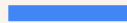


Les aspects concrets de l'ouverture des données



Saint Mandé
20 juillet 2017

Samuel Goëta
samuel@dataactivi.st

Objectifs de l'après midi

- Connaitre quelques “grands principes” de l’open data et leur rôle dans le processus concret de l’ouverture des données
- Comprendre comment concrètement on ouvre des données de leur localisation à leur publication
- Comprendre les obstacles et les frictions qui préviennent l’ouverture de données
- Comprendre les principales étapes du traitement et de la publication des données

1.

L'identification des données

De l'exploration des services à l'inventaire

Les données comme nouveau pétrole



Data in the 21st Century is like oil in the 18th Century: an immensely, untapped valuable asset. Like oil, for those who see Data's fundamental value and learn to extract and use it there will be huge rewards.

We're in a digital economy where data is more valuable than ever. It's the key to the smooth functionality of everything from the government to local companies. Without it, progress would halt.

WIRED

Une métaphore qui a ses limites...



- Avant de construire les pipelines, il faut savoir où forer...
- Or, on ne sait pas où sont les données
- On ne sait généralement pas non plus ce que sont les données
- Il faut donc les instaurer.

L'identification : un processus variablement équipé


Exploration

Injonction

Fouille des
bases de
données

Inventaire

Concertation
avec les
usagers

- 
- + équipé
 - + travail de l'organisation

L'exploration : une quête progressive et incertaine

Quelques stratégies :

- Parcourir les bureaux et l'organigramme à la recherche de données
- Parcourir les bases de données
- Regarder les données ouvertes par d'autres institutions
- Trouver les données brutes derrière les fichiers PDF déjà publiés



L'inventaire : un processus au long court

- L'utopie de l'inventaire exhaustif
- Une démarche progressive et exploratoire
- Question : comment qualifier les données ?
- A lier aux processus de plan d'occupation des sols des DSI

A	B	
Base/ Jeux de données	Description	
Détail des subventions de soutien à l'action culturelle française	D'où viennent les financements de la culture, comment sont-ils répartis au sein des directions et par projet, institution et manifestation	Minis
Effort financier de l'Etat dans le domaine de la culture et de la communication	Jaune associé au PLF	Minis
Base France Terme	Permet d'avoir la définition des termes des différents domaines scientifiques et techniques, recommandés au Journal officiel de la République française.	Minis
Base JOCONDE	base contenant près de 400 000 notices d'œuvres des musées nationaux	Minis
Base MERIMEE	permet d'identifier les immeubles protégés au titre des Monuments Historiques à travers un moteur de recherche multi critère	Minis
Adresses des bibliothèques municipales		Minis
Adresse des musées de France		Minis
Fréquentation des musées par région		Minis
Base histoire des arts	Notices et textes ce qui représente environ 4000 notices d'œuvres qui sont accompagnées des référentiels correspondants afin d'effectuer une indexation sémantiques des données.	Minis
Liste des établissements d'enseignement supérieur artistiques et culturels		Minis
Adresse des établissements publics culturels (BNF, école du Louvre, musée rodin, musée du Louvre, Opéra National de Paris, etc.)		Minis
Base EVE	Extrait de la base de données qui recense les événements culturels sur le territoire, organisée autour d'un référentiel des organismes culturels, événements et lieux d'accueil de ces événements. Certains événements (20%) sont géolocalisés.	Minis
Nuit des musées	Nom du lieu, Adresse, Pays, Région, Géolocalisation, Accès, Accès handicapés, Classé au titre des Monuments Historiques, Date de classement, etc	Minis
Rendez-vous aux jardins	Nom du lieu, Adresse, Pays, Région, Géolocalisation, Accès, Accès handicapés, Classé au titre des Monuments Historiques, Date de classement, etc.	Minis
Journée européenne du patrimoine	Nom du lieu, Adresse, Pays, Région, Géolocalisation, Accès, Accès handicapés, Classé au titre des Monuments	

De l'injonction à l'obligation d'ouverture

- Identification souvent nourrie par les injonctions politiques
- Loi dite “Lemaire”, obligation d'ouverture dans un standard ouvert :
 - Les « bases de données »
 - Les données « dont la publication présente un intérêt économique, social, sanitaire ou environnemental ».
 - Les documents communiqués suite à des demandes CADA.
 - Les documents qui figurent dans le répertoire des principaux documents administratifs

Données existantes au niv. Local, non disponibles au niveau national	
PV et Délibération (anonymisés)	Transparence
Subventions aux associations	Transparence
Services Urbains (Transport, Déchets, Eau)	Service
Equipements municipaux (bâtiments, espace public)	Service
Services municipaux (Cantine, Activité périscolaire, ...)	Consultation
Agenda municipal (culturel, manifestations, marché, ...)	Service
Etat Civil / Prénoms	Consultation
Etablissements Recevant du Public (ERP)	Service
Coûts des services...	Transp./Consultation
Autres...	

Passer d'une politique de l'offre à une politique de la demande

Comment faire ?

- Rechercher les applications déjà utilisées dans d'autres territoires
- Consulter les demandes de données sur cada.data.gouv.fr
- Constituer une communauté d'usagers potentiels et répondre à leurs demandes de données
- Ouvrir des données dans le cadre de concertations publiques et d'évènements

Données	Où la trouver	Pourquoi faire		Qui fait déjà ça
Enregistrements audio de tous les audio-guides de la ville	OT	Appli complète à tous les sites décrit par audio guide		Zévisit.com
Relatives au climat	Voir mail Plan climat			
les documents budgétaires de la ville de Montpellier (Budget primitif et supplémentaire, compte administratif...)				Rennes
(Animaux présents, localisation, plan du zoo...)	Zoo	mission pour téléphone mobile afin de se localiser pendant une visite dans le zoo		
Parking	Ville de Mip (Police) et aggio	Connaitre les lieux les plus occupés (à éviter) et les plus libres (pour trouver rapidement sa place)		

Passer d'une politique de l'offre à une politique de la demande



DODOdata

la plateforme qui réveille les données qui dorment

par datactivi.st

La philosophie du projet

La démarche

Faire une DODO

Aidez-nous !

RÉVEILLONS LES DONNÉES QUI DORMENT



✨ [Faites une DODO dès maintenant](#) ✨

La philosophie du projet



L'open data peut changer nos vies

Il faudra leur dire

2.

Les frictions de l'ouverture

**Principaux obstacles
organisationnels et
techniques**

La circulation des données provoque des frictions

Every movement of data across an interface comes at some cost in time, energy, and human attention. Every interface between groups and organizations, as well as between machines, represents a point of resistance where data can be garbled, misinterpreted, or lost. In social systems, data friction consumes energy and produces turbulence and heat – that is, conflicts, disagreements, and inexact, unruly processes.

Edwards et al. 2011



Rufus Pollock: Give Us the Data Raw, and Give it to Us Now

OPEN KNOWLEDGE INTERNATIONAL BLOG

[open knowledge home](#) [blog home](#) [latest posts](#) [donate](#) [about us](#)

[f](#) [t](#)

OPEN KNOWLEDGE INTERNATIONAL

Give Us the Data Raw, and Give it to Us Now

[Home](#) / [Ideas and musings](#) / Give Us the Data Raw, and Give it to Us Now


November 7, 2007, by [Rufus Pollock](#)


One thing I find remarkable about many data projects is how much effort goes into developing a shiny front-end for the material. Now I'm not knocking shiny front-ends, they're important for providing a way for many users to get at the material (and very useful for demonstrating to funders where all the money went). But shiny front ends (SFEs from now on) do have various drawbacks:

- They often take over completely and start acting as a restriction on the way you can get data out of the system. (A classic example of this is the Millenium Development Goals website which has lots of shiny ajax which actually make it really hard to grab all of the data out of the system — please, please just give me a plain old csv file and a plain old url).

About Rufus Pollock

Rufus Pollock is Founder and President of Open Knowledge.





This work is licensed under a [Creative Commons Attribution](#) license.

2007: Open Gov Data Principles

1. Complete

All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.

While non-electronic information resources, such as physical artifacts, are not subject to the Open Government Data principles, it is always encouraged that such resources be made available electronically to the extent feasible.

2. Primary

Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

If an entity chooses to transform data by aggregation or transcoding for use on an Internet site built for end users, it still has an obligation to make the full-resolution information available in bulk for others to build their own sites with and to preserve the data for posterity.

3. Timely

Data is made available as quickly as necessary to preserve the value of the data.

Hans Rosling: DbHd



Tim Berners-Lee : Raw Data Now



Quelques “bonnes raisons organisationnelles” de ne pas ouvrir des données

- Des données encastrées dans les systèmes d'information : explorer les bases de données, reconstruire les schémas et extraire les données
- Des données qui peuvent servir à des usagers malveillants : prévoir les risques et les dangers de l'ouverture
- Des données qui n'ont pas été pensées pour l'ouverture : améliorer leur qualité et leur intelligibilité
- Des données trop “sensibles” pour être ouvertes : la transparence, un mandat à obtenir

3.

Le traitement des données

**Du formatage à la
publication**

Formats ouverts et lisibles par les machines : une revendication essentielle

- Un **standard ouvert** est défini comme « tout protocole de communication, d'interconnexion ou d'échange et tout format de données interopérable et dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en œuvre ».
- « **Les formats lisibles par machine** permettent la réutilisation des données automatiquement et facilement (par des ordinateurs, sans qu'il faille les retranscrire "manuellement"). Par exemple, tous les fichiers au format Microsoft Excel ne sont pas lisibles par machine alors que les fichiers au format CSV (Comma-Separated Value) le sont car il s'agit d'un format informatique ouvert – compatible avec tous les logiciels de traitement de données – représentant des données sous forme de tableau structuré.»

2007: Open Gov Data Principles

5. Machine processable

Data is reasonably structured to allow automated processing.

The ability for data to be widely used requires that the data be properly encoded. Free-form text is not a substitute for tabular and normalized records. Images of text are not a substitute for the text itself. Sufficient documentation on the data format and meanings of normalized data items must be available to users of the data.








7. Non-proprietary

Data is available in a format over which no entity has exclusive control.

Proprietary formats add unnecessary restrictions over who can use the data, how it can be used and shared, and whether the data will be usable in the future. While some proprietary formats are nearly ubiquitous, it is nevertheless not acceptable to use only proprietary formats. Likewise, the relevant non-proprietary formats may not reach a wide audience. In these cases, it may be necessary to make the data available in multiple formats.

Open Data Index: machine readable

Company Register in Brazil

-  It's not openly licensed
-  It's not machine readable
-  It's not free
-  It's not available in bulk
-  It's up to date
-  It's not online
-  It's not digital
-  It's not public
-  Data does not exist

Access to company data in Brazil - on a federal level - is very limited. In order to access ...

[Read more »](#)

Tim Berners-Lee : modèle en 5 étoiles



Passage en CSV : bien plus que “enregistrer sous”

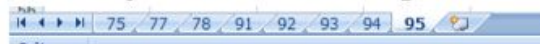


Bonnes pratiques sur Excel : structure

- Une feuille = un jeu de données
→ Un tableau par feuille

	A	B	C	D	E	F	G	H
1	ID	Année	Budget		Directions	1er semestre 2012	2e trimestre 2012	
2	CP-13	Année 2013	295 562 €		Service A	25 368 €	16 357 €	
3	CP-12	Année 2012	183 687 €		Service B		19 963 €	
4	CP-11	Année 2011	255 665 €		Service C	14 555 €	8 350 €	
5	CP-10	Année 2010	199 355 €					
6	CP-9	Année 2009	222 887 €		Directions	1er semestre 2011	2e trimestre 2011	
7	CP-8	Année 2008	231 300 €		Service A	25 368 €	16 357 €	
8					Service B	25 368 €	19 963 €	
9					Service C	14 555 €	8 350 €	
10								
11								
12								

- 1 onglet = un jeu de données
→ Ou 1 jeu = fusion des onglets



→ Exemples

- Recensement des équipements sportifs = 1 fichier redécoupé en 8 jeux de données (1 par département)
- Domaines d'intérêt majeur (DIM) : équipements mi-lourds financés en 2012 = 1 jeu de données reprenant l'ensemble des onglets

Passage en CSV : bien plus que “enregistrer sous”



Bonnes pratiques sur Excel : structure

- En-têtes sur la 1ère ligne (= titres de colonnes)

	A	B	C	D	E
1	Unité communication		Service innovation numérique		
2					
3		ID	Année	Budget	
4		CP-13	Année 2013	295 562 €	
5		CP-12	Année 2012	183 687 €	
6		CP-11	Année 2011	255 665 €	
7		CP-10	Année 2010	199 355 €	
8		CP-9	Année 2009	222 887 €	
9		CP-8	Année 2008	231 300 €	
10					
11					

- Pas de cellule vide dans les titres de colonnes

	A	B	C	D
1	ID	Année		Budget
2	CP-13	Année 2013	25 368 €	295 562 €
3	CP-12	Année 2012	35 987 €	183 687 €
4	CP-11	Année 2011	14 555 €	255 665 €
5	CP-10	Année 2010	16 357 €	199 355 €
6	CP-9	Année 2009	19 963 €	222 887 €
7	CP-8	Année 2008	8 350 €	231 300 €



Passage en CSV : bien plus que “enregistrer sous”



Bonnes pratiques sur Excel : structure

- Pas de cellule fusionnée (titres et contenu)

	A	B	C	D
1	Secteurs	Services	1er semestre 2012	
2	Secteur exemple 1	Service A	25 368 €	16 357 €
3		Service B	35 987 €	19 963 €
4		Service C	14 555 €	8 350 €
5		Service D	14 490 €	6 883 €
6	Secteur exemple 2	Service E	9 084 €	2 880 €
7		Service F	3 677 €	1 124 €
8		Service G	21 729 €	
9		Service H	7 136 €	9 131 €



- Attention aux lignes masquées !
→ elles s'affichent en CSV
- Éviter les lignes ou colonnes vides
- Attention aux données « orphelines » !

	A	B	C	D	E
1	Directions	1er semestre 2012	2e trimestre 2012		
2	Service A	25 368 €	16 357 €		26,03%
3	Service B	35 987 €	19 963 €		33,70%
4	Service C	14 555 €	8 350 €		
5					
6	NB. Suppression le 3/03/12 du dispositif...				



Passage en CSV : bien plus que “enregistrer sous”



Bonnes pratiques sur Excel : présentation

- Pas d'information transmise par la couleur

7753301	MUSEE DEPARTEMENTAL STEPHANE MALLARME	YVAINES-SUR-SEINE
7817201	MUSEE DE LA BATTELERIE	CONFLANS-STE.
7832201	MUSEE DE LA TOILE DE JOUY	JOUY-EN-JOSAS
9406801	MUSEE DE SAINT-MAUR - VILLA MEDICIS	LA VARENNE-SAINT-
9407301	MUSEE EMILE JEAN	VILLIERS SUR MARNE
9409301	MUSEE D'ART CONTEMPORAIN DU VAL-DE-MARNE	VITRY-SUR-SEINE
9501801	MUSEE D'ART ET D'HISTOIRE	ARGENTEUIL
9520501	MUSEE NATIONAL DE LA RENAISSANCE	ECOUEN
9529501	MUSEE ARCHEOLOGIQUE DU VAL D'OISE	GURRY-EN-VEUXIN
9531301	MUSEE LOUIS SENLECY	L'ISLE-ADAM
9535101	MUSEE INTERCOMMUNAL D'HISTOIRE ET D'ARCHEOLOGIE	LOUVRES
9542801	MUSEE JEAN-JACQUES ROUSSEAU	MONTMORENCY
9550001	MUSEE VITIS DEL LAGNE	PARIS

<http://www.data.gouv.fr/DataSet/313823877?xml=franceparis.fr+france+musees+de+france&xml=2>

Galeries nationales du Grand Palais
Musées nationaux
Musées de la ville de Paris - appellation en 2004
Musées du Muséum National d'Histoire Naturelle
Appellation M de F en 2004
Appellation M de F en 2006
Appellation M de F en 2007
Appellation M de F en 2009
Données confidentielles - Contacter le chef d'établissement
Fermeture du musée Picasso en septembre 2009 pour travaux pour
Chiffres d'entrées pour les expositions temporaires
Fermeture du musée le 10 janvier 2010 pour travaux. Réouverture p
Fermeture du musée de l'Homme (muséum) en mars 2009 pour trav
Collections permanentes gratuites depuis le 11 septembre 2007 à ce

→ Dans le format CSV, ces données sont supprimées !

7510103	MUSEE DE L'ORANGERIE DES T...	PARIS	Musée national	447 093	153 417	599 762	140 729
7510106	MUSEE DU LOUVRE	PARIS	Musée national	9 314 000	2 647 000	9 224 643	2 703 407
7510602	MUSEE EUGENE DELACROIX	PARIS	Musée national	34 044	9 890	38 425	9 916
7510305	MUSEE NATIONAL PICASSO	PARIS	Musée national	501 080	210 779	470 500	151 887
7510501	MUSEE NATIONAL DU MOYEN-AG.	PARIS	Musée national	289 958	124 048	293 075	129 570

http://data.iledefrance.fr/explore/dataset/frequeuntation_des_musees_francaiens_entre_2006_et_2010#?tab=table



L'édition des données : une étape inévitable

- *data editing* : les opérations par lesquelles les statisticiens traitent et transforment les données issues des sources administratives (Desrosières 2005)
- Pourquoi éditer des données ?
 - Anonymiser des données (loi CADA, un document administratif : non nominatif, non personnel)
 - Enlever la sensibilité de données qui ne pouvaient pas être publiées
 - Rendre intelligible les acronymes
 - Améliorer la lisibilité des données par les machines