MOD 2.1 - « DÉFIS INFORMATIQUE DU BIG-DATA »

Open Data et *Linked Open Data*

Stéphane Derrode, Dpt MI -
stephane.derrode@ec-lyon.fr

RGPD

Protection des données personnelles au niveau Européen

Protection des données personnelles

Loi de 1978 dite « Informatique et libertés ». Le président Valéry Giscard d'Estaing décide la création d'un organisme de contrôle des données personnelles dans la société de l'information : **la CNIL**, et avec elle est promulguée une loi majeure, la **loi de 1978 dite « Informatique et libertés ».**

Les droits

- le droit d'information : chacun peut être informé des traitements dont ses données font l'objet, cet article est applicable en toutes circonstances, même aux cas relevant de la sécurité nationale ;
- le droit d'accès (et d'effacement, droit à l'oubli) : plus complet que le droit d'information, il permet à chacun d'accéder aux informations qui sont conservées sur lui, il est toutefois interdit d'en faire usage dans certains cas ;
- le droit de rectification : chacun peut demander à faire corriger les données stockées le concernant;
- le droit d'opposition : chacun peut s'opposer à faire l'objet d'un traitement, pour un motif légitime (le démarchage commercial est reconnu par la loi comme motif légitime).

Ces droits se retrouveront, amplifiés et adjoints à d'autres, dans le RGPD.

Modification de 2004 : La loi de 2004 met aussi en place un nouveau système relatif à la déclaration des fichiers et traitements ; toute structure souhaitant effectuer un traitement de données à caractère personnel ou stocker ces données doit effectuer une déclaration à la CNIL.

Source: Zeste de savoir

RGPD (Loi Européenne, 2015) : Règlement Général de la Protection des Données

Concrètement, toute donnée concernant *un citoyen* européen, même traitée hors union, est dans le champ d'application du règlement ; c'est un cadre très large et protecteur pour les Européens.

"Le présent règlement protège les libertés et droits fondamentaux des personnes physiques, et en particulier leur droit à la protection des données à caractère personnel."

Pour les structures (y compris les administrations) traitant de la donnée personnelle, le principe de déclaration obligatoire à la CNIL est transformé en principe de responsabilité, permettant bien plus de souplesse, mais augmentant les sanctions (amendes dissuasives).

En ce qui concerne les utilisateurs, leurs droits sont renforcés, avec notamment l'arrivée du consentement « explicite » : le traitement des données personnelles est soumis à un consentement qui ne peut être forcé ; particulièrement, l'accès au service ne peut être conditionné à acceptation de traitement de données qui n'y seraient pas directement nécessaires.

Aux 4 droits de la lois Informatique et Liberté française, s'ajoute:

- le droit à la limitation du traitement (effacement partiel);
- le droit à la portabilité : permet à chaque citoyen de demander l'intégralité des données à caractère personnel les concernant.

2 MIN POUR COMPRENDRE LE RGPD : GALÈRE OU OPPORTUNITÉ ?

Source: Zeste de savoir

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle (texte final : fin 2023)

En avril 2021, la Commission européenne a proposé le premier cadre réglementaire de l'UE pour l'IA. Il propose que des systèmes d'IA qui peuvent être utilisés dans différentes applications soient **analysés et classés en fonction du risque qu'ils présentent pour les utilisateurs**. Les différents niveaux de risque impliqueront plus ou moins de réglementation. Une fois approuvées, ces règles seront les premières au monde sur l'IA.

La priorité du Parlement est de veiller à ce que les systèmes d'IA utilisés dans l'UE soient sûrs, transparents, traçables, non discriminatoires et respectueux de l'environnement.

Risque inacceptable : Les systèmes d'IA à risque inacceptable sont des systèmes considérés comme une menace pour les personnes et seront interdits. Ils comprennent:

- la manipulation cognitivo-comportementale de personnes ou de groupes vulnérables spécifiques : par exemple, des jouets activés par la voix qui encouragent les comportements dangereux chez les enfants
- un score social : classer les personnes en fonction de leur comportement, de leur statut socio-économique, de leurs caractéristiques personnelles
- des systèmes d'identification biométrique en temps réel et à distance, tels que la reconnaissance faciale

Certaines exceptions peuvent être autorisées : par exemple, les systèmes d'identification biométrique à distance "a posteriori", où l'identification se produit après un délai important, seront autorisés à poursuivre des crimes graves et seulement après l'approbation du tribunal

Source: Actualités du parlement Européen

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle

Risque élevé : Les systèmes d'IA qui ont un impact négatif sur la sécurité ou les droits fondamentaux seront considérés comme à haut risque et seront divisés en deux catégories.

1. **Les systèmes d'IA** qui sont utilisés dans les produits relevant de la législation de l'UE sur la sécurité des produits. Cela comprend les jouets, l'aviation, les voitures, les dispositifs médicaux et les ascenseurs.
2. **Les systèmes d'IA** relevant de huit domaines spécifiques qui devront être enregistrés dans une base de données de l'UE :
   - l'identification biométrique et la catégorisation des personnes physiques
   - la gestion et l'exploitation des infrastructures critiques
   - l'éducation et la formation professionnelle
   - l'emploi, la gestion des travailleurs et l'accès au travail indépendant
   - l'accès et la jouissance des services privés essentiels et des services et avantages publics
   - les forces de l'ordre
   - la gestion de la migration, de l'asile et du contrôle des frontières
   - l'aide à l'interprétation juridique et à l'application de la loi.

Tous les systèmes d'IA à haut risque seront évalués avant leur mise sur le marché et au long de leur cycle de vie.

L'IA générative, comme ChatGPT, devrait se conformer aux exigences de transparence :

- indiquer que le contenu a été généré par l'IA
- concevoir le modèle pour l'empêcher de générer du contenu illégal
- publier des résumés des données protégées par le droit d'auteur utilisées pour la formation

AI Act (Européen) : première réglementation (mondiale) de l'intelligence artificielle (texte final : fin 2023)

Risque limité :

Les systèmes d'IA à risque limité doivent respecter des exigences de transparence minimales qui permettraient aux utilisateurs de prendre des décisions éclairées. Après avoir interagi avec les applications, l'utilisateur peut alors décider s'il souhaite continuer à l'utiliser. Les utilisateurs doivent être informés lorsqu'ils interagissent avec l'IA.

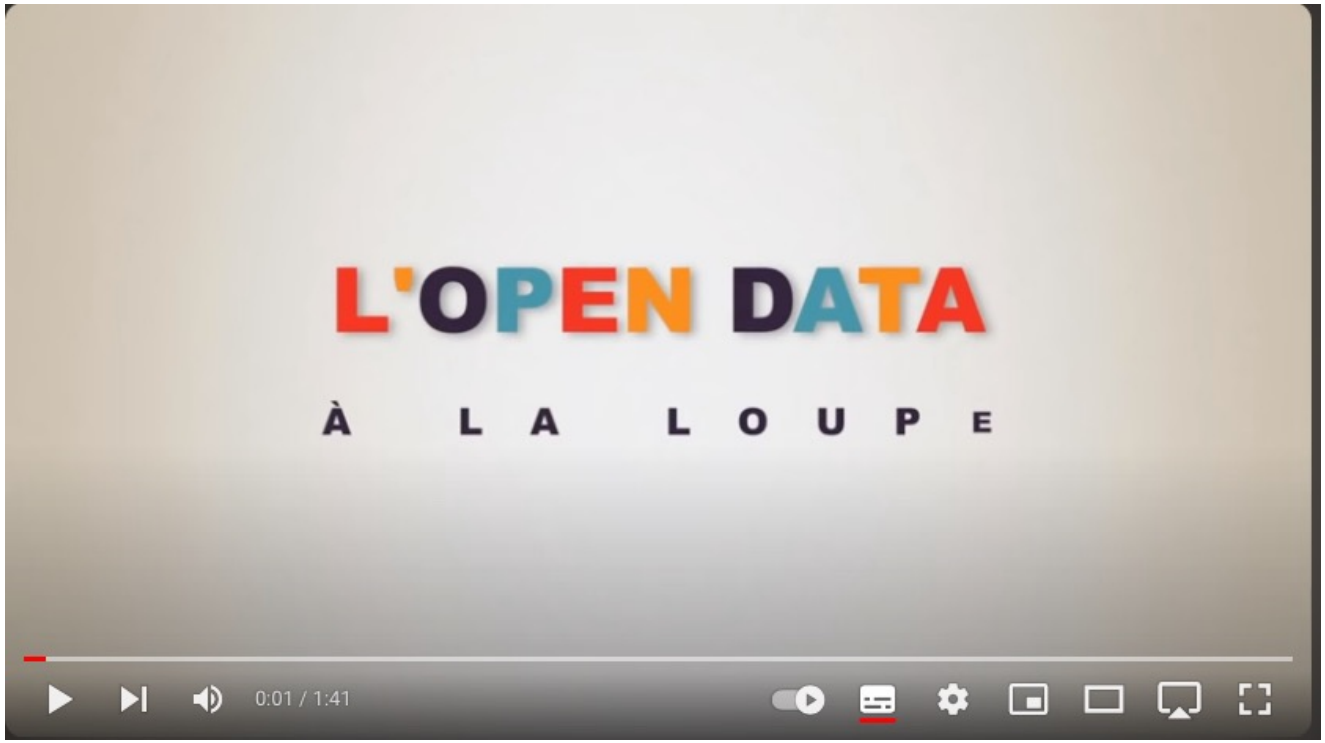
Cela inclut les systèmes d'IA qui génèrent ou manipulent du contenu image, audio ou vidéo (par exemple, les deepfakes, des contenus faux qui sont rendus crédibles par l'IA).

Source: Actualités du parlement Européen

OPEN DATA

1. Définition, enjeux
2. Les données, gisements
3. Licences, formats, outils d'exploitation

Open data à la loupe



Spot de présentation de la démarche d'ouverture des données numériques publiques, initié par LiberTIC, soutenu par Nantes Métropole et réalisé par A2B Production en licence Art Libre.

Open Data c’est quoi?

- **Définition :** L'***open data*** ou **donnée ouverte** est une donnée numérique dont l'accès et l'usage sont laissés libres aux usagers. Elle peut être **d'origine publique ou privée**, produite notamment par une collectivité, un service public ou une entreprise. Elle est diffusée de manière structurée selon une méthode et une **licence ouverte** garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

- Ces droits d'accès et de réutilisation s'inscrivent dans la pensée qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général.
- L'ouverture des données (*open data*) est à la fois **un mouvement**, **une philosophie d'accès à l'information** et une **pratique de publication de données** librement accessibles et exploitables.
- Le mouvement s'est étendu notamment sous l'impulsion d'ONG comme l'Open Knowledge Foundation (OKFN) et le Partenariat pour un gouvernement ouvert (PGO).

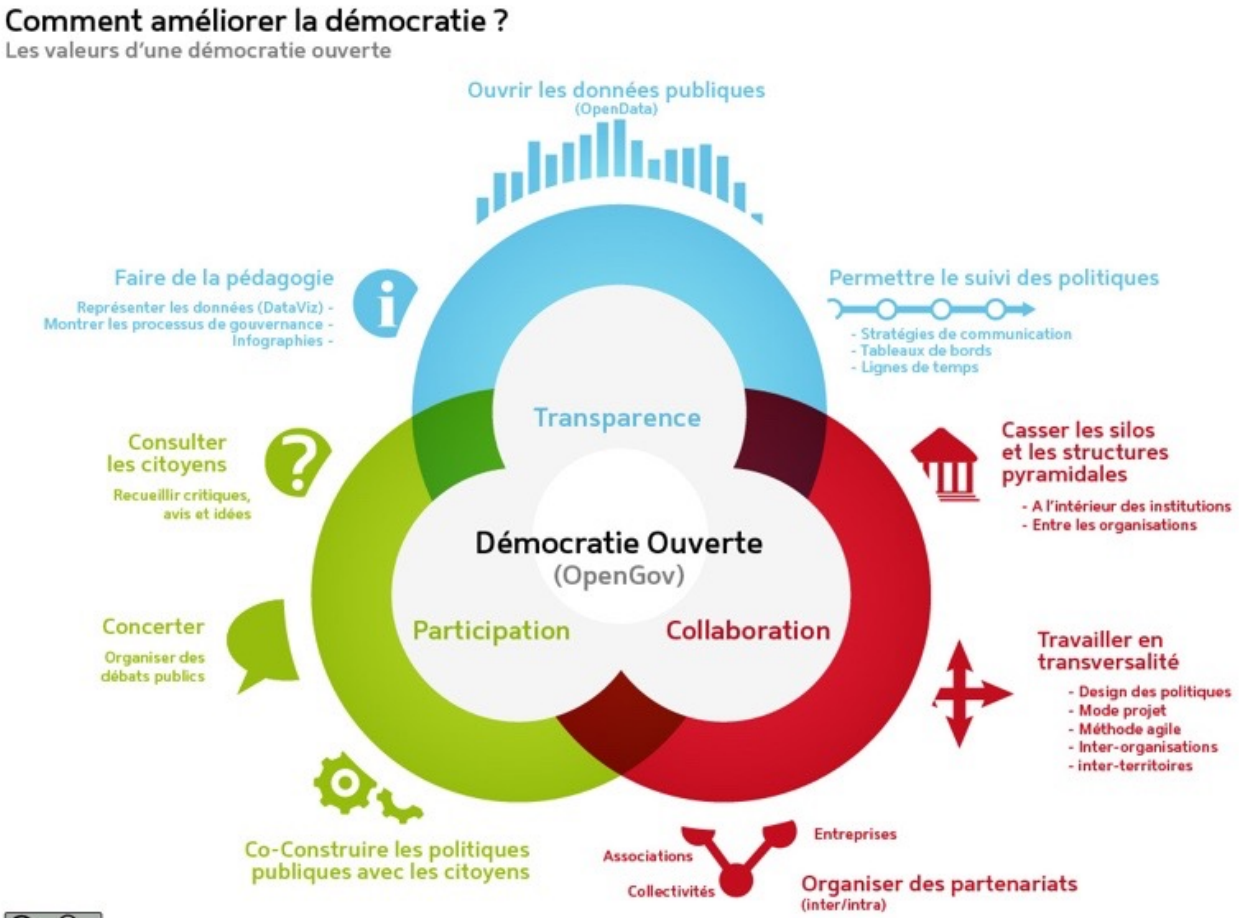
Wikipédia

Open knwoledge fundation

- L'**Open Knowledge Foundation** est une association à but non lucratif de droit britannique promouvant la culture libre, en particulier les contenus libres et l'*open data* (données ouvertes). Elle a été créée le 24 mai 2004 à Cambridge.

- 3 groupes de travail :
  - *Open research* groupe de travail dédié au domaine de la recherche ouverte et mettant l'accent sur le degré d'ouverture des données et sur comment Internet et les technologies numériques peuvent soutenir de nouvelles formes de collaboration et de partage dans la recherche scientifique.
  - *Open culture* groupe de travail voués à l'étude des moyens par lesquels des logiciels open source et les contenus culturels ouverts peuvent accroître l'accès à notre patrimoine culturel et forger de nouveaux processus de création et de collaboration.
  - *Open government* il s'agit de groupes de travail qui étudient le domaine des données publiques ouvertes, et qui sont tous engagés à construire des communautés qui rendront les **gouvernements plus transparents et responsables**.

Une démarche openGov



Une démarche openGov

https://www.democratieouverte.org

- **donner accès** à une meilleure connaissance du territoire : espace public et naturel, patrimoine, équipements, voirie, cartographie, information socio-démographique, etc.
- **renforcer la transparence** de l'action publique : budgets, dépenses, subventions, coûts des prestation, sécurité publique, etc.
- **développer les services** et de soutenir l'économie locale, améliorer la qualité des services rendus aux publics, participer aux actions en faveur du développement durable et par une dynamique territoriale, renforcer l'attractivité du territoire.
- **Mutualisation, amélioration** et efficacité des échanges entre services

Exemple : le projet Kawaa

L'évolution des lois sur les données

(focus sur les données publiques)

- **Loi du 17 juillet 1978**
  (CADA : Commission d'accès aux documents administratifs) :
  - Toute personne a le droit d'obtenir des documents détenus par une administration (indépendamment de leur forme ou de leur support).
  - Obligation pour toutes les administrations publiques ainsi que tous les organismes privés chargés d'une mission de service public.
  - "Les informations … peuvent être utilisées … à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été produits ou reçus".

- **Directive européenne de 2003**
  (transposée dans le droit français en 2005)
  - Le droit d'accès devient une obligation de publication.
- **Décret de 2011** : Principe de la gratuité du droit à réutilisation des documents et données publiques.

L'évolution des lois sur les données

- **La Loi pour une République numérique (2016)**
  - Cette loi vise à favoriser l'ouverture et la circulation des données et du savoir, à garantir un environnement numérique ouvert et respectueux de la vie privée des internautes et à faciliter l'accès des citoyens au numérique.
  - Tout le monde peut désormais lire, réutiliser librement et gratuitement les **données de l'administration. Elle consacre la conversion de l'administration française au mouvement de l'open data.** Plus grande transparence dans la gestion du budget et des finances publiques, services publics optimisés grâce aux données ouvertes, villes intelligentes, meilleur fonctionnement du marché du travail, intelligence artificielle éthique…

  Les possibilités que laisse entrevoir l'ouverture des données publiques à travers le monde sont nombreuses et promettent de transformer la société dans les décennies à venir.

- La loi sur la transition énergétique prévoit à partir de 2016 de progressivement rendre les données énergétiques (production, consommation par immeuble, quartier, par ville…) disponibles en ligne pour une libre réutilisation par toute personne intéressée (open data). Les gestionnaires de réseaux (électricité, gaz, réseau de chaleur et de froid) et les fournisseurs de produits pétroliers doivent fournir certaines données au service statistique du ministère de l'énergie.

Quelles données ?

Toutes les données structurées et produites par des acteurs publics ou privés dans le cadre d'une mission de service public. Cette définition s'étend aux données décrivant l'espace, les services publics et l'exercice politique, produites par des acteurs non institutionnels : associations, citoyens engagés, acteurs économiques ou académiques.

Sont exclus de l'opendata

• Les données à caractère personnel

« Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée directement ou indirectement par référence à un ou plusieurs éléments qui lui sont propres » (la loi 78-17 du 6 janvier 1978 art 2)

Ex : nom, adresse, photographie, num. tél., num. d'identifiant, adresse IP, adresse électronique…

• les documents administratifs dont la consultation ou la communication porterait atteinte :

A-6a) Au secret des délibérations du Gouvernement et des autorités responsables relevant du pouvoir exécutif ;

A-6d) A la sûreté de l'Etat, à la sécurité publique ou à la sécurité des personnes;

A-10c) Ou sur lesquels des tiers détiennent des droits de propriété intellectuelle.

Données publiques versus Big Data

Données publiques : services publics, intérêt général, ouvertes

- Effectifs: d'agents municipaux, d'élèves
- Horaires : bus et tram en temps réel,
- Patrimoine : arbres, équipements sportifs, points lumineux, bornes fontaines ...
- Éléments liés à l'exploitation : livres empruntés dans les bibliothèques, budget et compte administratif, marchés publics
- Description du territoire : photo aérienne, altimétrie, description des rues,
- Reflet de la vie locale : mariage, naissance, prénoms, élections, agendas...

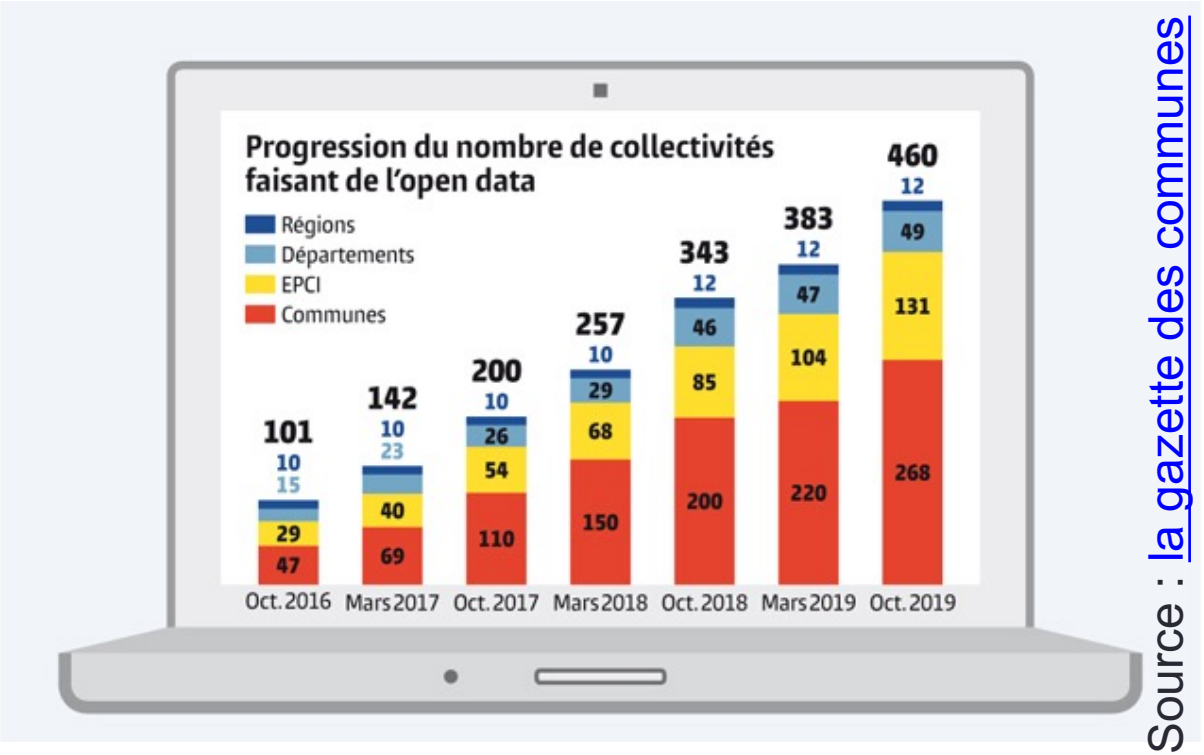
Big data: données personnelles, comportement individuel, données privées

- Traces (téléphone mobile, connexion)
- Navigation internet, réseaux sociaux, messagerie,
- Achats (CB), déplacements individuels ou collectifs (badge, carte)
- Consommations individuelles (eau, électricité, ..)
- Objets connectés

Gisement de données

- **Collectivités**
  - Accompagnement pédagogique : DataLab, InfoLab
  - OpenDataSoft

Youtube : Collectivités et open data : comment, pourquoi, pour qui ?



Source : la gazette des communes

- **Etat**
  - EtatLab : Missions
  - data.gouv.fr
    - 39 316 jeux de données ouvertes
    - 2792 applications ré-utilisatrices

Gisement de données

- **Citoyen**
  - Regards Citoyens,
  - DataPublica,
  - Data4Citizen
- **Instituts**
  - INSEE, INPI, LEGIFRANCE
  - IDG (infrastructure de données géographiques)
- **Agences de l'air, de l'eau**
  - Laboratoire Central de Surveillance de la Qualité de l'Air, Carte mondiale, IQAir,
  - Données publiques sur l'eau en France
- **Mondial**
  - Global open data index (*Open Knowledge Foundation*)
  - La Banque Mondiale

Licences

Licences Opendata stabilisées en France autour de :

- **ODbL : Open Data Base Licence**, édité par *OpenKnowledge Foundation*

- **Licence Ouverte** : édité par Etalab

- Elles autorisent la réutilisation y compris à des fins commerciales.
- Elles imposent des conditions de Mention et d'Intégrité
- Elles n'imposent pas d'obligation pour le producteur
- ODbL est plus restrictive car elle impose la notion de « Partage à l'Identique »

Licences ODbL

Vous êtes libres :

De partager : copier, distribuer et utiliser la base de données.

De créer : produire des créations à partir de cette base de données.

D'adapter : modifier, transformer et construire à partir de cette base de données.

Aussi longtemps que :

Vous mentionnez la paternité : Vous devez mentionnez la source de la base de données pour toute utilisation publique de la base de données, ou pour toute création produite à partir de la base de données, de la manière indiquée dans l'ODbL. Pour toute utilisation ou redistribution de la base de données, ou création produite à partir de cette base de données, vous devez clairement mentionner aux tiers la licence de la base de données et garder intacte toute mention légale sur la base de données originaire.

Vous partagez aux conditions identiques : si vous utilisez publiquement une version adaptée de cette base de données, ou que vous produisiez une création à partir d'une base de données adaptée, vous devez aussi offrir cette base de données adaptée selon les termes de la licence ODbL.

Gardez ouvert : si vous redistribuez la base de données, ou une version modifiée de celle-ci, alors vous ne pouvez utiliser de mesure technique restreignant la création que si vous distribuez aussi une version sans ces restrictions.

Formats

- Pour être Opendata, les données doivent être :
  - au format électronique,
  - accessible sur internet,
  - structurées,
  - à un format ouvert (et de préférence non propriétaire).

- **Types de format des fichiers disponibles**
  - Données alpha-num. : csv, ods, xls, doc, xml, JSON…
    - Un document au format pdf, même full text –et pas image- est lisible par une machine mais n'est ouvrable que par Acrobat/Adobe, donc non Opendata
  - Données géographiques : Shape, GéoJSON, KML, KMZ,
    - Format propriétaires non ouverts : DGW et DXF (Autocad)
  - Format spécifiques :
    - Domaine Transport : GTFS, Chouette, NetEX, Neptune…
    - GPX (ouvert), …

Outils d'exploitation

Données Alphanumériques

– Suites bureautiques habituelles : OpenOffice, Libre Office,
  - Et les outils commerciaux habituels (Google, MS, …)

– De nombreux outils en ligne : ConversionTools Services (xml<>csv…)

– Les langages de programmation habituels : C, Python…

– *OpenRefine* (previously Google Refine) pour l'exploration et le retraitement des données

Données géographiques

– QGIS : vraie suite logicielle de géomatique (SIG), convivial

– UMAP (openstreetmap), GeoServer : manipulation des cartes

– Non OpenSource : ARCGis, Autocad, MapInfo, GéoMap, GooglePro&Co

Exemples de réutilisations

- Handicap : Handimap, wheelmap
- Train temps réel : raildar, sncf geolocalisation
- Parking temps réel : Nantes, Rennes
- Transport en commun multi-modal : Toulouse, Lyon
- Développement logicels libres : makina corpus

Pire place de parking de NY !

Les agences municipales ont accès à une richesse de données et de statistiques décrivant chaque aspect de la vie urbaine. Mais comme l'analyste Ben Wellington le suggère dans cette conférence amusante, elles ne savent parfois pas quoi faire avec. Il montre comment une combinaison de questions inattendues et un traitement intelligent des données peut produire des idées étrangement utiles, et partage des conseils sur comment publier ces énormes jeux de données pour que n'importe qui puisse les utiliser.

Des sources d'info sur l'open data

Documents

- http://opendatahandbook.org/guide/fr/
  (*Open Knowledge Foundation*)

Youtube

- L'Open Data à la Loupe. Yann Bresson
- Quel est l'impact de l'Open Data ? Charles RUELLE, 2014
- Open Data explained in a nutshell, Simpleshow foundation, 2016
- Comment nous avons trouvé la pire place de parking de NY ?, Ben Wellington, TEDxNewYork
- L'Open Data, Avenir des Big Data, Jean Marc LAZAR, TEDxUTCompiègne