhaustive. Il n'existe pas en son sein de catalogue[11] recensant l'ensemble des données traitées par les différents services et directions[12]. Dès lors, comme l'affirme un urbaniste du système d'information : « *aujourd'hui, on ne sait pas ce que l'on possède* ». En effet, contrairement aux applications, les données n'ont pendant longtemps pas été perçues comme un actif stratégique du système d'information.

Pour ouvrir une donnée, il faut dès lors, en premier lieu, savoir ce qu'est une donnée, quelles données existent et qui les détient. Cette étape cruciale d'identification des données est effectuée par les responsables du projet *open data*. A la manière d'explorateurs, ils s'appuient sur différents outils, qu'ils perfectionnent peu à peu, pour tenter de s'orienter au sein de l'espace informationnel de l'agglomération : des cartes, des acteurs « référents » au sein des institutions, des réseaux de producteurs de données, la connaissance d'utilisateurs externes, etc. Ces outils leur offrent différentes pistes qu'ils vont suivre de manière décousue afin d'identifier les données candidates à l'ouverture. Comme lors de toute démarche exploratoire, l'incertitude règne, et les découvertes sont parfois le fruit du hasard de la sérendipité. Ce travail exploratoire ne se déroule pas uniquement au début du projet, mais il est permanent et toujours actualisé afin de continuer à enrichir le portail de mise à disposition de données.

L'exploration de l'espace des données de l'agglomération lyonnaise repose sur une démarche pragmatique et opportuniste. Elle ne repose sur aucune stratégie déterminée préalablement. En l'absence de cadre planificateur, l'identification du patrimoine informationnel est loin d'être exhaustif. Seule une minorité de donnée est identifiée à l'issue de cette épreuve. L'épreuve d'identification peut ainsi marquer la fin du processus d'ouverture pour certaines données. C'est le cas de celles dont les porteurs de projet n'ont pas connaissance, ou de celles identifiées, mais qui ne sont pas associées à la plateforme par manque de volonté, oubli ou d'autres priorités. A l'issue de l'épreuve d'identification, certaines données sont écartées de la démarche d'ouverture de la communauté urbaine. Seule une minorité de données accèderont à l'épreuve suivante de la publicisation.

Publiciser

Etape phare de l'ouverture des données, la publicisation est l'association des données à un nouvel usage et un nouveau public d'utilisateurs. De la même manière que les ingénieurs mettent en scène des utilisateurs tout au long de la phase de conception des objets techniques[13], les producteurs préfigurent des usages potentiels à partir desquels ils jugent de l'opportunité de mettre à disposition leurs données. Ils avancent un certain nombre d'hypothèses sur les éléments qui composent le monde dans lequel la donnée ouverte doit prendre place. Ils

10. Jérôme Denis et Samuel Goëta, « La fabrique des données brutes. Le travail en coulisses de l'open data. » dans Clément Mabi, Jean-Christophe Plantin et Laurence Monnoyer-Smith (eds.), *Ouvrir,partager, réutiliser. Regards critiques sur les données numériques*, Paris, Editions de la Maison des Sciences de l'Homme, 2017, p.
11. La loi CADA impose pourtant aux administrations un catalogue de leurs informations publiques. « Les administrations qui produisent ou détiennent des informations publiques tiennent à la disposition des usagers un répertoire des principaux documents dans lesquels ces informations figurent. » (Article 17 de la loi n°78-753)
12. Ce projet de catalogage exhaustif des données de l'institution est régulièrement annoncé comme indispensable par les acteurs au sein du Grand Lyon, mais également des villes américaines étudiées (New York, Philadelphie, Chicago).
13. Akrich M. (1993), Les objets techniques et leurs utilisateurs, de la conception à l'action. dans Bernard Conein, Nicolas Dodier et Laurent Thévenot (eds.), Les objets dans l'action, Raisons Pratiques., Paris, Editions de l'EHESS, p. 35-57.

élaborent des scénarios – ou scripts – mettant en scène des acteurs et l'espace dans lequel ils vont évoluer. Ces représentations varient en fonction des producteurs et des données. Selon une direction, la mise à disposition des données peut mener à une modification des rapports de pouvoir, à une remise en cause de l'action de la collectivité, ou encore à des usages malveillants pour les citoyens. D'autres producteurs de données peuvent être davantage attentifs à des aspects économiques (crainte d'espionnage industriel), juridiques (responsabilité du producteur engagée en cas de mauvais usage), ou sécuritaire (risque pour la sécurité publique et l'intérêt général). Enfin, pour certaines données spécifiques aux processus métiers internes à la collectivité, les producteurs n'imaginent aucun usage potentiel par des acteurs externes et ne voient donc pas l'intérêt de les mettre à disposition.

Afin d'intéresser les producteurs à l'*open data*, l'enjeu est alors de pondérer ces risques perçus par les avantages potentiels à l'ouverture de leurs données. La donnée est considérée comme publiable quand le producteur estime que les avantages à associer la donnée à de nouveaux utilisateurs sont plus grands que les risques inhérents. La publicisation de la donnée consiste ainsi à représenter un « public fantôme[14] » et à le traduire en une multitude d' « êtres intermédiaires[15] », dont on ne sait s'ils seront les utilisateurs réels des données mises à disposition, mais que l'on mobilise pour déterminer le caractère de diffusibilité des données.

L'épreuve de publicisation souligne le lien étroit entre la donnée et son usage. Contrairement aux discours des militants de l'*open data* qui incitent les producteurs à diffuser leurs données sans se préoccuper des usages qui en découleront, en pratique, le processus d'ouverture des données se caractérise par la construction d'utilisateurs imaginés. Ces « êtres intermédiaires » prennent des formes multiples selon les données et les enjeux auxquels elles sont associées. Cette phase de publicisation nous révèle, dans la continuité des travaux des pragmatistes américains[16], qu'il n'existe pas un public homogène. Il y a plusieurs publics « fantômes » pour reprendre l'expression de Lippmann qui sont rassemblés autour d'un problème (*issue*) spécifique matérialisé ici par une donnée. Dès lors, dans l'ouverture des données publiques, la notion du « public » est tout autant à interroger que celle de la « donnée ». Plutôt que l'ouverture des données publiques, il faut ainsi questionner l'ouverture des données aux publics.

En corollaire de la mise en visibilité des données sur une plateforme publique, l'épreuve de publicisation révèle également les dynamiques inverses de constitution d'opacité autour des données non publiées. L'ensemble des questionnements auxquelles sont soumises les données, leur association à de nouveaux usages et utilisateurs, les incertitudes des producteurs, les scènes de négociation et d'intéressement constituent un processus de mise en visibilité tout autant qu'ils rendent invisibles les données non publicisées à l'issue de cette épreuve. Solidement attachées à leurs producteurs, ces dernières n'acquièrent pas les caractéristiques de données diffusibles.

Extraire

La dernière épreuve de diffusibilité de la donnée est son extraction. Elle correspond à l'ensemble des opérations nécessaires pour extraire la donnée de son environnement initial et la mettre à disposition sur une infrastructure de diffusion. Les données ne sont en effet jamais indépendantes d'une infrastructure technique au sein de laquelle elles sont produites et utilisées. Ces systèmes d'information n'ont pas toujours été conçus pour permettre l'extraction et la mise à disposition des données qu'ils contiennent. Une transformation de l'infrastructure

14. Lippmann W. (1927), The *phantom public*, New York, Simon & Schuster.
15. Boullier D. (2010) Le client du poste téléphonique : archéologie des êtres intermédiaires, dans *Débordements. Mélanges pour Michel Callon*., Paris, Presses de l'Ecole des Mines, p. 41-61.
16. Lippmann W., *The phantom public*, op. cit. ; Dewey J. (2010) *Le public et ses problèmes*, Paris, Gallimard.

est souvent réalisée pour permettre la mise à jour automatique des données sur le portail, un changement de format de données ou la diffusion de données volumineuses et en temps réel, tout en assurant la sécurité du système d'information de l'institution. Ces opérations mettent à l'épreuve les liens qui associent la donnée à un système technique et participent au travail de détachement de la donnée de son environnement initial et de son attachement à un système externe. Cette épreuve doit permettre de stabiliser l'identité des données comme « données ouvertes ».

A l'issue de cette troisième épreuve de diffusibilité, certaines données identifiées et publicisées ne sont pas mises à disposition du fait de contraintes techniques ou organisationnelles. Ce dernier point est particulièrement important dans la mesure où la diffusion des données est souvent une tâche qui incombe aux producteurs alors qu'elle est chronophage et peu valorisée par leur encadrement. Plus généralement, cette épreuve précise encore le public associé aux données. Certains attributs sont sélectionnés ou modifiés, des formats sont identifiés, des modes d'accès et d'actualisation sont mis en place en fonction d'usages préfigurés des données[17]. Un certain type d'utilisateur est ancré dans l'infrastructure de diffusion au travers des formats, des métadonnées, des modalités d'accès, etc.

Conclusion : de la « donnée » à la « donnée ouverte »

Au cours des différentes épreuves du processus d'ouverture, les acteurs définissent successivement ce qu'est, ou n'est pas, une donnée ouverte. L'épreuve d'identification catégorise la « donnée candidate ». La publicisation fait émerger la « donnée publiable ». Enfin, l'extraction précise ce qu'est une « donnée ouverte ». La donnée est modifiée tout au long de la chaîne de diffusion pour finir par se stabiliser comme une entité diffusable. Ce processus est réversible : des données non catégorisées comme telles peuvent le devenir, et, inversement, des données ouvertes peuvent être redéfinies comme non-diffusables.

La donnée ouverte n'est ainsi pas pré-existante à son ouverture. Au début du processus de diffusibilité, la donnée n'existe pas. Dès lors, il est impossible d'affirmer que la donnée a une essence, c'est-à-dire certaines propriétés desquelles, il serait possible de déterminer, dès le début de ce processus, sa diffusibilité. Loin d'être joué à l'avance, le processus d'ouverture des données est le résultat d'une série d'épreuves, au résultat toujours incertain, au cours desquelles les caractéristiques des données, des producteurs, des utilisateurs, sont jugées et redéfinies par les acteurs. Ces différentes épreuves de diffusibilité conduisent ainsi à recomposer le réseau des données afin qu'elles soient considérées comme « diffusables ».

17. Ce que Jérôme Denis et Samuel Goëta nomment la "brutification".
Denis J. et Goëta S. (2017), Rawification and the careful generation of open government data, *Social Studies of Science*, SAGE Publications, 47 (5), pp.604 - 629.