PROPOSITIONS POUR LA PUBLICATION DES DONNÉES OUVERTES PUBLIQUES: WORKING PAPER DE L'IDHEAP

redigé par Auriane Marmier & Tobias Mettler

UNIL | Université de Lausanne
Institut de hautes études
en administration publique

Remerciements

Les auteurs tiennent à remercier les membres du Service d'organisation et d'informatique ainsi que ceux du Service du logement et des gérances de la ville de Lausanne pour leur travail et leur collaboration, et tout particulièrement Mr Schläppy, sans qui ce projet n'aurait pu avoir lieu.

Contact

Unité Management de l'Information
IDHEAP, Université de Lausanne
Rue de la Mouline 28
CH-1022 Chavannes-près-Renens

Téléphone : +41 21 692 69 50
E-mail : im-idheap@unil.ch
Web : https://www.unil.ch/idheap

L'unité management de l'information

Les technologies de l'information sont devenues incontournables dans la société aujourd'hui. Un management de l'information performant se révèle une compétence essentielle mais aussi une préoccupation constante pour les administrations publiques.
La mission de l'unité management de l'information de l'IDHEAP étudie et explore comment les nouvelles technologies impactent le futur rôle et changement actuel du fonctionnement de l'administration publique.

TABLE DES MATIÈRE

Introduction

Les données ouvertes

Définition des OD

Au cours de discussions, dans les journaux ou dans les agendas politiques, on ne parle plus seulement des concepts *d'open sources* et des *open mouvements* mais de plus en plus *d'open data* (OD) (i.e. données ouvertes). Selon le site Open Knowledge Fundation (2018b) le concept d'ouverture signifie une liberté d'accès, d'utilisation et de partage de contenu par quiconque et dans n'importe quels buts. Ainsi, toutes données qualifiées d'ouvertes doivent donc être librement utilisables, modifiables et partageables, par n'importe qui et à n'importe quelle fin (Open Knowledge Fundation, 2018b). Afin d'optimiser la réutilisation des OD, les défenseurs des mouvements ouverts pensent que ces données devraient être régies par des licences ouvertes, libres de droits, et elles devraient également provenir de leurs sources originelles, être fréquemment mises à jour et de manière automatisée (Open Data Charter, 2018; Open Knowledge Fundation, 2014; Sunlight Foundation, 2018).

Quatre aspects majeurs illustrent alors le concept d'open data

- L'aspect **légal**, bien souvent traduit par l'existence *d'open licences*,
- L'aspect **technique** recouvrant diverses exigences telles que leur compatibilité-machine et les formats requis (Open Knowledge Fundation, 2018a),
- L'aspect **accessibilité** des données englobant des exigences de gratuité de l'information pour tous, et
- L'aspect **temporel** avec l'idée d'une mise à jour régulière des données et un accès illimité

En résumé, les OD sont des données qui peuvent être:

- Copiées
- Utilisées
- Réutilisées
- Modifiées
- Partagées
- Distribuées
- Valorisées

Définition des OGD

La stratégie *Open Governement Data* (OGD) 2014-2018, décrit les OGD comme la croisée de 3 principes : celui *l'open governement* (i.e action gouvernementale ouverte), de l'*open data* (i.e. libre accès aux données) et celui des *governement data* (i.e. données publiques) (Federal Council, 2014).

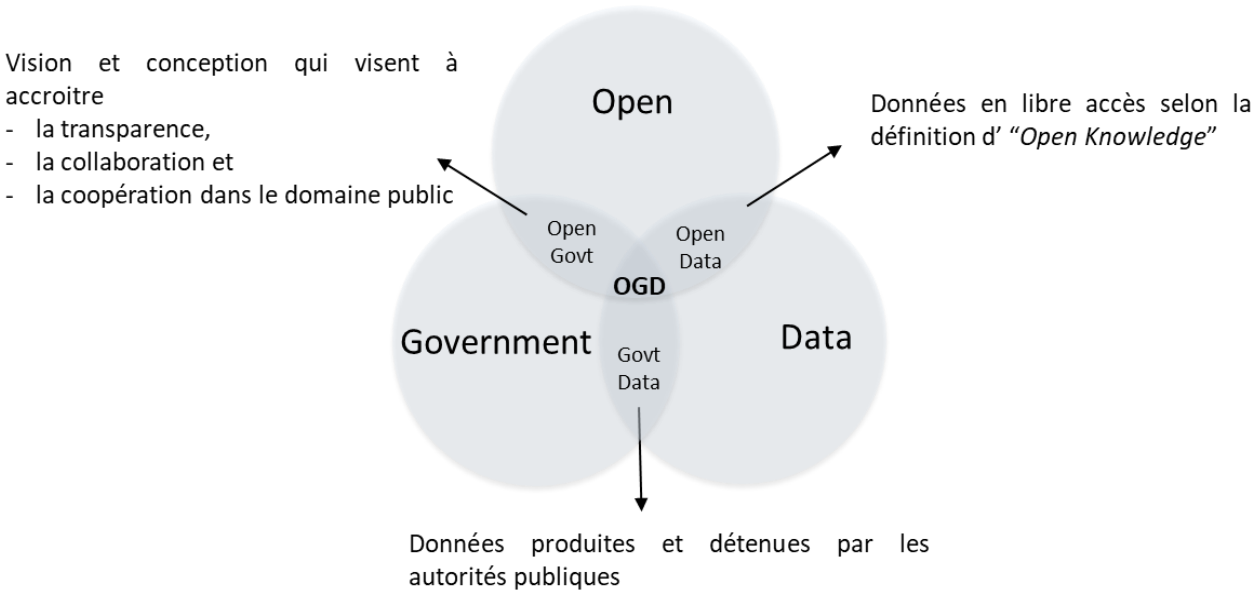

Figure 1 : Les OGD, adapté de (Gonzalez-Zapata and Heeks 2015) et (Conféderation Suisse 2014)

La stratégie définit les données publiques comme des données étant **collectées**, **établies**, **gérées**, **traitées** et **sauvegardées** par les autorités. Les OGD sont donc des données produites par l'administration, accessibles et réutilisables par n'importe qui pour n'importe quel but (Federal Council, 2014).

De par leurs différentes activités, les pouvoirs publics recèlent tous types de données. Les données publiques peuvent aussi bien être de nature

- Administrative
- Culturelle
- Statistique
- Légale
- Géospatiale
- Politique
- Mobilité et transport
- Construction et logement, etc.

Au sein des différents types de données détenues par les organisations publiques, on peut distinguer des données historiques, mais pas seulement. De plus en plus de données générées en temps réels (i.e. géospatiale, météorologique, énergie, etc.) émanent des activités de l'administration publique. Actuellement de nombreuses villes, à l'étranger mais également en Suisse, se sont déjà tournées vers l'usage des OGD. C'est par exemple le cas de la ville de Zurich, qui à l'aide des données sur les « systèmes de stationnement » et des « emplacements et informations sur le parking » a pu développer une application – *Liveparking* - permettant à ses citoyens d'identifier des places de parking disponibles (Marius B, 2019) ou à Genève d'offrir aux utilisateurs des Transports Publics Genevois (TPG) un chatbot permettant de connaître les horaires des prochains départs (Transports Publics Genevois, 2019).

Différences entre les OD et OGD :
Contrairement aux GAFAM, les données gouvernementales n'ont pas été pensées comme un actif à part entière, avec comme premier objectif leurs valorisations, mais bien comme des

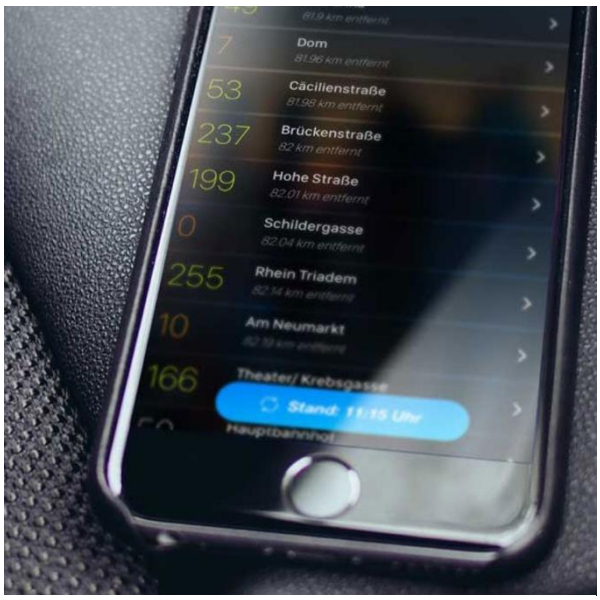

Figure 2 : Application Liveparking

ressources « en plus », disponibles et potentiellement réutilisables.
Voici quelques différences cruciales à prendre en compte lors de la mise en place d'une politique OGD :

Open Data

- Données naturellement issues du monde digital
- Volonté des propriétaires de les utiliser et de les publier

Open Governement Data

- Données qui en partie ne sont pas encore digitalisées
- Certaine réticence de la part des « propriétaires » à les publier
- Données personnelles et potentiellement sensibles

Pourquoi publier ?

Dans le contexte économique actuel, les données apparaissent comme le moteur de l'innovation et de nombreux gouvernements perçoivent leurs données comme une ressource d'importance stratégique (Bates, 2014; Munné, 2016). Il pourrait résulter de l'ouverture de ces données publiques diverses sortes de réutilisations, qui permettraient d'avoir de nombreux avantages (p.ex. économiques), pas seulement pour les

administrations publiques et le secteur privé (p.ex. applications, outils pour les smart city) mais aussi pour les citoyens (p.ex. plus de transparence, de participation) (Gascó-Hernández, Martin, Reggi, Pyo, & Luna-Reyes, 2018; Vickery, 2011).
Les politiques d'ouverture des données publiques poursuivent donc trois objectifs majeurs (CNIL, Cada, & Etalab, 2019) :

- Renforcer la **transparence** des administrations,
- Améliorer l'organisation et de la **gestion publique**,
- Accroître l'**innovation économique** via le développement de nouveaux services

Contexte international

Motivés par l'économie numérique et ces objectifs de transparence, d'ouverture et d'innovation, de nombreux États ont commencé, depuis près de dix ans, à prendre des mesures en vue de faciliter la réutilisation et la valorisation de leurs données restées jusque-là inexploitées (CNIL et al., 2019).
De plus en plus de plateformes gouvernementales permettant aux partenaires publiques de publier leurs données en libre accès voient le jour et avec elles le développement de diverses pratiques, recommandations et le développement de nouvelles lois à l'égard des gouvernements désireux de rendre publiques leurs données .Les États-Unis ont vu en janvier 2019 l'adoption de *l'OGD Act* , rendant obligatoire la publication des données publiques par les organismes fédéraux et la nomination d'un *Chief Data Officer* chargé de la mise en place de la loi et de cataloguer l'ensemble des données (The Senate and House of Representatives of the United States of America in Congress assembled, 2018).

Contexte national

En Suisse, c'est en avril 2014, dans le cadre de sa politique en cyberadministration, que le Conseil Fédéral a adopté la « Stratégie de libre accès aux données publiques suisses ». Plus connue sous le nom de Stratégie Open Governement Data, celle-ci a pour objectifs de libérer les données publiques, les rendre accessibles et d'établir une culture du libre accès. Bien qu'il s'agisse de l'objectif principal de ce premier opus, la stratégie OGD 2014-2018 apparaît néanmoins plus comme un état des lieux de l'OD en Suisse et fait le point sur les choses à mettre en place. Valeur d'obligation pour les administrations fédérales, la stratégie OD 2014-2018 n'a cependant que valeur de recommandation pour les administrations cantonales et communales (Conseil fédéral, 2014).
Poursuivant des buts de transparence, de participation et d'innovation, le Conseil fédéral dans son deuxième opus sur la stratégie en matière de libre accès aux données publiques 2019-2023 annonce que toutes les données publiées par ses services fédéraux devront être librement accessibles, gratuites et exploitables par un ordinateur et ce d'ici 2020 (Conseil fédéral, 2018). La stratégie distingue cette fois-ci les données déjà existantes (souvent pas encore numérisées), des nouveaux jeux de données (récoltés dans des formats plus facilement compatibles et exploitables par des ordinateurs), prônant dans les deux cas des publications dans un format, dès le départ, adaptées à la demande et effectuées le plus rapidement possible.
Afin de renforcer le caractère obligatoire de cette stratégie pour les institutions fédérales et l'importance de participation d'autres institutions administratives (i.e. cantonales, communales, paraétatiques ou particuliers), le Département fédéral de l'intérieur (DFI) a été chargé d'examiner le bienfondé de l'inscription des principes open data dans la loi. L'idée que la stratégie OGD soit déclarée force obligatoire pour toutes les institutions de tous les niveaux de l'administration suisse n'apparaît pas impossible (Conseil fédéral, 2018).

Où publier ?

En vue d'atteindre les objectifs des deux stratégies OGD, le Conseil fédéral a développé une infrastructure centralisée : le portail *opendata.swiss*. Anciennement administré par les Archives fédérales, il est à ce jour (août 2019) confié à l'Office fédéral de la Statistique (OFS)

(Conseil fédéral, 2018). La plateforme permet aux autorités Suisses de publier leurs données de manière à ce qu'elles soient plus facilement accessibles et réutilisables par d'autres organisations publiques mais aussi par les citoyens.
La publication des données publiques peut également s'effectuer au niveau d'infrastructure plus locales, sur des plateformes développées par les cantons (p.ex. Genève), villes (p.ex. Zurich) ou autres institutions publiques (p.ex. SIG, Bern Mobil).

Opportunités des OGD

Pour Janssen, Charalabidis, and Zuiderwijk (2012) la publication d'OGD pourrait générer plus de valeur que la vente de données publiques elles-mêmes. Les OGD pourraient stimuler l'innovation et favoriser le développement économique en permettant aux entreprises la création de nouveaux services ainsi que de nouveaux modèles d'économie digitale (A Zuiderwijk, Janssen, van de Kaa, & Poulis, 2016). Selon une étude européenne, les OGD pourraient générer plus de 40 milliards d'euro par ans (European Commission, 2014). Pour les citoyens, les OGD pourraient se traduire par plus de transparence des activités des gouvernements (Charalabidis et al., 2018), une augmentation de la participation citoyenne (Lourenço, 2015) et un meilleur contrôle des données par les citoyens (CNIL et al., 2019). Les OD seraient p.ex. pour les journalistes un outil de plus, venant renforcer la loi sur la transparence. En ce qui concerne les administrations publiques, l'ouverture de leurs données pourrait selon Toots et al. (2017) participer à l'amélioration de la responsabilité des administrations et à l'augmentation de l'efficacité des services publics. En normalisant la publication des données publiques, cela pourrait permettre aux différents services d'une même ville ou de différents cantons d'éviter la création de doublons et ainsi augmenter l'efficience des services et limiter certaines dépenses.

Risques des OGD

Malgré la passion que les OGD peuvent entrainer chez certains, la mise à disposition des données publiques n'est pas dépourvue de tous risques. Dans leur article, Martin, Foulonneau, Turki, and Ihadjadene (2013) identifient sept catégories de risques liés aux OD – la gouvernance, les questions économiques, le cadre juridique, les caractéristiques des données, les métadonnées, les accès et finalement les compétences. Ces différentes catégories de risques peuvent toucher différents acteurs tels que les citoyens, les entreprises privées, les chercheurs ou encore les administrations elles-mêmes. Elles pourraient publier, sans le vouloir de données personnelles ou à caractère sensible (Joo, Yoon, Kwon, & Lim, 2018), il pourrait y avoir une recrudescence du nombre de demandes de modification des données par les citoyens, mais aussi des dépenses inutiles liées à l'inefficience de la mise en place des OD (Anneke Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012). Selon Bozeman and Kingsley (1998) les plus gros risques pour les OGD proviennent de la culture de l'aversion aux risques. La publication des OGD pourrait avoir comme conséquences la violation de la vie privée des citoyens (Thurston, Childs, McLeod, Lomas, & Cook, 2014), leurs profilages (p.ex. avant un entretien d'embauche, avant d'effectuer un prêt bancaire, etc.) (Charalabidis et al., 2018) ou encore une meilleure personnalisation et ciblage des publicités. Quant aux entreprises, une mauvaise qualité des données publiées pourrait avoir pour effet la construction de services faussés et non-utilisables. L'utilisation de données personnelles à mauvais escient pourrait également avoir des effets néfastes pour les entreprises (Reale, 2014).

L'état des données publiées sur la plateforme *opendata.swiss*

Les bonnes pratiques pour publier des données ouvertes

Malgré la mise en place d'une infrastructure spécialisée et du développement de plateformes OGD, la publication de données permettant leur réutilisation n'est pas chose aisée. C'est pourquoi nombreux sont les défenseurs des OD à promouvoir l'application de bonnes pratiques pour améliorer l'ouverture des données publiques. La *Sunlight Fundation* (SF), entre autres, s'est penchée sur le sujet et à l'issue de plusieurs discussions a élaboré une liste de 10 principes à appliquer aux données publiques pour faciliter leur diffusion et améliorer leur réutilisation.

Les données doivent être (Sunlight Foundation, 2018):

Complètes : chaque données ou set de données devraient contenir toutes informations nécessaires à sa compréhension la plus détaillée et précise qu'il soit. Ces informations peuvent contenir tout type de renseignements bruts, des métadonnées (i.e. qui expliquent les données bruts), ou encore des formules expliquant l'agrégation des données dérivées.

Originales : Il s'agit des données de sources primaires comprenant les informations et les documents recueillis durant la collecte des données. L'idée derrière ce principe est d'assurer aux utilisateurs finaux l'exactitude des données recueillies.

Publiées de manière ponctuelle : pour une utilisation optimale les données devraient être publiées en temps réel ou à défaut, le plus rapidement possible, surtout en ce qui concerne les données sensibles au facteur temps.

Facilement accessibles : l'accès aux données ne devrait n'être entravé ni par des barrières d'ordre physique (i.e. déplacement sur place, en personne), ni d'ordre électronique (i.e. création de profil, formulaires etc.).

Lisibles par des machines : seul des formats permettant le traitement automatique des données devraient être utilisés sur les portails open data. Les notes manuscrites et fichiers PDF sont difficilement exploitable par des ordinateurs alors que des fichiers de type .json ou .csv, beaucoup plus.

Zéro discrimination : nulle identification ou justification ne devrait être requise pour accéder à des OD.

Format communément utilisé ou ouvert : l'utilisation de format de type open source (i.e. .txt, .csv) sont préférable à des formats propriétaires (i.e. .doc, .xls,).

Octroi de licence : étiqueter clairement chaque set de données avec ce qu'il est possible d'en faire (i.e. utilisation libre, utilisation libre avec obligation d'indiquer la source, utilisation libre mais utilisation à des fins commerciales uniquement avec l'autorisation du fournisseur de données, etc.).

Permanentes : C'est à la dire la possibilité d'avoir accès aux données au fils du temps. Cela nécessite la mise en place d'un système d'archivage accessible et la mise la disposition des informations en cas de modification, mise à jour et même suppression des données.

Gratuites : le coût des données est le facteur limitant le plus puissant dans la réutilisation des données. C'est pourquoi la gratuité des OD est inévitable pour garantir leur diffusion.

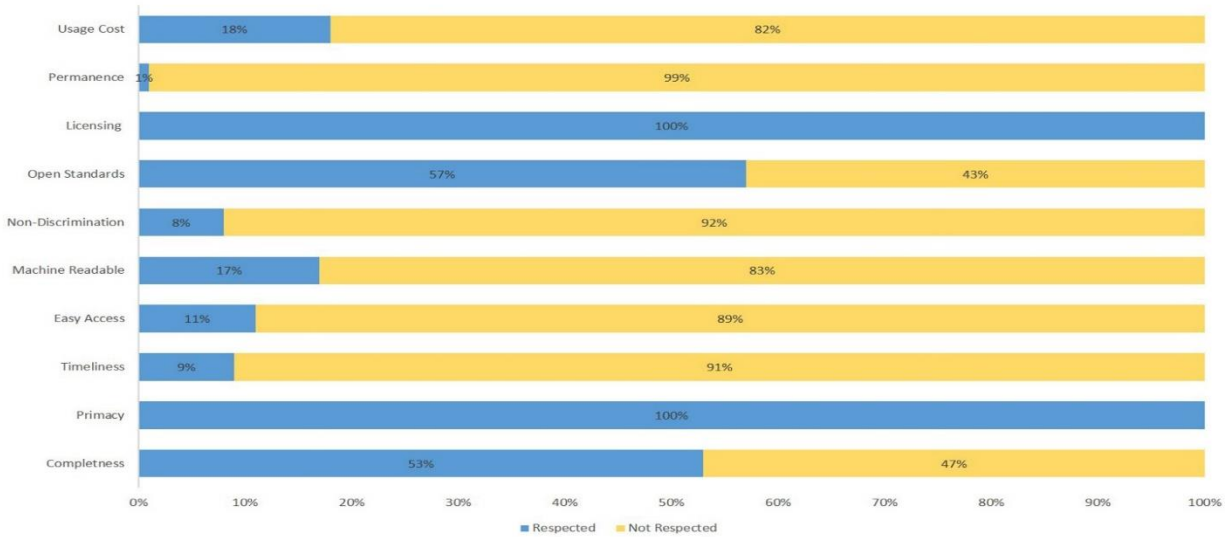

Figure 3. Adhérences aux bonnes pratiques de publication

Ce que nous avons observé

Suite à l'analyse de la plateforme *opendata.swiss*, nous avons développé un index mesurant le respect des 10 principes développés par la SF, à appliquer lors de la publication d'OGD (voir Figure 3). Nous avons pu observer que parmi les 10 principes destinés à améliorer la publication des données du gouvernement, seulement deux des principes (i.e. données originales et License) étaient respectés par l'ensemble des métadonnées publiées (i.e. cela signifie qu'il s'agissait de source des données primaires et que le type de licence utilisée était précisé). Deux des dix principes (i.e. données complètes et format communément utilisé ou ouvert) concernent uniquement la moitié des métadonnées disponibles sur la plateforme. Quant au reste des métadonnées, seul 1/5 respectent les six principes restants.

De plus nous avons pu observer durant cette étude que actuellement la plateforme *opendata.swiss* ne semblait pas publier de données de type personnel ou sensible (i.e. selon notre connaissance actuelle), mais plutôt des données sur la topographique, météorologique ou de géolocalisation.

Ce que l'on ne sait pas encore

Hormis le problème de non-application des principes recommandés pour l'ouverture de données publiques, l'autre problème majeur des OGD est celui de la mise à disposition de données sensibles ou personnelles. Leurs diffusions peuvent engendrer de graves conséquences et ne peuvent donc pas être publiées en l'état par l'administration publique. Actuellement en Suisse, outre les stratégies OGD, il n'existe pas de loi propre aux OGD. Le cadre légal concernant le respect de la vie privée et la protection des données ne fait pour le moment aucunement référence aux cas spécifiques des OGD. Les administrations qui ambitionnent de publier leurs données s'en remettent alors aux différentes lois sur des sujets proches tel que :

- La loi fédérale sur la protection des données (LPD)
- La loi sur la Transparence (LTrans),
- La loi sur le droit à l'information (LInfo)

Bien que des discussions soient en cours pour le développement, de nouvelles régulations concernant la publication de données gouvernementales, la crainte des villes et cantons de publier des données personnelle ou sensibles restreint et ralentit les initiatives d'OGD en Suisse.

Données personnelles et sensibles?

La directive européenne relative à la protection des données à caractère personnel (Le Parlement Européen et le Conseil de l'Union Européene, 2018) définit les données personnelles (ou *Personal Information Identifier* (PII) selon le *National Institut of Standards and Technology* (NIST)), comme toutes informations permettant d'identifier, directement ou indirectement un individu soit par référence à :

- Un numéro d'identification propre (i.e. numéro AVS) soit par,
- Un ou plusieurs éléments propres à l'identité physique (p.ex. signe distinctif), physiologique, psychique (p.ex. ADN), économique (p.ex. numéro carte de crédit), culturelle (p.ex. religion) ou sociale (p.ex. études, état civile).

Ces informations (i.e. les PII) sont séparées en deux catégories :

- Les identifiants directs ou *identifiers* (i.e. qui permettent l'identification direct d'un individu) et,
- Les identifiants indirects ou *quasi-identifiers* (i.e. qui permettent l'identification indirectes d'un individu, par combinaison ou croisement (Raghunathan, 2013; Simson, 2016)

Or les organisations publiques de par leurs activités, détiennent non seulement des données statistiques, météorologiques ou encore financières mais également de nombreux PII. Cependant, ni les données personnelles ni les données sensibles ne peuvent être publiées sur des plateforme OGD (p.ex. Site *opendata.swiss*) (Confédération Suisse; L'Assemblée fédérale de la Confédération suisse, 2011; Le Grand Conseil du Canton de Vaud, 2007).

Les administrations qui souhaitent publier leurs données doivent donc les dépersonnaliser avant de les proposer en libre accès (CNIL et al., 2019; L'Assemblée fédérale de la Confédération suisse, 2011).

Recommandations pour la dépersonnalisation des données ouvertes

De-identifier, anonymiser, pseudonymiser, qu'est-ce que cela signifie ?

La de-identification ou dépersonnalisation de données est un terme générique qui souvent fait référence aux processus d'anonymisation et de pseudonymisation des données permettant l'identification d'individu (Nguyen, 2015). On entend par anonymiser ou rendre anonyme, toutes démarches de dépersonnalisation d'identifiants ou de quasi-identifiants visant à empêcher l'identification des d'individus ou à ne rendre celles-ci possible qu'au prix d'efforts démesurés (Raghunathan, 2013; Simson, 2016).
L'**anonymisation** consiste donc à supprimer définitivement tous les identifiants et quasi-identifiants d'une base de données. Les données sont considérées comme étant anonymisées lorsqu'elles ne peuvent plus être attribuées à une personne (Gillieron, Desmarest, & Thévenet, 2018; Simson, 2016).
La **pseudonymisation** consiste à remplacer l'ensemble des données identifiables par un identifiant neutre (pseudonyme). Elle sert à limiter l'accès et la connaissance de ces données,

uniquement aux entités autorisées (p.ex. interdépartementale, entre chercheurs d'une même unité, etc.). Contrairement à l'anonymisation, cette approche est réversible et doit permettre à la donnée originale d'être retrouvée (p.ex. tables de correspondance, clés, etc.) (Simson, 2016).

À la différence de la pseudonymisation, l'anonymisation est utilisée dans les cas où les informations personnelles sont partagées avec des tiers. Cette technique trouve alors toute son utilité lors de partage de données publiques envers des tiers (i.e. secteur privé, recherche, citoyen etc.) (Personal Data Protection Commission Singapore, 2018; Raghunathan, 2013).

Pourquoi anonymiser des données ?

Les organisations publiques possèdent tous types de données et lors de la publication de données personnelles sous forme d'OGD, l'anonymisation des données peut permettre de protéger les propriétaires des données mises en ligne sur les plateformes. Cette technique offre plusieurs avantages et permet notamment de (Domingo-Ferrer, Sánchez, & Soria-Comas, 2016; Raghunathan, 2013):

- Respecter certaines obligations légales (i.e. LPD, LTrans, LInfo, etc.),
- Prévenir les violations et l'utilisation abusive des données (p.ex. transparent citizen),
- Publier les données,
- Protéger les citoyens et leurs données sensibles,
- Augmenter la réutilisation des données.

Quand anonymiser des données ?

La publication de données personnelles n'est possible en libre accès que si elle ne permet pas l'indentification du propriétaire des données (CNIL et al., 2019; L'Assemblée fédérale de la Confédération suisse, 2011). Cela signifie que seules les données qui ne sont plus considérées comme personnelles peuvent être publiées sur des plateformes d'OGD pour ensuite être réutilisées. Actuellement, l'anonymisation est la seule approche qui permet de répondre à ces attentes (i.e. changer des données personnelles en des données non-personnelles) (Domingo-Ferrer et al., 2016; Raghunathan, 2013).

Quels sont les risques ?

Les données anonymisées peuvent être vulnérables aux attaques de « corrélation » (Sedayao, Bhardwaj, & Gorade, 2014). Le croisement avec d'autres sources de données, plus connu sous le nom du « *mosaic effect* » (i.e. la possibilité que des ensembles de données et des informations variées puissent être combinés pour révéler des informations sensibles) (The Center for Open Data Entreprise, 2016) peut permettre d'identifier des individus, même si les données permettant leur identification (i.e. identifiant et quasi-identifiants) ont été anonymisées. Pour différentes techniques d'anonymisation, il existe différents risques (Information Commissioner's Office, 2012). Les plus traités dans la littérature sont les risques de divulgation des données (Domingo-Ferrer et al., 2016; Information Commissioner's Office, 2012; Raghunathan, 2013; Simson, 2016). Le Personal Data Protection Commission Singapore (2018) distingue trois types de risques de divulgation des données:

- *Identity discosure* : violation de la vie privée d'un individu lorsqu'un intrus est capable d'associer un document de l'ensemble des données divulguées à la personne qui en est à l'origine (Domingo-Ferrer et al., 2016).
- *Attribute disclosure* : lorsque l'accès aux données divulguées permet à l'intrus de déterminer la valeur d'une information confidentielles d'une personne avec

suffisamment de précision (Domingo-Ferrer et al., 2016).

- *Inference dislosure* : faire des inférences statistiques, concernant un individu alors qu'il ne fait pas partie de la base de données, pour en déduire certaines informations à son sujet (Personal Data Protection Commission Singapore, 2018).

L'anonymisation est-elle suffisante?

Le risque zéro n'existe pas de même qu'il n'existe pas de techniques infaillibles qui garantissent une parfaite protection des données personnelles (The Center for Open Data Entreprise, 2016). Les techniques d'anonymisation tendent à se concentrer plus sur la protection des risques de type *identity disclosure*, que sur les autres types des risques tels que *attribute disclosure* et *inference disclosure* (Raghunathan, 2013), mais l'anonymisation reste l'option la plus sûre (CNIL et al., 2019).

Comment minimiser le risque de divulgation de données sensibles?

Bien qu'il soit difficile d'écarter tout risques de ré-identification possible (i.e. *identity disclosure*) (Information Commissioner's Office, 2012), certaines solutions existent pour se prémunir, minimiser et se préparer aux risques liés à la divulgation de données personnelles (Personal Data Protection Commission Singapore, 2018). Une fois que les buts et avantages de la publication d'OGD ont bien été intégrés, la mise en œuvre d'une **solution d'anonymisation** pour les données publiques peu commencer. Elle comprend (Domingo-Ferrer et al., 2016; Personal Data Protection Commission Singapore, 2018; Raghunathan, 2013):

- Une phase d'analyse de l'architecture (i.e. les données destinées à être publiées en open data sont analysées, leurs type, source etc.)
- Une Phase d'analyse de sensibilité (i.e. les différentes catégories de données sensibles, personnelles, non-personnelles sont identifiées.
- Une phase d'évaluation des risques (i.e. évaluation du risque tels que *identity disclosure, attribute disclosure, inferential disclosure*)
- Une phase pour déterminer les objectifs en matière de protection de la vie privée, d'utilisabilité des données et d'accès (i.e. utilité Vs. risque encouru, plus le niveau de risque est important plus l'anonymisation doit être forte)
- Une phase de design de l'anonymisation (i.e. choix des techniques et modèles utilisés pour le traitement des données à risques)
- Une phase d'implémentation, de test et de déploiement des techniques d'anonymisation choisie
- Une phase d'exploitation (i.e. inclusion de la solution d'anonymisation lors de la publication des données sur la plateforme OGD)

Selon (Raghunathan, 2013; Simson, 2016) , certaines conditions devraient également être implémentées lors d'un projet d'OGD:

- Le cadre légal auquel l'organisation répond doit être clair et compris,
- Mise en place d'une équipe en charge de la gouvernance de la protection de la vie privée et de l'anonymisation des données, qui comprenne les enjeux et les risques,
- Mise en place de politique de sécurité et de confidentialité des données,
- Mise en place d'une équipe de gestions des risques et incidents de divulgation des données,
- Déterminer le niveau de risque qu'une organisation est prête à accepter,

- Mise en place d'un suivi et d'une évaluation des techniques utilisées.

En plus de la mise en place d'une solution d'anonymisation, Il est également possible pour les gouvernements de réfléchir à d'autres solutions :

- Licences d'exploitations sur les différentes bases de données (i.e. conditions d'utilisation des données mises à disposition : utilisation libre mais obligation d'utiliser la source, utilisation à des fins commerciales uniquement avec autorisation du fournisseur de données, etc.)
- Consentement général d'utilisation (p.ex. personnes volontaires qui acceptent que leurs données soient réutilisées, p.ex. consentement générale du CHUV)
- OGD semi-open: diffèrent degrés d'ouverture selon l'utilisation (i.e. chercheur, citoyens, à des fins commercial, etc.) (Open Knowledge Fundation, 2017) ou selon le degré d'ouverture souhaité (i.e. *open data*, *public data*, *access-controled data* and *data gasp*) .

Techniques d' anonymisation

Une des complexités de l'anonymisation réside dans le fait qu'il faut non seulement choisir une technique appropriée au contexte d'utilisation des données (i.e. données individuelles, agrégées), mais aussi en fonction du type de données (i.e. discrètes, continu). Cela signifie qu'il faut non seulement comprendre et connaître les données (i.e. être en mesure de définir s'il s'agit d'identifiant, de quasi-identifiants ou de données sans risque, comprendre leur définition, leurs fonctionnements, etc.) mais également identifier leurs probables cas d'utilisation. Le challenge avec les OD et les OGD c'est que même si les bonnes pratiques de publications sont appliquées et que l'on possède suffisamment d'informations à leurs égards, les cas d'utilisations ou d'application des données restent inconnus.

Autre complexité, les techniques d'anonymisations nécessitent des connaissances profondes et précises des statistiques, de l'informatique, du juridique et des domaines des données que l'on cherche à anonymiser. Ce qui engendre une collaboration entre différents acteurs pour garantir la mise en place d'une solution d'anonymisation efficace.

Différentes sortes de données requiert différentes sortes de techniques (Simson, 2016). Dans la littérature, apparait donc diverses techniques d'anonymisation, présentées sous différent noms et souvent classées selon différente approches, selon si l'on s'intéresse :

- Au type de donnée (i.e. discrète vs. continu) (Raghunathan, 2013)
- Au niveau de granularité requise pour l'utilisation (i.e. perturabtive vs. non-perturabtive) (Domingo-Ferrer et al., 2016)
- Au niveau d'efficacité des techniques (i.e. simple à complexe) (Raghunathan, 2013), ou
- Au identifiants et quasi-identifiants (Simson, 2016)

Présenté ci-après une liste, non exhaustive des techniques d'anonymisation les plus utilisées pour dépersonnaliser des données publiques.

Suppression/ Nulling out

Suppression d'une partie ou de l'entier des données. Cette technique est utilisée dans le cas où l'information n'est pas nécessaire ou lorsque que l'information ne peut pas être anonymisées de façon efficace avec d'autres techniques (Domingo-Ferrer et al., 2016; Personal Data Protection Commission Singapore, 2018; Simson, 2016).

Il existe plusieurs types de suppression :

- *Attribut suppression* : Suppression de certaine valeur de données, aberrantes et/ou unique,
- *Local suppression* : Suppression de valeurs extrême.

Base de données avant data suppression

Étudiants	Professeurs	Résultats
Mia	Guliano	A
Pim	Jean-Loup	B
Jérôme	Jean-Loup	D
Stefan	Tobias	A
Laura	Pirmin	C

Base de données après data suppression:

Professeurs	Résultats
Jean-Loup	B
Jean-Loup	D
Guliano	A
Tobias	A
Pirmin	C

Shuffling/ Swapping

Réarrangement des données au sein d'une même colonne ou d'un set de données. Cette technique peut être utilisé dans le cas où les valeurs des données individuelles doivent toujours être accessible sans pour autant être directement rattaché à leur propriétaire (Raghunathan, 2013) (Personal Data Protection Commission Singapore, 2018).

Adresses immeubles avant shuffling

ID	Rue	Ville	Canton	NPA
48	Avenue du mont d'or 32	Lausanne	Vaud	1007
49	Rue de la côte, 16	Montreux	Vaud	1820
50	Rue de la gare, 42	St Maurice	Valais	1890
51	Rue du Simplon, 18	Versoix	Genève	1290

Adresses immeubles après shuffling

ID	Rue	Ville	Canton	NPA
48	Avenue du mont d'or 32	Versoix	Genève	1290
49	Rue de la côte, 16	St Maurice	Valais	1890
50	Rue de la gare, 42	Lausanne	Vaud	1007
51	Rue du Simplon, 18	Montreux	Vaud	1820

Generalisation/ Noise addition

Réduction de la précision des données par combinaison des données ou par addition de « bruit ». Cette technique est plutôt utilisée pour les données discrètes, et lorsque que la généralisation des données ne pose pas de problème lors de l'utilisation des données.

Il existe différent type de « addition » telles que (Personal Data Protection Commission Singapore, 2018):

- Addition de bruit non corrélé / Addition de bruit corrélé
- Addition de bruit et transformation linéaire / Addition de bruit et transformation non linéaire
- Data perturbation

Base de données avant « noise additon »

ID	Personne	Age	Adresse
1	673298	24	Rue de l'Ale
2	098762	31	Rue de Bourg
3	415426	44	Avenue Floréal
4	162739	29	Rue du Pont
5	092673	23	Rue du Valentin
6	738298	75	Avenue des Alpes
7	565443	28	Place de la Louve
8	906173	50	Escalier Hollard
9	124510	30	Rue Neuve
10	096617	37	Rue Voltaire

Base de données après « noise additon »

ID	Personne	Age	Adresse
1	673298	20-24	Rue de l'Ale
2	098762	30-34	Rue de Bourg
3	415426	40-44	Avenue Floréal
4	162739	25-29	Rue du Pont
5	092673	20-24	Rue du Valentin
6	738298	>60	Avenue des Alpes
7	565443	25-29	Place de la Louve
8	906173	50-55	Escalier Hollard
9	124510	30-34	Rue Neuve
10	096617	35-39	Rue Voltaire

Masking techniques

Modification des caractères d'une valeur de données par l'ajout d'un symbole. Cette technique est utilisée lorsqu'il n'est pas nécessaire d'anonymiser tous les caractères d'une valeur de données (Personal Data Protection Commission Singapore, 2018; Raghunathan, 2013).

- Caracter masking
- Social security number masking technique
- String format masking
- Numeric format masking
- Format-Specific Credit Card Masking Technique
- Phone Number Masking Technique
- E-mail ID Masking Technique

NPA avant caracter masking

ID	*Rue*	*Ville*	*Canton*	*NPA*
48	Avenue du mont d'or 32	Lausanne	Vaud	1007
49	Rue de la côte, 16	Montreux	Vaud	1820
50	Rue de la gare, 42	St Maurice	Valais	1890
51	Rue du Simplon, 18	Versoix	Genève	1290

NPA après caracter masking

ID	*Rue*	*Ville*	*Canton*	*NPA*
48	Avenue du mont d'or 32	Lausanne	Vaud	10XX
49	Rue de la côte, 16	Montreux	Vaud	18XX
50	Rue de la gare, 42	St Maurice	Valais	18XX
51	Rue du Simplon, 18	Versoix	Genève	12XX

Aggregation

Conversion d'un ensemble de données d'un dataset en de valeurs compressées. Cette technique peut être utilisée dans les cas ou l'utilisation de donnée individuel n'est pas requise pour l'analyse.

- Data aggregation,
- Microagregation

Base de données avant agrégation :

Donneurs	*Revenu Chf*	*Donation Chf*
A	4000	210
B	4200	420
C	2200	150
D	4200	110
E	3500	260
F	2600	40
G	3300	130

Base de données après agrégation :

Revenus Chf	*Nb de donation*	*Somme de donations*
1000- 1999	0	0
2000-2999	2	190
3000-3999	2	390
4000-4999	3	740
5000-6000	0	0

Méthodes d'anonymisation

L'anonymisation des données personnelles est une étape nécessaire pour assurer le respect de la vie privée et permettre la publication des données publiques sur une plateforme OGD. Dans le cas d'identifiants directs (i.e. nom, numéro AVS, etc.), l'anonymisation est simple et peut se traduire par l'application de techniques de d'anonymisation présentées précédemment ou par leur retrait complet. Néanmoins, dans le cas des quasi-identifiants (i.e. l'âge, le sexe, le code postal, etc.), la mise en œuvre d'une solution d'anonymisation ne peut se limiter à une simple dépersonnalisation (Domingo-Ferrer et al., 2016). Elle nécessite une vision plus globale de la situation. (Personal Data Protection Commission Singapore, 2018). Les quasi-identifiants tels que l'âge, le sexe, le code postal, etc. de par leur combinaison rende la ré-identification possible (Domingo-Ferrer et al., 2016). Ils existent des modèles de protection de la vie privée, utilisant des techniques d'anonymisations, qui inclus la problématique des quasi-identifiants et qui tentent de limiter ses risques de ré-identification (Domingo-Ferrer et al., 2016).

- K-anonymity privacy model
- L-diversity model
- T-closeness model
- Differential Privacy

k-anonymity model

Le *k-anonymity* est un modèle de protection de la vie privée qui limite les attaques de ré-identification de données en rendant chaque quasi-identifiants impossible à distinguer au sein d'un groupe de k autres quasi-identifiants. Pour un groupe de deux quasi-identifiants âge / code postal l'idée est d'avoir au moins deux individus du même âge vivant au même endroit. Il devient par conséquent, plus difficile d'identifier un individu avec certitude, et ce même après le croisement de plusieurs informations (Domingo-Ferrer et al., 2016). Concrètement, le *k-anonymity model* garantit qu'il y a au moins k individus dans une base de données qui partageant un même ensemble de quasi-identifiants (Nguyen, 2015) (p.ex. k individus qui partagent le même code postal et le même âge).

l-diversity model

Le modèle *l-diversity* est une extension du modèle *k-anonymity*. Plus une base de données est conséquente, plus il devient difficile de garantir qu'il y a au moins k individus qui partagent les même quasi-identifiants. L'idée est donc de définir un niveau minimal de diversité l au sein de chaque groupe de quasi-identifiants. En d'autre termes, un groupe doit au moins contenir l quasi-identifiants différents pour limiter le manque de diversité (Source expert).

t-closeness model

Extension du modèle *k-anonymity*, le *t-closeness* model vise à réduire toujours plus l'information observable, en limitant le nombre de corrélations possibles. Pour ce faire, cette méthode cherche à répondre à cette question : Comment partitionner mes données de telle sorte que toutes les partitions se ressemblent en termes de distribution ? (Nguyen, 2015)

Differential Privacy

Le *Differential Privacy* est un des rare modèle qui limite mathématiquement les informations que l'on puisse obtenir sur un individu. Elle cherche à troubler la re-identification en introduisant un échantillonnage des vraies données (p.ex. avec une probabilité de x) et en générant des données fictives (p.ex. avec une probabilité de y) (Nguyen, 2015).

Afin d'appliquer ses différentes techniques, le *National Institute of Standards and Technology* met à disposition une liste des outils recommandés (The National Institute of Standards and Technology (NIST)).

Conclusion

L'ouverture et la diffusion des données publiques permettant leur réutilisation ne signifient pas seulement la digitalisation des archives du gouvernement pour ensuite les publier sur une plateforme. L'OGD est une révolution au long court qui demande de réels investissements, qu'il s'agisse de la mise à jour du cadre légal, du changement de mentalité des administrations ou de celle des citoyens. Cette révolution a déjà commencé et nombreux sont les pays à l'avoir compris. L'*open data act* adopté début 2019 par les États-Unis, en est la preuve et nombreux sont les citoyens, organisations et adeptes des OD à promouvoir l'utilité, le bien-fondé et la nécessité des OGD. La Suisse tout comme la France ou le Royaume-Uni a commencé l'implémentation de stratégie OGD, via de nouvelles lois, plateformes de diffusion ou encore grâce à des activités de type *hackaton* contribuant au changement de mentalité. Mais pour que les OGD atteignent leurs objectifs, les efforts ne doivent pas s'arrêter. C'est en prenant conscience de la valeur que représentent les OGD, en standardisant les processus de publication, et en définissant les risques (i.e. données privées et sensibles) que les gouvernements seront à même de définir les solutions les plus adaptées et ainsi que l'amélioration de la publication et donc la réutilisation de OGD aura lieu.

Bibliographie

Bates, J. (2014). The strategic importance of information policy for the contemporary neoliberal state: The case of open government data in the United Kingdom. *Government Information Quarterly, 31*(3), 388-395.

Bozeman, B., & Kingsley, G. (1998). Risk culture in public and private organizations. *Public administration review, 58*(2), 109-118.

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). Open data evaluation models: Theory and practice. In *The World of Open Data* (pp. 137-172): Springer International Publishing.

CNIL, Cada, & Etalab. (2019). *Guide pratique de la publication en ligne et de la réutilisation des données publiques (open data)*. Retrieved from https://www.cnil.fr/sites/default/files/atoms/files/guide_open_data.pdf

Confédération Suisse. Handbook Open Data Swiss. Retrieved from https://handbook.opendata.swiss/en/publish/swiss.html

Conseil fédéral. (2014). *Strategie en matiere de libre accès aux données publiques en Suisse pour les années 2014 à 2018*. Retrieved from https://www.admin.ch/opc/fr/federal-gazette/2014/3347.pdf

Conseil fédéral. (2018). *Stratégie en matière de libre accès aux données publiques en Suisse pour les années 2019 à 2023 (Stratégie open government data, OGD)*. Retrieved from https://www.admin.ch/opc/fr/federal-gazette/2019/855.pdf

Domingo-Ferrer, J., Sánchez, D., & Soria-Comas, J. (2016). Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy & Trust, 8*(1), 1-136.

European Commission. (2014). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/communication-commission-european-parliament-council-european-economic-and-social-committee-a-0

Federal Council. (2014). Open Government Data Strategy for Switzerland 2014 - 2018. Retrieved from https://www.egovernment.ch/index.php/download_file/force/761/3631/

Gascó-Hernández, M., Martin, E. G., Reggi, L., Pyo, S., & Luna-Reyes, L. F. (2018). Promoting the use of open government data: Cases of training and engagement. *Government Information Quarterly, 35*(2), 233-242.

Gillieron, K., Desmarest, Y., & Thévenet, B. (2018). Pseudonymisation et anonymisation. Retrieved from https://www.e-xpertsolutions.com/pseudonymisation-et-anonymisation/

Information Commissioner's Office. (2012). *Anonymisation: managing data protection risk code of practice*. Information Commissioner's Office Retrieved from https://ico.org.uk/media/1061/anonymisation-code.pdf

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management, 29*(4), 258-268.

Joo, M., Yoon, S., Kwon, H., & Lim, J. (2018). De-identification policy and risk distribution framework for securing personal information. *Information Polity, 23(1)*, 1-25.

Loi fédérale sur la protection des données (LPD) (2011).

Loi sur la protection des données personnelles (LPrD), (2007).

Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (2018).

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly, 32*(3), 323-332.

Marius B. (2019). Live Parking Application. Condacore, Montabaur (D). Retrieved from https://www.stadt-zuerich.ch/portal/de/index/ogd/anwendungen/2017/Liveparking.html

Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). *Open data: Barriers, risks and opportunities.* Paper presented at the Proceedings of the 13th European Conference on eGovernment: ECEG, England.

Munné, R. (2016). Big data in the public sector. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe* (pp. 195-208): Springer.

Nguyen, B. (2015). Techniques d'anonymisation. *Statistique et Société, 2*(4), 43-50.

Open Data Charter. (2018). Open Data Principles. Retrieved from https://opendatacharter.net/

Open Knowledge Fundation. (2014). What is Open? Retrieved from https://okfn.org/opendata/

Open Knowledge Fundation. (2017). Global open data index methodology. Retrieved from https://index.okfn.org/methodology/

Open Knowledge Fundation. (2018a). Defining Open Data. Retrieved from https://blog.okfn.org/2013/10/03/defining-open-data/

Open Knowledge Fundation. (2018b). The Open definition. Retrieved from http://opendefinition.org/

Personal Data Protection Commission Singapore. (2018). *Guide to basic data anonymisation techniques*. Retrieved from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

Raghunathan, B. (2013). *The complete book of data anonymization: from planning to implementation*: Auerbach Publications.

Reale, G. (2014). Opportunities and differences of open government data policies in Europe. *Athens Journal of Social Sciences, 1*(3), 195-205.

Sedayao, J., Bhardwaj, R., & Gorade, N. (2014). *Making big data, privacy, and anonymization work together in the enterprise: experiences and issues.* Paper presented at the 2014 IEEE International Congress on Big Data, Anchorage, AK, USA.

Simson, G. (2016). *De-Identifying Government Datasets*. Retrieved from https://csrc.nist.gov/publications/detail/sp/800-188/draft

Sunlight Foundation. (2018). Ten principles for opening up government information. Retrieved from https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

The Center for Open Data Entreprise. (2016). *Briefing Paper on Open Data and Privacy*. Retrieved from http://reports.opendataenterprise.org/BriefingPaperonOpenDataandImprovingDataQuality.pdf

The Senate and House of Representatives of the United States of America in Congress assembled. (2018). Foundations for Evidence-Based Policymaking Act of 2018. Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/4174/text

Thurston, A., Childs, S., McLeod, J., Lomas, E., & Cook, G. (2014). Opening research data: Issues and opportunities. *Records Management Journal, 24*(2), 142-162.

Toots, M., McBride, K., Kalvet, T., Krimmer, R., Tambouris, E., Panopoulou, E., Tarabanis, K. (2017). *A framework for data-driven public service co-production*. Paper presented at the International

Conference on Electronic Government, St. Petersburg, Russia.

Transports Publics Genevois. (2019). Application Tipigee. Retrieved from http://www.tpg.ch/web/7289503/34

Vickery, G. (2011). *Review of recent studies on PSI re-use and related market developments*. Paris: Information Economics.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government, 10*(2), 156-172.

Zuiderwijk, A., Janssen, M., van de Kaa, G., & Poulis, K. (2016). The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments. *Information Polity, 21*(3), 223-236.

UNIL | Université de Lausanne

Institut de hautes études
en administration publique

Contact

Unité Management de l'Information
IDHEAP, Université de Lausanne
Rue de la Mouline 28
CH-1022 Chavannes-près-Renens

Téléphone : +41 21 692 69 50
E-mail : im-idheap@unil.ch
Web : https://www.unil.ch/idheap