Université de Montréal

Comprendre et manipuler les données ouvertes de l'administration publique

La situation au Gouvernement du Québec et à la Ville de Montréal

par

Nicolas Dickner

Faculté des arts et des sciences

École de bibliothéconomie et des sciences de l'information

Mémoire présenté à l'École de bibliothéconomie et des sciences de l'information
en vue de l'obtention du grade de maîtrise
en sciences de l'information

avril 2017

© Nicolas Dickner, 2017

Résumé

Ce mémoire cherche à faire le point sur l'ouverture des données de l'administration publique, qui s'est généralisée depuis de 2009. Les données ouvertes s'inscrivent dans le mouvement du droit d'accès à l'information, mais se caractérisent par leur caractère proactif : plutôt que d'être diffusée à la demande, les données ouvertes sont divulguées en ligne, généralement regroupées sur un portail. L'ouverture des données vise plusieurs objectifs, dont notamment l'instauration d'un régime de transparence au sein de l'administration publique, et la stimulation de l'activité économique et de la participation citoyenne. Les applications des données ouvertes ont surtout été logicielles, mais nous avons repéré plusieurs sources qui démontrent le potentiel analytique du phénomène. Pour ce faire, les données doivent néanmoins répondre à plusieurs conditions : format, qualité et couverture appropriés, licence adéquate, etc. Nous avons examiné les politiques et pratiques sur deux sites québécois — *Données Québec* et le portail de données ouvertes de la ville de Montréal — afin de voir si ces conditions étaient respectées. Bien que la situation soit essentiellement convenable, nous avons noté certaines pratiques susceptibles de nuire à la réutilisation des données. Afin d'exposer ces problèmes et de proposer des stratégies pour les résoudre, nous avons procédé à des opérations de nettoyage et d'intégration de données. Nous expliquerons enfin l'intérêt analytique du croisement de plusieurs sources de données, en dépit des difficultés que présente cette approche.

Mots-clés : données ouvertes, administration publique, ouverture des données, sécurité routière, Montréal, Gouvernement du Québec

Abstract

The goal of this masters thesis is to assess the opening of public sector data, a phenomenon that became widespread since 2009. Open data stem from the freedom of information movement, with however a proactive dimension : rather than being provided on demand, open data are published online and usually centralized on a portal. Open data have several goals, in particular the promotion of transparency within the public sector, and the stimulation of both economic activity and civic participation. Open data have been mostly used to create software applications, but we found several sources that demonstrate the analytic potential of the phenomenon. However, to realize this potential, open data must comply with several conditions, such as appropriate format, quality and coverage, adequate user license, etc. We looked into the policies and practices of two Quebec portals — *Données Québec* and Montreal City open data portal — to see if these conditions were met. While the overall situation was acceptable, we noted some practices that could be detrimental to the reuse of data. In order to illustrate these problems and offer possible strategies to solve them, we performed data cleaning and integration. Finally, we explain the analytic gain of the data integration, despite the difficulties of the operation.

Keywords : open data, public sector, road traffic safety, Montreal, Quebec Government

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

API : Application programming interface

BAnQ : Bibliothèque et archives nationales du Québec

CKAN : Comprehensive Knowledge Archive Network

CSV : Comma Separated Values

DIP : dirigeant principal de l'information

ISQ : Institut de la statistique du Québec

JSON : Javascript Object Notation

KML : Keyhole Markup Language

OCDE : Organisation de coopération et de développement économiques

OKF : Open Knowledge Foundation

SAAQ : Société de l'assurance automobile du Québec

SPVM : Service de police de la ville de Montréal

SSSQ : Santé et services sociaux du Québec

UPAC : Unité permanente anticorruption

W3C : World Wide Web Consortium

XML : Extensible Markup Language

Je dédie ce mémoire

à

Ennio Morricone

Remerciements

Isabelle Bastien, pour m'avoir secouru avec ses excellentes archives personnelles.

Dominic Forest, pour ses conseils et recommandations.

Anne-Marie Provost, pour m'avoir orienté vers les données de la SAAQ.

Marie Wright-Laflamme, pour ses encouragements, ses conseils méthodologiques, et nos salutaires séances de kempo.

1. Introduction

Dans cette section, nous proposerons une définition des données ouvertes (1.1), un bref survol historique du phénomène (1.2) et le but de notre recherche (1.3).

1.1. Données ouvertes : une définition des termes

La distribution de données gouvernementales n'est pas un phénomène nouveau. Dans certains domaines — en géomatique, en météorologie et en sciences sociales, par exemple —, la diffusion de données à grande échelle remonte à plusieurs décennies. Le phénomène a cependant pris un essor particulier depuis 2009, notamment grâce aux initiatives de diverses administrations publiques américaines et britanniques, et l'on peut désormais parler d'un véritable enthousiasme pour les données ouvertes.

Par donnée ouverte, on entend une donnée qui peut être « librement utilisée, réutilisée et redistribuée par quiconque — sujette seulement, au plus, à une exigence d'attribution et de partage à l'identique » (Open Knowledge Foundation, 2012). Les données libérées par l'administration publique se sont considérablement diversifiées au cours des dernières années : salaires des élus, subventions accordées, géolocalisation de bâtiments ou de zones, résultats d'élection, pluviométrie, etc. Les données ouvertes peuvent, en somme, décrire toutes les dimensions d'une administration publique — sauf, naturellement, lorsque la sécurité publique, le respect de la vie privée ou les revenus de l'État sont en jeu.

L'ouverture des données inspire une comparaison naturelle avec les lois sur l'accès à l'information (voir section 1.2.1). Ces phénomènes partagent en effet les mêmes racines idéologiques, mais diffèrent sur deux aspects importants. D'une part, les données ouvertes sont divulguées de manière proactive, tandis que l'accès à l'information fonctionne à la demande. D'autre part, les termes *donnée* et *information* ne désignent pas la même chose : tandis que les données sont des éléments informationnels bruts tels que des chiffres ou des faits (Office québécois de la langue française, 2002), l'information résulte d'une interprétation de ces données (Gorunescu, 2011, p. 45; Hill, 2005; Kitchin, 2014; Rubin, 2010). Des mesures

de qualité de l'air, par exemple, constituent des données, cependant qu'un rapport faisant l'analyse de ces mesures relève de l'information.

Il en résulte que l'information s'adresse essentiellement à un lecteur humain, tandis que la lecture et la manipulation des données gagnent en général à être effectuées par le truchement de logiciels — à la condition, bien entendu, que ces données soient enregistrées dans des formats aisément manipulables par logiciel, une caractéristique que certains nomment *machine-friendliness* (Sunlight Foundation, 2010).

Ces deux aspects — le caractère brut des données et leur orientation machine — constituent les éléments cardinaux du mouvement des données ouvertes gouvernementales : ils permettent d'une part de présenter un portrait factuel des activités d'une administration publique, et d'autre part de faciliter la création d'applications dérivées à partir des ensembles de données. Nous nous pencherons sur ces questions aux sections 2.2 (« Objectifs des données ouvertes ») et 2.3 (« Utilisation des données ouvertes »).

Bien que ces considérations occupent une place importante au sein de la littérature, elles ne semblent pas toujours guider clairement les pratiques et politiques. Si au niveau fédéral, par exemple, on a séparé les « données ouvertes » et l'« information ouverte » (Gouvernement du Canada, s.d.-b), à la ville de Montréal on a en revanche décidé de considérer le terme *données* comme un « équivalent du terme *document* » (Groupe de travail sur les données ouvertes, 2011, p. 9)[1] — un choix qui se reflète forcément dans l'offre de données du portail (voir à ce sujet la section 3.2).

De manière générale, l'étiquette *donnée ouverte* nous apparaît plus ou moins fiable : de nombreux documents disponibles sur les portails de données ouvertes ne contiennent pas réellement de données (procès-verbaux, rapports ou photos d'archives), tandis que certains documents publiés par des institutions publiques comme l'Observatoire de la culture du

[1] La seconde mouture de cette politique repose sur la même idée, formulée de manière indirecte (Ville de Montréal, 2016d).

Québec ou le Directeur général des élections constituent des données ouvertes *de facto*, bien qu'elles ne soient pas diffusées sous ce nom.

Précisons enfin que les données ouvertes ne sont pas un phénomène exclusivement gouvernemental. De nombreux domaines et disciplines font une part grandissante aux données ouvertes, à commencer par les sciences pures où, depuis plusieurs décennies, on utilise la libération des données afin de stimuler la collaboration, de favoriser la diffusion des méthodes, d'améliorer la reproductibilité des expériences ou d'aider à l'identification des cas de fraude (Hernandez-Perez et Garcia-Moreno, 2013; Molloy, 2011; Reichman, Jones et Schildhauer, 2011). On pratique également l'ouverture des données au sein d'initiatives communautaires, tels le système de cartographie collaborative *Open Street Map*, ou la base de données *DBpedia* fondée sur l'extraction des données structurées de *Wikipédia*. Certains chercheurs prennent même en considération certaines données gérées par des entreprises privées comme Twitter (Kalampokis, Tambouris et Tarabanis, 2013).

Le paysage des données ouvertes est en somme fort complexe, et nous n'avons pas l'ambition de couvrir son entièreté dans le cadre de cette recherche. Par souci de concision, et en l'absence d'indication contraire, nous utiliserons l'expression *données ouvertes* afin de désigner les données ouvertes issues de l'administration publique, c'est-à-dire les ministères et institutions publiques, mais aussi les agences et organismes parapublics. Cette restriction peut sembler discutable, notamment parce qu'elle donne une connotation politique au phénomène, mais les différentes catégories de données ouvertes ne sont pas toujours perméables, et leurs contours peuvent se déplacer selon le contexte de diffusion, le public cible et le type d'utilisation envisagé[2].

Ces mises au point terminologiques étant complétées, nous présenterons maintenant un bref historique des données ouvertes.

---

[2] À titre d'exemple, le gouvernement du Canada diffuse des données climatologiques et géospatiales sous l'étiquette de données ouvertes gouvernementales, et non de données ouvertes scientifiques (Gray, 2014).

1.2. **Origine et évolution des données ouvertes**

Dans cette section il sera question de l'origine et de l'évolution des données ouvertes. Nous traiterons d'accès à l'information et de l'ouverture de données comme modèle de développement (section 1.2.1), de l'importance des réseaux pour les données ouvertes (1.2.2), et du rôle des portails (1.2.3).

1.2.1. **Accès à l'information et modèle de développement**

De nombreuses sources (Département des affaires économiques et sociales, 2013; Fioretti, 2010; Gray et Darbishire, 2011; Ubaldi, 2013) situent les données ouvertes dans la lignée du droit d'accès aux documents administratifs apparu dans de nombreux pays au cours du 20e siècle, et qui a donné naissance à plusieurs lois, telles que le *Freedom Information Act* (*The Freedom of Information Act*, 1966), la *Loi sur l'accès à l'information* (*Loi sur l'accès à l'information*, 1985) ou la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels* (*Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, 1982), ainsi que des amendements consécutifs à l'apparition du Web, comme l'*Electronic Freedom of Information Act Amendments of 1996* (*Electronic Freedom of Information Act Amendments of 1996*, 1996).

Ces lois réglementent l'accessibilité aux documents produits ou gérés par des organismes publics, en détaillant les modalités d'accès, mais aussi certaines exceptions qui relèvent, notamment, de la sécurité publique ou du respect de la vie privée. De manière générale, elles visent à établir un climat de transparence et d'imputabilité au sein des institutions démocratiques. De nombreux documents publiés sur les portails de données ouvertes reflètent de telles préoccupations politiques : liste des contrats octroyés par une administration, dépenses et rémunérations des élus, etc.

En dépit de ces similitudes, l'accès à l'information et l'ouverture des données diffèrent sur plusieurs aspects, en particulier le mode de diffusion. Tandis que l'accès à l'information consiste à rendre disponibles des documents sur demande, ponctuellement, les données

ouvertes se caractérisent par la divulgation proactive : elles sont publiées et mises à jour sans que les citoyens doivent en faire la requête[3], proactivité qui permet de joindre un plus vaste public que l'accès à l'information.

Il ne faut donc pas s'étonner que les données ouvertes soient devenues un élément rhétorique important dans le discours sur les transformations démocratiques. La nature exacte de ces transformations ne semble toutefois pas créer consensus.

À une extrémité du spectre, on tient un discours pro-institutionnel : les données ouvertes permettent de mettre en place une *vitrine électronique* (Misuraca, 2012, p. 25) afin d'améliorer et de diversifier les services aux citoyens (Kitchin, 2014). Selon cette approche, la libération des données ne constitue qu'un outil supplémentaire pour le gouvernement qui, tout en faisant preuve de transparence, accroît son domaine d'activités et sa sphère d'influence.

À l'autre extrémité, on considère la libération des données comme un des mécanismes menant à l'effacement de l'institution politique : le gouvernement, affirme-t-on, devrait se contenter de publier des données et laisser l'entreprise privée en tirer des produits et des services (Gray, 2014; Robinson, Yu, Zeller et Felten, 2009) — une idée parfois renforcée par la conviction que l'entreprise privée peut agir plus rapidement que le gouvernement (Commissariat général du plan, 1999, p. 59).

Entre ces deux pôles, on trouve une variété de positions plus ou moins neutres, plus ou moins ouvertes à l'intervention gouvernementale, mais où l'on estime à tout le moins que l'utilisation des données ouvertes ne devrait pas être monopolisée par l'administration publique — à plus forte raison si l'on considère que ces données ont été générées avec l'argent des contribuables (Garriga-Portola, 2011, p. 299). Il s'agit notamment du point de vue défendu par l'Open Knowledge Foundation, une organisation sans but lucratif qui fait la promotion

[3] Cela n'exclut pas, du reste, de consulter les usagers afin d'identifier les ensembles de données dont la publication est prioritaire (Ren et Glissmann, 2012). Ces consultations prennent généralement la forme passive d'un simple formulaire, mais peuvent à l'occasion être proactives: c'est le cas d'*Open Toronto*, qui a fait circuler un sondage à l'intention de ses usagers à l'été 2015.

d'une licence de réutilisation peu contraignante pour les utilisateurs (Open Knowledge Foundation, 2005).

En fin de compte, et contrairement au mouvement de l'accès à l'information, il est difficile de circonscrire les données ouvertes d'un point de vue politique. Le discours des partisans et des activistes du phénomène n'est pas homogène, et ne se reflète pas toujours concrètement dans les initiatives des diverses administrations publiques. Plusieurs rappellent que la libération des données ne s'accompagne pas forcément d'une gouvernance ouverte, et que certains gouvernements peuvent considérer les données ouvertes comme un enjeu d'ordre purement technique (Martin, Foulonneau, Turki et Ihadjadene, 2013, p. 350; Yu et Robinson, 2012). Pour paraphraser Richard Stallman, les données ouvertes — à tout le moins dans leur état actuel — représentent peut-être moins un phénomène politique qu'un modèle de développement (Stallman, 2010, p. 99). Cela tend à expliquer pourquoi les données ouvertes ont donné lieu à des hackathons (ou marathons de programmation) et au développement d'un écosystème de logiciels, souvent portés par des plateformes collaboratives comme GitHub (GitHub, s.d.).

Ce modèle de développement décentralisé est révélateur : si l'accès à l'information pouvait être pratiqué dans un monde dominé par les supports et protocoles traditionnels, les données ouvertes n'auraient aucun sens en dehors des réseaux informatiques actuels.

1.2.2. **Un phénomène de réseau**

Afin d'illustrer l'importance des réseaux dans le phénomène des données ouvertes, il est intéressant de rappeler l'expérience du bureau du maire de New York, qui publiait dès 1993 un catalogue des données municipales disponibles en formats numériques (Commission on Public Information and Communication, 1993). L'initiative s'apparentait beaucoup à l'ouverture des données actuelle, à un détail près : la plupart des bases de données répertoriées n'étaient pas accessibles en ligne, et les usagers devaient effectuer leurs requêtes en télécopiant une requête d'accès à l'information conventionnelle. La lourdeur et la lenteur de ce mécanisme d'accès neutralisaient pratiquement les bénéfices de la diffusion proactive.

Cette situation a changé quelques années plus tard, lorsque le Web s'est imposé comme plateforme de diffusion numérique globale. L'idée de mettre en ligne des données brutes ne s'est cependant pas manifestée dès les premières années : le Web n'était encore qu'un réseau au premier degré — ce qu'on a appelé, a posteriori, le *Web des documents*, ou *Web 1.0* —, où la diffusion se faisait par le truchement de pages statiques, écrites en HTML.

Le Web des documents a rapidement montré ses limites pour la diffusion des données : les premières versions du HTML, contrairement au XML, n'avaient pas été conçues dans une perspective d'interopérabilité, et encore moins afin de structurer leur contenu de manière sémantique. Les éléments d'une page Web étaient délimités par des balises parfois sémantiques, parfois stylistiques, et cette hétérogénéité compliquait la tâche — déjà fastidieuse — d'extraire les données sous une forme utilisable. Le Web des documents, en somme, n'était pas orienté machine.

La solution généralement adoptée, et qui prévaut encore aujourd'hui, a consisté à offrir le téléchargement des ensembles de données en formats natifs, c'est-à-dire dans les formats des logiciels ayant servi à créer les données. Cette approche règle les problèmes dont nous venons de parler. D'une part, elle rend inutile l'extraction des données par l'usager. D'autre part, elle permet une meilleure normalisation que le HTML : les documents en formats CSV, XML ou JSON respectent des règles formelles qui assurent la représentation correcte des données.

Ce compromis n'est toutefois pas sans inconvénient, le principal étant d'ordre fonctionnel : les documents en formats natifs ne peuvent pas tous être lus ou manipulés en ligne; certains doivent être téléchargés et interprétés en dehors du navigateur.

En outre, l'approche par document — qu'il s'agisse de documents en HTML ou en formats natifs — complique les mises à jour, non seulement du côté serveur, où il faut modifier ou téléverser des documents, mais aussi du côté client. La prolifération des téléphones intelligents a en effet donné naissance à un nouvel écosystème logiciel et à de nouveaux besoins, qui se caractérisent par une grande consommation de données, souvent en temps réel. Or, si les administrations publiques ont fait la promotion de leurs portails de données ouvertes auprès des développeurs, les efforts d'automatisation n'ont pas toujours été

conséquents : tous les ensembles de données ne sont pas générés dynamiquement et, pour l'essentiel, on recourt encore au téléversement des documents à la manière traditionnelle, avec les délais que cela suppose.

1.2.3. **Le portail : un outil de décloisonnement**

Bien que certaines entités gouvernementales diffusent leurs données ouvertes de manière indépendante, il s'agit d'initiatives exceptionnelles. De manière générale, l'ouverture des données a été accompagnée à peu près systématiquement par la mise en place de portails permettant la diffusion — voire l'entreposage — centralisée des documents.

S'il est possible que ces portails constituent en partie des dispositifs de relation publique, qui visent à magnifier l'importance et la visibilité des données ouvertes, on a vite vu le potentiel de telles plateformes pour favoriser la découvrabilité des données et briser la culture du cloisonnement (« *silo mentality* »).

Les analystes s'accordent pour affirmer que le cloisonnement est contre-productif (Dyson et Goldstein, 2013, p. 82) et qu'il s'agit d'une attitude obsolète héritée d'une époque révolue (Bauer et Kaltenböck, 2011, p. 22). Ce discours trouve un écho parmi les citoyens, et au sein de l'appareil administratif : en effet, lors de l'élaboration de la *Stratégie ville intelligente et numérique*, les participants des causeries citoyennes ont souligné l'importance de « briser les silos et les structures pyramidales et travailler en transversalité » (Bureau de la ville intelligente et numérique, 2014, p. 32). En outre, le rapport Gautrin, consacré au rôle du Web 2.0 dans l'amélioration des services aux citoyens, recommande au gouvernement de « se donner une stratégie des ressources humaines qui lui permettra de remplacer la culture des silos par une culture de collaboration et d'engagement » (Gautrin, 2012, p. 117), une idée reprise quelques années plus tard dans la *Stratégie gouvernementale en technologies de l'information* (Conseil du trésor, 2015, p. 39). On perçoit d'ailleurs l'influence de ce discours sur le portail Données Québec, dont l'un des objectifs vise à « simplifier […] le croisement des données » (Gouvernement du Québec, s.d.).

On conviendra sans peine que le rassemblement opéré grâce aux portails peut contribuer à transcender l'effet de vase clos. Paradoxalement, l'organisation de ces portails

obéit à des catégories thématiques ou organisationnelles qui peuvent perpétuer le cloisonnement. Nous avons également constaté la sous-utilisation de certaines métadonnées qui visent précisément à mettre en relation les données (voir section 3.2.3). En outre, l'intégration des données montre que des embûches de nature plus technique — telles que des incompatibilités formelles ou des valeurs erronées — peuvent nuire à l'utilisation transversale des données.

En résumé, bien que l'ouverture de données par l'administration publique ait progressé rapidement au cours des dix dernières années, l'analyse des pratiques suggère que le phénomène n'a pas atteint une phase de maturité, et plusieurs faiblesses importantes peuvent encore être observées sur le terrain. Dans la section suivante, nous examinerons deux de ces faiblesses qui ont particulièrement attiré notre attention, et nous exposerons nos objectifs de recherche.

1.3. **But de la recherche**

Dans cette section, nous expliquerons comment la qualité variable et l'offre parfois déficiente des données constituent des obstacles majeurs (section 1.3.1.), et nous présenterons les étapes de notre recherche (section 1.3.2).

1.3.1. **Offre de données et gestion de la qualité : des obstacles majeurs**

En dépit de l'enthousiasme que suscitent les données ouvertes, plusieurs auteurs ont souligné les nombreux obstacles qui empêchent encore le phénomène d'atteindre son plein potentiel (Braunschweig, Eberius, Thiele et Lehner, 2012; CEFRIO, 2016; Cole, 2012; M. Janssen, Charalabidis et Zuiderwijk, 2012; Zuiderwijk, Janssen, Choenni, Meijer et Alibaks, 2012; Zuiderwijk et al., 2012). Nous n'entendons pas dresser ici une liste exhaustive de ces obstacles, mais plutôt nous pencher sur deux situations souvent observées, et qui sont à l'origine de notre recherche : la pertinence et la qualité inégales des données.

Pertinence des données

Il va sans dire que les portails de données ouvertes ne diffusent qu'une partie des données générées par l'administration publique, et la sélection de ces données dépend avant tout de facteurs administratifs et légaux, tels que les brèches de sécurité potentielles, la résistance aux tâches supplémentaires, les lois sur le respect de la vie privée, et les négociations ou désaccords au sein d'une organisation (Cole, 2012; Martin, Foulonneau, Turki et Ihadjadene, 2013). L'accès à certaines sources de données peut également être restreint afin de préserver des revenus associés à leur exploitation (CEFRIO, 2016, p. 16).

Outre ces contraintes, les ressources financières imposent des limites pratiques : la gestion d'un portail de données ouvertes n'implique pas uniquement des coûts d'implantation, mais des frais systémiques de production, de traitement et de mise à jour des données (Bannister et Connolly, 2011; Cole, 2012). Or, si le coût marginal de la diffusion tend à diminuer, le coût marginal de manipulation des données demeure constant[4]. Les administrateurs doivent par conséquent prioriser les ensembles de données susceptibles de générer les meilleures retombées. Il s'agit cependant d'un processus complexe, qui suppose à la fois de définir la mission du portail de données ouvertes — incluant la nature des retombées attendues — et d'anticiper les besoins des utilisateurs. Nous verrons dans les chapitres 2 et 3 comment l'offre de données sur les portails de données ouvertes peut restreindre le champ des réutilisations possibles.

Qualité des données

La qualité inégale des données ouvertes a été plusieurs fois soulignée dans la littérature (Böhm et al., 2010; Hunnius, Krieger et Schuppan, 2014; Kuk et Davies, 2011; Pineau et Bacon, 2015; Ren et Glissmann, 2012; World Wide Web Foundation, 2015;

[4] L'automatisation de la collecte et de la mise en ligne des données pourrait diminuer le coût marginal de traitement (Bannister et Connolly, 2011, p. 23), en réduisant notamment le nombre d'heures-personnes que nécessitent ces opérations. L'automatisation n'est cependant pas encore répandues sur les portails québécois, où les pourcentages de documents mis à jour automatiquement oscilleraient entre 6% et 15% (CEFRIO, 2017, p. 28).

Zuiderwijk-van Eijk et Janssen, 2015). Parmi les causes possibles, mentionnons le manque d'uniformité causé par l'absence de politiques informationnelles ou de processus normalisés, les ressources insuffisantes, l'utilisation de logiciels inadéquats, ainsi que le manque de formation ou les réticences des intervenants (Martin et al., 2013).

La gestion de la qualité est également liée à la fréquence de mise à jour : une diffusion précipitée peut affecter la qualité des données, cependant qu'une diffusion trop tardive compromettra leur pertinence. Plusieurs documents légaux et autres *vade-mecum* placent la rapidité de diffusion parmi les nécessités de premier ordre (Commissariat général du plan, 1999; Executive Office of the President, 2013, p. 5; Open Government Working Group, 2007; Sunlight Foundation, 2010), et certains intervenants vont jusqu'à prioriser la vélocité plutôt que la qualité (Open Knowledge Foundation, 2012). Cette question a une influence sur les décisions de prétraitement à la source, comme nous le verrons à la section 2.1.4.

1.3.2. **Étapes de la recherche**

Notre recherche porte sur la pertinence et la qualité inégales des données ouvertes, lesquelles comptent, comme nous venons de le voir, parmi les principaux obstacles à l'utilisation des données ouvertes. Plus exactement, nous entendons mettre en lumière les stratégies qui peuvent être déployées par les utilisateurs afin de repérer, sélectionner et prétraiter des données ouvertes dans la perspective d'une analyse des données.

Nous étudierons d'abord la place qu'occupent ces obstacles dans la littérature, et leur prévalence dans le contexte québécois. Nous soulignerons en particulier la faiblesse de l'offre de données adaptées à des méthodes d'analyse complexes telles que la fouille de données. Puis, nous ferons la démonstration des méthodes nécessaires pour contourner ces obstacles : nous repérerons des ensembles de données présentant une complexité et un volume significatifs, et nous documenterons la chaîne de traitement permettant de convertir ces données brutes en un corpus dûment analysable. Cet exercice permettra non seulement d'illustrer le processus de prétraitement des données, mais de mettre en évidence certaines limites dans l'offre de données ouvertes au Québec, et de formuler des recommandations.

Afin d'atteindre ces buts, nous procéderons d'abord à une revue de la littérature afin de déterminer la nature et la fonction des données ouvertes (chapitre 2). Nous considérerons l'origine et la forme des données ouvertes (section 2.1), puis nous présenterons les principaux objectifs des initiatives de données ouvertes (section 2.2) et les types d'utilisations qui en découlent (section 2.3).

Dans un deuxième temps, nous proposerons un état de la situation des données ouvertes au Québec (chapitre 3). Nous décrirons en détail le contexte québécois des données ouvertes (section 3.1), l'offre de données actuelle (section 3.2), ainsi que les applications et retombées de ces données (sections 3.3 et 3.4). Cette section permettra de déterminer comment les aspects considérés dans le chapitre 2 — formes, objectifs et utilisations des données — se manifestent dans le contexte québécois. Nous identifierons non seulement ce qui a été accompli, mais également ce qui reste à accomplir, ainsi que les limites et lacunes actuelles.

Dans un troisième temps, enfin, nous procéderons au traitement des données (chapitre 4) : nous décrirons les ensembles de données ouvertes sélectionnés pour notre recherche (section 4.1), et nous documenterons les manipulations effectuées en vue d'obtenir deux corpus analysables, ainsi que les démarches permettant de croiser ces deux corpus afin d'offrir de nouvelles opportunités analytiques (section 4.2). De plus, comme les données proviendront de deux unités administratives relevant de deux paliers de gouvernement, cette intégration permettra de démontrer les défis techniques du décloisonnement dont il a été question à la section 1.2.3.

Nous conclurons notre recherche en formulant des recommandations basées sur nos manipulations et observations, afin de favoriser une meilleure offre de données ouvertes (chapitre 5).

2. Nature et fonction des données ouvertes

Afin de bien illustrer la complexité des données ouvertes, il importe de dresser un état de la question qui prenne en compte non seulement la perspective académique, mais également celle des praticiens. Pour ce faire, nous avons considéré une variété de sources incluant non seulement des textes académiques, mais également des documents prescriptifs

(lois, règlements, avis, manifestes), des articles journalistiques, ainsi que divers textes d'opinion et d'analyse[5].

Nous avons en premier lieu repéré de nombreuses sources permettant de définir les données ouvertes, le contexte de leur ouverture, et leur portée sur les plans politique, social et économique. Bien que les racines du phénomène puissent être retracées sur plusieurs décennies, nous avons jugé inutile de remonter au-delà de l'apparition du Web, au milieu des années 90.

Les premières sources pertinentes et faisant autorité se composent essentiellement de documents légaux et de *vade-mecum*, qui permettent de formaliser et d'encadrer des pratiques naissantes. C'est le cas en particulier de l'*Electronic Freedom of Information Act Amendments* of 1996 (*Electronic Freedom of Information Act Amendments of 1996*, 1996) qui, inspiré par l'apparition du Web, modernise les lois d'accès à l'information en stipulant que les agences gouvernementales devront désormais rendre leurs informations disponibles en format numérique. Bien que l'amendement ne concerne que les demandes d'accès à l'information, il pose néanmoins des bases légales qui, au tournant du 21^{e} siècle, permettront l'apparition des données ouvertes.

En Europe, le *Plan d'action gouvernemental pour la société de l'information* (Commissariat général du plan, 1999) et le *Performance and Innovation Unit* (Performance and Innovation Unit, 2000) témoignent d'une réflexion bien entamée au sein des administrations nationales. On entrevoit cependant une autre réalité que l'unique transparence : alors que l'amendement de la constitution américaine de 1996 met l'accent sur l'identification de fraudes ou de failles de sécurité, la *Directive 2003/98/CE du Parlement*

[5] Malgré cette approche à large spectre, notre revue de la littérature a permis de constater que certaines approches demeurent négligées. Certains auteurs font notamment remarquer que le discours académique est très conceptuel (Hivon et Titah, 2015, p. 14), et une revue de 143 articles effectuée en 2012 a essentiellement révélé des « articles conceptuels et des descriptions de l'utilisation des données ouvertes ou des technologies employés », cependant que le recours à la théorie afin d'analyser des tendances ou des patrons recevait « considérablement moins d'attention » (M. Janssen, Charalabidis et Zuiderwijk, 2012)

européen s'attarde davantage aux bénéfices économiques de l'information (Parlement européen et Conseil de l'Union européenne, 2003).

L'émergence du phénomène ne se note pas uniquement dans les documents législatifs: en 2006, le *Guardian*, de concert avec Tim Berners-Lee, fait campagne pour que soient libérées les données géospatiales de l'Ordnance Survey (Cross et Mathieson, 2006). Quelque mois plus tard, l'Open Knowledge Foundation (OKF) diffuse la première mouture de l'*Open Definition*, en cours de préparation depuis un an (Open Knowledge Foundation, 2005). Ce texte définit assez largement ce que sont les données ouvertes, et s'attarde surtout aux conditions d'ouverture et aux meilleures pratiques, en posant la licence d'utilisation comme prérequis *sine qua non*.

L'année suivante, en Californie, des gens issus du milieu des affaires et de la société civile font le point sur le thème du gouvernement ouvert. Malgré le mandat assez vaste de cette rencontre, il en ressort un document à la portée très spécifique : *The 8 Principles of Open Government Data* (Open Government Working Group, 2007). On y déclare que les données ouvertes gouvernementales doivent être intégralement libérées dans des délais rapides, telles que collectées à la source, et dans un format orienté machine et non propriétaire. Le processus de libération doit viser l'accessibilité maximale et permettre la plus grande palette possible d'utilisations. Enfin, les données doivent être libérées sous une licence aussi permissive que possible. Ce document aux accents de manifeste deviendra une référence importante dans de nombreuses sources.

Puis, au début de 2009 paraît aux États-Unis ce que plusieurs considèrent comme un jalon dans le mouvement des données ouvertes : le mémorandum présidentiel sur la transparence et le gouvernement ouvert (Executive Office of the President, 2009). Ce document, publié peu après l'élection de Barack Obama, déclenche un important mouvement d'ouverture des données aux États-Unis et donnera une visibilité sans précédent au phénomène. Le mémorandum mentionne trois axes — 1) transparence, 2) participation des citoyens et 3) collaboration entre agences gouvernementales, secteur privé et société civile — que l'on retrouvera par la suite dans de nombreuses définitions des données ouvertes. Ces concepts n'étaient pas entièrement nouveaux, mais leur endossement par l'administration américaine leur donne une portée plus formelle.

Quelques mois plus tard, un groupe de travail du W3C — organisme dont le rôle de prescripteur est notoire — propose de bonnes pratiques pour publier des données ouvertes gouvernementales (Bennett et Harvey, 2009). Ces lignes directrices abordent les questions habituelles (sélection et formatage des données, normalisation, documentation, licence, etc.) et les mettent en relation avec les normes développées au cours des années précédentes en matière de balisage sémantique (XML) et de données liées (RDF).

À partir de 2009, on note une spécialisation et une multiplication des textes académiques (Zuiderwijk, Helbig, Gil-Garcia et Janssen, 2014). Les premiers portails de données ouvertes étant désormais en service — notamment ceux des États-Unis et du Royaume-Uni —, on voit apparaître des études de cas et des bilans d'initiatives. Un nombre croissant de sources posent également un regard critique sur l'ouverture des données : passé l'enthousiasme initial les observateurs découvrent en effet que la mise en œuvre des programmes est souvent déficiente, ou que les retombées ne sont pas forcément aussi importantes ou positives qu'on l'avait anticipé.

2.1. **Origine et forme des données ouvertes**

Dans les sections qui suivent, nous procéderons à un état de la littérature et des pratiques afin de décrire le phénomène depuis la création et la sélection des données jusqu'à leur utilisation finale. Pour ce faire, nous tâcherons de répondre aux trois questions suivantes : *que sont les données ouvertes?* (section 2.1), *quels sont les objectifs des données ouvertes ?* (section 2.2) et *comment atteint-on ces objectifs ?* (section 2.3).

2.1.1. **Formats des données**

Les données libérées par les administrations publiques sont notoirement hétérogènes, et cette diversité se manifeste dans les nombreux formats des documents diffusés. Nous ne procéderons pas ici à un inventaire détaillé des pratiques, mais plutôt à une présentation des différents formats recommandés.

Ces formats répondent à une exigence centrale : ils doivent être manipulables par une grande variété de logiciels et de systèmes d'exploitation, afin de favoriser l'accessibilité et la

réutilisation. Les formats ouverts et normalisés sont privilégiés, car leur documentation est aisément accessible et ils ne limitent pas les usagers à un écosystème logiciel spécifique. Dans le même ordre d'idée, plusieurs observateurs ont souligné le rôle que devait jouer le logiciel libre dans les données ouvertes (Miller, Styles et Heath, 2008; Open Government Working Group, 2007; Open Knowledge Foundation, 2005; Parlement européen et Conseil de l'Union européenne, 2003). Malgré ces recommandations, certains formats propriétaires occupent une place importante. C'est le cas en particulier pour les données géospatiales, qui demeurent en grande partie diffusées dans le format Shapefile développé par la compagnie Esri, et considéré comme une norme *de facto* (Library of Congress, 2011).

Mais l'exigence formelle la plus importante consiste sans doute à publier les données ouvertes dans des formats orientés machine (Braunschweig et al., 2012; Obama, 2013; Open Government Working Group, 2007). On entend par là des formats structurés « de telle manière que des applications logicielles puissent facilement identifier, reconnaître et extraire des données spécifiques, notamment chaque énoncé d'un fait et sa structure interne » (Parlement européen et Conseil de l'Union européenne, 2003, p. 10).

Dans la pratique, cela signifie généralement d'écarter les fichiers binaires — c'est-à-dire tous les documents qui ne sont pas constitués de texte ASCII (images matricielles, documents vidéo, applications, etc.) — et de favoriser plutôt les documents textuels, qui sont plus portables et se prêtent mieux au développement logiciel (Braunschweig et al., 2012; Ordnance Survey, s.d.). En règle générale, les documents binaires se prêtent mal à l'analyse automatisée du contenu; il est difficile, par exemple, d'extraire automatiquement des données spécifiques d'un tableau en format PDF ou JPEG. Certains documents binaires ou composites conçus pour le transfert de données, comme les feuilles de calcul (Excel, ODS), peuvent être aisément convertis en formats orientés-machine à la condition de n'avoir pas été mis en page à l'intention d'un lecteur humain, en utilisant notamment des cellules fusionnées.

Nous présenterons maintenant les trois formats textuels les plus répandus sur les portails de données ouvertes : CSV, XML et JSON.

Un document CSV (Comma Separated Values) est l'équivalent en format texte d'une feuille de calcul : chaque ligne y représente un enregistrement, dont les champs (ou colonnes)

sont séparés par un caractère désigné, en général la virgule. La première ligne contient les en-têtes de chacun de ces champs. Le CSV constitue donc un fichier à plat, et la représentation de données hiérarchisées se traduira donc par des valeurs redondantes (voir figure 1).

```
Nom,Prénom,NoTéléphone
Wozniak,Steve,(888)888-8888
Jobs,Steve,(123)000-1234
Jobs,Steve,(000)123-1234
```

Figure 1. Échantillon de format CSV

Le XML (Extensible Markup Language) est un langage de balisage sémantique dérivé du SGML, qui se prête bien à la représentation de données hiérarchisées (voir figure 2). Les balises du XML — contrairement à celles du HTML — ne sont pas prédéterminées : l'usager peut concevoir des balises *ad hoc*. Si nécessaire, les règles de validation de ces balises (contenu, position, format, etc.) peuvent être inscrites dans un document annexe : la DTD (Document Type Definition). Plusieurs implémentations normalisées du XML sont utilisées en données ouvertes, notamment les documents RDF servants aux données liées, et les documents géospatiaux KML popularisés par Google Earth.

```
<?xml version="1.0" encoding="UTF-8"?>
<carnet_adresses>
 <personne prenom="Steve" nom="Wozniak">
      <numero>(888)888-8888</numero>
 </personne>
 <personne prenom="Steve" nom="Jobs">
      <numero>(123)000-1234</numero>
      <numero>(000)123-1234</numero>
 </personne>
</carnet_adresses>
```

Figure 2. Échantillon de XML

Le JSON (JavaScript Object Notation) est un format de données basé sur des paires attribut/valeur, utilisant la syntaxe des tableaux et objets de JavaScript (voir figure 3). Le

JSON permet d'offrir une représentation hiérarchisée des données, ce qui en fait un concurrent populaire du XML. Parmi les implémentations normalisées de JSON, signalons GeoJSON, un format de document géospatial utilisé par plusieurs portails de données ouvertes.

```
{
  "carnet_adresses": {
    "personnes": [
      {
        "prenom": "Steve",
        "nom": "Steve",
        "numero": [
          "(123)000-1234",
          "(000)123-1234"
        ]
      },
      {
        "prenom": "Steve",
        "nom": "Wozniak",
        "numero": [
          "(888)888-8888"
      ]
      }
    ]
        }
      }
```

Figure 3. Échantillon de JSON

Enfin, nous ne pouvons parler de formats de données sans aborder la question des données liées. Ce concept consiste, en quelque sorte, à utiliser le Web à la manière d'une base de données dont le contenu serait accessible par le truchement d'URI — une adresse permanente désignant la ressource —, ce qui permet notamment de mettre les données en relation de manière automatisée. Les données liées ont été popularisées par Tim Berners-Lee (Berners-Lee, 1998, 2006, 2009; Bizer, Heath et Berners-Lee, 2009), pour qui elles représentent le mécanisme idéal du Web des données, ou Web sémantique. La structure des données ouvertes est décrite par le modèle RDF (Resource Description Framework), qui peut être implémenté dans n'importe quel format, le XML étant une option répandue.

Les données liées ont connu un développement relativement rapide, au cours des dernières années. Parmi les projets phares, on trouve DBpedia, une initiative communautaire

qui consiste à extraire des données structurées de Wikipédia. Des entreprises comme Google ont également adopté les données liées (Steiner, Troncy et Hausenblas, 2010), et un groupe de travail du W3C a développé le DCAT (Data Catalog Vocabulary), une spécification basée sur RDF qui permet d'assurer l'interopérabilité entre des catalogues en ligne (W3C, 2014).

Malgré ces initiatives, les données liées demeurent peu utilisées par les administrations gouvernementales (World Wide Web Foundation, 2015, p. 21). Le portail de données ouvertes du Royaume-Uni a été un précurseur dans le domaine (Shadbolt et al., 2012), mais à ce jour, seule une fraction des données ouvertes britanniques sont disponibles en RDF. Il s'agit en somme d'un phénomène qui n'a pas encore donné sa pleine mesure, mais qu'il convient de surveiller.

2.1.2. **Sélection des données**

Comme nous l'avons mentionné en introduction, aucune administration publique ne peut libérer la totalité de ses données : de nombreux facteurs légaux, administratifs et budgétaires restreignent la publication de certains documents (voir section 1.3.1). Qui plus est, les administrations publiques pourront hésiter à ouvrir certains ensembles de données dont l'accès est traditionnellement tarifé, afin de ne pas subir de pertes de revenus (CEFRIO, 2016; Plamondon Émond, 2016)[6]. Le défi de l'ouverture des données consiste donc à identifier et prioriser les ensembles susceptibles d'entraîner le meilleur retour sur investissement. Il s'agit d'un exercice par définition difficile, comme nous le verrons dans les sections consacrées à l'évaluation des retombées dans la littérature (2.2.4) et à l'examen du contexte québécois (3.2.5).

---

[6] L'*Ordnance Survey* constitue à cet égard un cas de figure : tenue par l'administration britannique de générer des revenus au même titre qu'une entreprise privée, l'agence de cartographie refusait d'offrir gratuitement ses données aux citoyens. La situation a inspiré le *Guardian* à lancer la campagne *Free Our Data*, qui a joué un rôle important dans l'ouverture des données en Grande-Bretagne (Cross et Mathieson, 2006).

La pertinence d'effectuer une sélection ne fait cependant pas consensus, et certains observateurs plaident plutôt pour la publication de la totalité des données gouvernementales (Open Government Working Group, 2007). Cet objectif peut sembler improbable sur les plans pratique, juridique et financier; il existerait pourtant des bénéfices économiques et logistiques chiffrables à cette culture de l'ouverture (Tait, 2011), et un nombre croissant d'administrations adoptent (ou envisagent d'adopter) des politiques d'ouverture par défaut des données. C'est notamment le cas des gouvernements du Canada (Gouvernement du Canada, s.d.-a) et de l'Ontario (Gouvernement de l'Ontario, 2015), et de la ville de Montréal (Ville de Montréal, 2016d)[7].

Si cette approche permet de faire l'économie du processus de sélection, elle pose en revanche plusieurs défis sur les plans de la sécurité et du droit à la vie privée. Plusieurs sources soulignent notamment les dangers de l'effet mosaïque, cette situation où des documents sans relations apparentes deviennent des sources d'information sensible une fois combinés (Executive Office of the President, 2013; M. Janssen et al., 2012; Lavoie, 2014; Zarsky, 2011). Caviarder les documents ne suffit pas toujours : les mesures d'anonymisation minimisent les risques, mais l'application de technique de fouilles ou de rétro-ingénierie peut permettre de désanonymiser des données ou de créer des connaissances inédites à partir de sources en apparence anodines (Bannister et Connolly, 2011; Czajka, Schneider, Sukasih et Collins, 2014; Hern, 2014; M. Janssen et Hoven, 2015; Kulk et Van Loenen, 2012; Srihari et Voeller, 2008).

Il va sans dire qu'une métropole ou un État qui se livreraient à une réelle ouverture par défaut publieraient des quantités substantielles de données. L'exemple du très volumineux portail du gouvernement ouvert du Canada, qui regroupe près de 118 000 ensembles de données, illustre l'importance de ne pas uniquement faciliter la recherche des données, mais

[7] Notons cependant qu'en l'absence de répertoires de données exhaustifs, il demeurera impossible de déterminer avec certitude si ces administrations pratiquent réellement une ouverture par défaut. Nous traiterons de cette question dans la section 2.2.1, consacrée à la transparence.

aussi leur découvrabilité. La prochaine section portera sur deux facteurs essentiels de cette découvrabilité : la description et l'organisation des documents.

2.1.3. **Métadonnées et ensembles de données**

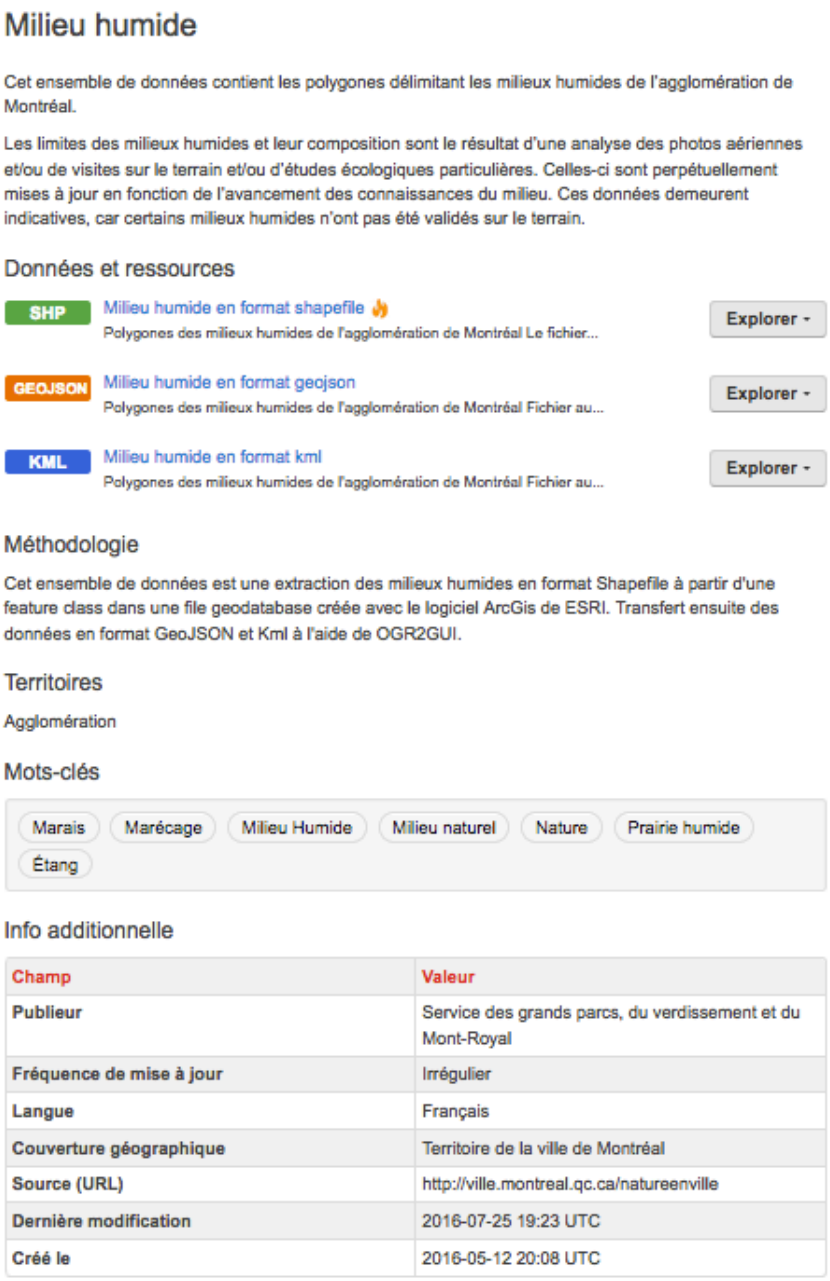
La publication d'un grand volume de données ouvertes implique des stratégies d'organisation dont la plus importante est assurément le regroupement des documents. Ces ensembles peuvent inclure des données disponibles sous plusieurs formats ou divisées en plusieurs segments (par territoires, périodes, unités administratives, etc.), ainsi que des documents complémentaires (dictionnaire de données, etc.).

Bien que la plupart des portails de données ouvertes pratiquent ce type de regroupement, il n'existe pas de norme de classement établie. Notre analyse de deux portails québécois (section 3.2) suggère même qu'il est difficile d'adopter et d'appliquer des règles de regroupement cohérentes au sein d'une même administration.

La constitution de ces ensembles est essentielle, puisqu'elle permet aux utilisateurs de repérer efficacement la totalité des données relatives à un même phénomène ou une même réalité. Le regroupement de nombreux documents au sein d'ensembles très généraux peut cependant nuire au repérage de données spécifiques, et doit donc s'accompagner d'une description rigoureuse. La littérature témoigne d'ailleurs de l'importance de mettre en place des métadonnées adéquates (Dawes et Helbig, 2010; Ubaldi, 2013; Whitmore, 2014; Zuiderwijk et Janssen, 2014), qui contribuent à l'organisation, à la recherche et la découvrabilité des ressources (Braunschweig et al., 2012). Les métadonnées permettent également d'évaluer rapidement les caractéristiques d'un ensemble de données, voire sa fiabilité. À titre d'exemple, un document censé être mis à jour hebdomadairement, mais dont la dernière publication remonterait à plusieurs mois pourrait être considéré d'une valeur discutable.

Malgré de légères variations entre les portails, on y retrouve les mêmes métadonnées de base : description, service ou individu responsable des données, couverture temporelle et/ou territoriale, fréquence de publication et date de dernière mise à jour, mots-clés, catégories et permaliens des ressources, format des documents (voir figure 4). Cette relative uniformité peut

s'expliquer par l'important taux d'adoption de CKAN, la plateforme de gestion des données ouvertes développée par l'Open Knowledge Foundation (Open Knowledge Foundation, s.d.-a).



Milieu humide

Cet ensemble de données contient les polygones délimitant les milieux humides de l'agglomération de Montréal.

Les limites des milieux humides et leur composition sont le résultat d'une analyse des photos aériennes et/ou de visites sur le terrain et/ou d'études écologiques particulières. Celles-ci sont perpétuellement mises à jour en fonction de l'avancement des connaissances du milieu. Ces données demeurent indicatives, car certains milieux humides n'ont pas été validés sur le terrain.

Données et ressources

SHP Milieu humide en format shapefile
Polygones des milieux humides de l'agglomération de Montréal Le fichier... Explorer

GEOJSON Milieu humide en format geojson
Polygones des milieux humides de l'agglomération de Montréal Fichier au... Explorer

KML Milieu humide en format kml
Polygones des milieux humides de l'agglomération de Montréal Fichier au... Explorer

Méthodologie

Cet ensemble de données est une extraction des milieux humides en format Shapefile à partir d'une feature class dans une file geodatabase créée avec le logiciel ArcGis de ESRI. Transfert ensuite des données en format GeoJSON et Kml à l'aide de OGR2GUI.

Territoires

Agglomération

Mots-clés

Marais | Marécage | Milieu Humide | Milieu naturel | Nature | Prairie humide | Étang

Info additionnelle

Champ	Valeur
Publieur	Service des grands parcs, du verdissement et du Mont-Royal
Fréquence de mise à jour	Irrégulier
Langue	Français
Couverture géographique	Territoire de la ville de Montréal
Source (URL)	http://ville.montreal.qc.ca/natureenville
Dernière modification	2016-07-25 19:23 UTC
Créé le	2016-05-12 20:08 UTC

Figure 4. Exemple de métadonnées (Ville de Montréal, 2016c)

2.1.4. **Prétraitement**

Une description formelle des données ouvertes doit prendre en compte la question du prétraitement, qui suppose une transformation directe des données. En effet, la nature des données ouvertes dépendra en grande partie des altérations que l'on aura fait subir ou non aux données, et de l'étape à laquelle interviendront ces altérations.

Le terme prétraitement ne désigne pas une opération unique, mais plutôt un ensemble de manipulations qui varient en fonction de l'utilisation envisagée. Pour les besoins de notre recherche, nous distinguerons ici trois catégories de prétraitements, qui visent à :

1. rendre les données légalement publiables;
2. améliorer la qualité formelle des données;
3. formater les données pour une méthode d'analyse particulière.

Dans cette section — et lors des manipulations du chapitre 4 —, nous nous intéresserons essentiellement aux deux premières catégories d'opérations, qui jouent un rôle crucial dans le cadre des données ouvertes.

Rendre les données légalement publiables

La première catégorie de prétraitements répond à des impératifs tels que la protection de la vie privée ou de la sécurité publique, comme nous l'avons vu dans la section 2.1.2. Certaines opérations de prétraitement visent donc à exercer une rétention ciblée afin de permettre la publication des ensembles de données sensibles. Cette rétention peut se faire par soustraction — en retirant certains champs d'un document, par exemple —, mais aussi par substitution.

Le recours à la substitution plutôt qu'à la soustraction pure et simple est un compromis intéressant, qui vise à rendre les données légales ou sécuritaires tout en préservant leur potentiel analytique. À titre d'exemple, les données sur les introductions par effraction publiées par le SPVM (Ville de Montréal, 2016a) ont été anonymisées en remplaçant chaque adresse par les coordonnées de l'intersection la plus proche, et l'heure exacte par le quart de travail. Ce prétraitement empêche — en théorie du moins — l'identification de résidences

spécifiques, mais permet tout de même de procéder à une analyse temporelle ou géospatiale des données.

Améliorer la qualité formelle des données

La seconde catégorie de prétraitements consiste à nettoyer les données, c'est-à-dire à assurer un niveau de qualité adéquat pour le traitement automatisé. Ce type d'opérations est particulièrement important dans le cadre des données ouvertes gouvernementales, où l'on a souvent constaté une qualité inégale des ensembles de données (voir section 1.3.1).

Ce type de prétraitement a fait l'objet d'une attention particulière bien avant l'apparition des données ouvertes : en 2000, Erhard Rahm et Hon Hai Do ont dressé un inventaire exhaustif des différentes opérations de nettoyage possibles dans le cadre de la constitution d'un entrepôt de données (Rahm et Do, 2000), dans le but de corriger une grande variété d'erreurs, dont :

- les valeurs manquantes, non documentées ou invalides;
- les valeurs multiples saisies dans un seul champ, ou les valeurs saisies dans le mauvais champ;
- les enregistrements dupliqués;
- les erreurs de référencement;
- les erreurs factuelles ou les coquilles.

Deux facteurs, selon les auteurs, exercent une grande influence sur les opérations de nettoyage : la présence ou non de règles structurelles, et l'intégration de plusieurs sources de données.

Par *règles structurelles*, nous entendons les formats de documents comportant un schéma de données (XML avec DTD, par exemple) ou les bases de données. Dans ces deux cas, les données doivent obéir à des règles de validation qui assurent *de facto* une meilleure qualité — pourvu que ces règles soient suffisamment restrictives — et font en sorte que les erreurs se trouvent uniquement au niveau des instances elles-mêmes : coquilles, enregistrements dupliqués, erreurs factuelles, etc. Cela explique en partie la qualité inégale des

données ouvertes, qui se composent essentiellement de documents sans schéma : CSV, feuilles de calcul, XML sans DTD, etc[8].

L'intégration de plusieurs sources de données pose des problèmes particuliers, puisqu'elle peut générer des erreurs aussi bien structurelles qu'au niveau des instances : duplication ou conflit d'instances, incompatibilité de formats de champs ou de données, etc. L'un des principaux problèmes, selon Rahm et Do, consiste à gérer les instances qui sont à la fois partiellement redondantes et susceptibles de se compléter.

Une telle situation peut se produire, dans le contexte des données ouvertes, du fait que les subdivisions d'un ensemble de données — les années ou les arrondissements, par exemple — sont souvent publiées dans des documents distincts. L'usager qui voudra regrouper ces données devra possiblement composer avec des incohérences entre ces différentes sources, certaines de ces incohérences pouvant être résolues de manière automatisée, d'autres nécessitant une intervention manuelle. Quoi qu'il en soit, il importe de procéder à une analyse préliminaire pour identifier et résoudre ces problèmes.

Nous proposerons, au chapitre 4, un cas d'analyse qui illustre les deux facteurs mis en évidence par Rahm et Do : le nettoyage de documents de données sans schéma, et l'intégration de plusieurs sources de données.

Nettoyage à la source : pour et contre

Il existe une importante différence entre les deux catégories de prétraitement : tandis que les opérations comme l'anonymisation sont par définition effectuées à la source — c'est-à-dire par l'organisation responsable de la diffusion —, le nettoyage des données peut se faire aussi bien à la source que lors de l'utilisation. La revue de la littérature révèle des positions contrastées sur cette question.

---

[8] Les documents générés par les systèmes d'information géographique (SIG), très nombreux sur les portails de données ouvertes, sont susceptibles de présenter peu d'erreurs. Ils peuvent néanmoins nécessiter d'autres types de prétraitement, comme nous le verrons à la section 4.2.3.

En effet, certains auteurs craignent que la neutralité, l'intégrité ou la fiabilité des données ne soient compromises par un nettoyage à la source (Braunschweig et al., 2012; Ceolin et al., 2013, 2014). Des données nettoyées, en somme, ne seraient plus des données brutes.

Plusieurs sources évoquent également les éventuels délais de diffusion que pourraient causer les opérations de nettoyage. L'OKF exprime sans doute la position la plus catégorique à ce sujet, lorsqu'elle affirme qu'il est préférable de « publier des données brutes maintenant que des données parfaites dans six mois » (Open Knowledge Foundation, 2012)[9]. Cette position repose essentiellement sur la péremption de l'information : certaines données perdent en valeur lorsqu'elles tardent à être diffusées (Open Government Working Group, 2007), et cette dévaluation peut être fulgurante : à titre d'exemple, un délai de quelques heures pourrait réduire les possibilités d'utilisation de données sur la qualité de l'air (Ochando, Julián, Ochando et Ferri, s.d.).

Enfin, les coûts du prétraitement peuvent s'avérer considérables pour une administration qui libérerait d'importants volumes de données à grande fréquence. Externaliser le nettoyage des données aux utilisateurs ou aux organisations civiles (*crowdsourcing*) constituerait une technique de contournement du problème (Maguire, 2011, p. 523). Une administration pourrait même soutenir que cette externalisation est compatible avec la stimulation de la participation citoyenne, qui constitue un des objectifs cardinaux des données ouvertes (voir à ce sujet la section 2.2.3).

Ces arguments ne font pas l'unanimité, cependant, et certains observateurs préconisent plutôt le prétraitement des données ouvertes à la source. Une étude commandée par l'OCDE inclut notamment le prétraitement dans la chaîne de valeur des données ouvertes gouvernementales (Ubaldi, 2013).

---

[9] Cette conception des données ouvertes diverge considérablement de l'entrepôt de données que décrivent Rham et Do, où le nettoyage s'inscrit dans une opération appelée Extract-Tranform-Load (ETL) au terme duquel les données nettoyées remplacent les données brutes.

Bien que notre recherche présente un exemple de nettoyage de données par l'utilisateur (voir chapitre 4), il ne s'agit pas d'une recommandation de notre part : la multiplicité des facteurs en jeu — qu'ils soient techniques, politiques ou financiers — empêche de se prononcer catégoriquement sur ce qui constitue la meilleure pratique entre le prétraitement à la source ou le prétraitement lors de l'utilisation.

Par ailleurs, il est possible qu'une réduction du prétraitement à la source soit compatible avec les objectifs de nature politique (voir section 2.2.1), tandis que la publication de données déjà nettoyées favorise plutôt les objectifs de nature économiques (voir section 2.2.2). Dans la prochaine section, nous examinerons non seulement ces différents types d'objectifs, mais également l'évaluation de l'atteinte de ces objectifs.

2.2. **Objectifs des données ouvertes**

L'ouverture des données ouvertes répond à des mandats très variés, dont la portée et les objectifs finaux s'avèrent parfois imprédictibles. Cela explique la grande diversité de points de vue qui règne chez les partisans et les détracteurs de l'ouverture des données.

Bien que les sources consultées permettent d'identifier un certain nombre d'idées récurrentes, nous avons remarqué que les objectifs des données ouvertes demeuraient souvent ambigus[10]. Les différentes sources décrivent ces concepts en utilisant une terminologie changeante, où l'on perçoit parfois une certaine confusion entre le mandat d'un programme de données ouvertes, et les objectifs des données ouvertes. La stimulation de l'innovation, par exemple, est un mandat dont l'un des objectifs — ou effets souhaités — serait la création de richesses au sein de la société. Les deux concepts ont des liens de causalité, mais ne constituent pas pour autant une seule et même chose : une entreprise pourrait être innovante et

[10] Certains auteurs ont tenté d'inventorier et de catégoriser ces objectifs (Gray, 2014; M. Janssen et al., 2012). Janssen, Charalabidis et Zuiderwijk qui dressent par ailleurs un inventaire assez exhaustif, où les bénéfices des données ouvertes sont regroupés en trois catégories: 1) politiques et sociaux 2) économiques 3) opérationnels et techniques. Il s'agit cependant d'un exercice difficile, et on peut repérer des chevauchements dans ces catégories.

néanmoins déficitaire. Cette nuance explique certaines dissonances dans la littérature. À titre d'exemple, la plupart des observateurs présentent la transparence comme une manière d'assurer l'imputabilité des élus et des fonctionnaires, tandis que d'autres y voient un droit du citoyen, une fin en soi (Zuiderwijk et Janssen, 2014).

Afin de ne pas alourdir inutilement notre propos, nous utiliserons ici le terme *objectif*, tout en gardant à l'esprit qu'un objectif peut contenir une arborescence de sous-objectifs. Cette précision semble anodine, mais elle permet d'expliquer comment l'ouverture des données peut à l'occasion produire des effets indésirables (voir section 2.2.4).

Mentionnons également que certaines sources mentionnent des notions relativement vagues, susceptibles de désigner plusieurs objectifs à la fois. C'est le cas en particulier de l'efficacité, qui peut s'appliquer à la délivrance des services publics, à l'administration publique en général (European Commission, 2011; Janssen et al., 2012), à l'échange d'information entre les organisations gouvernementales (Executive Office of the President, 2009; K. Janssen, 2012; Open Knowledge Foundation, 2012; Ville de Montréal, 2016d) ou à l'évaluation des politiques publiques (Margetts, 2011; Open Knowledge Foundation, s.d.-c).

Nous nous contenterons ici de mettre l'emphase sur les trois objectifs les plus fréquemment invoqués dans les sources consultées (voir par exemple Département des affaires économiques et sociales, 2013; Executive Office of the President, 2009; Hellberg et Hedström, 2015; Open Knowledge Foundation, s.d.-c; Ren et Glissmann, 2012; Secrétariat du Conseil du Trésor du Canada, 2014; Ubaldi, 2013; Zuiderwijk et Janssen, 2014) :

- gouverner avec transparence (section 2.2.1);
- stimuler l'activité économique (section 2.2.2);
- stimuler la participation citoyenne (section 2.2.3).

Pour chacun de ces trois objectifs, nous présenterons une synthèse du discours et nous examinerons comment il s'incarne — ou non — dans la réalité. Nous conclurons avec un bref survol de la question des retombées, qui couvre dans une certaine mesure la difficile question des effets recherchés.

2.2.1. **Gouverner avec transparence**

Bien que sa publication soit relativement tardive par rapport aux politiques et initiatives européennes, le mémorandum présidentiel signé par Barack Obama en février 2009 (Executive Office of the President, 2009) est souvent mentionné parmi les événements fondateurs de l'ouverture des données. Ce document couvre certes plusieurs aspects de la question — dont la participation citoyenne, et la collaboration entre le gouvernement et la société civile —, mais il accorde une importance centrale à la question de la transparence.

Ce choix est représentatif de la littérature : de nombreuses sources rappellent en effet que l'ouverture des données doit notamment permettre d'assurer la confiance des citoyens et l'imputabilité des élus.

Bien que ces objectifs semblent reposer sur le sens commun, plusieurs observateurs apportent néanmoins des nuances : la libération des données ne garantirait pas forcément une administration plus transparente ou imputable, comme en témoignent les cas de Singapour (évoqué dans Margetts 2006, cité par Bannister 2011) ou de la Hongrie (Yu et Robinson, 2012). Dans son troisième rapport annuel sur les données ouvertes, la World Wide Web Foundation dénonce l'*open-washing*, c'est-à-dire l'utilisation d'un programme de données ouvertes aux seules fins de l'image et des relations publiques (World Wide Web Foundation, 2015).

L'idée même que la transparence soit source de confiance ne fait pas l'unanimité : les attentes à l'endroit des données ouvertes sont élevées, mais ne reposent pas toujours sur des preuves concrètes (Bannister et Connolly, 2011). En outre, la confiance est un phénomène complexe, qui ne se produit pas dès l'instant où l'on fait preuve de transparence, et certains soutiennent que la transparence ne peut produire l'imputabilité escomptée qu'à certaines conditions, notamment la mise en place de mécanismes participatifs (Peixoto, 2013).

Les données ouvertes posent par ailleurs un problème similaire à celui de l'accès à l'information, qui permet d'obtenir de nombreuses réponses à condition de poser les bonnes questions. Dans le cas des données ouvertes, on ne peut déterminer qu'un ensemble de données n'a *pas* été libéré que si l'on connaît préalablement l'existence de cet ensemble. Kieron O'Hara souligne l'importance de connaître l'existence des données non libérées afin de

créer la confiance; en d'autres mots, « le processus de transparence doit être lui-même transparent » (O'Hara, 2012, p. 4).

Enfin, on ne peut parler de transparence que dans la mesure où les citoyens sont aptes à interpréter les données libérées. Cet enjeu ne repose pas uniquement sur des prérequis techniques — tels que posséder un ordinateur ou avoir accès à Internet —, mais également sur des compétences générales comme la littératie numérique et la numératie[11], la connaissance du cadre légal ou technique, etc. (Bannister et Connolly, 2011). Les données ouvertes posent à l'évidence des défis considérables pour le citoyen lambda, et de nombreuses sources soulignent l'importance des intermédiaires — ou « infomédiaires » (Baack, 2015; Commissariat général du plan, 1999, p. 26-27; Margetts, 2011; McClean, 2011; O'Hara, 2012). Les opérations de nettoyage et d'intégration de données que nous présenterons au chapitre 4 illustrent l'ampleur de ces défis.

2.2.2. **Stimuler l'activité économique**

Le potentiel économique des données ouvertes a suscité l'intérêt des administrations publiques dès les années 1990, notamment en Union européenne où l'on considérerait déjà l'ouverture des données comme une source de richesse (Commissariat général du plan, 1999). La *Directive 2003/98/CE du Parlement européen et du Conseil* a rapidement défini un cadre légal pour les usages économiques de ce qu'on n'appelait pas encore données ouvertes (Gray, 2014; K. Janssen, 2011; Parlement européen et Conseil de l'Union européenne, 2003).

Cet objectif dépasse toutefois les simples transactions de données : de nombreux observateurs affirment que les données ouvertes doivent surtout favoriser l'innovation (Chan, 2013; European Commission, 2011; Groupe de travail sur les données ouvertes, 2011; Jetzek, Avital et Bjorn-Andersen, 2014; McLeod, 2012; Parycek, Hochtl et Ginner, 2014; Valentin, 2012; Zuiderwijk et al., 2014). Dans l'espace médiatique, cet objectif est souvent associé aux

[11] La numératie désigne le fait d'être fonctionnel dans le domaine des mathématiques, cependant que la littératie numérique désigne une connaissance adéquate des technologies de l'information.

hackathons citoyens — mais tandis que le hackathon vise un simple prototypage ou une preuve de concept (Hivon et Titah, 2015), l'entreprise privée cherchera plutôt à développer et maintenir un produit, ce qui constitue une forme d'innovation plus durable.

Afin d'assurer l'atteinte des objectifs économiques, les portails de données ouvertes doivent respecter un certain nombre de conditions, notamment l'utilisation de formats orientés-machine et de plateformes de diffusion appropriées. En outre, dans le cas où une administration ne libère que partiellement ses données, il est crucial de sélectionner les jeux qui présentent le meilleur potentiel de retour sur investissement (Ren et Glissmann, 2012).

Soulignons enfin l'importance de libérer les données sous une licence appropriée, qui établit un cadre d'opération clair pour les utilisateurs en ce qui a trait à la disponibilité, à la réutilisabilité et à la dérivation des données. En théorie, la licence d'utilisation permet aux administrateurs d'établir un équilibre entre la fermeture complète des données et leur versement pur et simple dans le domaine public (Miller et al., 2008), mais plusieurs observateurs préconisent une licence aussi permissive que possible (Open Knowledge Foundation, 2005; Sunlight Foundation, 2010). À titre d'exemple, les lignes directrices de l'*Open Definition* (Open Knowledge Foundation, 2005) spécifient qu'une telle licence devrait autoriser la redistribution — y compris pour les usagers tiers —, la modification, la segmentation et la compilation, et ce, pour tous les types d'usages et d'usagers, sans frais d'utilisation. On concède cependant que certaines conditions raisonnables peuvent s'appliquer, telles que la déclaration d'intégrité, le partage sous licence identique ou l'attribution des sources.

2.2.3. **Stimuler la participation citoyenne**

La participation citoyenne constitue possiblement le plus récent des objectifs que nous avons identifiés : tandis que le désir de transparence s'est manifesté dès le milieu du 20^{e} siècle et que la stimulation économique a été l'un des premiers mandats des données ouvertes en Europe à la fin des années 90, le discours sur la participation semble être un effet du Web social.

Les termes de cet objectif sont souvent formulés de manière assez vague, et si les sources se font parfois plus spécifiques sur le but à atteindre — par exemple, « donner des opportunités aux citoyens de participer à la prise de décision » (Executive Office of the President, 2009) —, elles n'expliquent pas comment les données ouvertes doivent mener à cet objectif. Chose certaine, de nombreuses sources soulignent l'influence de la participation citoyenne sur le modèle de gouvernance, et on ne se surprendra pas qu'elle soit associée à l'activisme politique et au milieu communautaire (Baack, 2015).

Bien que la participation citoyenne soit souvent mentionnée parmi les principaux objectifs de l'ouverture des données, il ne s'agit pas pour autant d'un phénomène de masse : beaucoup d'utilisateurs précoces des données ouvertes proviennent de la frange technophile de la population (Hivon et Titah, 2015). Plusieurs de ces technophiles semblent d'ailleurs estimer que l'administration publique devrait non seulement libérer des données, mais s'impliquer activement afin d'accroître la littératie numérique (Hivon et Titah, 2015) ou stimuler l'expérimentation, ce qui suggère que « les données ouvertes ne débouchent pas spontanément sur une participation accrue » (Baack, 2015). Certaines propositions visent même à obtenir la « contribution par inadvertance » d'un nombre plus important d'usagers en combinant, par exemple, les données ouvertes et les médias sociaux, ce qui témoigne du faible degré d'implication naturel des citoyens (Kalampokis, Hausenblas et Tarabanis, 2011). Il va de soi que la médiation joue ici aussi un rôle capital : on ne saurait concevoir une participation citoyenne à grande échelle sur la base d'ensembles de données que la moyenne des citoyens ne sont pas habiletés à interpréter ou utiliser.

En matière de participation citoyenne, il importe de séparer la contribution des données ouvertes en deux pôles principaux, qui correspondent d'une part aux modes de participation traditionnels, et d'autre part à ce que l'on pourrait appeler la *participation 2.0*.

Participation traditionnelle

La participation traditionnelle désigne les pratiques qui existaient avant la vague de libération des données : assistance aux assemblées municipales, expression d'opinion, rétroaction. Parmi les éléments significatifs de ce pôle, on trouve des outils développés à partir des données ouvertes, et qui servent à la mobilisation et à l'organisation de la société civile.

De nombreux cas intéressants sont basés sur la syndication ou la consolidation de données ouvertes issues de l'écosystème démocratique. À cet égard, le Royaume-Uni a été avant-gardiste : apparus respectivement en 2003 et 2004, des sites comme *The Public Whip* (Bairwell Ltd, s.d.) et *They Work For You* (mySociety, s.d.) permettent non seulement de trouver des députés à partir d'un code postal et de regrouper des informations à leur sujet (parti d'appartenance, profil, votes, statistique sur les interventions en chambre, etc.), mais aussi de se renseigner sur les débats, les travaux des comités et les projets de loi. Au Canada, *Open Parliament* (Mulley, s.d.) joue le même rôle, avec des fonctionnalités semblables. Il est difficile d'évaluer exactement combien d'outils similaires utilisent les données ouvertes. Une étude de 2011 recense néanmoins 191 organisations qui surveillent quelque 80 parlements (Mandelbaum, 2011), ce qui permet de supposer l'existence d'un nombre considérable de projets apparentés.

On a également vu apparaître des outils permettant d'élargir la portée d'activités déjà existantes, notamment la distribution des tâches de repérage sur le mode participatif (*crowdsourcing*). Le projet Open311, développé par OpenPlans, se veut un « modèle normalisé et ouvert, en lecture et écriture, pour le signalement par les citoyens de problèmes non urgents » (OpenPlans, s.d.). Des applications comme *FixMyStreet* ou *SeeClickFix* ont été conçues sur la base d'Open311, qui permettent non seulement de signaler des problèmes dans l'espace urbain, mais également de consulter les plaintes déjà enregistrées dans le système. Ce genre d'initiatives a été adopté par plusieurs administrations municipales, qui l'ont couplé à leurs plateformes de données ouvertes.

Participation 2.0

La participation 2.0 englobe les pratiques d'implication citoyenne qui seraient impensables sans données ouvertes — et qui, par la force des choses, impliquent une manipulation quelconque des données : développement d'une application, analyse et visualisation avancée de données, créations d'une API, etc. Il s'agit d'un type de participation spécialisé, dont la manifestation la plus médiatisée — voir la plus emblématique — est le hackathon citoyen (*civic hackathon*).

Le hackathon est un événement de développement d'application collaboratif qui se déroule durant une période assez courte (typiquement, quelques jours) et souvent dans un lieu aménagé pour l'occasion. Cet événement incarne l'esprit du développement ouvert : travail en équipes souvent formées sur place, division spontanée des tâches, diffusion publique du code source, etc. Souvent présenté comme une opportunité de développement d'applications, le hackathon s'apparente plutôt à un « incubateur » (Johnson et Robinson, 2014), où les participants veulent apprendre et expérimenter, prouver un concept, créer un prototype (Hivon et Titah, 2015). On ne s'étonnera donc pas que la pérennité des applications développées soit limitée (Johnson et Robinson, 2014).

Un hackathon citoyen est généralement organisé autour d'un thème, ou d'ensembles de données ouvertes sélectionnés à l'avance : durabilité urbaine (ÉcoHackMtl, s.d.), géomatique (« Défi Géohack 2014 », s.d.), corruption (« Hackons la corruption 2012 », s.d.), transport (Hackworks, s.d.-b; Ville de Montréal, 2012) ou sécurité publique (Schmite, 2017). Le hackathon peut se limiter à un portail de données ouvertes en particulier ou se dérouler à l'échelle nationale (Hackworks, s.d.-a), voire internationale (« International Open Data Hackathon », s.d.).

Enfin, il convient de mentionner le rôle central qu'occupent les processus consultatifs dans le fonctionnement des données ouvertes (Hivon et Titah, 2015; Open Government Working Group, 2007), et l'importance par conséquent d'installer des outils de rétroaction sur les portails (Zuiderwijk-van Eijk et Janssen, 2015). Bien que ces observations décrivent moins un objectif des données ouvertes qu'un prérequis ou un facteur de bonification, elles illustrent néanmoins l'importance culturelle de la participation citoyenne au sein de l'ouverture des données.

En outre, selon la fondation W3C, « engager de bonnes relations avec les usagers avant de mettre en place ou d'ajuster des politiques » constitue une manière d'évaluer les retombées des programmes de données ouvertes (World Wide Web Foundation, 2015, p. 13).

2.2.4. **Évaluation des retombées**

La littérature a souvent souligné l'évaluation lacunaire des retombées économiques et sociales des programmes de données ouvertes, et l'insuffisance de preuves ou de données empiriques (T. Davies, 2013; Ubaldi, 2013).

En effet, aucun gouvernement national ne semble avoir mesuré avec précisions les retombées de l'ouverture des données. Il n'existe apparemment pas de consensus sur les méthodes permettant d'évaluer le retour sur investissement (Martin et al., 2013), et on a parfois recouru à des études macro-économiques mal adaptées pour décrire des phénomènes locaux (Huijboom et Van den Broek, 2011). On a observé à l'occasion des mesures d'évaluation du degré d'ouverture des données ou de mise en œuvre des politiques de données ouvertes, mais aucune quantification des apports effectifs de ces politiques (Bertot, McDermott et Smith, 2012; Veljkovic, Bogdanovic-Dinic et Stoimenov, 2014).

Parmi les études dignes de mention, signalons *Measuring European Public Sector Information Resources* (MEPSIR) publiée en 2006 (Dekkers, Polman, te Velde et de Vries, 2006), qui faisait le suivi socio-économique de la directive de 2003 sur la réutilisation des informations du secteur public (Parlement européen et Conseil de l'Union européenne, 2003). Malgré l'ampleur de cette étude, qui couvre les vingt-cinq états membres de l'Union européenne, les résultats demeurent généraux ou approximatifs, et reposent sur des estimations fournies par divers répondants, des études de cas et des calculs d'équivalences. Qui plus est, les études de cas effectuées se limitaient essentiellement aux grands ensembles de données traditionnellement libérés : météo, géomatique, information juridique.

Dans une perspective moins eurocentrique, l'Open Data Research Network s'est penché sur les retombées dans les pays en voie de développement, notamment en publiant un cadre d'évaluation (Davis, Perini et Alonso, 2013) et en soutenant des études de cas (Open Data for Developpement Network, 2016, p. 31). L'intérêt de ces démarches, outre leur dynamisme, consiste notamment à diversifier le discours sur les données ouvertes.

Malgré un certain nombre de travaux sur l'évaluation des retombées, les normes d'évaluations brillent par leur absence. Voilà pourquoi des intervenants du milieu, à l'instigation de la World Wide Web Foundation (Web Foundation) et du Governance Lab

(GovLab[12]), ont entrepris de créer un cadre d'évaluation commun à partir de dix cadres d'évaluation existants ou projetés (Caplan et al., 2014)[13] :

1. l'*Open Data Barometer* (ODB), un index multidimensionnel développé par la World Wide Web Foundation et le réseau Omidyar, et qui fait l'objet de publications annuelles depuis 2013;
2. l'*Open Data Index & Open Data Census*, une évaluation périodique de la disponibilité de certains ensembles de données essentiels au niveau national;
3. l'*UN EGovernment Survey*, une enquête de l'ONU sur la gouvernance numérique, et comprenant des questions sur les données ouvertes;
4. l'*Open Data Monitor*, un projet d'évaluation automatisée des données ouvertes de l'Union européenne;
5. l'*Open Data Certificate*, un programme d'auto-évaluation permettant aux responsables de données ouvertes de mesurer la qualité de leurs données;
6. les propositions et lignes directrices quantitatives et qualitatives pour un futur cadre d'évaluation de l'OCDE;
7. l'*Open Data 500*, une étude exhaustive visant à identifier les entreprises privées utilisant les données ouvertes, et les données dont elles font usage;
8. l'*Open Data Era in Health and Social Care*, un document préparatoire conçu par le National Health Service (NHS England) et GovLab, où l'on discute de plusieurs approches et aspects méthodologiques;

---

[12] The GovLab (thegovlab.org) est un centre de recherche-action (« action research center ») associé à la New York University, et ne doit pas être confondu avec le *think thank* GovLab de la firme Deloitte ou avec le groupe de recherche MIT GOV/LAB.

[13] L'ébauche de ce cadre commun servira de base pour notre évaluation de la situation au Québec et à Montréal (chapitre 3).

9. l'*European PSI Scoreboard*, une enquête sur les politiques et pratiques en matière d'informations du secteur public en Europe;
10. l'*Open Data Compass*, une évaluation de la disponibilité des données publiques sur les entreprises et les litiges.

De manière générale, ces cadres accordent beaucoup d'attention au choix des données et à la manière dont elles devraient être libérées, mais nettement moins à leur impact. Sur les dix cadres considérés, quatre seulement abordent l'évaluation des retombées :

- l'*Open Data Barometer* (ODB), dont le tiers est consacré à parts égales aux retombées sociales, politiques et économiques; l'approche consiste à traiter les publications citoyennes, médiatiques et académiques comme des indices de l'impact des données ouvertes. (World Wide Web Foundation, 2015);
- l'*Open Data 500*, qui cherche à établir un mode d'évaluation des retombées économiques (GovLab, s.d.-b);
- l'*Open Data Era in Health and Social Care*, qui formule des recommandations très élaborées, mais orientées sur les questions de santé (Verhulst, Noveck, Caplan, Brown et Paz, 2014);
- les propositions et lignes directrices quantitatives et qualitatives pour un futur cadre d'évaluation de l'OCDE, qui soulignent notamment l'importance de développer des méthodes de mesure et d'analyse empiriques (Ubaldi, 2013).

L'examen de ces cadres et des travaux de GovLab permet d'identifier plusieurs difficultés sous-jacentes à la constitution d'un cadre d'évaluation. L'une de ces difficultés — pour le moins fondamentale — consiste à définir la nature et la portée des retombées.

Comme le fait remarquer Barbara Ubaldi, on peut « évaluer la portée démocratique des données ouvertes de manière intrinsèque (sur la base des valeurs qu'elles promeuvent) ou

instrumentale (en fonction des améliorations qu'elles permettent d'apporter aux politiques et aux services publics) » (Ubaldi, 2013, p. 44, notre traduction). La distinction n'a rien d'abstrait : il existe des risques réels à considérer certains mandats comme des fins en soi, sans égard pour leurs conséquences. Rien n'exclut que des mesures en apparence parfaitement légitimes puissent demeurer sans effets perceptibles, voire produire des effets indésirables. On pourrait supposer, par exemple, que la divulgation du salaire des élus permettra d'éviter toute rémunération excessive. Pourtant, l'examen des faits révèle qu'une telle transparence ne donne pas forcément le contrôle escompté (Worthy, 2014) et pourrait même se solder par un effet inflationniste (Gomez et Wald, 2010; Schmidt, 2012).

L'évaluation quantitative — et notamment financière — des retombées pose également des difficultés considérables. On peut supposer que les administrations disposent d'une capacité d'analyse adéquate sur les phénomènes qui relèvent de leur propre périmètre comptable. Un gouvernement pourrait aisément comparer, par exemple, les sommes engagées pour la libération des données et celles associées à la gestion des demandes d'accès à l'information[14]. Il s'avère néanmoins plus difficile pour un gouvernement d'obtenir de telles données — à plus forte raison des données exhaustives — de la part des entreprises ou de la société civile.

Conséquence de cette difficulté, les analystes semblent avoir privilégié jusqu'à présent l'étude de cas[15]. Cette approche permet assurément de recueillir de précieuses données et de présenter un portrait multidimensionnel des retombées d'un programme de données ouvertes. Les résultats de telles études s'avèrent toutefois, par définition, plus ou moins généralisables.

Signalons enfin un paradoxe méthodologique intéressant sur lequel se conclut le rapport de GovLab : l'évaluation des retombées implique généralement d'évaluer des cibles

---

[14] À cet effet, plusieurs auteurs affirment qu'il est moins coûteux de publier proactivement des documents que de répondre à des requêtes individuelles (Granickas, 2013, p. 16), des économies qui pourraient toutefois être annulées si le volume d'information devenait trop important (Bannister et Connolly, 2011, p. 23; Tait, 2011).

[15] GovLab a notamment diffusé les résultats de 25 études de cas menées en collaboration avec le réseau Omidyar (GovLab, s.d.-a).

spécifiques. Or, comme nous l'avons vu en introduction du chapitre 2, l'un des intérêts des données ouvertes consiste à couvrir un large spectre d'usages, dont plusieurs ne sont pas forcément anticipés. On ne s'étonnera donc pas qu'il soit difficile d'évaluer l'atteinte d'une cible par définition mouvante.

En conclusion, il est non seulement difficile de définir les principaux objectifs des données ouvertes, mais il apparaît que l'atteinte de ces objectifs ne dépend pas exclusivement de l'ouverture des données elles-mêmes : des mesures connexes — intervention d'infomédiaires, mise en place de mécanismes participatifs, transparence du processus d'ouverture — sont nécessaires afin que les données ouvertes produisent les effets désirés.

Dans la section suivante, nous examinerons la manifestation concrète de ces objectifs, en présentant une typologie des utilisations des données ouvertes.

2.3. **Utilisation des données ouvertes**

Cette section ne constitue pas une typologie exhaustive des applications[16] développées à partir des données ouvertes. En effet, il ne nous a pas semblé pertinent de procéder à la description définitive d'un paysage en constante transformation. Nous avons plutôt cherché à rassembler un échantillon diversifié à partir de sources qui représentent différentes perspectives (administrative, académique, citoyenne), différents paliers gouvernementaux (fédéral, provincial et municipal) et différentes fréquences de mise à jour (page statique, médias sociaux, etc.). Ces sources sont les suivantes :

- **Galeries d'applications.** Les galeries (ou répertoires) d'applications des portails de données ouvertes de Montréal, de New York, du Québec, du Canada, de la France et du Royaume-Uni nous ont révélé l'existence de diverses applications de type logiciel;

[16] Sauf mention contraire, nous utiliserons le terme *application* dans son acception large — l'utilisation pratique d'une ressource —, et non seulement au sens d'application logicielle.

- **Littérature spécialisée.** Notre revue de littérature multidisciplinaire a permis de repérer des articles qui constituent en soi des applications analytiques, ou qui font référence à des applications de type logiciel;
- **Médias d'information.** Les médias d'information offrent des exemples d'applications analytiques sous la forme de journalisme de données; nos principales sources ont été Radio-Canada/CBC, le *Guardian* et le *New York Times*.
- **Médias sociaux.** Les médias sociaux sont utilisés par de nombreux développeurs et médias pour publiciser leurs applications. Nous avons mené une veille sur Twitter et sur le groupe de discussion Google consacré aux données ouvertes de Montréal (https://groups.google.com/forum/#!forum/open-data-montreal).

Sur la base des applications repérées, nous avons procédé à une catégorisation sommaire afin de mettre en évidence les principales tendances. Notre typologie repose sur deux grandes catégories d'applications : d'une part les interfaces (section 2.3.1), qui permettent aux usagers d'interroger et d'analyser par eux-mêmes les données ouvertes, et d'autre part les analyses à proprement parler (section 2.3.2).

Dans les sections suivantes, nous décrirons ces catégories plus en détail et nous fournirons quelques exemples représentatifs.

2.3.1. **Interfaces**

Les interfaces constituent le type d'utilisation le plus courant. Elles prennent généralement la forme d'une application mobile ou Web, et permettent une interaction avec les données ouvertes en temps réel ou en différé. Bien qu'il s'agisse souvent d'interfaces visuelles, elles font rarement l'économie d'un formulaire de recherche, lequel se résume souvent à un champ unique, mais peut à l'occasion se déployer en mode avancé.

Bien que la visualisation constitue une pratique de médiation importante, tant pour les usagers (Graves et Hendler, 2013) que pour les administrateurs (Mercier, 2014), l'inventaire

des galeries d'applications révèle une diversité assez faible. La carte géographique demeure sans conteste l'interface visuelle dominante : qu'il s'agisse de représenter des espaces de stationnement, des trajets de transports en commun, des points d'intérêts ou des centres de service, des événements culturels, des temps de déplacement ou des indicateurs de qualité de vie, les développeurs optent presque invariablement pour l'interface cartographique. Cette prédominance reflète non seulement la proportion importante de données géospatiales sur les portails de données ouvertes, mais sans doute aussi une tendance générale du Web, accentuée par l'utilisation *in situ* des applications mobiles.

Si la plupart des applications que nous avons repérées permettent simplement d'afficher les résultats d'une requête sur une carte, certaines interfaces offrent un degré de complexité plus élevé. C'est le cas du site *Indices of Deprivation Explorer* (figure 5), où l'on compile plusieurs indicateurs et indices (niveau d'éducation, criminalité, indice de défavorisation des aînés, etc.) afin de générer un indice de défavorisation général que l'on peut afficher par code postal ou zone administrative. Un système de filtrage par facette permet également de décomposer l'indice.

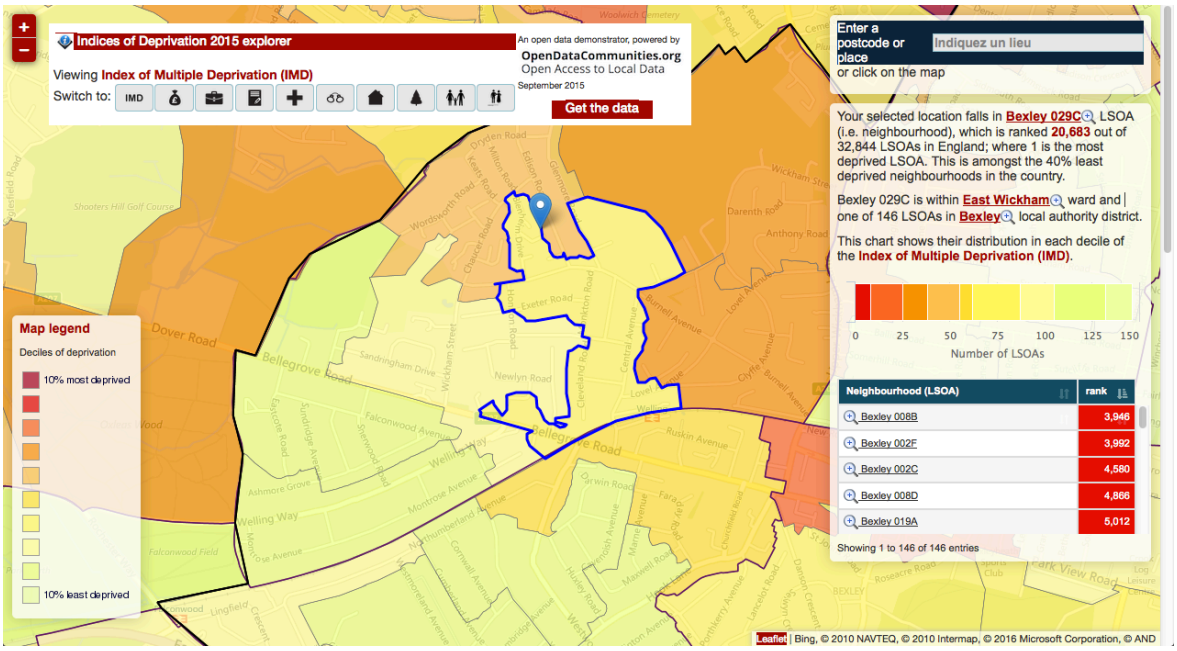


Figure 5. *Indices of Deprivation Explorer* (Peters, 2015)

Mentionnons également TaxiVis, une application qui permet d'analyser les données de taxi de New York : plutôt que d'utiliser les outils de recherche traditionnels (champs textuels, menus, facettes, etc.), l'application permet de faire des requêtes avec une interface schématique et cartographique, notamment en désignant des zones *ad hoc* à l'aide de vecteurs et de polygones (figure 6).

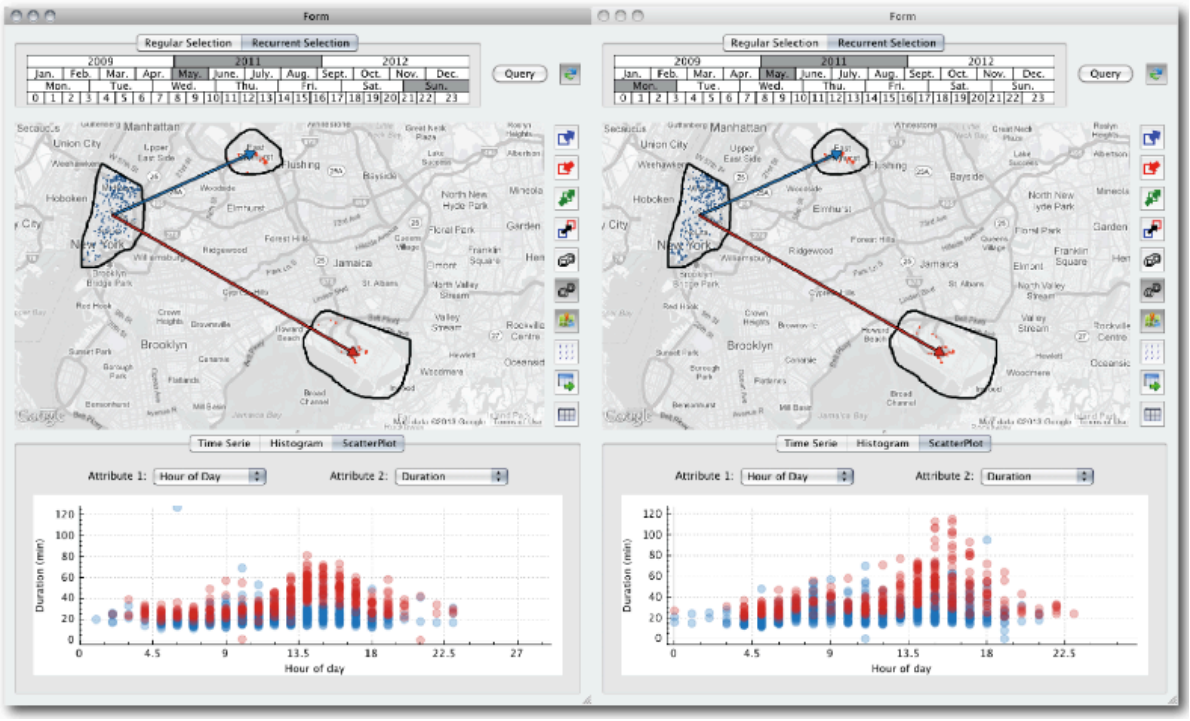


Figure 6. *TaxiVis* (Ferreira, Poco, Vo, Freire et Silva, 2013)

Les interfaces de recherche basées sur une visualisation non cartographique sont nettement moins fréquentes. Les applications les plus intéressantes permettent de fureter dans des ensembles de données hiérarchisées. Parmi les exemples intéressants, mentionnons *Vue sur les contrats* (figure 7), qui permet d'explorer l'arborescence des subventions attribuées par la ville de Montréal, et *Where does my money go?* (figure 8), qui présente un organigramme interactif du budget du Royaume-Uni.

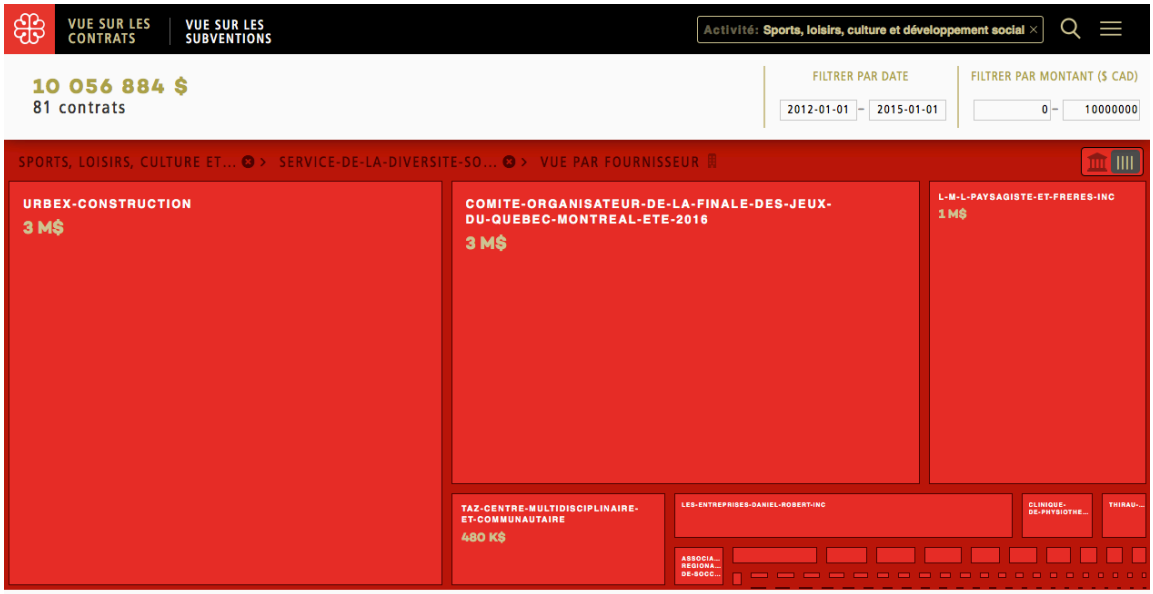


Figure 7. *Vue sur les contrats* (FFunction, s.d.)

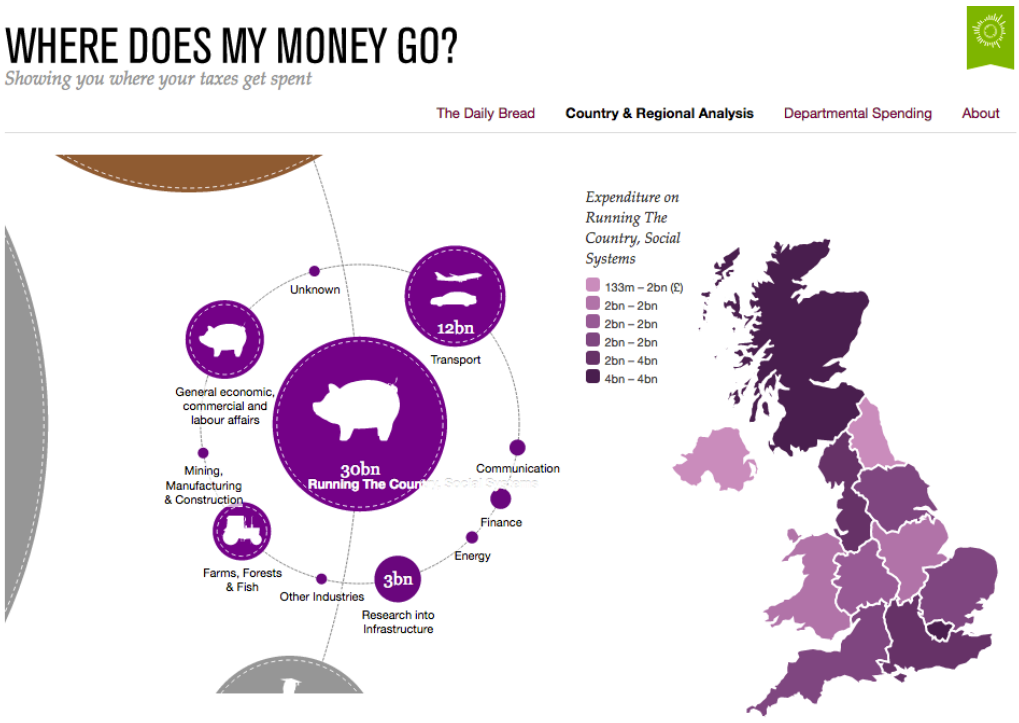


Figure 8. *Where does my money go?* (Open Knowledge Foundation, s.d.-b)

Certains outils de visualisation empruntent la forme de graphiques traditionnels, auxquels on ajoute des fonctionnalités qui facilitent la consultation ou la comparaison. Signalons les cas suivants :

- l'*Interactive Visualization of NYC Street Trees*, un graphique à barres empilées classique où l'on peut néanmoins comparer les valeurs entre différents arrondissements (figure 9);
- le *UK Energy Consumption Guide*, où les graphiques interactifs empruntent une facture quasi narrative (Evoenergy, s.d.);
- l'analyse des données du service 311 de Montréal sous forme de tableaux croisés dynamiques avec colorisation par densité (Kruchten, 2015).

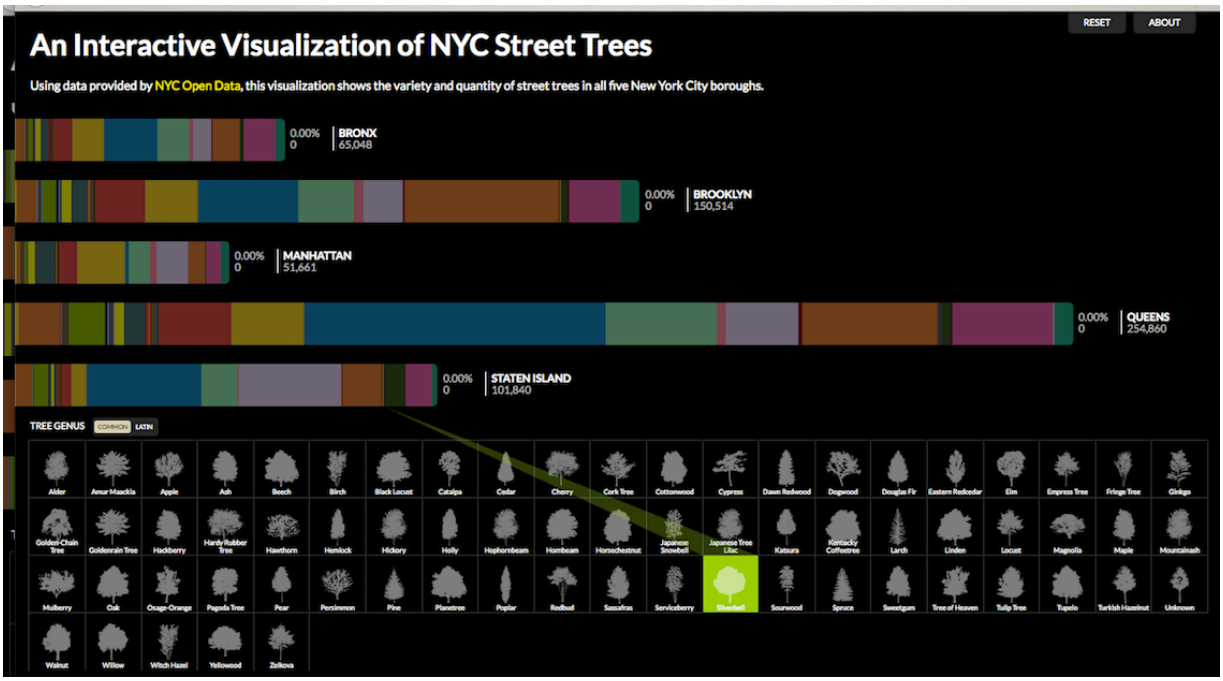


Figure 9. *Interactive Visualization of NYC Street Trees* (Cloudred Multimedia, s.d.)

On note aussi un recours assez fréquent aux tableaux de bord (*dashboards*), qui regroupent plusieurs sources de données afin de donner une vue d'ensemble d'un territoire, d'un système, d'une offre de services, etc. Ces tableaux de bord peuvent être de simples

plateformes de syndication — comme *They Work for you* (mySociety, s.d.) ou *Open Parliament* (Mulley, s.d.) —, mais il arrive que différentes sources de données y soient combinées afin de créer des indices ou de dégager des tendances.

En conclusion, on peut supposer que la prévalence et la faible diversité des interfaces visuelles sont en quelque sorte inévitables : la notion même d'interface repose sur la convivialité, et les développeurs recourent sans doute à une grammaire visuelle familière afin d'éviter toute confusion chez les usagers.

2.3.2. **Analyses**

Nous avons distingué, parmi les cas d'analyses repérées, deux grandes catégories méthodologiques : l'approche statistique classique, et les méthodes d'extraction avancée comme la fouille de données.

L'analyse statistique constitue la méthode prédominante, et elle s'accompagne fréquemment de visualisation afin de faciliter la présentation des résultats et l'interprétation les données. On y remarque d'ailleurs les deux mêmes axes visuels que dans les interfaces dont nous avons parlé en section 2.3.1 :

- les visualisations cartographiques, telles que la distribution par sexe des usagers de vélos (Kaufman, 2014), l'utilisation des vélos durant une journée (Ferzoco, 2014) ou l'orientation politique des quartiers lors des élections (Kruchten, 2014);
- les représentations « classiques », comme les graphiques (Rocha, 2015), les tableaux colorisés selon la densité (Clark, 2013; Friedman et DeBold, 2015; Tétreault-Pinard, 2014) ou les nuages de mots.

Cette prédominance de l'approche statistique s'explique par la contribution du journalisme de données — un phénomène notamment popularisé par le *Guardian* et le *New York Times* —, mais aussi par la portée souvent limitée des données ouvertes : la complexité de l'analyse ne dépend pas seulement des méthodes employées, mais également de la nature des données disponibles. Certains auteurs font remarquer que les données ouvertes se prêtent peu

aux techniques de fouille, en particulier à cause de la rareté des ensembles de données assez volumineux, présentant une dimension diachronique, ou comportant des traits suffisamment hétérogènes et interreliés (Helbing et Balietti, 2011, p. 7; Pineau et Bacon, 2015).

Nous avons en effet constaté, à l'instar de ces auteurs, l'importante présence de nombreux ensembles de données qui reflètent ces limites. Des ressources telles que la liste des piscines municipales (Ville de Montréal, 2013c), le plan de classement de BAnQ (Gouvernement du Québec, 2012) ou le répertoire des fromages canadiens (Gouvernement du Canada, 2015b) ne se prêtent pas à une analyse complexe. En outre, plusieurs ensembles se composent de données synthétiques plutôt que de données brutes. De tels ensembles de données comportent peu d'instances et peu de traits discriminants, et leur principal intérêt analytique consiste à être croisés avec d'autres ensembles de données.

Mentionnons enfin que les ensembles de données ouvertes diachroniques disponibles couvrent souvent des périodes relativement courtes, dont le début coïncide avec l'adoption des premières politiques de libération de données ou la création des portails de données ouvertes, c'est-à-dire cinq ou six ans tout au plus (voir 4.3.3. *couverture temporelle*). Cette couverture relativement brève diminue l'intérêt analytique des données, puisqu'elle empêche de déduire des tendances à long terme.

Signalons enfin que l'implantation de systèmes de captation de données afin de surveiller divers aspects d'une communauté — qu'il s'agisse du comptage de véhicules (Ville de Montréal, 2013d), de trajets de taxis (New York City, s.d.) ou de données captées par des véhicules utilitaires (Walcott, 2015) — laissent entrevoir une éventuelle multiplication d'ensembles de données diachroniques présentant un grand degré de granularité, et qui se prêteront peut-être mieux à des analyses poussées. Les ensembles de données issus de telles technologies n'acquerront toutefois leur pleine maturité qu'après plusieurs années d'opération.

En dépit de ces lacunes, nous avons repéré un certain nombre de travaux qui illustrent l'intérêt de soumettre les données ouvertes aux techniques de fouille, aussi bien à des fins d'analyse exploratoire que prédictive :

- Les données d'utilisation des systèmes de vélos en libre-service ont notamment inspiré plusieurs travaux de classification : identification de patrons d'activité

ou de profils d'usagers, parfois mis en relation avec des territoires donnés (O'Brien, Cheshire et Batty, 2014; Vogel, Greiser et Mattfeld, 2011; Zimmermann, Kaytoue, Plantevit, Robardet et Boulicaut, 2015).

- Les données sur la qualité de l'air ont été analysées afin de dresser un portrait précis de la pollution au niveau national (Li et Shue, 2004) ou d'améliorer les processus de prédiction permettant de cartographier la présence de polluants en temps réel (Ochando et al., s.d.).
- La notion d'accessibilité suscite beaucoup d'intérêt, qu'il s'agisse d'évaluer l'accessibilité réelle de points de service (Salonen et Toivonen, 2013), de comparer les patrons de déplacement pour différents modes de transport (Jäppinen, Toivonen et Salonen, 2013) ou encore d'évaluer l'influence de l'accessibilité sur le marché immobilier (Zliobaite, Mathioudakis, Lehtiniemi, Parviainen et Janhunen, 2015).

En conclusion, l'examen des applications analytiques met en lumière l'importance du volume, de la couverture et de la qualité des données libérées, ainsi que des possibilités offertes par le croisement ou la consolidation de plusieurs sources de données.

L'état des lieux effectué dans cette section nous a permis d'établir les différents mécanismes qui président à la création, au traitement et à la description des données ouvertes, les principaux objectifs de la libération de données, ainsi que les types d'utilisation de ces données. Dans la section suivante, nous verrons si les pratiques au gouvernement du Québec et à la ville de Montréal reflètent les préoccupations identifiées dans la littérature.

3. Données ouvertes au Québec : un état des lieux

Dans ce chapitre, nous présenterons le contexte des données ouvertes québécoises (section 3.1), une description de l'offre de données actuelle (section 3.2), ainsi qu'un survol des applications et retombées des données ouvertes dans la province (sections 3.3 et 3.4).

Nous n'entendons pas classer ici le Québec au sein d'un quelconque palmarès des données ouvertes : l'ampleur d'un tel exercice dépasserait largement le cadre et les objectifs de notre recherche. Il est néanmoins important d'examiner la situation québécoise à la lumière des enjeux et pratiques internationales identifiés au chapitre 2, afin de mettre en contexte le prétraitement que nous effectuerons dans le chapitre 4.

Pour ce faire, nous nous sommes inspirés du cadre commun en cours de développement chez GovLab (Caplan et al., 2014), qui fait la synthèse de dix cadres d'évaluation majeurs (voir section 2.2.4). Ce cadre commun se compose de quatre volets, qui constitueront le contenu des quatre prochaines sections :

- contexte/environnement (section 3.1) : le contexte de diffusion des données ouvertes, en particulier le cadre légal ou réglementaire, et le niveau d'engagement politique et de littératie numérique;
- données (section 3.2) : la nature et la qualité des ensembles de données, incluant le degré d'ouverture et la pertinence des données;
- utilisation (section 3.3) : les catégories d'utilisateurs et d'utilisations, ainsi que leurs objectifs. Il s'agit en somme du « qui, quoi et comment » des données ouvertes;
- retombées (section 3.4) : les impacts bénéfiques[17] sur les plans social, environnemental, politique et économique.

Nous avons légèrement adapté le contenu de ce cadre pour les besoins de notre recherche : nous avons écarté certaines facettes redondantes — la dimension organisationnelle, par exemple, se retrouvait à la fois dans les sections *contexte* et *utilisation* — et nous avons ajouté une facette historique à la section *contexte*.

[17] Mentionnons que le cadre d'évaluation de GovLab est idéologiquement orienté, en ce sens qu'on n'y propose pas d'identifier d'éventuels impacts négatifs (voir à sujet la section 2.2.4).

En outre, malgré la participation de six administrations au portail *Données Québec*, nous avons concentré notre examen sur les données provenant de la ville de Montréal et du gouvernement provincial, deux sous-ensembles qui totalisent 67 % des données disponibles, et d'où proviennent par ailleurs les ensembles de données choisis pour notre prétraitement.

Enfin, certaines organisations paragouvernementales opèrent leurs propres portails, sans être référencées sur *Données Québec* ou sans utiliser l'expression *données ouvertes*. Nous avons choisi de les ignorer, puisque les données ouvertes devraient par définition être centralisées et clairement identifiées afin de favoriser la découvrabilité (voir 4.3.3).

3.1. **Contexte**

Dans cette section, nous présenterons en détail le contexte des données ouvertes au Québec : après un bref historique (section 3.1.1), nous présenterons le cadre légal (section 3.1.2), le contexte organisationnel (section 3.1.3) et le cadre technique (section 3.1.4) de l'ouverture des données.

3.1.1. **Bref historique**

Au Canada, c'est au niveau municipal que l'on a assisté aux premières initiatives officielles en matière de données ouvertes. La création du portail de données ouvertes de Vancouver, en septembre 2009, a inauguré une période d'effervescence à l'échelle nationale : au cours des six années suivantes, on a vu apparaître quelque cinquante portails municipaux de données ouvertes, ce qui représente l'apparition d'un portail toutes les cinq semaines en moyenne[18].

[18] Cette croissance reflète évidemment l'évolution aux États-Unis et en Union européenne. Il est néanmoins possible que l'adoption rapide des données ouvertes au Canada ait été favorisée par un certain climat politique. Les scandales de corruption ont non seulement mené à la création de l'Unité permanente anticorruption (2011), à la *Commission d'enquête sur le programme de commandites et les activités publicitaires* (2004) et à la *Commission d'enquête sur l'octroi et la gestion des contrats publics dans l'industrie de la construction* (2011), mais ils ont aussi mis à l'avant-plan médiatique les

Le portail de la ville de Montréal a été créé en novembre 2011, suite aux recommandations d'un groupe de travail mandaté pour étudier la question huit mois plus tôt (Groupe de travail sur les données ouvertes, 2011). Le portail a connu dès sa création une croissance rapide, quoique parfois désorganisée, comme en témoigne la couverture temporelle ou spatiale discontinue de certains ensembles de données, ou certaines incohérences dans les formats (voir à ce sujet les sections 3.2.1 et 3.2.2)

Au niveau provincial, l'ouverture des données s'est produite un peu plus tardivement. Nous devons le premier portail de données ouvertes provincial à la Colombie-Britannique : mis en ligne à l'été 2011, il a été suivi par les portails du Québec et de l'Ontario (2012), puis par celui de l'Alberta (2013). Aujourd'hui, toutes les administrations provinciales et territoriales possèdent, sous une forme ou une autre, un portail de données ouvertes. La moitié de ces portails remplissent toutefois des mandats exclusivement (et explicitement) géospatiaux.

Au Québec, la réflexion sur les données ouvertes a eu lieu dès 2006, comme en témoigne un amendement à la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels* (voir section 4.1.2.). En outre, des consultations sur le Web social entamées en 2010 mèneront à la publication du rapport Gautrin (Gautrin, 2012), qui recommandera, parmi les processus de mise en œuvre du « gouvernement 2.0 », la création d'un portail de données ouvertes. Ce portail sera disponible à l'adresse www.donnees.gouv.qc.ca dès l'été 2012.

Bien que ces initiatives datent de quelques années, elles ne semblent pas avoir atteint une forme parfaitement stable. En effet, au cours du premier trimestre de 2016, Montréal a publié une refonte de sa politique de données ouvertes, et on a assisté à la fermeture du portail de données exclusivement provincial, remplacé désormais par un portail qui regroupe à la fois

---

notions de transparence et d'imputabilité, deux raisons d'être récurrentes dans le discours sur les données ouvertes.

les données du gouvernement du Québec et de cinq municipalités. Ces municipalités continuent néanmoins à opérer leurs propres portails.

3.1.2. **Cadre légal**

Bien que les premières initiatives québécoises concrètes en matière de données ouvertes aient eu lieu au niveau municipal — notamment avec l'ouverture des portails de Montréal et de la ville de Québec —, les prémices légales du mouvement se situent au niveau de l'administration provinciale (A. Davies et Lithwick, 2010).

La notion même de diffusion proactive par le gouvernement et ses agences date de l'amendement de 2006 de la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels* (*Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, 1982, art. 16.1). Ces dispositions ne seront toutefois implémentées qu'en novembre 2009, avec le *Règlement sur la diffusion de l'information et sur la protection des renseignements personnels* (*Règlement sur la diffusion de l'information et sur la protection des renseignements personnels*, 2008). Alysia Davies et Dara Lithwick soulignent à juste titre la nature très inclusive de ce règlement, qui touche toutes les entités publiques régies par les lois d'accès et la province, incluant « non seulement le gouvernement provincial, mais également les municipalités québécoises, les commissions scolaires, et les institutions de santé et de services sociaux. » (A. Davies et Lithwick, 2010).

Pour sa part, l'administration montréalaise a adopté en 2012 une première politique de données ouvertes, puis une seconde en 2016, qui sont « subordonnées aux dispositions de la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels* » (Ville de Montréal, 2016d). Cette politique s'applique à toutes les unités administratives de la ville, incluant les arrondissements. On recommande en outre son adoption « aux organisations paramunicipales ou faisant partie de son périmètre comptable. »

Montréal s'est par ailleurs dotée d'une *Directive sur la gouvernance des données de la Ville de Montréal*, un document « dont la portée est principalement interne [et vise] à clarifier les responsabilités vis-à-vis des données » (Ville de Montréal, s.d.-d).

3.1.3. **Contexte organisationnel**[19]

Politiques et normes

Outre les politiques et directives dont il a été question plus haut, la ville de Montréal et le Gouvernement du Québec ont élaboré deux documents afin d'encadrer les intervenants :

- *Lignes directrices sur la diffusion de données ouvertes* (Gouvernement du Québec, 2016)
- *Guide des publieurs de données* (Ville de Montréal, s.d.-b)

Bien que les documents des deux administrations soient très similaires — on y trouve de nombreux passages identiques —, il existe un certain nombre de différences plus ou moins notables.

Ainsi, les lignes directrices provinciales couvrent principalement des questions formelles : on y traite des types de données (tabulaires, hiérarchiques et géomatiques), des formats de données et de documents, ainsi que des métadonnées. À ces éléments s'ajoutent quelques considérations sur le cadre réglementaire, ainsi que sur la création et la mise à jour d'ensembles de données.

Le guide montréalais, pour sa part, se compose d'un ensemble de documents de référence. Outre le *Guide des formats* et le *Guide des métadonnées*, qui couvrent les mêmes aspects formels que les lignes directrices provinciales, on y trouve des documents comme le *Guide général des données ouvertes* ou la *Politique de données ouvertes*, ainsi que des guides et formulaires de publication détaillés. Les documents équivalents au niveau provincial n'ont pas été diffusés.

La question de la granularité des données constitue l'une des importantes différences entre les documents des deux administrations : les lignes directrices provinciales n'en parlent

[19] Afin d'éviter la redondance, nous avons détaillé les organisations participantes à la section 3.2.1.

pas du tout, cependant que le *Guide des métadonnées* de Montréal donne des lignes directrices sommaires. Nous examinerons cette importante question dans la section 3.2.1.

Responsabilités

Les programmes de données ouvertes sont par nature des entreprises transorganisationnelles qui nécessitent un effort de coordination. Aux États-Unis et en Grande-Bretagne, ce rôle a été confié à un *chief data officer*, dont le mandat consiste non seulement à valoriser les données et à gérer leur utilisation, mais également à assurer l'application des politiques.

Au Québec, le gouvernement provincial a créé le poste de dirigeant principal de l'information (DPI), qui répond directement du président du Conseil du Trésor. Son mandat est plus large que le *chief data officer*, et couvre l'ensemble des technologies de l'information. Le DPI est notamment « chargé de mettre en œuvre les politiques et les directives établies conformément à la Loi, d'en surveiller l'application et d'en coordonner l'exécution » (*Loi sur la gouvernance et la gestion des ressources informationnelles des organismes publics et des entreprises du gouvernement*, 2011). Il coordonne en outre quelque 120 dirigeants de l'information (Conseil du trésor, 2015)

À Montréal, les données ouvertes relèvent du Bureau de la ville intelligente et numérique, où l'on trouve, à défaut d'un dirigeant des données, un chef d'équipe en données ouvertes.

Au quotidien, l'édition et la publication des données sont déléguées à une variété d'intervenants qui occupent, de toute évidence, des situations très variables au sein de leurs organisations respectives. L'examen des documents laisse d'ailleurs deviner une difficulté à désigner précisément les intervenants.

En effet, dans les *Lignes directrices* du portail québécois, on emploie le terme *diffuseur*, tout en indiquant que les métadonnées du portail comportent deux catégories distinctes — *organisation participante* et *organisation responsable* — dont le niveau de précision peut « varier selon l'organisation » (Gouvernement du Québec, 2016). Un examen attentif des métadonnées par le truchement de l'API confirme cette ambivalence : si, dans les

faits, le contenu des métadonnées *organisation principale* et *organisation participante* est assez cohérent, la métadonnée responsable (*author*) peut cependant contenir, selon le contexte, une organisation principale (« Ville de Laval »), une organisation participante (« Service de la gestion des immeubles »), voire un individu (« Marc Lebel »).

La situation n'est guère plus simple à Montréal : la *Directive sur la gouvernance des données* parle de *dépositaire de données*, de *fiduciaire de données*, de *répondant de contenu*, de *répondant technique* et de *coordonnateur de données* (Ville de Montréal, 2015). Dans la *Politique de données ouvertes*, on parle simplement du *responsable d'une ressource informationnelle*, que l'on définit comme le « gestionnaire de l'unité de la Ville responsable de l'acquisition et de l'intégrité d'une ressource informationnelle » (Ville de Montréal, 2016d). Ces subtilités sémantiques sont cependant absentes du portail, où l'on se borne à parler de *publieur*, que l'on définit comme un « répondant de contenu de l'ensemble de données ouvertes ou du coordonnateur aux données » (Ville de Montréal, s.d.-e).

Cette confusion terminologique reflète la complexité organisationnelle sous-jacente aux données ouvertes, et la difficulté de parvenir à une application uniforme des politiques. Dans la section 3.2, nous verrons comment cette situation peut donner lieu à des incohérences, notamment en ce qui a trait à la granularité des ensembles de données et au contenu des métadonnées.

3.1.4. **Cadre technique**

Degré d'ouverture

Il est impossible d'évaluer le degré de numérisation et d'ouverture des données du gouvernement du Québec ou de la ville de Montréal, car bien que les deux administrations aient adopté des politiques d'ouverture extensives — Montréal a notamment opté pour une politique d'ouverture « par défaut » —, il n'existe aucun répertoire de l'entièreté des données produites ou gérées à l'interne, et qui servirait de point de comparaison.

Nous avons cependant remarqué que la couverture temporelle des données se limite souvent à la durée d'existence des portails de données, et qu'aucune mesure notable de

numérisation ou de publication rétroactive ne semble avoir été mise en place. Pour les utilisateurs qui désirent effectuer une analyse historique, l'ouverture des données comporte des angles morts évidents.

Mode de publication

Le gouvernement du Québec et la ville de Montréal utilisent tous deux le système de gestion de données CKAN, dont l'interface Web permet aux intervenants de téléverser directement les données au portail, avec ou sans validation des administrateurs. La décentralisation des tâches de mises en ligne permet d'accélérer la publication, mais peut également augmenter les incohérences.

3.2. **Données**

Dans cette section, nous présenterons une vue d'ensemble des données ouvertes actuellement offertes au Québec (section 3.2.1), puis nous examinerons certains aspects plus spécifiques de la question : les formats (section 3.2.2), l'accès et l'accessibilité (3.2.3), les licences (3.2.4) et les pratiques d'évaluation des données (3.2.5).

3.2.1. **Vue d'ensemble**[20]

Volume de données

Le portail de *Données Québec* recense 761 ensembles de données[21] qui totalisent 2982 documents. Ces deux chiffres doivent être interprétés avec prudence.

---

[20] Tous les chiffres de cette section résultent d'observations faites en février 2017.

[21] La terminologie employée sur les portails ou dans la littérature est parfois flottante, et les termes *données*, *jeu de données*, *ensemble de données* et *package* sont employés de manière plus ou moins interchangeables. Dans le cadre de notre recherche, nous opterons pour *ensemble de données*. Chaque

D'une part, le découpage et la structuration des jeux semblent laissés à la discrétion des organisations. Le *Guide des métadonnées* du portail de Montréal formule des recommandations générales au sujet de la granularité des jeux, et aucun document de ce type n'a été rendu public au niveau provincial. Quoi qu'il en soit, les administrateurs de portails ne semblent procéder à aucune uniformisation *a posteriori*. Les données peuvent donc se voir regroupées, par exemple, sur une base annuelle (« Dépenses par programme et par centre d'activités 2008-2009 ») ou historique (« Ventes au détail de l'industrie bioalimentaire 2004-2011 »). De telles disparités nuisent naturellement au calcul de la couverture temporelle moyenne des ensembles de données (voir section 3.2.1 *couverture temporelle*) ou de la contribution respective des organisations (voir section 3.2.1 *contributeurs*).

D'autre part, le nombre de documents publiés par une organisation ne constitue pas non plus une mesure fiable, puisque certaines organisations multiplient les formats de documents (voir à ce sujet la section 3.2.2).

Couverture temporelle et géographique des données

Les portails de Montréal et du Québec possèdent tous deux la métadonnée *temporal*, qui permet d'indiquer la couverture temporelle d'un ensemble de données. Un examen de *Données Québec* révèle cependant que ce champ contient une valeur dans 271 des cas à peine, contre 490 champs vides. Les contributeurs du niveau provincial n'utilisent que rarement la métadonnée *temporal*, et la majorité des valeurs proviennent des portails de Montréal (157) et Gatineau (91)[22].

---

ensemble de données peut contenir plusieurs documents (données en plusieurs parties ou formats, documentation, etc.).

[22] Il est possible que cette faible utilisation de la métadonnée *temporal* reflète simplement la nature non-diachronique des données. Signalons qu'environ 25% des documents disponibles sur *Données Québec* sont en format géospatial. De plus, la recherche de divers patrons de couverture temporelle dans les titres et descriptions ne nous a permis de repérer que 82 ensembles de données diachroniques dont la métadonnée *temporal* ne contenait aucune valeur.

Après avoir sélectionné les instances provenant de Montréal et du Gouvernement du Québec, nous avons retranché les occurrences qui ne représentaient pas des durées calculables[23], afin de conserver 109 valeurs qui indiquent une couverture temporelle moyenne de 2,59 années. Cette valeur est bien inférieure à la date de création des portails, ce qui tend à confirmer que la publication rétroactive de données est un phénomène marginal.

Cela étant dit, plusieurs organisations ont décidé de recourir à la syndication plutôt que de publier des données historiques. Parmi les cas notables, signalons l'état des stations BIXI, la situation des salles d'urgence du Québec[24], ainsi que les acquisitions de BAnQ. Ce choix favorise le développement logiciel au détriment de l'analyse diachronique : une application de localisation de vélos en libre service bénéficiera à l'évidence d'un flux de données en temps réel, mais les chercheurs ne pourront pas effectuer une analyse historique de l'utilisation des vélos à partir de ces données sans cesse fluctuantes.

La couverture géographique, quant à elle, n'est réellement consignée que sur le portail de Montréal, où elle prend généralement la forme d'acronymes contrôlés dans le champ *territoire*, mais aussi, à l'occasion, de mots-clés — ce qui atténue légèrement la précision des résultats. En outre, une requête par l'API sur le champ *territoire* révèle une utilisation plus ou moins cohérente des termes « Montréal » et « agglomération ». Si l'on ignore ces deux termes englobants, il est possible d'obtenir une vue d'ensemble de la représentation des arrondissements (tableau 1).

---

[23] Ces valeurs pouvaient être une date unique (« 2017-03-12 »), une année unique (« 2015 ») ou des valeurs ambigues (« 1987-1988 / 2007 / 2015 / 2016 », « 02-06-05 »).

[24] Il existe des données historiques des urgences pour certaines régions administratives, mais ces données ne sont pas répertoriées sur le portail *Données Québec*, et ne sont pas fournies en formats orientés machine.

Tableau 1. Représentation des arrondissements de Montréal par métadonnée *territoire*

Ahuntsic-Cartierville	72
Anjou	66
Côte des Neiges	86
L'Île-Bizard-Sainte-Geneviève	64
Lachine	70
LaSalle	73
Le Sud-Ouest	68
Mercier–Hochelaga-Maisonneuve	77
Montréal-Nord	70
Outremont	63
Pierrefonds-Roxboro	61
Plateau Mont-Royal	74
Rivière-des-Prairies–Pointe-aux-Trembles	79
Rosemont–La Petite-Patrie	77
Saint-Léonard	71
St-Laurent	70
Verdun	64
Ville-Marie	79
Villeray–Saint-Michel–Parc-Extension	73
moyenne	**71**

Bien que la répartition des arrondissements semble assez régulière, il faut garder à l'esprit que ces chiffres peuvent cacher des déséquilibres que seule une analyse détaillée permettrait de repérer. À titre d'exemple, mentionnons les données sur les documents des bibliothèques de Montréal (Ville de Montréal, s.d.-a) et l'historique des conditions des patinoires (Ville de Montréal, s.d.-c), deux ensembles auxquels les arrondissements n'ont pas commencé à contribuer simultanément. La couverture géographique varie donc en fonction de la période considérée, et cette variation devra être prise en compte afin d'éviter les distorsions lors de l'analyse des données.

Contributeurs

Les données publiées sur *Données Québec* proviennent de neuf organisations principales, qui contribuent au portail dans des proportions inégales (voir tableau 2).

Tableau 2. Contributions des organisations principales au portail *Données Québec*

Gouvernement du Québec	301	37,8 %
Montréal	234	29,4 %
Gatineau	92	11,6 %
Ville de Québec	53	6,7 %
Sherbrooke	41	5,2 %
Longueuil	30	3,8 %
Laval	19	2,4 %
Rimouski	17	2,1 %
Blainville	9	1,1 %
	796	**100 %**

L'administration provinciale constitue le plus important contributeur du portail, avec 301 ensembles provenant de trente organisations (voir tableau 3).

Il nous a été impossible de quantifier avec précision la contribution des différentes entités administratives au portail de Montréal. En effet, la métadonnée *organization* contient presque systématiquement la valeur « Ville de Montréal », et les métadonnées *author* et *maintainer* peuvent contenir indifféremment une unité administrative (« Service du greffe »), un individu (« Marc Lebel ») ou un territoire (« Arrondissement LaSalle »). Quant à la métadonnée *territoire*, elle ne s'avère pas plus utile : chaque occurrence comporte entre une et vingt et une valeurs, les plus fréquentes étant simplement « Montréal » ou « Agglomération »[25].

[25] L'hétérogénéité de ce champ découle possiblement des épisodes de centralisation et de décentralisation qui se sont produits à Montréal : les données sont parfois publiées par les arrondissements eux-mêmes (sous forme d'ensembles indépendants) et parfois par la ville centrale (sous une forme regroupée).

Tableau 3. Contributions des organisations provinciales au portail *Données Québec*

Institut de la statistique du Québec	41	13,6 %
Société de l'assurance automobile du Québec (SAAQ)	40	13,3 %
Agence métropolitaine de transport	21	7,0 %
Transports Québec	20	6,6 %
Ministère des Forêts, de la Faune et des Parcs	19	6,3 %
Bibliothèque et Archives nationales du Québec	17	5,6 %
Agriculture, Pêcheries et Alimentation	16	5,3 %
Ministère de la Culture et des Communications	16	5,3 %
Ministère de la Sécurité publique	15	5,0 %
Ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques	12	4,0 %
Santé et services sociaux	12	4,0 %
Ministère du Travail, de l'Emploi et de la Solidarité sociale	10	3,3 %
Financière agricole du Québec	9	3,0 %
Ministère de l'Énergie et des Ressources naturelles	8	2,7 %
Secrétariat du Conseil du trésor	8	2,7 %
Ministère du Tourisme	7	2,3 %
Affaires municipales et Occupation du territoire	4	1,3 %
Commission de toponymie	4	1,3 %
Commission de protection du territoire agricole	3	1,0 %
Musée de la civilisation	3	1,0 %
Musée national des beaux-arts du Québec	3	1,0 %
Curateur public du Québec	2	0,7 %
Ministère de l'Éducation et de l'Enseignement supérieur	2	0,7 %
Musée d'art contemporain de Montréal	2	0,7 %
Retraite Québec	2	0,7 %
Institut national de santé publique du Québec (INSPQ)	1	0,3 %
Le Musée des beaux-arts de Montréal	1	0,3 %
Registraire des entreprises	1	0,3 %
Société du Grand Théatre	1	0,3 %
Sûreté du Québec	1	0,3 %
	301	**100 %**

3.2.2. **Formats**[26]

Lisibilité machine

Le gouvernement du Québec et la ville de Montréal utilisent essentiellement le même cadre de référence, qui accorde la priorité aux documents orientés-machines (section 2.1.1), généralement en format textuel (voir tableau 4). Sauf exception, notamment en géomatique, on ne recommande pas la publication de documents compressés (Ville de Montréal, 2016b).

Tableau 4. Formats de données ouvertes recommandés au gouvernement du Québec et à Montréal (Ville de Montréal, 2016b)

Type de données	Directives
Données tabulaires	Recommandé : CSV Acceptables sous condition : formats Excel ou ODS
Hiérarchiques	Recommandé : JSON Acceptable sous condition : XML
3D	Recommandé : CityGML
Géomatiques	Recommandé : GeoJSON Acceptable : CSV, KML, Shapefile
Images géoréférencées	Recommandé : GeoTIFF

On précise aussi que les images et les textes ne constituent pas des données, mais qu'en cas de nécessité — pour les documents complémentaires, en particulier — on acceptera certains formats d'image (JPG, PNG et TIFF) ou de texte (PDF, DOCX et ODT). Il est difficile de mesurer avec précision l'application de cette directive, puisque la métadonnée

[26] On trouvera un lexique des formats de données et de documents à l'annexe 2.

resource_type, qui permettrait de distinguer les documents complémentaires, est vide dans 95 % des cas, et ne contient que les valeurs « file » ou « file.upload » dans les 5 % restants.

L'ensemble des documents aux formats recommandés ou acceptables totalise 60 % des ressources disponibles sur Données Québec, et les documents en format texte (CSV, XML, KML, JSON et GeoJSON) représentent 41 %. Ces faibles proportions s'expliquent par la présence de quatre formats non recommandés — DOC, ODT, PDF et ZIP — qui comptent pour 26 % des documents. L'importante présence de documents compressés est en partie associée aux documents géomatiques, et celle des documents DOC, ODT et PDF s'explique par les nombreux rapports et procès-verbaux publiés par la ville de Montréal, ainsi que par un certain nombre de documents complémentaires. Si l'on retranche ces documents du calcul, le taux de formats recommandés ou acceptables grimpe à 73 %, et celui des formats textuels à 56 %.

Dans l'ensemble, en somme, les portails de Montréal et du gouvernement du Québec répondent aux besoins communément admis en matière de lisibilité-machine (section 2.1.1). Rappelons néanmoins que cette lisibilité ne dépend pas seulement du format des documents, mais aussi de la validité et du format des données qu'ils contiennent.

Formats ouverts

Les lignes directrices de *Données Québec* n'énoncent aucune recommandation explicite par rapport aux formats ouverts, bien que le Directeur principal de l'information supervise le Centre d'expertise en logiciel libre, dont le rôle est de promouvoir l'utilisation des logiciels libres au sein de la fonction publique. La situation est similaire à Montréal, où le *Guide des formats* stipule même que le format propriétaire XLSX est recommandable, cependant qu'ODS est simplement acceptable.

Une analyse détaillée de *Données Québec* et du portail de Montréal dévoile une utilisation plus ou moins uniforme des formats ouverts. En géomatique, le format propriétaire Shapefile occupe une place importante (220 documents), mais est largement dépassé par les formats ouverts KML, KMZ et GeoJSON (317 documents). Quant aux feuilles de calcul, elles sont en majeure partie en formats propriétaires : on trouve, sur le portail de Montréal, 143 documents XLS ou XLSX, contre 100 documents ODS. À l'échelle de la province, cette

proportion est de 281 contre 100 — ce qui permet de déduire que, dans le domaine des données ouvertes à tout le moins, le format ODS n'a été adopté qu'à Montréal, et que de manière partielle.

On constate en somme un certain écart entre l'importance des formats ouverts dans le discours (Open Knowledge Foundation, 2012; Sunlight Foundation, 2010; Ville de Montréal, 2016b) et dans la pratique.

3.2.3. Accès et accessibilité

Découvrabilité et organisation

Données Québec recense les données de huit villes et de trente organisations, mais son rôle demeure essentiellement celui d'un catalogue : les données en tant que telles sont hébergées sur les serveurs des contributeurs. Nous n'avons pas remarqué de mesures notables afin de mettre en relation les différents ensembles de données, et qui auraient apporté une valeur ajoutée : même regroupées sur un portail commun, les données demeurent organisées en silos thématiques ou organisationnels.

Les portails du Québec et de Montréal proposent un moteur de recherche similaire, assez limité, composé d'un champ de requête unique dont les résultats peuvent être raffinés à l'aide de différentes facettes : thématique, format, mots-clés et organisation (voir figure 10). Parmi les limites notables de cette interface, signalons que les facettes peuvent être cumulées, mais pas exclues.

Figure 10. Interface de recherche du portail de Montréal.

Les usagers qui désirent effectuer des recherches plus complexes doivent utiliser l'API CKAN, qui permet d'envoyer des requêtes par le truchement de la barre d'adresse du navigateur ou en utilisant un langage de programmation. Cette approche est relativement peu documentée, notamment en ce qui a trait à la structure de la base de données sous-jacente. S'il est relativement facile d'obtenir la structure par défaut de CKAN ou de DCAT, ni Québec ni Montréal n'ont en revanche diffusé les altérations locales aux métadonnées. Certaines requêtes par l'API peuvent donc exiger un certain tâtonnement qui ajoute à la complexité de la manœuvre.

Métadonnées

La structure de métadonnées des portails du Québec et de Montréal a été développée de manière collaborative, et s'inspire de la norme DCAT, que nous avons décrite à la section 2.1.1 (Ville de Montréal, s.d.-e). Il s'agit donc d'une structure normalisée, adaptée aux données ouvertes. Cela étant dit, la qualité des métadonnées ne dépend pas uniquement de la

structure, mais aussi du contenu. À ce chapitre, l'examen des deux portails permet de repérer plusieurs faiblesses et incohérences.

Si certaines métadonnées — le classement thématique par exemple — se composent d'expressions contrôlées, le contenu d'autres métadonnées est en revanche laissé à la discrétion des publieurs. C'est visiblement le cas pour les mots-clés, où l'on trouve plusieurs dédoublements causés par l'absence d'uniformisation ou par des coquilles. Les mots-clés sont particulièrement hétérogènes sur *Données Québec*, ce qui s'explique sans doute par la multiplication des organisations d'où proviennent les données. À titre d'exemple, nous avons pu repérer les mots-clés « Procès-verbal », « procès-verbal », « Procès verbaux », « procès-verbaux », « Procès-verval » et « pv ».

Nous avons également signalé, à la section 3.2.1 les problèmes occasionnés par des champs contenant des valeurs incohérentes — tels que *organization*, *author*, *maintainer* — ou par des champs sous-utilisés — tels que *territoire* ou *resource_type* —, et qui peuvent s'expliquer par une documentation incomplète ou imprécise.

Dans certains cas, la documentation n'est pas en faute : le contenu et le format du champ sont bien spécifiés, mais les valeurs saisies ne sont pas validées lors de la publication. C'est le cas du champ *temporal*, qui doit normalement contenir deux dates au format ISO8601 séparées par une barre oblique (Ville de Montréal, s.d.-e). Dans les faits, plusieurs enregistrements ne contiennent qu'une seule date, avec ou sans barre oblique, ou encore des dates incomplètes ou ambiguës (voir note 23, p. 57). Dans certains cas, les publieurs ont plutôt choisi de consigner ces informations dans le titre ou la description des ensembles de données, ce qui se traduit par un appauvrissement des métadonnées.

Nous avons par ailleurs constaté une perte de précision lors de la syndication des métadonnées qui décrivent la couverture géographique : le champ *territoire* qui, à Montréal, décrits les ensembles de données au niveau de l'arrondissement, est uniformisé au terme « Montréal » sur *Données Québec*.

Ces incohérences dans la saisie ou le traitement des métadonnées découlent vraisemblablement du grand nombre de contributeurs, comme nous l'avons vu à la section 3.2.1, et de la distribution des tâches de publication.

Documentation, outils de rétroaction et de support

Les portails de Montréal et du Québec ne se démarquent pas en ce qui a trait aux relations avec l'usager. Dans les deux cas, on recourt aux pratiques de base : documentation sommaire, formulaire ou adresse courriel afin de communiquer avec les administrateurs, fil d'actualité, médias sociaux. L'API est peu ou pas documentée : on se contente de renvoyer à la documentation de CKAN.

Les dispositifs de rétroaction sont un peu plus développés sur le portail de Montréal, où un outil de discussion permet aux usagers de commenter les ensembles de données. Cette fonctionnalité est toutefois peu utilisée : l'activité entre janvier 2014 et février 2017 ne se compose que de 26 discussions totalisant 110 commentaires, soit 3 commentaires par mois en moyenne (Disqus, s.d.). Par ailleurs, l'équipe des données ouvertes a pris contact avec ses usagers à l'occasion d'événements publics (« Journée internationale des données ouvertes 2016 : autoportrait du milieu montréalais des données ouvertes », 2016) ou en recrutant des citoyens-testeurs (Guidoin, 2015).

Sur les deux portails, enfin, on invite les usagers à suggérer des ensembles de données qu'ils souhaiteraient voir libérer. Montréal divulgue la liste de ces demandes, mais il s'agit d'une simple compilation automatisée, sans indication de statut : bien que les demandes les plus anciennes remontent à 2013, on ne précise pas si les données ont déjà été libérées ou non, quel est l'échéancier de publication estimé, ou si un quelconque motif pourrait empêcher ou limiter la libération.

3.2.4. **Licence**

Toutes les organisations qui participent au portail Données Québec ont libéré leurs données sous la licence *Attribution 4.0 International* de Creative Commons (CC-BY), ce qui constitue une pratique normale dans le domaine des données ouvertes.

CC-BY est une licence très peu contraignante, qui permet aux usagers de « copier, distribuer et communiquer le matériel par tous moyens et sous tous formats », de « remixer, transformer et créer à partir du matériel », pour toutes les formes d'utilisation, y compris commerciales (Creative Commons, s.d.). L'usager doit cependant respecter des conditions

d'attribution, qui consistent à créditer l'organisation qui a libéré les données, faire un lien vers le texte de la licence CC-BY, et indiquer les modifications apportées au matériel.

Signalons enfin que, dans le cas du portail de Montréal, la licence CC-BY se double d'une autorisation à verser des données dans *OpenStreetMap*.

3.2.5. **Évaluation des données**

L'évaluation de la qualité des données et des processus de publication semble relever de l'évidence. Plusieurs intervenants ont notamment fait valoir l'importance de déterminer des ensembles de données clés, que toute administration devrait libérer par défaut (Caplan et al., 2014; Gouvernement du Canada, 2015a; Ubaldi, 2013). Pourtant ni Montréal ni Québec n'ont divulgué clairement leurs objectifs à ce chapitre. Dans la politique de données ouvertes de Montréal, on se contente d'affirmer que la ville s'engage à « prendre les moyens qu'elle juge raisonnables afin d'adhérer aux principes de transparence et de qualité tels qu'énoncés par la Sunlight Foundation[27] » (Ville de Montréal, 2016d).

Le CEFRIO, l'organisme « mandaté par le gouvernement du Québec afin de contribuer à l'avancement de la société québécoise par le numérique » (CEFRIO, s.d.), dirige depuis 2015 un projet de recherche qui vise notamment à l'élaboration d'un « baromètre quantitatif qui donnerait une appréciation annuelle des progrès accomplis dans l'ouverture des données et des questions connexes à cette ouverture » (CEFRIO, 2016). Le rapport final met en lumière un avancement très inégal des données ouvertes, qui s'explique par de multiples « freins » — manque de budgets, d'expertise, de cadre de gouvernance, de données adéquates, etc. —, et qui révèle « une réticence [des organisations] à libérer des données gouvernementales, en contraste avec la volonté gouvernementale » (CEFRIO, 2017).

---

[27] La *Sunlight Foundation* est une organisation à but non lucratif américaine fondée en 2006, et qui a notamment énoncé dix principes pour guider l'ouverture des données gouvernementales (Sunlight Foundation, 2010).

Le rapport propose notamment une évaluation de la qualité des données libérées par les organisations et municipalités du Québec, à partir de critères qui recoupent partiellement ceux de la *Sunlight Foundation*, et qui laisse entrevoir des lacunes notamment dans la mise à jour des données, dans l'utilisation des métadonnées et dans la diversité des formats. Ce bilan doit néanmoins être considéré avec prudence, puisque les chiffres ont été recueillis grâce à un formulaire d'auto-évaluation complété par les dirigeants de l'information et les responsables des ressources informationnelles.

En somme, la situation québécoise semble similaire à celle qui prévaut ailleurs : comme on l'a vu à la section 2.2.4, un certain nombre de cadres d'évaluation existent — ou sont en cours de développement —, mais aucun ne s'est encore imposé comme norme d'évaluation répandue (Caplan et al., 2014).

3.3. **Applications**

Il existe peu d'études sur la variété et le nombre d'applications des données ouvertes. Notre revue de la littérature a surtout permis de repérer des études assez générales basées sur des échantillons d'applications (Kassen, 2013) ou sur des témoignages d'intervenants du milieu (Dyson et Goldstein, 2013). On ne s'étonnera guère de l'absence presque totale d'études du genre portant sur la situation québécoise.

Ce faible nombre d'études est sans doute imputable à la grande dispersion des applications, qui complique la tâche de recensement. Les galeries d'applications des portails de données ouvertes, qui constituent la première étape obligée d'un inventaire, n'ont en effet rien d'exhaustif. Les licences d'utilisation exigent des développeurs qu'ils attribuent la source des données utilisées, mais ne les contraignent pas à se manifester auprès des administrateurs d'un portail. En outre, les licences ne spécifient pas un format de citation normalisé qui permettrait un repérage systématique des applications (Verhulst et al., 2014, p. 53). Les galeries d'applications résultent, en somme, d'un processus de divulgation purement volontaire, et elles s'avèrent par conséquent lacunaires. Pour recenser les applications de manière plus poussée, il faut faire preuve de créativité. La méthodologie d'une étude effectuée par Karen Okamoto est, à cet égard, révélatrice : après avoir entamé ses recherches avec la

galerie d'applications et le compte Tumblr du portail de données ouvertes de New York, l'auteure a étendu ses recherches à Twitter et, par extension, aux sites Web d'organisations spécialisées dans les données ouvertes (Okamoto, 2016).

Enfin, les galeries d'applications souffrent d'un biais méthodologique : on n'y recense que les applications Web ou mobiles. Les applications analytiques (journalisme de données, recherche académique, etc.) ne sont pas prises en considération, ce qui donne un portrait des lieux non seulement incomplet, mais orienté.

Les applications Webs et mobiles développées à partir des données ouvertes de Montréal ou du gouvernement du Québec constituent essentiellement des applications d'interface (voir section 2.3), où l'on observe les mêmes pôles d'intérêt que dans les autres portails au Canada et à l'étranger :

- la mobilité : stationnements, transports en commun, partage de vélos, entraves à la circulation, etc. ;
- infrastructures et services : installations sportives ou culturelles, arbres, etc. ;
- transparence et imputabilité : octroi de contrats, budgets, appels d'offres, etc.

Au-delà de ces applications conventionnelles et relativement bien connues, nous nous sommes intéressés aux applications analytiques telles que le journalisme de données et la recherche académique.

Les médias d'information québécois, en particulier, font un usage croissant des données ouvertes, comme en témoigne une récente étude où les auteurs ont repéré quelque 178 occurrences de journalisme de données entre 2011 et 2013 (Tabary, Provost et Trottier, 2016). L'étude souligne non seulement le rôle prédominant des données institutionnelles dans ce phénomène, mais en particulier celui des données ouvertes, sans toutefois fournir de proportions chiffrées. Les auteurs signalent par ailleurs divers problèmes vécus par les journalistes qui utilisent les données ouvertes, en particulier les mises à jour trop peu fréquentes, une granularité insuffisante, ainsi que la sélection par les administrations de données « bon enfant », susceptibles de ne créer aucun remous (Roy, 2013).

Le repérage des utilisations des données ouvertes dans les médias d'information permet de constater l'importante présence des représentations cartographiques, qui reflète assurément la place des données géolocalisées sur les portails de données ouvertes. Les textes repérés portent par exemple sur la distribution des accidents de la route impliquant des cyclistes (Shiab, 2015), des piétons (Provost, 2016) ou des autobus (Shiab, 2016), sur la localisation des entrées par effraction (Cameron et Halsey-Watkins, 2016) ou sur la distance entre les axes routiers et les institutions éducatives pour enfants (Yates et Shiab, 2016).

Les données ouvertes ne constituent pas toujours la ressource principale de l'analyse : elles servent parfois à organiser d'autres données. C'est le cas d'une analyse démographique faite par Radio-Canada à partir des données de l'Enquête nationale auprès des ménages (ENM), où les résultats ont été organisés en fonction des stations de métro (Tremblay, Rocha, Salcido, Julien et Guiamaraes, 2016). Cette utilisation des données géospatiales permet non seulement de structurer le reportage, mais aussi, par ricochet, de formuler des hypothèses au sujet du rôle des transports en commun dans la distribution des populations, hypothèses qui n'auraient pu être faites sur la seule base de l'ENM.

Repérer les applications académiques des données ouvertes pose plusieurs problèmes : alors que la majorité des occurrences de journalisme de données au Québec se retrouvent dans une dizaine de médias, les études académiques sont publiées sur de nombreuses plateformes (périodiques, actes de colloques, sites Web, etc.) qui relèvent d'une grande variété de disciplines (géographie, urbanisme, informatique, santé publique, environnement, etc.). La nature disciplinaire des bases de données académiques et les performances insatisfaisantes des moteurs de métarecherche compliquent considérablement le repérage des sources qui nous intéressent.

Après plusieurs recherches préliminaires dans différentes bases de données, nous avons opté pour Google Scholar qui, en dépit de son interface limitée et de son corpus parfois hétéroclite, offre à la fois un large spectre disciplinaire et la recherche en plein texte. Inspirés par les travaux de GovLab (Verhulst et al., 2014), nous avons utilisé les URL des portails de données ouvertes en guise de citations normalisées, ce qui nous a permis de repérer 12 documents pertinents (voir annexe 1). Il s'agit d'un ensemble de résultats non seulement très restreint, mais partiellement représentatif puisque les données ouvertes y occupent une place

variable : il s'agit parfois de la source principale de données, et parfois d'une simple source complémentaire.

Ces résultats mitigés s'expliquent peut-être par l'aspect trop restrictif de notre méthode de recherche : les auteurs ont pu citer leurs sources sans utiliser d'URL, ou encore utiliser l'URL d'un site miroir. Des requêtes avec un plus grand taux de rappel nous ont en effet permis de repérer des études dignes d'intérêt, basées sur des ensembles de données issus des administrations montréalaise et provinciale, et obtenus par d'autres moyens que les portails de données ouvertes (numérisation, rapports, demandes d'accès à l'information, etc.). Bien que ces études illustrent le potentiel analytique des données de l'administration publique, elles ne témoignent pas de la portée des initiatives d'ouverture des données.

Parmi les cas dignes d'intérêt, signalons les travaux associés à la chaire de recherche Mobilité de l'École Polytechnique de Montréal. La liste des mémoires, thèses et publications de ce groupe révèle une utilisation importante de données issues d'organisations comme la STM ou l'AMT (École polytechnique de Montréal, s.d.), dont certaines sont publiées (ou devraient être publiées) sur les portails de données ouvertes. Les données ouvertes de la ville de Montréal ont également fait l'objet d'un cours d'apprentissage machine du département informatique de l'Université McGill (Pineau et Bacon, 2015). L'initiative ne semble pas avoir été répétée, ce qui s'explique peut-être par les limites des ensembles de données, dont il a été question dans la section 2.3.2.

3.4. **Retombées**

Comme nous l'avons vu à la section 2.2.4, il n'existe pas de méthode ou de norme établie afin d'évaluer les retombées d'un programme d'ouverture de données. Nous avons tout de même voulu déterminer si Montréal ou Québec avaient développé leurs propres cadres d'évaluation, sans résultat.

Non seulement la Ville de Montréal n'a diffusé aucun cadre d'évaluation, mais la *Directive sur la gouvernance* ne précise pas à quel service incombe le mandat d'évaluation. Selon le *Plan d'action 2015-2017 ville intelligente et numérique*, Montréal doit mettre en place des indicateurs de performance dont plusieurs devraient normalement toucher les

données ouvertes. On ne précise pas, cependant, si ces indicateurs permettront seulement d'évaluer l'atteinte d'objectifs de gestion, ou s'ils serviront aussi à évaluer le retour sur investissement.

Il a également été impossible de repérer un cadre d'évaluation au gouvernement du Québec. Le plus récent rapport de gestion annuel du secrétariat du Conseil du trésor (Secrétariat du Conseil du trésor et Gouvernement du Québec, 2015) souligne la réalisation d'un certain nombre d'objectifs, notamment la « libération d'ensembles de données à haute valeur », mais ne fait aucune mention d'indicateurs de performance ou de cadre d'évaluation de retombées. Les rédacteurs n'ont cependant pas défini l'expression « haute valeur », ce qui trahit l'idée discutable selon laquelle la valeur est inhérente aux données et non à l'utilisation qu'on en fait. Le rapport mentionne que ces jeux doivent être identifiés à l'aide d'un « plan de diffusion de données ouvertes » en cours d'élaboration, mais nous n'avons pas pu déterminer de quel document il s'agissait, ni si l'on avait achevé sa rédaction depuis.

L'organisme mandaté par le gouvernement du Québec, le CEFRIO, a publié à l'hiver 2017 un rapport intitulé *Mesure de l'apport et de l'évolution du gouvernement ouvert au Québec*. En dépit de son titre, ce rapport ne propose aucune véritable mesure des retombées : les auteurs se limitent à faire la synthèse des retombées potentielles évoquées dans la littérature, et à relayer une étude de cas portant sur les données ouvertes de Londres (CEFRIO, 2017).

En conclusion, la mise en œuvre des données ouvertes au Québec s'est visiblement inspiré des initiatives internationales et des bonnes pratiques reconnues dans le milieu, autant sur les plans technique et juridique que politique. Les portails ont connu une croissance rapide — à tout le moins dans leurs premières années d'opérations —, mais la qualité des données demeure cependant inégale, et une mesure adéquate des retombées reste à faire. Cette situation se traduit par des ensembles de données qui ne sont pas toujours optimalement utilisables, et à un nombre restreint d'ensembles de données se prêtant à des analyses poussées.

Le chapitre suivant découle directement de ces constats, et vise à identifier et démontrer les manipulations qui permettent aux usagers de corriger certaines lacunes observées dans les données ouvertes. Nous procéderons notamment à l'identification d'un ensemble de données adéquat, que nous soumettrons à un processus de prétraitement dans l'objectif d'obtenir un corpus dûment analysable. Nous effectuerons également la consolidation de deux corpus de données complémentaires, dans le but de générer un corpus plus volumineux et plus complexe, offrant de meilleures opportunités analytiques.

4. Traitement des données

Dans ce dernier volet de notre recherche, nous poursuivrons deux objectifs : 1) nettoyer et traiter deux ensembles de données afin d'améliorer leur potentiel analytique, et 2) croiser ces ensembles afin de générer un corpus ayant une valeur analytique accrue.

Le premier de ces objectifs procède de la qualité notoirement inégale des données ouvertes (section 1.3.1) et de la pratique de publication de données brutes (section 2.1.4). Ces deux phénomènes ont une incidence directe sur les activités des utilisateurs, comme nous le démontrerons avec diverses opérations de nettoyage et de traitement. En outre, les ensembles de données sélectionnés permettront de comparer deux tendances en matière de données ouvertes : la captation automatisée et la saisie manuelle.

Le second objectif nous a été inspiré par les considérations de Pineau et Bacon sur le croisement des ensembles de données (Pineau et Bacon, 2015), et par une analyse journalistique effectuée par Anne-Marie Provost sur les accidents impliquant des piétons sur l'île de Montréal (Provost, 2016).

L'analyse en question reposait sur des données compilées par la SAAQ à partir de rapports de police produits entre 2009 et 2015. À l'aide d'opérations statistiques, l'auteure mettait en évidence les intersections et les périodes où s'étaient produits les plus grands nombres d'accidents. Malgré l'intérêt évident des données utilisées, nous avons noté l'absence de données précises sur le volume de circulation aux intersections. Pour estimer le rôle de cette variable sur la fréquence des accidents, l'auteure devait s'en remettre à des généralités : circulation plus intense au centre-ville, trafic accru à l'heure de pointe en direction des ponts,

etc. Ces données existent pourtant : on les trouve dans le document *Feux de circulation – comptage des véhicules et des piétons aux intersections munies de feux*, disponible sur le portail de données ouvertes de Montréal (Ville de Montréal, 2013a). Nous avons par conséquent supposé qu'en croisant ces données avec celles de la SAAQ, nous pouvions générer un corpus qui décrirait de manière plus riche les intersections où les piétons sont victimes (ou non) d'accidents de la route.

4.1. **Description des données**

Afin d'évaluer le potentiel analytique des données sélectionnées, nous proposerons dans cette section une méthode de repérage des traits discriminants (section 4.1.1), que nous mettrons en application grâce à une description détaillée des données (section 4.1.2 et 4.1.3).

4.1.1. **Repérage des traits discriminants**

Afin de mesurer le potentiel analytique des données, nous tâcherons d'identifier leurs traits discriminants, c'est-à-dire les attributs ayant un potentiel de différenciation lors de l'analyse (Gorunescu, 2011, p. 13). Si, par exemple, nous voulons analyser les accidents aux intersections, les attributs temporels (heure, date, etc.) présenteront assurément un potentiel de différenciation, puisqu'ils permettront de dégager les patrons de variation pour les heures de la journée, les jours de la semaine, les saisons, et ainsi de suite.

Dans cette section, nous avons voulu identifier les attributs qui pourraient avoir un intérêt analytique de manière générale, indépendamment des objectifs d'analyse. Nous avons donc procédé par élimination, et avons identifié quatre caractéristiques qui permettent d'écarter les attributs incontestablement non discriminants :

1. Trop de valeurs manquantes : si un attribut comporte trop de valeurs manquantes, il ne se prêtera pas à l'analyse.
2. Dispersion trop faible : si un attribut comporte systématiquement la même valeur, il ne permet pas de différencier les instances de manière significative. À

titre d'exemple, le champ CD_MUNCP (voir tableau 5) contient le code correspondant à Montréal dans 95,4 % des instances.

3. Valeur redondante : un attribut peut à l'occasion contenir des valeurs déductibles d'un autre attribut. Par exemple, dans les données d'accidents (voir tableau 5) le contenu du champ ANNEE (« 2014 ») peut être extrait du champ DT_ACCDN (« 2014-05-12 »). Le champ ANNEE peut s'avérer utile lors des manipulations — en évitant une étape d'extraction —, mais il ne contribue pas au potentiel analytique des données.
4. Valeur non fiable : la documentation (ou l'absence de documentation) ne permet pas d'interpréter l'attribut.

Ces quatre caractéristiques nous ont permis de repérer les attributs non discriminants parmi les ensembles de données sélectionnés (voir tableaux 6 et 8), et d'identifier par élimination les attributs ayant un intérêt analytique potentiel.

4.1.2. **Accidents**

Les données fournies par la SAAQ incluent tous les accidents impliquant au moins un piéton et un véhicule autorisé à circuler sur la voie publique dans la région 06 (Montréal). Diffusées sous la forme d'un document XSLX, elles se composent de 21 156 instances, réparties entre janvier 2000 et octobre 2014. Chaque instance comporte 27 attributs, qui décrivent l'accident et son contexte (voir tableau 5).

Ces données ont été compilées à partir de rapports d'accident complétés par les policiers : outre d'éventuelles erreurs, on y trouvera des approximations (notamment dans les distances), des abréviations (« ND » au lieu de « Notre-Dame »), ainsi que des coquilles (« CAHTEAUBRIAND ») qui ont pu se produire lors de la saisie initiale des données ou de leur compilation. Dans certains cas, la conception inadéquate des formulaires a pu inciter les policiers à saisir des données dans des formats ou des champs imprévus. En outre, certains changements dans les données permettent de supposer que les formulaires ont fait l'objet de modifications au cours de la période visée. Ces nombreuses erreurs et incohérences compliqueront notamment l'intégration des données, comme on le verra à la section 4.2.4.

Par ailleurs, l'examen des données démontre que sur 27 attributs, à peine 8 présentent un potentiel de différenciation (voir tableau 6). Les deux principaux problèmes sont les valeurs manquantes ou insuffisamment dispersées.

Tableau 5. Attributs des données sur les accidents

Attribut	Description	Format
ANNEE	année de l'accident	numérique
DT_ACCDN	date de l'accident	aaaa-mm-jj
HR_ACCDN	heure de l'accident	hh:mm
ID_GR	gravité de l'accident	code numérique
NB_VEH_ACCDN_C	nombre de véhicules impliqués	numérique
VITSS_AUTOR	vitesse autorisée	numérique
CD_ENVRN_ACCDN	activité dominante du secteur	code numérique
CD_LOCLN_ACCDN	localisation longitudinale de l'accident	code numérique
CD_MUNCP	code de la municipalité	code numérique
IN_ACCDN_CHMIN_PBL	accident sur chemin public	oui/non/null
NO_ARRND_MUNCP	numéro d'arrondissement	code numérique
NO_ROUTE	numéro de route	numérique
BORNE_KM_ACCDN	borne kilométrique	numérique
CD_PNT_CDRNL_ROUTE	direction de la route numérotée	N/S/E/O/null
NO_CIVIQ_ACCDN	numéro civique	numérique
SFX_NO_CIVIQ_ACCDN	suffixe du numéro civique	numérique
RUE_ACCDN	rue	alphanumérique
TP_REPRR_ACCDN	type de repère	code numérique
ACCDN_PRES_DE	repère	alphanumérique
CD_PNT_CDRNL_REPRR	position par rapport au repère	N/S/E/O/null
NB_METRE_DIST_ACCD	distance du repère/no civique/borne	numérique
CD_CATEG_ROUTE	catégorie de route	code numérique
CD_CONFG_ROUTE	configuration de la route	code numérique
PIET_MORT	nombre de piétons décédés	numérique
PIET_GRAVE	nombre de piétons blessés gravement	numérique
PIET_LEGER	nombre de piétons blessés légèrement	numérique
TOTAL_PIET	nombre total de piétons victimes	numérique

Tableau 6. Élimination des attributs non discriminants parmi les données sur les accidents

Attribut	Description	Critères d'élimination
ANNEE	année de l'accident	Redondant.
DT_ACCDN	date	
HR_ACCDN	heure	
ID_GR	gravité de l'accident	Indice non documenté.
NB_VEH_ACCDN_C	nombre de véhicules impliqués	
VITSS_AUTOR	vitesse autorisée	
CD_ENVRN_ACCDN	activité dominante du secteur	Dispersion faible, documentation peu fiable.
CD_LOCLN_ACCDN	localisation longitudinale de l'accident	Dispersion faible.
CD_MUNCP	code de la municipalité	Dispersion faible.
IN_ACCDN_CHMIN_PBL	accident sur chemin public	Trop de valeurs manquantes (73 %).
NO_ARRND_MUNCP	numéro arrondissement	Trop de valeurs manquantes (96 %).
NO_ROUTE	numéro de route	Trop de valeurs manquantes (99 %).
BORNE_KM_ACCDN	borne kilométrique	Trop de valeurs manquantes (100 %).
CD_PNT_CDRNL_ROUTE	direction de la route numérotée	Trop de valeurs manquantes (100 %).
NO_CIVIQ_ACCDN	numéro civique	Trop de valeurs manquantes (100 %).
SFX_NO_CIVIQ_ACCDN	suffixe du numéro d'immeuble	Trop de valeurs manquantes (100 %).
RUE_ACCDN	rue	Redondant[28].
TP_REPRR_ACCDN	type de repère	Trop de valeurs manquantes (74 %).
ACCDN_PRES_DE	repère	Redondant.
CD_PNT_CDRNL_REPRR	position par rapport au repère	Trop de valeurs manquantes (87 %).
NB_METRE_DIST_ACCD	distance repère/no d'immeuble/borne	Dispersion faible.
CD_CATEG_ROUTE	catégorie de route	Dispersion faible, documentation peu fiable.
CD_CONFG_ROUTE	configuration de la route	Trop de valeurs manquantes (72 %).
PIET_MORT	nombre de piétons décédés	
PIET_GRAVE	nombre de piétons blessés gravement	
PIET_LEGER	nombre de piétons blessés légèrement	
TOTAL_PIET	nombre total de victimes piétonnes	Redondant.

Total des attributs éliminés : 20 sur 27

[28] Les attributs *rue_accdn* et *accdn_pres_de* sont considérés redondants par rapport à leurs coordonnées géographiques, que nous calculerons dans la section 4.2.3.

4.1.3. **Circulation**

Les données sur la circulation décrivent les passages de piétons et de diverses classes de véhicules à 1628 intersections. Ces données sont diffusées par la ville de Montréal en format CSV et contiennent 606 657 instances réparties entre mars 2001 et juin 2012. Chaque instance comporte 28 attributs (tableau 7).

Ces données ont été générées par des appareils de comptage, dans le cadre d'une mise aux normes des feux de circulation des 19 arrondissements de la ville de Montréal. Les observations portaient essentiellement sur des intersections dotées de feux de circulation, et à l'occasion sur des intersections où l'on envisageait l'installation de feux. Les périodes d'observation duraient typiquement entre 3 et 8 heures par jour, et se déroulaient en général aux heures de pointe. Les données se caractérisent par une grande cohérence formelle : la génération automatique implique *de facto* un vocabulaire contrôlé, des mesures précises et une absence de valeurs nulles.

Nous estimons que 22 des 28 attributs présentent un potentiel de différenciation (voir tableau 8). Signalons cependant que certains attributs discriminants doivent être considérés avec prudence, en particulier les champs indiquant le type de trafic (CODE_BANQUE et NOM_BANQUE), dont la couverture temporelle est inégale : les vélos, à titre d'exemple, n'ont été ajoutés qu'en 2009.

Tableau 7. Attributs des données sur la circulation

Attribut	Description	Format
UFEFFIDREFERENCE	numéro de session	code numérique
IDCARREFOUR	code de l'intersection	code numérique
NOMINTERSECTION	noms des rues de l'intersection	alphanumérique
LAT	latitude	NAD83 MTM 8
LONG	longitude	NAD83 MTM 8
DATE	date	aaaa-mm-jj
PÉRIODE	heure complète	hh:mm:ss
HEURE	heure	h
MINUTE	minute	m
SECONDE	seconde	s
CODE_BANQUE	type de trafic	code numérique
NOM_BANQUE	type de trafic	alphanumérique
NBLT	direction nord, virage à gauche	numérique
NBT	direction nord, tout droit	numérique
NBRT	direction nord, virage à droite	numérique
SBLT	direction sud, virage à gauche	numérique
SBT	direction sud, tout droit	numérique
SBRT	direction sud, virage à droite	numérique
EBLT	direction est, virage à gauche	numérique
EBT	direction est, tout droit	numérique
EBRT	direction est, virage à droite	numérique
WBLT	direction ouest, virage à gauche	numérique
WBT	direction ouest, tout droit	numérique
WBRT	direction ouest, virage à droite	numérique
APPROCHENORD	traverse nord	numérique
APPROCHESUD	traverse sud	numérique
APPROCHEEST	traverse est	numérique
APPROCHEOUEST	traverse ouest	numérique

Tableau 8. Élimination des attributs non discriminants parmi les données sur la circulation

Attribut	Description	Critères d'élimination
UFEFFIDREFERENCE	Numéro de session	Redondant.
IDCARREFOUR	Code de l'intersection	Redondant.
NOMINTERSECTION	Noms des rues de l'intersection	Redondant.
LAT	Latitude	Redondant[29].
LONG	Longitude	Redondant.
DATE	Date	
PÉRIODE	Heure complète	
HEURE	Heure	Redondant.
MINUTE	Minute	Redondant.
SECONDE	Seconde	Redondant.
CODE_BANQUE	Type de trafic	
NOM_BANQUE	Type de trafic	Redondant.
NBLT	Direction nord, virage à gauche	
NBT	Direction nord, tout droit	
NBRT	Direction nord, virage à droite	
SBLT	Direction sud, virage à gauche	
SBT	Direction sud, tout droit	
SBRT	Direction sud, virage à droite	
EBLT	Direction est, virage à gauche	
EBT	Direction est, tout droit	
EBRT	Direction est, virage à droite	
WBLT	Direction ouest, virage à gauche	
WBT	Direction ouest, tout droit	
WBRT	Direction ouest, virage à droite	
APPROCHENORD	Traverse nord	
APPROCHESUD	Traverse sud	
APPROCHEEST	Traverse est	
APPROCHEOUEST	Traverse ouest	

Total des attributs éliminés : 9 sur 28

[29] Les champs LAT et LONG sont considérés redondants par rapport aux attributs géospatiaux que nous calculerons dans la section 4.2.3.

4.2. **Manipulation des données**

Nous entamerons cette section avec diverses considérations méthodologiques (section 4.2.1). Nous soumettrons ensuite nos deux ensembles de données à diverses opérations de traitement : le nettoyage (section 4.2.2), la création d'attributs (section 4.2.3) et l'intégration (ou jonction) des deux ensembles afin de former un corpus analysable unique (section 4.2.4).

4.2.1. **Considérations méthodologiques**

Reformatage des documents

Les données sur les accidents étant fournies au format Excel à raison d'une feuille par année, nous avons consolidé l'ensemble des données dans un seul document au format CSV, plus approprié pour les manipulations.

Outils

Étant données la taille considérable des documents et la nature de certaines manipulations, nous avons décidé d'effectuer le nettoyage des données et les manipulations géospatiales à l'aide de Python. La visualisation des données, particulièrement utile lors des phases de débogage, a été effectuée sur CARTO (CARTO, s.d.), un logiciel de cartographie en ligne. Enfin, l'intégration des données a été réalisée avec RapidMiner (RapidMiner, s.d.).

Ajout d'attributs

Toute manipulation de données pose des risques de corruption : il est possible d'effacer des valeurs par inadvertance, ou d'y introduire des erreurs. Par mesure de prudence, nous avons choisi de ne pas altérer les valeurs déjà présentes, mais plutôt de consigner les valeurs modifiées dans de nouveaux champs. En plus de préserver l'intégrité des données originales, cette méthode permet de comparer aisément les input et output de chaque opération.

4.2.2. Nettoyage

Normalisation typographique

Lors du reformatage des données sur les accidents, nous avons constaté quelques conflits de conversion causés par quatre instances, où certains attributs contenaient des nombres avec une virgule comme séparateur décimal. La virgule étant également le caractère séparateur de nos documents CSV, ces nombres étaient interprétés comme deux champs plutôt qu'un seul.

Vu le petit nombre de valeurs conflictuelles, nous avons simplement remplacé les virgules par des points. Si de trop nombreuses instances avaient causé des conflits, nous aurions plutôt produit nos fichiers CSV en utilisant un caractère séparateur différent.

Correction des erreurs linguistiques

Comme nous l'avons mentionné plus haut, les données sur les accidents contiennent un certain nombre de coquilles et d'erreurs. Le grand nombre d'instances et la variété des cas rendent pratiquement impossible toute forme de correction, qu'elle soit manuelle ou automatique. Nous avons cependant identifié certaines pratiques récurrentes — telles que l'utilisation de l'abréviation « ND » afin de désigner « Notre-Dame » — que nous avons pu corriger en série.

Sélection des instances pertinentes

Afin de créer un attribut de jointure (section 4.2.3), il nous faudra distinguer, parmi les 21 156 instances de départ, celles qui désignent des accidents s'étant produits à des intersections.

Selon la documentation, ces instances seraient identifiées grâce à la valeur 1 dans le champ TP_REPRR_ACCDN (type de repère) ou grâce à la valeur 32 dans le champ CD_LOCLN_ACCDN (localisation longitudinale de l'accident). Aucun de ces deux champs ne nous permet cependant d'identifier les intersections de manière satisfaisante : le champ

TP_REPRR_ACCDN n'a été utilisé qu'à partir de 2010[30], et le champ CD_LOCLN_ACCDN est constitué à 42 % de la valeur 21, qui est non documentée.

L'identification des intersections de manière analytique nous apparaît plus fiable. Pour ce faire, nous avons utilisé les champs suivants :

- RUE_ACCDN (rue sur laquelle a eu lieu l'accident);
- ACCDN_PRES_DE (rue à proximité de l'accident);
- NO_CIVIQ_ACCDN (numéro d'immeuble vis-à-vis duquel l'accident a eu lieu).

Lorsqu'un accident s'est produit à une intersection, l'instance doit normalement avoir une valeur dans les champs RUE_ACCDN et ACCDN_PRES_DE. En outre, aucun numéro d'immeuble ne devrait normalement être fourni, ce qui permet d'utiliser le champ NO_CIVIQ_ACCDN pour des fins d'exclusion.

L'examen de ces attributs permet toutefois de repérer un grand nombre de notations ambiguës qui trahissent une apparente confusion lors de la saisie des données (voir tableau 9).

Tableau 9. Exemples de notation ambiguë des intersections dans les données sur les accidents

	RUE_ACCDN	ACCDN_PRES_DE	Nbre de cas	% du total
A	St-Laurent et St-Joseph		599	2,8 %
B		St-Laurent et St-Joseph		
C	St-Laurent et St-Joseph	St-Laurent	36	0,2 %
D	St-Laurent	St-Laurent et St-Joseph		
E	6000 St-Laurent	St-Joseph	2740	13,0 %
F	6000 St-Laurent		813	2,9 %
G	Face au 6000 St-Laurent	St-Joseph	613	3,8 %
H	Face au 6000 St-Laurent		208	1,0 %
			5009	**23,7 %**

[30] Cette distribution déficiente pourrait être attribuable à un changement de formulaire.

Considérant le pourcentage élevé de l'ensemble de ces incohérences, nous avons soumis les données à trois séries de manipulations dans Python afin d'identifier les instances pertinentes, et de transférer les valeurs nettoyées dans deux nouveaux champs : AXE_0 et AXE_1. Ces manipulations sont les suivantes :

1. Nous avons identifié les instances où les deux valeurs recherchées étaient saisies dans un seul champ (tableau 9, exemples A, B, C et D). Nous avons ignoré les occasionnelles valeurs redondantes (tableau 9, exemples C et D), scindé les paires de valeurs pertinentes, et copié les valeurs résultantes dans les champs AXE_0 et AXE_1.

2. Partant du principe qu'un accident localisé grâce à une adresse d'immeuble ne s'est pas produit à une intersection, nous avons utilisé des requêtes en expressions régulières afin d'exclure du prétraitement toutes les instances où un numéro d'immeuble avait été erronément consigné dans le champ RUE_ACCDN plutôt que dans le champ NO_CIVIQ_ACCDN (tableau 9, exemples E, F, G et H). Lors de la construction de nos requêtes, nous nous sommes assurés de ne pas interpréter les voies numérotées (« 10e avenue ») comme des numéros d'immeuble.

3. Pour toutes les instances où les attributs désignaient sans ambiguïté une intersection, nous avons copié les valeurs de RUE_ACCDN et ACCDN_PRES_DE dans les attributs AXE_0 et AXE_1.

Ces trois séries d'opérations permettent de repérer 15 086 instances. Afin d'assurer un degré de précision supplémentaire, nous avons utilisé un quatrième attribut, NB_METRE_DIST_ACCD, qui indique la distance de l'accident par rapport au repère. Nous avons exclu toutes les instances où ce champ contient une distance supérieure à 5 mètres, ce qui correspond à la largeur approximative d'un corridor piétonnier; au-delà de cette distance, on ne peut avoir la certitude qu'un accident s'est effectivement produit à une intersection. Cette requête permet d'éliminer 12,8 % des instances repérées lors des étapes précédentes (voir tableau 10). Notre corpus final se compose donc de 13 148 instances, soit 62 % du corpus initial.

Tableau 10. Synthèse des valeurs du champ NB_METRE_DIST_ACCD.

distance	nombre d'instances	
valeur nulle	5803	38,5 %
0 mètre	6124	40,6 %
1 à 5 mètres	1222	8,1 %
plus de 5 mètres	1937	12,8 %
	15 086	**100 %**

4.2.3. **Création d'attributs**

L'intégration de données implique de fusionner les instances de deux corpus grâce à des paires d'attributs identiques. Il convient donc d'examiner chaque ensemble de données afin de repérer — ou de générer, comme nous le verrons — des attributs susceptibles de décrire, sous la même forme, une entité commune.

Après avoir étudié la possibilité de joindre nos données par le truchement d'attributs temporels (période quotidienne, hebdomadaire ou saisonnière), nous avons décidé d'utiliser plutôt les attributs géographiques, qui nous semblaient susceptibles d'indiquer des variations propres au contexte montréalais. Deux groupes d'attributs permettent d'effectuer une telle jointure : les coordonnées géographiques et les odonymes (noms propres de voie de circulation). Ces deux approches présentent chacune leurs difficultés propres.

Dans un premier temps, l'association des odonymes se heurte non seulement aux coquilles et erreurs de transcription, mais aussi aux très nombreuses variations de graphie : ponctuation, espaces, signes diacritiques, abréviations, orientation géographique, convention de notation et particules sont autant d'éléments susceptibles de varier (voir tableau 11). Dans certains cas, différents odonymes désigneront une même entité — lorsque, par exemple, une voie changera de nom à l'intersection. Effectuer une jointure fiable sur les odonymes impliquerait par conséquent un traitement complexe, qui devrait prendre en compte tous les types de variations possibles.

Tableau 11. Exemples de variations dans les odonymes

NOMINTERSECTION	RUE_ACCDN	ACCDN_PRES_DE
Décarie/Vézina inter. Est	Bd Decarie	Vezina
Cathédrale/Metcalfe /René-Lévesque	Cathedrale	Bd R Levesque O
Grandes-Prairies /Pie-IX	Gdes Prairies	Pie Ix
Crémazie/Saint-Denis inter. Sud	Cremzaie	St Denis
Lucien-L'Allier /Saint-Jacques	Lucien Lalier	St Jacques
Cathédrale/Notre-Dame	La Cathedrale	Notre-Dame
Champlain/Église de l'	Lesage	De L Eglise
L-H-Lafontaine (dir.Nord)/Yves-Prévost	L H Lafontaine	Yves Prevost
15 e Avenue/Jean-Talon	15E Av	Jean Talon

L'utilisation des coordonnées géographiques, d'autre part, pose un problème tout différent, puisque les données sur les accidents ne comportent aucun attribut géospatial. Cette jointure implique donc d'effectuer un géocodage, c'est-à-dire d'extraire des coordonnées géographiques à partir des odonymes.

Nous avons opté pour cette option qui, bien qu'assez lourde sur le plan de la programmation, a toutefois comporté un avantage inattendu : les algorithmes de géocodage de l'API Googlemaps (Google Maps, 2016) composent de manière très satisfaisante avec les coquilles et variations de graphie des odonymes. Cette approche permet, en d'autres mots, d'externaliser le nettoyage des données.

Afin de tirer parti de cette flexibilité d'interprétation, nous avons soumis à l'API des requêtes en langage naturel (« St-Laurent et Jarry, Montréal ») combinées avec le paramètre national (« 'country' : 'CA' »). Cette méthode a permis de résoudre les coordonnées de la quasi-totalité des intersections. Les coordonnées obtenues ont été consignées dans deux nouveaux attributs, LAT et LONG, et nous avons confirmé la qualité des valeurs fournies par Google Maps en les visualisant sur CARTO. Cette étape nous a en outre permis de repérer et éliminer quelques anomalies manifestes.

Dans un second temps, nous avons modifié les coordonnées des données de comptage, qui étaient consignées au format NAD83 MTM8. Cette projection n'étant pas supportée par la plupart des outils que nous entendions utiliser, nous l'avons converti à la projection WGS84 à l'aide du module *pyproj* (Whitaker, 2016). Les coordonnées résultantes ont été ajoutées dans deux nouveaux attributs : LAT_WGS84 et LONG_WGS84.

En théorie, les deux manipulations précédentes génèrent des paires d'attributs comparables. En pratique, cependant, ces coordonnées s'avèrent inutilisables telles quelles : leur degré de précision est de sept décimales — une résolution de l'ordre du centimètre. Cela signifie, concrètement, que les paires de coordonnées suivantes :

45.4571962, -73.8729735

45.4572034, -73.8729812

sont numériquement différentes, mais désignent en réalité la même intersection. La comparaison des valeurs à sept décimales obtenues lors des opérations précédentes donne donc un taux de rappel nul.

On peut remédier à la situation en arrondissant les coordonnées jusqu'à obtenir des correspondances. Une précision trop faible peut en revanche rendre les coordonnées non représentatives en désignant un point qui ne correspond à aucune intersection. La visualisation cartographique a permis de déterminer empiriquement que des coordonnées à quatre décimales constituaient le niveau résolution idéal pour joindre nos attributs : sous les quatre décimales, la précision s'avère insuffisante, et au-dessus de quatre décimales le taux de rappel est insignifiant (figure 11).

Figure 11. Géolocalisation des intersections à trois décimales (haut) et à quatre décimales (bas).

4.2.4. **Intégration**

Grâce aux multiples opérations de nettoyage, de reformatage et de géocodage décrites dans les sections précédentes, nous obtenons deux ensembles de données prêtes à être fusionnés. Avant de procéder à l'intégration, il importe toutefois de prendre en considération une différence cruciale entre les deux ensembles de données : le sens exact d'une valeur manquante.

En effet, les données sur la circulation ne contiennent aucune valeur nulle. S'il n'existe aucune instance avec la banque *piétons* pour la session de comptage *x*, cela ne signifie pas qu'aucun piéton n'est passé à cette intersection, mais simplement que l'on n'a pas comptabilisé les piétons durant cette session. L'absence de donnée n'équivaut donc pas à zéro. Une absence comptabilisée de piétons se traduirait par la présence d'une instance avec la banque *piétons*, et comportant la valeur 0 (voir tableau 12).

Tableau 12. Exemple d'une absence comptabilisée de piétons

UFEFFIDREFERENCE	5315
IDCARREFOUR	2
NOMINTERSECTION	Docteur-Penfield /Peel
LAT	298588.536
LONG	5040384.56
DATE	12-06-06
PÉRIODE	00:00:00
HEURE	0
MINUTE	30
SECONDE	0
CODE_BANQUE	10
NOM_BANQUE	Piéton
NBLT	0
NBT	0
NBRT	0
SBLT	0
SBT	0
SBRT	0
EBLT	0
EBT	0
EBRT	0
WBLT	0
WBT	0
WBRT	0
APPROCHENORD	0
APPROCHESUD	0
APPROCHEEST	0
APPROCHEOUEST	0

Le document d'accidents, au contraire, constitue un portrait complet de la réalité, et non un simple échantillonnage : il décrit présumément la totalité les accidents avec piétons s'étant produits durant la période et sur le territoire visés. Cela signifie que, pour toutes les intersections absentes de cet ensemble, nous pouvons déduire certaines valeurs : les attributs qui décrivent les nombres de victimes (PIET_LEGER, PIET_MORT et PIET_GRAVE) auront nécessairement une valeur de 0.

Selon le type d'analyse projeté, l'intégration de nos données pourra donc reposer sur deux types de jointures (figure 12). La jointure interne regroupera uniquement les instances qui décrivent des intersections communes aux deux ensembles. Le corpus final sera plus restreint, mais chacune des instances contiendra tous les attributs présents dans les ensembles de départ.

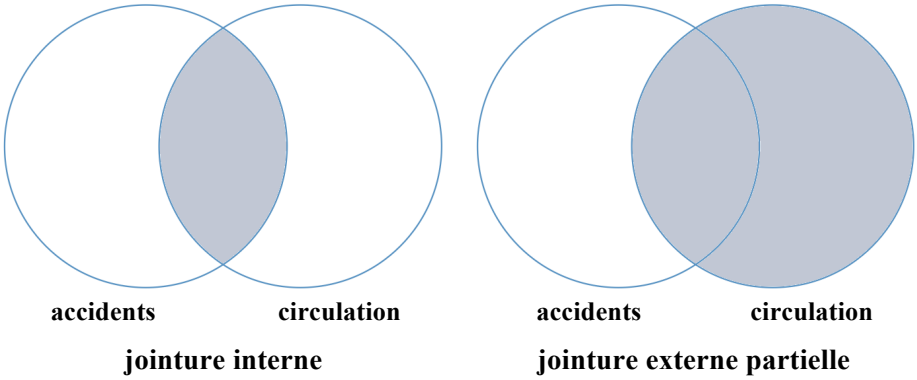


Figure 12. Types de jointures envisagées lors de l'intégration.

La jointure externe partielle, pour sa part, regroupera non seulement les instances qui décrivent des intersections communes, mais inclura également toutes les intersections orphelines de l'ensemble de données sur la circulation. Le corpus obtenu de la sorte sera plus volumineux, mais contiendra moins de traits discriminants (voir tableau 12) puisque les seuls attributs des données sur les accidents seront les nombres (réels ou extrapolés) de victimes.

Tableau 13. Comparaison des corpus

	corpus *accidents*	**corpus *circulation***	**corpus intégré**	
			jointure interne	jointure externe partielle
intersections uniques	4 813	1 628	540	1 628
instances totales	13 149	606 657	1 108 602	1 532 233
attributs discriminants	3 ou 7[31]	21[32]	28	24

Concrètement, cela signifie que la jointure interne permettra d'analyser, par exemple, l'effet combiné du volume de circulation et de la limite de vitesse sur les accidents, mais pas de formuler des conclusions à propos des intersections — indubitablement dignes d'intérêt — où aucun accident ne s'est produit. Ces intersections seraient bien sûr présentes dans un corpus créé par jointure externe partielle, mais elles ne présenteraient qu'un ensemble réduit d'attributs.

Cela étant dit, il demeure possible de pallier ce problème en générant par géocodage de nouveaux attributs qui s'appliqueraient à toutes les intersections avec ou sans accidents. Des ensembles de données telles que les limites administratives des arrondissements (Ville de Montréal, 2013b) ou l'affectation du sol au plan d'urbanisme (Ville de Montréal, 2014) constitueraient des avenues d'expérimentation intéressantes.

Nous avons proposé, dans cette section, une méthode sommaire permettant de décrire des ensembles de données afin d'y repérer les attributs discriminants. Nous avons également démontré comment nettoyer et reformater les données, dans le but de joindre les ensembles de

[31] Selon le type de jointure, le corpus *accidents* comprendra les 7 attributs repérés par élimination (voir tableau 6) ou seulement les 3 attributs représentant les victimes estimées (0) aux intersections sans accidents. Nous n'avons pas inclus les deux attributs géospatiaux créés à la section 4.2.3, puisque ceux-ci sont en quelque sorte annulés par la jointure des corpus.

[32] Ce nombre d'attributs a été estimé par élimination (voir tableau 8), en ajoutant les deux attributs géospatiaux créés à la section 4.2.3.

données en un corpus analysable unique, dont le volume plus important permet d'envisager des analyses plus complexes. La combinaison de manipulations que nous venons d'effectuer constitue cependant un simple aperçu des possibilités offertes par le croisement d'ensembles de données ouvertes, ainsi que des problèmes qui doivent être résolus au préalable : la variété de traitements possibles sera assurément proportionnelle à la grande hétérogénéité des données ouvertes diffusées par les administrations publiques.

5. Conclusion

Si elles demeurent un phénomène relativement méconnu du grand public, il n'en demeure pas moins que les données ouvertes se sont rapidement développées en quelques années à peine. Notre revue de littérature (voir chapitre 2) a permis de constater un enthousiasme généralisé, non seulement chez les acteurs de la société civile (citoyens utilisateurs, militants, prosélytes, entrepreneurs, etc.), mais également au sein des administrations publiques, où la classe politique en a rapidement fait l'un des rouages de la gouvernance numérique et des villes intelligentes.

Cet enthousiasme engendre des attentes élevées à plusieurs égards. Sur le plan politique, les données ouvertes conduiraient à plus de transparence et d'imputabilité, et augmenteraient, à terme, la confiance des électeurs à l'endroit des dirigeants (section 2.2.1). Sur le plan économique, elles stimuleraient l'innovation et contribueraient à la création de richesses (section 2.2.2). Sur le plan social, elles encourageraient la participation citoyenne grâce à de nouveaux modes d'interaction avec l'administration publique (section 2.2.3). Elles permettraient d'évaluer et d'améliorer l'efficacité des politiques, de favoriser l'échange d'information entre les différents secteurs d'une administration. En bref, le discours ambiant présente souvent l'ouverture des données comme une forme de panacée.

Nous avons cherché à estimer l'écart entre le discours et la réalité en dressant un portrait de l'ouverture des données au gouvernement du Québec et à la ville de Montréal (voir chapitre 3). Divers documents divulgués par les deux administrations révèlent que la réflexion sur les objectifs et la mise en œuvre des données ouvertes ne s'est pas déroulée en vase clos, mais qu'elle s'est au contraire inspirée de diverses initiatives internationales. Les politiques québécoises en matière de données ouvertes se révèlent par conséquent, dans leur ensemble, en phase avec les recommandations des organismes reconnus. Nous avons cependant noté un écart — sans doute prévisible — entre ces politiques et les pratiques sur le terrain, en particulier en ce qui a trait à la couverture, à la granularité et la qualité des données. La génération et la publication automatisées des données seraient susceptibles d'apporter une plus

grande cohérence formelle et une meilleure granularité (section 4.1.3), mais demeurent des pratiques marginales.

Par ailleurs, certaines faiblesses nous ont semblé s'expliquer par un biais en faveur des applications logicielles, au détriment des applications analytiques poussées (section 3.3). Nous n'avons remarqué, en particulier, aucun effort significatif afin de publier rétroactivement les données générées avant l'ouverture des portails de données, ce qui permettrait de mettre à la disposition des utilisateurs des ensembles de données plus volumineux, couvrant des périodes plus longues.

Enfin, nous n'avons trouvé mention d'aucune procédure précise pour évaluer les retombées de l'ouverture des données, en particulier de manière quantitative — et sur ce point, les portails québécois ne se distinguent pas de leurs homologues internationaux. Il est donc difficile, ici comme à l'étranger, d'établir si l'enthousiasme qui entoure les données ouvertes se justifie par les résultats obtenus.

Lors de notre revue de la littérature, nous n'avons trouvé aucun plaidoyer en faveur de la culture des vases clos — sauf, peut-être, sous la forme d'un appel à la prudence face aux résultats parfois imprévisibles des croisements de données (voir section 2.1.2). Dans la littérature spécialisée aussi bien que dans les documents administratifs, le discours dominant semble préconiser le décloisonnement — et l'agrégation des données ouvertes par le truchement de vastes portails participe sans conteste à ce mouvement (voir section 1.2.3).

Nos recherches ont cependant révélé que la mise en place d'un portail n'entraîne pas *de facto* une mise en relation transversale des données. Nous avons notamment constaté que les pratiques en matière de métadonnées ne favorisaient pas toujours le décloisonnement (section 3.2.3), et que l'association d'ensembles de données se bornait souvent à la création de catégories thématiques ou organisationnelles. En outre, le contenu même des documents peut compromettre les objectifs transversaux des portails : comme l'a démontré notre exercice d'intégration (sections 4.2.3 et 4.2.4), le croisement de données issues de diverses unités administratives, voire de divers paliers de gouvernement, peut être affecté par le format ou la qualité des données.

Pour résoudre ce problème, il ne suffit pas de rédiger des politiques de données précises afin de normaliser les pratiques : il faut surtout assurer l'application rigoureuse de ces politiques.

En outre, il importe que les portails publient des métaensembles de données non pas uniquement sur la base thématique — comme c'est déjà le cas —, mais en fonction du potentiel de croisement. Il s'agit d'un travail de recherche complexe, que l'on réalise déjà pour les hackathons, à l'occasion desquels les organisateurs sélectionnent des ensembles de données susceptibles d'être utilisés conjointement. Un tel travail de défrichage et de sélection devrait constituer une démarche permanente des portails de données.

Il s'agit d'une entreprise considérable, certes, et pour laquelle la contribution des chercheurs, journalistes et citoyens experts apparaît non seulement souhaitable, mais inévitable. Pour mener une telle recherche à terme, ces intervenants ont besoin d'outils et de ressources adéquates — dont, notamment, des inventaires de données exhaustifs. Afin de repérer des données complémentaires, susceptibles de s'éclairer mutuellement, composer avec les données disponibles ne suffit pas toujours : il serait avantageux de compter sur l'inventaire des données *potentiellement* disponibles, c'est-à-dire l'ensemble des données — ouvertes ou non — de l'administration publique. Il ne s'agit pas uniquement d'un critère de transparence, comme nous l'avons souligné à la section 2.2.1, mais d'un impératif fonctionnel.

Enfin, ces pratiques doivent reposer sur une définition plus large des applications : les portails de données ouvertes ne doivent plus simplement signaler les applications logicielles basées sur leurs données, mais effectuer une veille informationnelle afin de repérer systématiquement les occurrences d'analyses citoyennes, de journalisme de données et de recherches académiques. En outre, ces inventaires devraient être eux-mêmes publiés sous forme de données, afin de permettre l'analyse des tendances analytiques.

L'innovation en somme ne consiste pas à publier des données comme s'il s'agissait d'un matériau inerte. L'ouverture des données implique un travail d'analyse et de mise en valeur constant. L'innovation doit se refléter non seulement dans les politiques et dans la culture organisationnelle, mais dans les budgets alloués aux portails pour mener leurs activités

quotidiennes. Nous croyons que ces changements seront cruciaux pour permettre à l'ouverture des données de dépasser la simple phase d'émergence, et d'atteindre la maturité.

Bibliographie

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*, *2*(2). doi:10.1177/2053951715594634

Bairwell Ltd. (s.d.). The Public Whip. Repéré à http://www.publicwhip.org.uk

Bannister, F. et Connolly, R. (2011). The Trouble with Transparency: A Critical Review of Openness in e-Government. *Policy & Internet*, *3*(1), 1-30.

Bauer, F. et Kaltenböck, M. (2011). *Linked open data: The essentials*. Vienne : Edition mono/monochrom.

Bennett, D. et Harvey, A. (2009). Publishing open government data. *W3C Working Draft*. Repéré à http://www.w3.org/TR/gov-data

Berners-Lee, T. (1998). What a semantic can represent. Repéré à http://www.w3.org/DesignIssues/RDFnot.html

Berners-Lee, T. (2006). Linked Data - Design Issues. Repéré à http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2009). *The Next Web*. Communication présentée au TED 2009. Repéré à https://www.ted.com/talks/tim_berners_lee_on_the_next_web

Bertot, J. C., McDermott, P. et Smith, T. (2012). Measurement of open government: Metrics and process. Dans *Proceedings of the 2012 45th Hawaii International Conference on System Sciences* (p. 2491-2499). IEEE Computer Society.

Bizer, C., Heath, T. et Berners-Lee, T. (2009). Linked data - the story so far. Dans *Semantic Services, Interoperability and Web Applications : Emerging Concepts* (p. 205-227). IGI Global.

Böhm, C., Naumann, F., Freitag, M., George, S., Höfler, N., Köppelmann, M. et Schmidt, T. (2010). Linking open government data: what journalists wish they had known. Dans *Proceedings of the 6th International Conference on Semantic Systems* (p. 34-36).

Braunschweig, K., Eberius, J., Thiele, M. et Lehner, W. (2012). The state of open data - limits of current open data platforms. Communication présentée au www2012, Lyon, France.

Repéré à http://www2012.org/proceedings/nocompanion/wwwwebsci2012_braunschweig.pdf

Bureau de la ville intelligente et numérique. (2014). *Montréal ville intelligente - stratégie montréalaise 2014-2017*. Repéré à http://villeintelligente.montreal.ca/sites/villeintelligente.montreal.ca/files/strategie-montrealaise-2014-2017-ville-intelligente-et-numerique-fr-amendee.pdf

Cameron, D. et Halsey-Watkins, R. (2016, 28 avril). Secteur de l'UdeM : 143 vols sur 8 coins de rue. *La Presse*. Repéré à http://www.lapresse.ca/actualites/montreal/201604/27/01-4975873-secteur-de-ludem-143-vols-sur-8-coins-de-rue.php

Caplan, R., Davies, T., Wadud, A., Verhulst, S., Alonso, J. et Farhan, H. (2014). Towards common methods for assessing open data : workshop report & draft framework. Repéré à http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf

CARTO. (s.d.). CARTO — Predict through location. Repéré à https://carto.com

CEFRIO. (2016). *L'ouverture des données dans les municipalités québécoises : Enjeux et degré d'avancement*. Repéré à http://www.enap.ca/cerberus/files/nouvelles/documents/La_recherche/NetGouv_Ouvertureddesdonneesdanslesmunicipalites_2016.pdf

CEFRIO. (2017). *NETGouv données ouvertes : Mesure de l'apport et de l'évolution du gouvernement ouvert au Québec*. Repéré à http://www.cefrio.qc.ca/media/uploader/RAPPORTNETGOUVDONNESOUVERTES-Final.pdf

CEFRIO. (s.d.). Le CEFRIO. *CEFRIO L'expérience du numérique*. Repéré à http://www.cefrio.qc.ca/cefrio/

CEFRIO. (2016). *Les données ouvertes dans l'administration publique québécoise : Utilités, freins et pistes de solution*. Repéré à http://www.enap.ca/cerberus/files/nouvelles/documents/La_recherche/NetGouv_Lesdonneesouvertesdansadministrationquebecoise_2016.pdf

Ceolin, D., Moreau, L., O'Hara, K., Fokkink, W., Van Hage, W. R., Maccatrozzo, V. Shadbolt, N. (2014). Two Procedures for Analyzing the Reliability of Open Government

Data. Dans *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (p. 15-24). France : Springer International Publishing.

Ceolin, D., Moreau, L., O'Hara, K., Schreiber, G., Sackley, A., Fokkink, W. et Shadbolt, N. (2013). Reliability analyses of open government data. Dans *Proceedings of the 9th International Conference on Uncertainty Reasoning for the Semantic Web - Volume 1073* (p. 34-39). Sydney, Australie : CEUR-WS.org.

Chan, C. M. (2013). From open data to open innovation strategies: Creating E-services using open government data. Dans *Proceedings of the 46th Annual Hawaii International Conference on System Sciences* (p. 1890-1899). IEEE Computer Society.

Clark, J. (2013). Toronto 311 - Service Requests, 2012. Repéré à http://neoformix.com/Projects/Toronto311/

Cloudred Multimedia. (s.d.). An Interactive Visualization of NYC Street Trees. Repéré à http://www.cloudred.com/labprojects/nyctrees

Cole, R. J. (2012). Some observations on the practice of « open data » as opposed to its promise. *The Journal of Community Informatics*, *8*(2).

Commissariat général du plan. (1999). *Diffusion des données publiques et révolution numérique*. Paris : La Documentation française.

Commission on Public Information and Communication. (1993). *Public Data Directory*. New York : Office of the Mayor. Repéré à http://fr.scribd.com/doc/18121479/1993-NYC-Public-Data-Directory

Conseil du trésor. (2015). Rénover l'état par les technologies de l'information. Repéré à https://www.tresor.gouv.qc.ca/fileadmin/PDF/ressources_informationnelles/strategie_ti/strategie_ti.pdf

Creative Commons. (s.d.). Attribution 4.0 International — CC BY 4.0. Repéré à https://creativecommons.org/licenses/by/4.0/deed.fr

Cross, M. et Mathieson, S. (2006, 23 mars). Free our data: Ordnance Survey challenged to open up. *The Guardian*. Repéré à http://www.theguardian.com/society/2006/mar/23/epublic.technology

Czajka, J., Schneider, C., Sukasih, A. et Collins, K. (2014). *Minimizing Disclosure Risk in HHS Open Data Initiatives*. Washington DC : Department of Health and Human Services.

Davies, A. et Lithwick, D. (2010). Government 2.0 and Access to Information : 1. Recent Developments in Proactive Disclosure and Open Data in Canada. Bibliothèque du Parlement. Repéré à https://lop.parl.ca/Content/LOP/ResearchPublications/2010-15-e.pdf

Davies, T. (2013). Open Data Barometer : 2013 Global Report. World Wide Web Foundation and Open Data Institute. Repéré à http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf

Davis, T., Perini, F. et Alonso, J. M. (2013). Researching the emerging impacts of open data - ODDC conceptual framework. World Wide Web Foundation. Repéré à http://www.opendataresearch.org/sites/default/files/posts/Researching%20the%20emerging%20impacts%20of%20open%20data.pdf

Dawes, S. S. et Helbig, N. (2010). Information strategies for open government: Challenges and prospects for deriving public value from government transparency. Dans *Electronic Government, 6228* (vol. 2456, p. 50-60). Springer. Repéré à https://hal.archives-ouvertes.fr/file/index/docid/1056566/filename/paper_107.pdf

Défi Géohack 2014. (s.d.). Repéré à http://defigeohack.org/

Dekkers, M., Polman, F., te Velde, R. et de Vries, M. (2006). MEPSIR: Measuring european public sector information resources. Final report of study on exploitation of public sector information. Repéré à http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=1198

Département des affaires économiques et sociales. (2013). *Open Government Data for Citizen Engagement in Managing Development Guidance Toolkit*. New York : Nations Unies. Repéré à http://www.unpan.org/DPADM/ProductsServices/OpenGovernmentDataandServices/tabid/1536/language/en-US/Default.aspx

Disqus. (s.d.). Portail Données ouvertes · Conversations. Repéré à https://disqus.com/home/forums/ckanmtldata/

Dyson, L. et Goldstein, B. (dir.). (2013). *Beyond transparency: Open data and the future of civic innovation*. Code for America Press. Repéré à http://beyondtransparency.org/pdf/BeyondTransparency.pdf

ÉcoHackMtl. (s.d.). ÉcoHackMtl – Hackathon montréalais sur la durabilité urbaine. Repéré à http://ecohackmtl.org

École polytechnique de Montréal. (s.d.). Publications - rapports - notes de recherche. *Chaire de recherche sur l'évaluation et la mise en œuvre de la durabilité en transport*. Repéré à http://www.polymtl.ca/mobilite/pub

Electronic Freedom of Information Act Amendments of 1996, 5 U.S.C. section 552 tel qu'amendé par la loi No 104-231, 110 Stat. 3048 (1996).

European Commission. (2011). Open data: An engine for innovation, growth and transparent governance. Repéré à http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:HTML

Evoenergy. (s.d.). The UK Energy Consumption Guide. Repéré à http://www.evoenergy.co.uk/uk-energy-guide

Executive Office of the President (2013). Open Data Policy-Managing Information as an Asset. Repéré à https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf

Executive Office of the President (2009). President's Memorandum on Transparency and Open Government - Interagency Collaboration. Repéré à https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2009/m09-12.pdf

Ferreira, N., Poco, J., Vo, H. T., Freire, J. et Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2149-2158.

Ferzoco, J. (2014). Citi Bike Rides: September 17th & 18th, 2013. *Line Point Path*. Repéré à https://vimeo.com/89305412

FFunction. (s.d.). Vue sur les contrats. Repéré à http://ffctn.com/a/contrats

Fioretti, M. (2010). Open Data, Open Society. A research project about openness of public data in EU local administration. DIME network. Repéré à http://www.dime-eu.org/files/active/0/ODOS_report_1.pdf

Friedman, D. et DeBold, T. (2015, 11 février). Battling Infectious Diseases in the 20th Century: The Impact of Vaccines. *Wall Street Journal*. Repéré à http://graphics.wsj.com/infectious-diseases-and-vaccines

Garriga-Portola, M. (2011). Open Data? Yes, But in a Sustainable Way. *Profesional de la*

Informacion, *20*(3), 298-303.

Gautrin, H.-F. (2012). *Gouverner ensemble : comment le Web 2.0 améliorera-t-il les services aux citoyens?* Québec : Secrétariat du Conseil du trésor.

GitHub. (s.d.). The world's leading software development platform · GitHub. *GitHub*. Repéré à http://github.com

Gomez, R. et Wald, S. (2010). When public-sector salaries become public knowledge: Academic salaries and Ontario's Public Sector Salary Disclosure Act. *Canadian Public Administration*, *53*(1), 107-126. doi:10.1111/j.1754-7121.2010.00114.x

Google Maps. (2016, 30 juin). googlemaps 2.4.4. *Python Package Index*. Repéré à https://pypi.python.org/pypi/googlemaps/2.4.4

Gorunescu, F. (2011). *Data mining : concepts, models and techniques*. Berlin : Springer-Verlag

Gouvernement de l'Ontario. (2015). Accès aux données gouvernementales. *Ontario.ca*. Repéré à https://www.ontario.ca/fr/page/acces-aux-donnees-gouvernementales

Gouvernement du Canada. (2015a). G8 Open Data Charter – Canada's Action Plan. Repéré à http://open.canada.ca/en/g8-open-data-charter-canadas-action-plan

Gouvernement du Canada. (2015b). Répertoire des fromages canadiens. *Gouvernement ouvert*. Repéré à http://ouvert.canada.ca/data/fr/dataset/3c16cd48-3ac3-453f-8260-6f745181c83b

Gouvernement du Canada. (s.d.-a). Des données et des renseignements ouverts par défaut, dans des formats modernes et conviviaux. *Gouvernement ouvert*. Repéré à http://ouvert.canada.ca/fr/consultation/des-donnees-et-des-renseignements-ouverts-par-defaut-dans-des-formats-modernes-et-conviviaux

Gouvernement du Canada. (s.d.-b). Gouvernement ouvert. Repéré à http://ouvert.canada.ca/fr

Gouvernement du Québec. (2012). Plan de classification de BAnQ. Repéré à http://www.donnees.gouv.qc.ca/?node=/donnees-details&id=cf8e1a11-e2f7-4864-ad4b-1ea20abdaf2c

Gouvernement du Québec. (2016). Lignes directrices sur la diffusion de données ouvertes. Repéré à https://www.donneesquebec.ca/wp-content/uploads/2016/08/Lignes-directrices-2016-08-26.pdf

Gouvernement du Québec. (s.d.). Foire aux questions. *Données Québec*. Repéré à

https://www.donneesquebec.ca/fr/faq/

GovLab. (s.d.-a). Open Data's Impact. Repéré à http://odimpact.org

GovLab. (s.d.-b). The OD500 Global Network. Repéré à http://www.opendata500.com

Granickas, K. (2013). *Understanding the impact of releasing and re-using open government data. Topic Report n^{o} 2013/08*. European Public Sector Information Platform. Repéré à https://www.europeandataportal.eu/sites/default/files/2013_understanding_the_impact_of_releasing_and_re_using_open_data.pdf

Graves, A. et Hendler, J. (2013). Visualization Tools for Open Government Data. Dans *Proceedings of the 14th Annual International Conference on Digital Government Research* (p. 136-145). New York, NY, USA : ACM. doi:10.1145/2479724.2479746

Gray, J. (2014). Towards a Genealogy of Open Data. Communication présentée au General Conference of the European Consortium for Political Research, Glasgow. doi:10.2139/ssrn.2605828

Gray, J. et Darbishire, H. (2011). *Beyond Access : Open Government Data & the Right to (Re)use Public Information*. Madrid : Access Info.

Groupe de travail sur les données ouvertes. (2011). *Rapport sur l'ouverture des données de la Ville de Montréal : Un capital numérique générateur d'innovation et de participation*. Repéré à http://www1.ville.montreal.qc.ca/banque311/webfm_send/1453

Guidoin, S. (2015, 8 février). Données ouvertes à Montréal et groupe de citoyens-testeurs. *Données ouvertes Montréal*. Repéré à https://groups.google.com/forum/#!topic/open-data-montreal/UunCCZxRN24

Hackons la corruption 2012. (s.d.). *Québec Ouvert*. Repéré à http://quebecouvert.org/events/hackonslacorruption/

Hackworks. (s.d.-a). Canadian Open Data Experience. Repéré à https://www.canadianopendataexperience.ca

Hackworks. (s.d.-b). TrafficJam. Repéré à http://trafficjam.to

Helbing, D. et Balietti, S. (2011). From social data mining to forecasting socio-economic crises. *The European Physical Journal Special Topics*, *195*(1), 3-68. doi:10.1140/epjst/e2011-01401-8

Hellberg, A.-S. et Hedström, K. (2015). The story of the sixth myth of open data and open government. *Transforming Government: People, Process and Policy*, *9*(1), 35-51.

Hern, A. (2014, 27 juin). New York taxi details can be extracted from anonymised data, researchers say. *The Guardian*. Repéré à http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

Hernandez-Perez, T. et Garcia-Moreno, M.-A. (2013). Open Data and Data Repositories: A New Challenge for Librarians. *Profesional de la Informacion*, *22*(3), 259-263.

Hill, M. W. (2005). *The impact of information on society: an examination of its nature, value and usage* (2e éd.). Munich : K. G. Saur Verlag.

Hivon, J. et Titah, R. (2015). Citizen Participation in Open Data Use at the Municipal Level. Communication présentée au Open Data Research Symposium, Shaw Center, Ottawa. Repéré à http://www.opendataresearch.org/dl/symposium2015/odrs2015-paper37.pdf

Huijboom, N. et Van den Broek, T. (2011). Open data : an international comparison of strategies. *European journal of ePractice*, *12*(1), 4–16.

Hunnius, S., Krieger, B. et Schuppan, T. (2014). Providing, Guarding, Shielding: Open Government Data in Spain and Germany. Communication présentée au 2014 EGPA Annual Conference, Speyer, Allemagne.

International Open Data Hackathon. (s.d.). Repéré à http://opendataday.org

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, *28*(4), 446-456.

Janssen, K. (2012). Open Government Data : Right to Information 2.0 or its Rollback Version? *ICRI Research Paper No. 8*. Repéré à https://ssrn.com/abstract=2152566

Janssen, M., Charalabidis, Y. et Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258-268. doi:10.1080/10580530.2012.716740

Janssen, M. et Hoven, J. van den. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly*, *32*(4), 363-368. doi:10.1016/j.giq.2015.11.007

Jäppinen, S., Toivonen, T. et Salonen, M. (2013). Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach. *Applied Geography*, *43*, 13-24. doi:10.1016/j.apgeog.2013.05.010

Jetzek, T., Avital, M. et Bjorn-Andersen, N. (2014). Data-driven innovation through open government data. *Journal of theoretical and applied electronic commerce research*, *9*(2), 100-120.

Johnson, P. et Robinson, P. (2014). Civic Hackathons : Innovation, Procurement, or Civic Engagement? *The Review of Policy Research*, *31*(4), 349-357. doi:10.1111/ropr.12074

Journée internationale des données ouvertes 2016 : autoportrait du milieu montréalais des données ouvertes. (2016, 12 février). *Open Data Day Wiki*. wiki. Repéré à http://wiki.opendataday.org/MONTREAL2016

Kalampokis, E., Hausenblas, M. et Tarabanis, K. (2011). Combining Social and Government Open Data for Participatory Decision-Making. Dans E. Tambouris, A. Macintosh et H. de Bruijn (dir.), *Proceedings of the Third IFIP WG 8.5 international conference on Electronic participation* (vol. 6847, p. 36-47). Berlin : Springer. doi:10.1007/978-3-642-23333-3_4

Kalampokis, E., Tambouris, E. et Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544-559. doi:10.1108/IntR-06-2012-0114

Kassen, M. (2013). A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly*, *30*(4), 508-513. doi:10.1016/j.giq.2013.05.012

Kaufman, S. (2014). Citi Bike and Gender. *NYU Rudin Center for Transportation*. Repéré à http://wagner.nyu.edu/rudincenter/2014/05/citi-bike-and-gender/

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Londres : Sage.

Kruchten, N. (2014). Zoomable Map for Montreal Election Results. Repéré à http://nicolas.kruchten.com/content/2014/01/mtlelection-zoomable-map/

Kruchten, N. (2015). Montreal 311 Service Requests, an Analysis. Repéré à http://nicolas.kruchten.com/content/2015/06/montreal-311/

Kuk, G. et Davies, T. (2011). The roles of agency and artifacts in assembling open data complementarities. Communication présentée au ICIS 2011, Shanghai. Repéré à https://eprints.soton.ac.uk/273064/

Kulk, S. et Van Loenen, B. (2012). Brave new open data world? *International Journal of*

Spatial Data Infrastructures Research, *7*, 196-206.

Lavoie, A. (2014). *Open Data (donnée ouverte) : doit-on tout publier? Pourquoi pas? Et l'archiviste, il fait quoi dans tout ça?* Communication présentée à la Conférence des milieux documentaires 2014, Montréal.

Li, S.-T. et Shue, L.-Y. (2004). Data mining to aid policy making in air pollution management. *Expert Systems with Applications*, *27*(3), 331-340. doi:10.1016/j.eswa.2004.05.015

Library of Congress. (2011). ESRI Shapefile. *Sustainability of Digital Formats Planning for Library of Congress Collections*. Repéré à http://www.digitalpreservation.gov/formats/fdd/fdd000280.shtml

Loi sur la gouvernance et la gestion des ressources informationnelles des organismes publics et des entreprises du gouvernement, L.R.Q. ch. G-1.03 (2011). Repéré à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cs/G-1.03

Loi sur l'accès à l'information, L.R.C. ch. A-1 (1985). Repéré à http://laws-lois.justice.gc.ca/fra/lois/a-1/page-1.html

Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels, L.R.Q. ch. A-2.1 (1982). Repéré à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cs/A-2.1

Maguire, S. (2011). Can Data Deliver Better Government? *The Political Quarterly*, *82*(4), 522-525. doi:10.1111/j.1467-923X.2011.02249.x

Mandelbaum, A. G. (2011). *Strengthening Parliamentary Accountability, Citizen Engagement and Access to Information: A Global Survey of Parliamentary Monitoring Organizations*. Washington DC : National Democratic Institute and World Bank Institute. Repéré à https://www.ndi.org/sites/default/files/governance-parliamentary-monitoring-organizations-survey-september-2011.pdf

Margetts, H. (2011). The Internet and Transparency. *The Political Quarterly*, *82*(4), 518-521. doi:10.1111/j.1467-923X.2011.02253.x

Martin, S., Foulonneau, M., Turki, S. et Ihadjadene, M. (2013). Risk analysis to overcome barriers to open data. *Electronic Journal of e-Government*, *11*(1), 348-359.

McClean, T. (2011). Not with a Bang but a Whimper: The Politics of Accountability and Open Data in the UK. Dans *APSA 2011 Annual Meeting Paper*. Repéré à

https://ssrn.com/abstract=1899790

McLeod, J. (2012). Thoughts on the opportunities for records professionals of the open access, open data agenda. *Records Management Journal*, *22*(2), 92-97. doi:10.1108/09565691211268711

Mercier, D. (2014). *Données ouvertes, qualité et visualisation*. Communication présentée au MTL Data Meetup, Montréal. Repéré à http://dianemercier.quebec/donnees-ouvertes-qualite-et-visualisation/

Miller, P., Styles, R. et Heath, T. (2008). Open Data Commons, a License for Open Data. Dans *CEUR Workshop Proceedings* (vol. 369). Repéré à http://wtlab.um.ac.ir/images/e-library/linked_data/2008/08-miller-styles-open-data-commons_.pdf

Misuraca, G. (2012). Renouveler la gouvernance à l'ère du numérique. *Télescope*, *18*(1-2), 21-43. doi:10.7202/1009253ar

Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS biology*, *9*(12), e1001195.

Mulley, M. (s.d.). Open Parliament. Repéré à https://openparliament.ca

mySociety. (s.d.). They Work For You. Repéré 7 janvier 2016, à http://www.theyworkforyou.com

New York City. (s.d.). 2014 Yellow Taxi Trip Data. *NYC OpenData*. Repéré à https://data.cityofnewyork.us/view/gn7m-em8n

Obama, B. (2013). Executive Order 13642 of May 9, 2013: Making open and machine readable the new default for government information. *Federal Register*, *78*(93), 28111-28113.

O'Brien, O., Cheshire, J. et Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, *34*, 262-273. doi:10.1016/j.jtrangeo.2013.06.007

Ochando, L. C., Julián, C. I. F., Ochando, F. C. et Ferri, C. (s.d.). AirVLC: An Application for Real-Time Forecasting Urban Air Pollution. Dans *Proceedings of the 2nd International Workshop on Mining Urban Data* (vol. 1392, p. 72-79). Repéré à http://ceur-ws.org/Vol-1392/#paper-10

Office québécois de la langue française. (2002). Donnée. *Le grand dictionnaire*

terminologique. Repéré à http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8363711

O'Hara, K. (2012). Transparency, open data and trust in government: shaping the infosphere. Dans *Proceedings of the 4th Annual ACM Web Science Conference* (p. 223-232). New York : ACM. Repéré à https://eprints.soton.ac.uk/337558/1/ohara_websci_2012_final.pdf

Okamoto, K. (2016). What is being done with open government data? An exploratory analysis of public uses of New York City open data. *Webology*, *13*(1). Repéré à http://www.webology.org/2016/v13n1/a142.pdf

Open Data for Developpement Network. (2016). *Open Data for Development. Building an Inclusive Data Revolution. Annual Report 2015*. Repéré à http://od4d.com/wp-content/uploads/2016/06/OD4D_annual_report_2015.pdf

Open Government Working Group. (2007). 8 Principles of Open Government Data. Repéré à https://public.resource.org/8_principles.html

Open Knowledge Foundation. (2005). Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. Repéré à http://opendefinition.org/od/

Open Knowledge Foundation. (2012). Manuel de l'Open Data. Repéré à http://opendatahandbook.org/guide/fr/

Open Knowledge Foundation. (s.d.-a). CKAN instances around the world. *CKAN - the open source data portal software*. Repéré à https://ckan.org/instances/

Open Knowledge Foundation. (s.d.-b). Where Does My Money Go? Repéré à http://wheredoesmymoneygo.org

Open Knowledge Foundation. (s.d.-c). Why Open Data? *Open Data Handbook*. Repéré à http://opendatahandbook.org/guide/en/why-open-data/

OpenPlans. (s.d.). About. *Open311 - A collaborative model and open standard for civic issue tracking*. Repéré à http://www.open311.org/about/

Ordnance Survey. (s.d.). Data and file formats. Repéré à https://www.ordnancesurvey.co.uk/support/understanding-gis/data-and-file-formats.html

Parlement européen et Conseil de l'Union européenne. Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public (2003). Repéré à http://eur-

lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:FR:HTML

Parycek, P., Hochtl, J. et Ginner, M. (2014). Open Government Data Implementation Evaluation. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*, 80-99.

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*. Repéré à https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2264369

Performance and Innovation Unit. (2000). *E.gov : Electronic Government Services for the 21st Century*. Londre : Cabinet Office.

Peters, S. (2015). Indices of Deprivation 2015 explorer. *OpenDataCommunities.org*. Repéré à http://dclgapps.communities.gov.uk/imd/idmap.html

Pineau, J. et Bacon, P.-L. (2015). Analyzing Open Data from the City of Montreal. Dans *Proceedings of the 2nd International Workshop on Mining Urban Data*. Repéré à http://www.cs.mcgill.ca/~jpineau/files/jpineau-icml15ws-mud2.pdf

Plamondon Émond, É. (2016, 25 mai). Des données ouvertes, mais peu accessibles. *Le Devoir*. Repéré à http://www.ledevoir.com/politique/quebec/471484/des-donnees-ouvertes-mais-peu-accessibles

Provost, A.-M. (2016, 8 février). Où sont les endroits les plus dangereux pour les piétons à Montréal? *ICI.Radio-Canada.ca*. Repéré à http://ici.radio-canada.ca/regions/montreal/2016/02/08/001-accidents-pietons-montreal-carte.shtml

Rahm, E. et Do, H. H. (2000). Data cleaning: Problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *23*(4), 3-13.

RapidMiner. (s.d.). RapidMiner. Repéré à https://rapidminer.com

Règlement sur la diffusion de l'information et sur la protection des renseignements personnels, L.R.Q. ch A-2.1, r. 2 (2008). Repéré à http://legisquebec.gouv.qc.ca/fr/ShowDoc/cr/A-2.1,% 20r.%202

Reichman, O. J., Jones, M. B. et Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, *331*(6018), 703-705. doi:10.1126/science.1197962

Ren, G.-J. et Glissmann, S. (2012). Identifying information assets for open data: the role of business architecture and information quality. Dans *IEEE 14th International Conference*

on Commerce and Enterprise Computing (CEC) (p. 94-100). IEEE. doi:10.1109/CEC.2012.23

Robinson, D. G., Yu, H., Zeller, W. P. et Felten, E. W. (2009). Government data and the invisible hand. *Yale Journal of Law & Technology*, *11*, 160.

Rocha, R. (2015, 6 mai). Montreal's 311 records shed light on residents' concerns — to a point. *Montreal Gazette*. Repéré à http://montrealgazette.com/news/local-news/311-calls-whats-bugging-montreal

Roy, J.-H. (2013). *Trois recommandations pour un gouvernement plus ouvert*. Communication présentée à la Consultation générale sur le rapport Technologies et vie privée à l'heure des choix de société, Assemblée nationale du Québec, Commission des institutions. Repéré à http://www.assnat.qc.ca/Media/Process.aspx?MediaId=ANQ.Vigie.Bll.DocumentGenerique_70541

Rubin, R. E. (2010). *Foundations of Library and Information Science.* (3e éd.). New York : Neal-Schuman Publishers.

Salonen, M. et Toivonen, T. (2013). Modelling travel time in urban networks: comparable measures for private car and public transport. *Journal of Transport Geography*, *31*, 143-153. doi:10.1016/j.jtrangeo.2013.06.011

Schmidt, C. (2012). Does transparency increase executive compensation? Communication présentée au 39th Annual Meeting of the European Finance Association, Frederiksberg, Danemark. Repéré à http://www.efa2012.org/papers/t4b3.pdf

Schmite, P. (2017). Retour sur le hackathon #HackRisques. *Le blog de la mission Etalab*. Repéré à https://www.etalab.gouv.fr/retour-sur-le-hackathon-hackrisques

Secrétariat du Conseil du Trésor du Canada. (2014). Plan d'action du Canada pour un gouvernement ouvert 2014-2016. Repéré à http://publications.gc.ca/site/fra/475592/publication.html

Secrétariat du Conseil du trésor et Gouvernement du Québec. (2015). *Rapport annuel de gestion 2014-2015*. Repéré à https://www.tresor.gouv.qc.ca/fileadmin/PDF/publications/rag_1415.pdf

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W. et schraefel, m c. (2012). Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent*

Systems, *27*(3), 16-24. doi:10.1109/MIS.2012.23

Shiab, N. (2015, 27 juillet). Où les cyclistes se font-ils percuter à Montréal? *Métro*. Repéré à http://journalmetro.com/actualites/montreal/813675/les-accidents-de-velos-cartographies/

Shiab, N. (2016, 15 février). Les accidents d'autobus en hausse à la STM. *Métro*. Repéré 22 septembre 2016, à http://journalmetro.com/actualites/montreal/916982/les-accidents-dautobus-en-hausse-a-la-stm/

Srihari, R. K. et Voeller, J. G. (2008). Finding Inadvertent Release of Information. Dans *Wiley Handbook of Science and Technology for Homeland Security*. John Wiley & Sons, Inc. doi:10.1002/9780470087923.hhs289

Stallman, R. M. (2010). *Free Software, Free Society: Selected Essays of Richard M. Stallman* (2e éd.). Free Software Foundation.

Steiner, T., Troncy, R. et Hausenblas, M. (2010). How Google is using Linked Data Today and Vision For Tomorrow. Dans *Proceedings of Linked Data in the Future Internet at the Future Internet Assembly (FIA 2010), Ghent, December 2010*. Repéré à http://CEUR-WS.org/Vol-700/Paper5.pdf

Sunlight Foundation. (2010). Ten Principles for Opening Up Government Information. Repéré à http://assets.sunlightfoundation.com.s3.amazonaws.com/policy/papers/Ten%20Principles%20for%20Opening%20Up%20Government%20Data.pdf

Tabary, C., Provost, A.-M. et Trottier, A. (2016). Data journalism's actors, practices and skills: A case study from Quebec. *Journalism*, *17*(1), 66-84.

Tait, J. (2011). Open Data in Manchester. *Greater Manchester – Open Data City*. Repéré à http://blog.okfn.org/2011/08/25/greater-manchester-open-data-city/

Tétreault-Pinard, É. (2014). Progression journalière de l'achalandage cycliste à Montréal en 2013. Repéré à http://etpinard.github.io/VisMTL/

The Freedom of Information Act, 5 U.S.C. § 552 (1966).

Tremblay, M.-È., Rocha, R., Salcido, S., Julien, M. et Guiamaraes, A. (2016). Les 68 stations de métro de Montréal vues autrement. *ICI.Radio-Canada.ca*. Repéré à http://ici.radio-canada.ca/nouvelles/special/2016/9/metro-montreal-68-stations-analyse/#

Ubaldi, B. (2013). Open Government Data - Towards Empirical Analysis of Open

Government Data Initiatives. *OECD Working Papers on Public Governance*, (22). doi:10.1787/5k46bj4f03s7-en

Valentin, J. (2012). La donnée ouverte et libre, facteur d'innovation urbaine. *Documentaliste - Sciences de l'Information*, *49*(4), 30-31.

Veljkovic, N., Bogdanovic-Dinic, S. et Stoimenov, L. (2014). Benchmarking open government : An open data perspective. *Government Information Quarterly*, *31*(2), 278-290. doi:10.1016/j.giq.2013.10.011

Verhulst, S., Noveck, B. S., Caplan, R., Brown, K. et Paz, C. (2014). The Open Data Era in Health and Social Care: A blueprint for the National Health Service (NHS England) to develop a research and learning programme for the open data era in health and social care. New York : The Governance Lab

Ville de Montréal. (2012). L'Autobus des créateurs roulera à Montréal ! *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/portail/autobus-des-createurs-roulera-a-montreal/

Ville de Montréal. (2013a). Feux de circulation – comptage des véhicules et des piétons aux intersections munies de feux. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/comptage-vehicules-pietons

Ville de Montréal. (2013b). Limite administrative de l'agglomération de Montréal (Arrondissement et Ville liée). *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/polygones-arrondissements

Ville de Montréal. (2013c). Piscines municipales. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/piscines-municipales

Ville de Montréal. (2013d). Vélos - comptage sur les pistes cyclables. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/velos-comptage

Ville de Montréal. (2014). Plan d'urbanisme - Affectation du sol. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/affectation-du-sol

Ville de Montréal. (2015). Directive sur la gouvernance des données de la Ville de Montréal. *Portail données ouvertes*.

Ville de Montréal. (2016a). Actes criminels - introductions par effraction. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/actes-criminels

Ville de Montréal. (2016b, février). Formats de données. *Portail données ouvertes*. Repéré à

http://donnees.ville.montreal.qc.ca/portail/formats-de-donnees/

Ville de Montréal. (2016c). Milieu humide. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/milieu-humide

Ville de Montréal. (2016d). Politique de données ouvertes de la Ville de Montréal. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/portail/politique-de-donnees-ouvertes/

Ville de Montréal. (s.d.-a). Bibliothèques de Montréal – informations sur les documents. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/catalogue-bibliotheques

Ville de Montréal. (s.d.-b). Guide des publieurs de données. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/portail/publieurs/

Ville de Montréal. (s.d.-c). Patinoires – historique des conditions. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/dataset/patinoires-historique

Ville de Montréal. (s.d.-d). Refonte de la Politique de données ouvertes : Montréal propose l'ouverture par défaut. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/portail/commentaires-sur-les-nouvelles-politiques-de-donnees-ouvertes/

Ville de Montréal. (s.d.-e). Structure de métadonnées. *Portail données ouvertes*. Repéré à http://donnees.ville.montreal.qc.ca/portail/structure-metadonnees/

Vogel, P., Greiser, T. et Mattfeld, D. C. (2011). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, *20*, 514-523. doi:10.1016/j.sbspro.2011.08.058

W3C. (2014). Data Catalog Vocabulary (DCAT). Repéré à https://www.w3.org/TR/vocab-dcat/

Walcott, A. (2015). Better roads are paved with big data analytics. *IBM Research*. Repéré à http://ibmresearchnews.blogspot.ca/2015/07/better-roads-are-paved-with-big-data.html

Whitaker, J. (2016). pyproj 1.9.5.1. *Python Package Index*. Repéré à https://pypi.python.org/pypi/pyproj/1.9.5.1

Whitmore, A. (2014). Using open government data to predict war: A case study of data and systems challenges. *Government Information Quarterly*, *31*(4), 622-630.

World Wide Web Foundation. (2015). *Open Data Barometer Global Report - Third Edition*.

Repéré à http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf

Worthy, B. (2014). What Happens When You Publish Salaries? *Open Data Study*. Repéré à https://opendatastudy.wordpress.com/2014/07/24/what-happens-when-you-publish-salaries/

Yates, J. et Shiab, N. (2016, 7 mars). Des garderies et des écoles à proximité des axes routiers malgré des recommandations. *Métro*. Repéré à http://journalmetro.com/actualites/montreal/926117/des-garderies-et-des-ecoles-a-proximite-des-axes-routiers-malgre-des-recommandations/

Yu, H. et Robinson, D. G. (2012). The New Ambiguity of « Open Government ». *UCLA Law Review*, *59*. Repéré à http://www.uclalawreview.org/the-new-ambiguity-of-%E2%80%9Copen-government%E2%80%9D/

Zarsky, T. Z. (2011). Governmental Data Mining and its Alternatives. *Penn State Law Review*, *116*, 285-330.

Zimmermann, A., Kaytoue, M., Plantevit, M., Robardet, C. et Boulicaut, J.-F. (2015). Profiling Users of the Velo'v Bike Sharing System. Dans *Proceedings of the 2nd International Workshop on Mining Urban Data* (vol. 1392, p. 63-64). Repéré à http://ceur-ws.org/Vol-1392/#paper-08

Zliobaite, I., Mathioudakis, M., Lehtiniemi, T., Parviainen, P. et Janhunen, T. (2015). Accessibility by Public Transport Predicts Residential Real Estate Prices: A Case Study in Helsinki Region. Dans *Proceedings of the 2nd International Workshop on Mining Urban Data* (vol. 1392, p. 65-71). Repéré à http://ceur-ws.org/Vol-1392/#paper-09

Zuiderwijk, A., Helbig, N., Gil-Garcia, J. R. et Janssen, M. (2014). Special Issue on Innovation through Open Data - A Review of the State-of-the-Art and an Emerging Research Agenda: Guest Editors´ Introduction. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2), 1-2. doi:10.4067/S0718-18762014000200001

Zuiderwijk, A. et Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17-29. doi:10.1016/j.giq.2013.04.003

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. et Alibaks, R. S. (2012). Socio-technical

impediments of open data. *Electronic Journal of eGovernment*, *10*(2), 156–172.

Zuiderwijk-van Eijk, A. M. G. et Janssen, M. (2015). Participation and data quality in open data use: open data infrastructures evaluated. Dans *Proceedings of the 15th European Conference on e-Government, Portsmouth, UK, 18-19 June 2015; Authors version*. ACPI.

Annexe 1 : analyses académiques basées sur les données ouvertes québécoises

Beaudoin, M. (2016). Faire d'une pierre deux coups : retombées positives d'actions contre les îlots de chaleur urbains. *Environnement, Risques & Santé*, *15*(4), 326-331.

Boisjoly, G. et El-Geneidy, A. (2016). Are we connected? Assessing bicycle network performance through directness and connectivity measures, a Montreal, Canada case study. Dans *Transportation Research Board 95th Annual Meeting*.

Castonguay, M.-J. (2013). *Les toits urbains, un gisement vert à exploiter* (Université de Sherbrooke). Repéré à http://savoirs.usherbrooke.ca/handle/11143/7097

Galvez-Cloutier, R., Guesdon, G. et Fonchain, A. (2014). Lac-Mégantic : analyse de l'urgence environnementale, bilan et évaluation des impacts. *Canadian Journal of Civil Engineering*, *41*(6), 531-539.

Mouchikhine, V. (2013). *Estimation des coûts indirects des bris d'infrastructures souterraines au Québec à travers 3 études de cas* (Mémoire de maîtrise, École Polytechnique de Montréal). Repéré à https://publications.polymtl.ca/1257/

Mulder, D. (2015). *Automatic repair of 3d city building models using a voxel-based repair method* (Mémoire de maîtrise, Université de technologie de Delft, Hollande). Repéré à https://repository.tudelft.nl/islandora/object/uuid%3A8ef4459d-b940-4007-bc3c-d87349015129?collection=education

Newell, E., Kia, E. et Wen, Y. (2014). Modeling imbalance in bike share networks. Repéré à http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_111.pdf

Noseworthy, M. et La Schiazza, B. (2014). Montreal Real Estate Pricing. Repéré à http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_89.pdf

Pow, N., Janulewicz, E. et Liu, L. D. (2014). Prediction of real estate property prices in Montréal. Repéré à http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf

Power, H. (2016). *Comparaison des biais et de la précision des estimations des modèles Artémis-2009 et Artémis-2014 pour la surface terrière totale des peuplements forestiers, avec et sans coupe partielle, sur une période de 40 ans* (vol. 143). Québec : Ministère des Forêts, de la Faune et des Parcs, Direction de la recherche forestière.

Scherrer, F. P., Meloche, J.-P., Morency, C., Boisjoly, G., Fortin, P. et Goulet-Beaudry, L. (2015). *Pour une connaissance et une gestion renouvelées du stationnement*. Montréal : CRE-Montréal. Repéré à http://www.cremtl.qc.ca/sites/default/files/upload/comprendre_le_stationnement_etude_cre_et_dsp.pdf

Wenger, R., Zheng, H. et Dimitrov, S. (2014). Biking Lane Usage Prediction. Repéré à http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_102.pdf

Annexe 2 : formats de documents et de données cités

CSV (Comma Separated Values) : format textuel de données à plat, similaire à une feuille de calcul.

GeoJSON : format géospatial basé sur JSON.

GeoTIFF : image TIFF incluant des métadonnées de géoréférencement.

JSON (JavaScript Object Notation) : format de données hiérarchisées basé sur JavaScript.

KML (Keyhole Markup Language) : format géospatial basé sur XML.

KMZ : version compressée d'un document KML.

NAD83 (North American Datum) : système géodésique utilisé pour l'Amérique du nord.

ODS (Open Document Spreadsheet) : format de feuille de calcul basé sur XML.

ODT (Open Document Text) : format de document de traitement de texte basé sur XML.

PDF (Portable Document Format) : format permettant de transmettre des documents en conservant leur mise en page.

SHP (Shapefile) : format géospatial développé par la compagnie Esri.

WGS84 : système géodésique utilisé par les appareils GPS.

XLS (Excel Spreadsheet) : format binaire de feuille de calcul.

XLSX (Excel Spreadsheet XML) : format de feuille de calcul basé sur XML.

XML (Extensible Markup Language) : langage de balisage sémantique hiérarchisé, dérivé du SGML.

ZIP : format de compression de document.