**Thèse de doctorat en sociologie**

Samuel Goëta, Telecom ParisTech
Sous la direction de Jérôme Denis

# Résumé du mémoire « Les coulisses de l'open data : sociologie de la production et de la libération de données publiques »

Aujourd'hui, plus de cinquante gouvernements dans le monde ont initié une démarche d'*open data*. Ces politiques, mises en œuvre aussi dans des collectivités locales et des institutions internationales, ont conduit à l'ouverture de portails publics contenant des données issues du travail des administrations qui peuvent être librement réutilisées. Plutôt que d'explorer les potentiels de la réutilisation comme la majorité des études actuelles sur l'*open data,* cette enquête entre dans les coulisses des administrations pour faire remonter à la surface le travail invisible de l'ouverture des données publiques. Elle s'appuie sur les nombreux travaux en *Science and Technology Studies* qui ont documenté de manière précise et approfondie les opérations concrètes qui constituent la fabrique des données en sciences. Une de ses branches en particulier, les *Infrastructures Studies,* s'est intéressée depuis plusieurs décennies, aux pratiques de partage de données scientifiques, aujourd'hui généralisées dans certaines disciplines. Ces travaux proposent des ressources essentielles pour mieux comprendre ce qui joue dans les administrations au moment où la législation tente de généraliser l'ouverture des données.

Le premier chapitre retrace les origines des grands « principes » de l'ouverture des données. Délimité en six épisodes, il s'appuie sur des sources publiques (archives du web, listes de diffusion, wikis) pour resituer l'élaboration de manifestes, d'outils de *benchmarking* et d'une déclaration diplomatique. Il révèle la coexistence de deux modèles divergents de l'ouverture des données : l'un réclamant l'ouverture de l'ensemble des données publiques, l'autre désignant des données essentielles ; l'un proposant des critères essentiellement techniques d'une donnée ouverte, l'autre s'intéressant au contenu même des données. Au-delà de ces divergences, ces moments de définition de *l'open data* répandent une nouvelle catégorisation des politiques de diffusion de l'information publique fondées sur la donnée. Elles s'intéressent 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.

Le deuxième chapitre montre comment ces grands principes ont été traduits en politiques publiques. Il décrit la trajectoire de la mission Etalab, le service en charge de l'ouverture des données de l'Etat français créé en 2011. D'une mission dédiée au départ uniquement à l'ouverture vers une administration des données, la trajectoire d'Etalab traduit un des effets de la catégorisation par la donnée des politiques informationnelles évoquée dans le chapitre précédent. L'attention portée aux données a conduit à la création de structures comme Etalab 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.

Après avoir dressé la généalogie des politiques d'open data, l'enquête entre dans les coulisses des administrations. Chaque chapitre qui suit retrace les grandes étapes du processus d'ouverture des données en s'appuyant sur l'analyse d'un corpus d'entretiens, d'observations et de documents internes constitué lors d'une enquête ethnographique dans une variété d'organisations situées en France. Il débute, dans le troisième chapitre, par 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 et se nourrit d'explorations progressives et incertaines. Pour ancrer l'identification dans l'organisation, des lieux et des personnes responsables de l'ouverture des données sont souvent désignés. L'identification constitue un geste d'instauration à part entière : 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 qui 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 » qu'ils invoquent. Les difficultés d'extraction constituent l'une d'entre elles. 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. Autre source de frictions, la qualité : les projets d'*open data* concernent souvent 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 d'autres cas, 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. Si elles sont jugées « sensibles », les données doivent obtenir l'approbation de la hiérarchie pour être ouvertes, passant à travers des circuits de validation plus ou moins formalisés.

Le cinquième chapitre décrit les transformations que subissent les données avant leur ouverture. Une première série de transformations est opérée par les formats de données préconisés par les politiques d'*open data*. Le chapitre s'arrête sur un format en particulier, le CSV (*Comma Separated Value*), qui réclame souvent une transformation des fichiers coûteuse en temps et en énergie. Certains standards peuvent aussi réclamer des transformations au niveau des termes, des catégories et des nomenclatures. Le cas de l'implémentation d'un standard, le GTFS (*General Transit Feed Specification)*, dans une entreprise de transport montre comment les données peuvent être configurées pour les développeurs et permet d'évaluer le coût de l'imposition de définitions communes aux objets que décrivent les données. Enfin, les transformations peuvent parfois porter sur le contenu même des données. Regroupées sous le terme d'édition, ces opérations peuvent servir à réduire les risques politiques ou juridiques de l'ouverture, mais aussi à garantir leur intelligibilité. Cette dernière doit souvent répondre à deux horizons divergents : une intelligibilité pour les machines, opérée notamment par les formats et les standards, et une intelligibilité pour les humains, mise en œuvre lors du travail d'édition. Ces cas interrogent la notion de données brutes dont l'intelligibilité résulte souvent des transformations évoquées précédemment. La lisibilité des données brutes par les machines, réclamée dans les politiques *d'open data*, reconfigure la répartition du travail de transformation des données des réutilisateurs vers les agents.

Le sixième chapitre s'intéresse aux instruments qui instaurent 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, configurent les publics de l'*open data*. J'en évoque trois en particulier : la visualisation des données, la production des métadonnées et les concours de réutilisation. 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. 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. 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 de services et d'application. 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*.

En conclusion, je reviens sur un des fils rouges qui traverse ce mémoire : qu'est ce qu'une donnée? Cette question fondamentale, qui peut trouver une réponse définitive et des caractéristiques stables dans certains travaux, se pose de manière saillante au moment où des acteurs politiques envisagent la généralisation de l'ouverture à l'ensemble des données publiques. En suivant la transmutation de fichiers de gestion en des données ouvertes, disponibles pour de nombreux traitements, cette enquête invite à se demander à quel moment ces fichiers sont considérés comme des données. Plutôt que de réduire la donnée à une catégorie relative, attribuée à toutes sortes de matériaux informationnels, les cas étudiés montrent qu'elle est généralement attribuée dès lors que ces ressources sont prises dans des réseaux sociotechniques dédiés à leur circulation, leur exploitation et leur mise en visibilité. Ce n'est que lorsque des publics, humains ou non-humains, internes ou externes, se lient à eux que les milliers de fichiers des portails d'*open data* deviennent des données.