

LE MOIS DE LA DATA

La donnée : de quoi parle-t-on ?

Intervenants : Marie-Anne Chabin, Jean-Marie Bourgogne

Animatrice : Magali Barnoin

#moisdeladata

Comment en est-on venu à échanger des données numériques ?

De la transmission orale des connaissances à la numérisation du savoir

Le dessin



-18 000 av J-C

L'écriture



- 3500 av J-C

L'imprimerie



1450

L'informatique



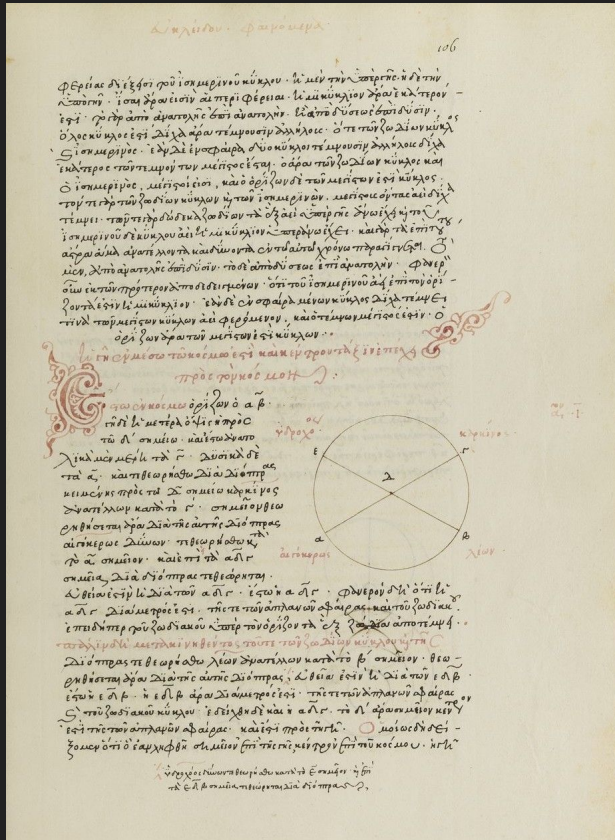
1950

L'internet (web)



De 1970 à aujourd'hui

Mais au fait, de quelle langue vient data ?



“L'un des premiers usages scientifiques connus du terme « data » (issu du verbe latin « dare » : donner) remonte au IIIe siècle avant J.-C. dans un traité de mathématiques d'Euclide où il était démontré des propriétés géométriques « étant donné » d'autres propriétés géométriques - autrement dit des axiomes. Son sens a ensuite évolué en deux branches.

Il a donné le mot « date », que l'on retrouve dans des documents médiévaux officiels pour signifier le « moment » où ceux-ci ont été produits

A compter du XVIIe siècle, le sens de « données » en référence à des « principes acceptés de tous », par l'intermédiaire du divin en théologie, puis par l'expérimentation en mathématiques.”

Retrouve-t-on dans cette définition des “data” le caractère de “gratuité” et d’immédiateté, ainsi que d’objectivité et de factuelité ?

Information, données, base de connaissance

Information (*Information*) : toute connaissance pouvant être échangée. Lors de l'échange, elle est représentée par des données.

Données (*Data*) : une représentation interprétable formalisée de l'information, adaptée à la communication, à l'interprétation ou au traitement.

Base de connaissance (*Knowledge Base*) : ensemble d'informations préalablement acquises par une personne ou un système lui permettant de comprendre l'information reçue.

Source: glossaire Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) ou norme ISO14721 (2003, 2012)

Exemple d'une recette de cuisine

Tarte aux pommes fine et caramélisée

 Tartes et Clafoutis  Aucun avis  Imprimer

Cette tarte aux pommes fine et croustillante, sera caramélisée en dessous grâce à un peu de sucre saupoudré sous la pâte. C'est un vrai régal.

Ingrédients

Pour 1 tarte de 10 parts environ

3 pommes acidulées, par exemple des Granny

1 pâte feuilletée ronde

1 peu de sucre

1 peu de beurre

1 pincée de cannelle (facultatif)

Matériel

1 couteau

1 économe ou 1 épluche-pomme

Du papier sulfurisé

Ingrédients

Pour 1 tarte de 10 parts environ

3 pommes acidulées, par exemple des Granny

1 pâte feuilletée ronde

1 peu de sucre

1 peu de beurre

1 pincée de cannelle (facultatif)

Préparation

1 Retire la peau des pommes à l'aide d'un économe. Coupe-les en deux et retire le trognon. Coupe-les en fines lamelles.

2 Allume le four à 180°C (thermostat 6) pour le faire préchauffer.

Préparation

1 Retire la peau des pommes à l'aide d'un économe. Coupe-les en deux et retire le trognon. Coupe-les en fines lamelles.

2 Allume le four à 180°C (thermostat 6) pour le faire préchauffer.

3 Pose une feuille de papier sulfurisé sur une plaque allant au four. Saupoudre la feuille avec quelques pincées de sucre.

Un document (contexte)

Des informations
(objectifs, méthodes et chiffres)

Des données
(précises, quantifiées)

Pommes	3
Pate feuilletée	1
Sucre (gramme)	10
Beurre (gramme)	20
Nb de personnes	10
Temp Four (°)	180
Durée cuisson (')	30

Autour du mot data

- **Les données personnelles (Self-data)**

Les informations personnelles (état civil) ou produites par notre activité personnelle (achats, web, ...)

- **L'open data**

Les données produites par les acteurs publics, non protégées par un secret et librement accessibles

- **Le big data**

Une masse complexe de données issues de multiples sources techniques (capteurs) ou administratives qui induit les notions de : vélocité, volume et variété (3V)

- **La smart data**

Les données nécessaires pour le pilotage technique des services urbains (ville intelligente/smart city)

Et puis aussi quelques disciplines :

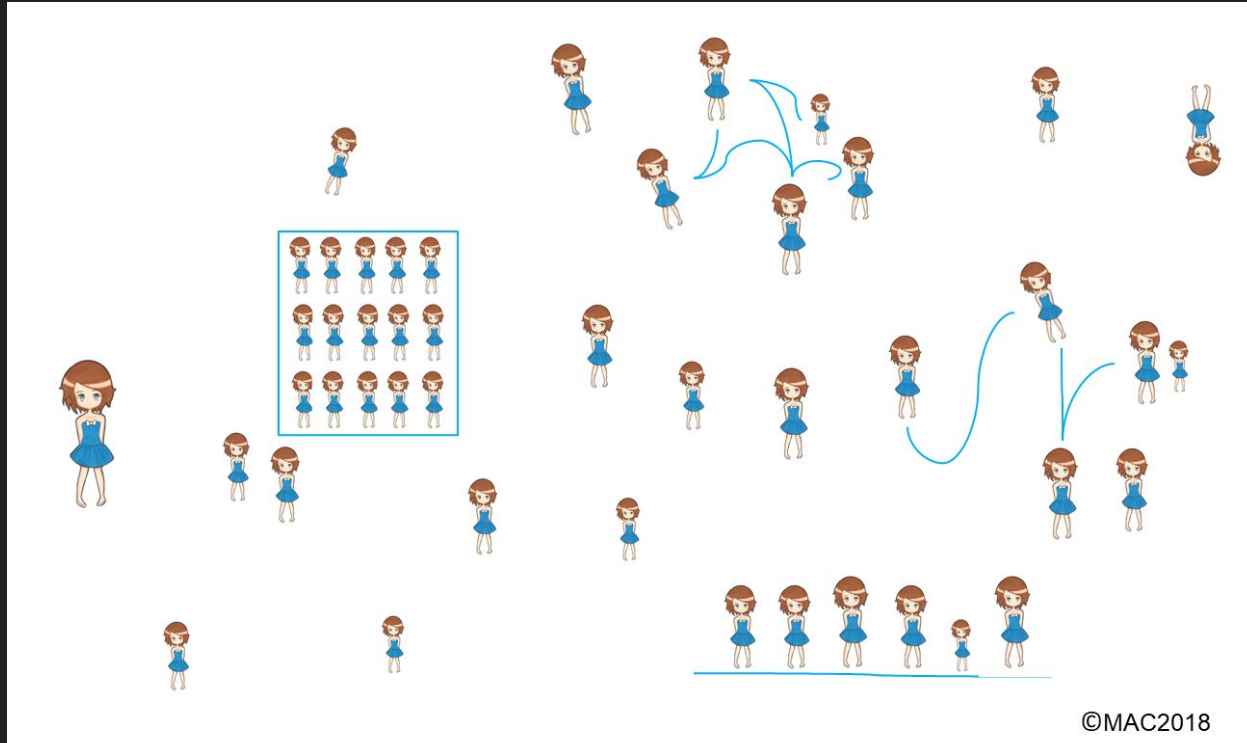
- **Data science et Data engineering**

compétences et métiers liés au traitement des données

- **DataViz**

représentation graphique des données (courbes, cartes, illustrations inspirées)

Les données, les documents, les liens interpersonnels



La donnée isolée ne signifie pas grand-chose. Les données d'aujourd'hui sont les documents d'hier

- une facture
- un flux entrant
- un dossier patient

Ce sont les interactions qui comptent, qui donnent du sens.

Une histoire française de la transparence administrative

Article 15 de la déclaration du 26 août 1789 des droits de l'homme et du citoyen:

« La société a le droit de demander compte à tout agent public de son administration »

De 1800 aux années 1960, chaque préfecture publiait un annuaire administratif, historique, statistique du département.

Ici, le Pas-de-Calais, 1810, (statistiques de décès)

Population. 1809				
Il est des états et relevés remis à la Préfecture qu'il est sent				
	En 1806.	En 1809.	En 1810.	Totals.
Am bassons d'un an... 2756	5108	2658	1551	9315
Du 1 à 5 ans... 1216	5114	5514	1474	12358
Du 6 à 10 ans... 586	911	524	1171	5992
Du 11 à 15 ans... 472	364	301	609	2746
Du 16 à 20 ans... 358	307	267	625	1557
Du 21 à 25 ans... 252	194	159	445	1090
Du 26 à 30 ans... 195	128	109	351	623
Du 31 à 35 ans... 130	103	84	261	378
Du 36 à 40 ans... 79	67	51	167	297
Du 41 à 45 ans... 50	43	32	105	125
Du 46 à 50 ans... 31	26	19	66	86
Du 51 à 55 ans... 19	16	11	46	56
Du 56 à 60 ans... 11	9	6	26	36
Du 61 à 65 ans... 6	5	3	14	18
Du 66 à 70 ans... 3	2	1	6	6
Du 71 à 75 ans... 1	1	0	2	2
Du 76 à 80 ans... 0	0	0	0	0
Du 81 à 85 ans... 0	0	0	0	0
Du 86 à 90 ans... 0	0	0	0	0
Du 91 à 95 ans... 0	0	0	0	0
Du 96 à 100 ans... 0	0	0	0	0
Total des décès... 1104	1294	1209	3607	6704

Le nombre des naissances dans l'année... 15611

Il est à remarquer que le nombre des naissances dans l'année 1809 n'est pas le même que celui de l'année 1808.

AN 012

Des deux morts dans le 1^{er} semestre... 1

Des deux premières années... 2

Des quatre premières années... 4

Des six premières années... 6

Des dix premières années... 10

Des vingt premières années... 20

Des dix morts avant d'avoir atteint l'âge de huit ans... 2

Cette observation n'est pas absolue. Il n'est pas certain que les ministères, à partir de ce moment, ne soient toujours en rapport continu avec les communes et les villes, par où il est prouvé que, dans ces trois années, le nombre d'enfants morts dans l'âge de 5 ans et au-dessous, est pas tout à fait égal au tiers du nombre des naissances qui ont eu lieu dans ces années.

Suivent les tableaux de la population en 1810

Janvier 1809 et du mouvement de la population pendant les années 1809 et 1810.

?

=



Open Data désigne des données auxquelles n'importe qui peut accéder, que tout le monde peut utiliser ou partager.

Les enjeux stratégiques du partage des données

TRANSPARENCE ET SOUVERAINETE

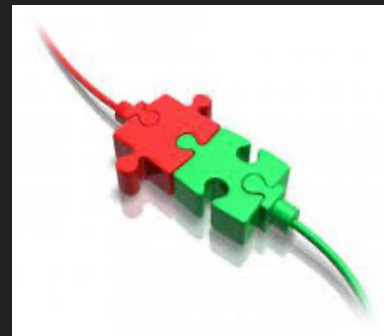
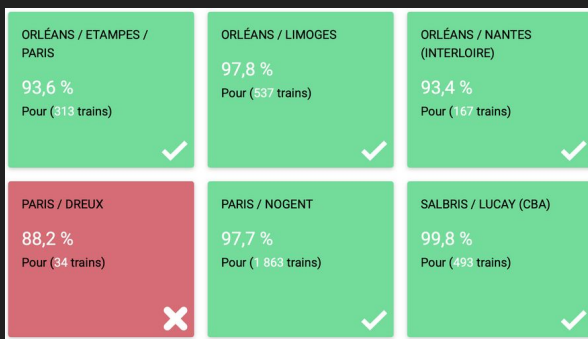
- Les citoyens ont accès à des données fiables et de qualité sur l'action de l'Etat et des organisations chargées d'une mission de service public
- Les citoyens sont informés des données produites et utilisées par l'administration dans le cadre des démarches administratives
- L'administration est en mesure de rendre compte des règles et des choix qui sous-tendent des décisions collectives et individuelles.
- Les acteurs publics ont la maîtrise des données essentielles à l'exercice du service public.



Les enjeux stratégiques du partage des données

PERFORMANCE DU SERVICE PUBLIC

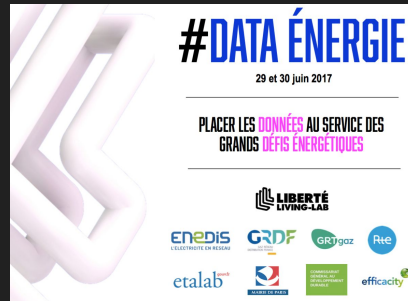
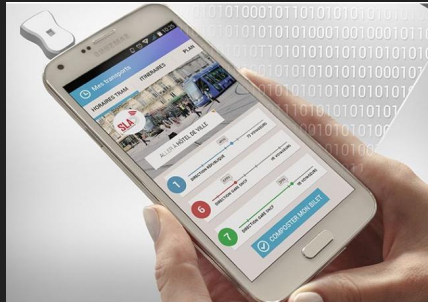
- Amélioration de l'expérience usager par un accès simple aux informations publiques : PLU, santé, FranceConnect
- Evaluation de la qualité du service public et satisfaction des usagers : enquêtes et tableau de bord
- Le décloisonnement et de le partage des données au sein et en entre administrations



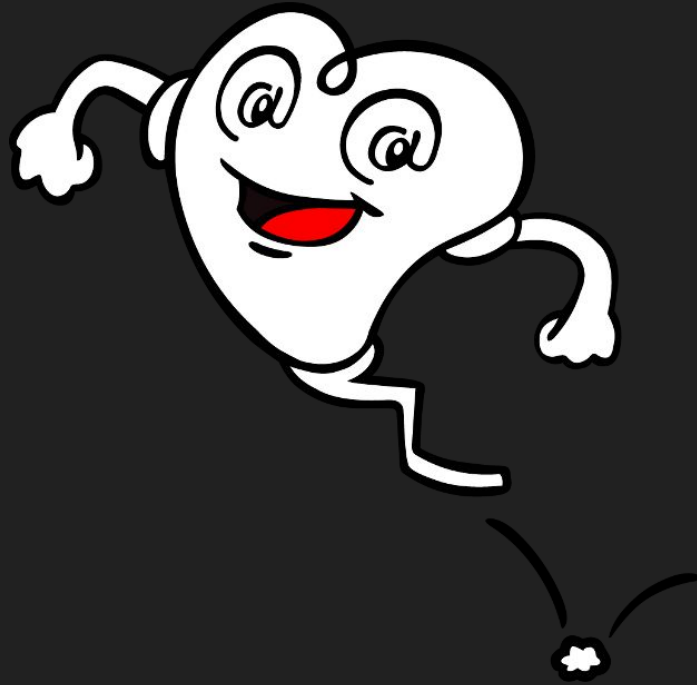
Les enjeux stratégiques du partage des données

INNOVATION

- Faciliter la création de nouveaux services en abaissant les barrières à l'innovation
- Possibilité de s'appuyer sur l'imagination et l'engagement de la société civile (ex COVID Tracker, OpenStreetMap)
- Promotion de biens communs informationnels (par exemple la météo, les données cartographiques (IGN), wikipédia)



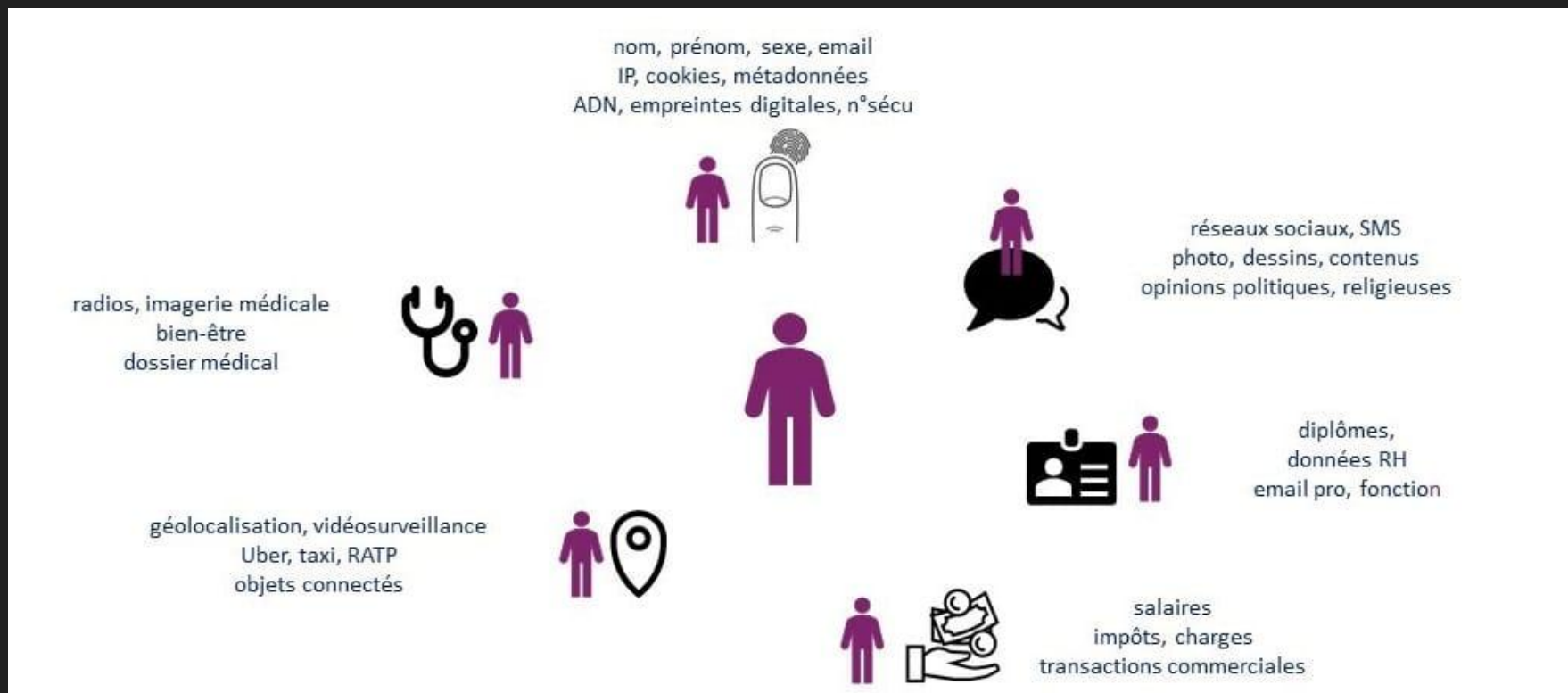
C'est l'heure du quiz....



Focus sur quelques sujets

- **Les différentes catégories de données**
- **De la nature des données...**
- **Les données de références**
- **La forme et les standards de données**
- **Les conditions juridiques de partage des données**

Qu'est-ce qu'une donnée personnelle ?



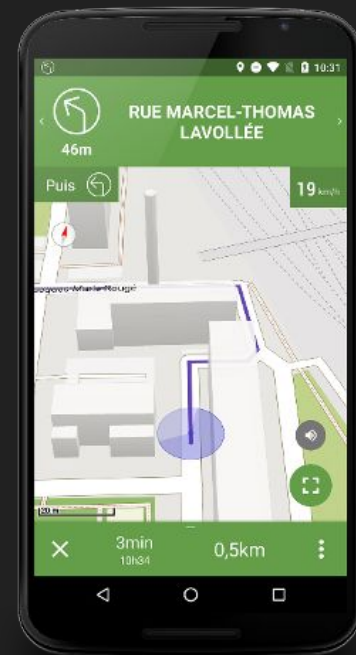
La protection des données personnelles est très strictement encadrée par la CNIL et le RGPD (règlement général de la protection des données)

À propos de l'opendata

« L'ouverture et le partage des données publiques, ou Open data, consistent à mettre à disposition de tous les citoyens, sur Internet, toutes les données publiques brutes qui ont vocation à être librement accessibles et gratuitement réutilisables. » - [Etalab, 2013]

- Donner accès à une meilleure connaissance du territoire
- Renforcer la transparence de l'action publique
- Développer les services, soutenir l'économie locale et favoriser l'innovation
- Améliorer les échanges entre/ au sein des administrations

➔ La loi le rend obligatoire pour les communes de + 3 500 hab



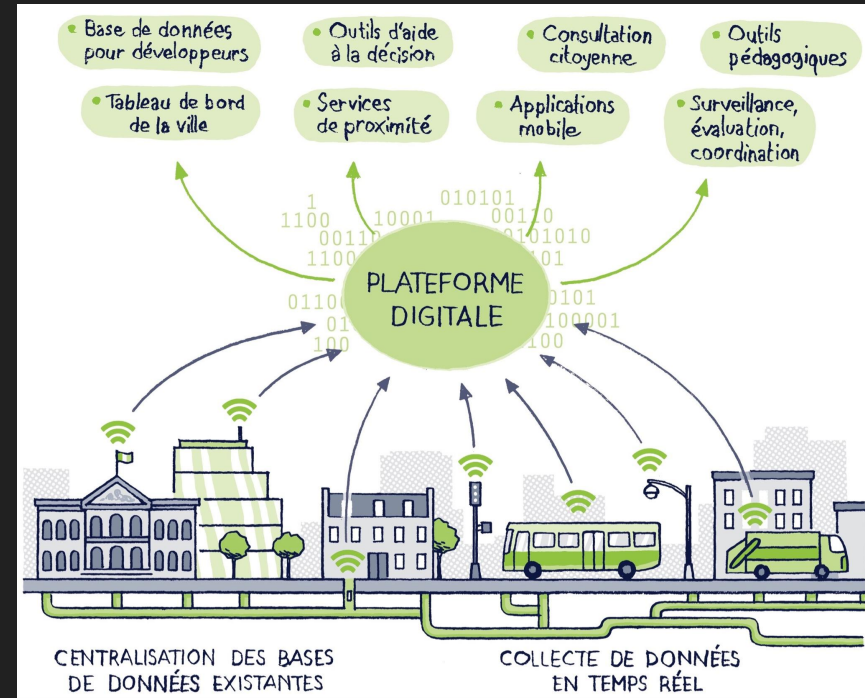
Smart data

La data, **liant d'un monde connecté**, pour une meilleure optimisation des services urbains

La Smart City, ou ville intelligente, est une ville dans laquelle de multiples capteurs, voire des données issues des téléphones des citoyens, remontent à un siège central des informations sur la ville : congestion routière, pollution de l'air, consommation énergétique, ...

On lui appose le préfixe *smart* car elle est censée s'adosser, pour diriger ces opérations, à la **synthèse de données** factuelles variées, présentes en **très grand nombre**.

Les données peuvent être produites et utilisées par les citoyens (en principe...)



Ex. de plateforme digitale de Saint Etienne

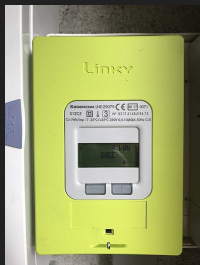
Big data

Le big data suit l'accroissement des capacités techniques :

- Flux (Mb/s)
- Stockage des données (181 mille milliards de Go)
- Puissance de calcul (MIPS)

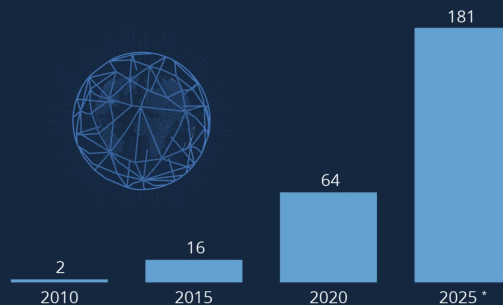
Internet, c'est 10% de la consommation électrique, près de 10% des émissions de gaz à effet de serre ! et beaucoup de matières premières...

Le stockage des données représente 5% de cette ébriété numérique.



Le Big Bang du Big Data

Estimation du volume de données numériques créées ou répliquées par an dans le monde, en zettaoctets



Un zettaoctet équivaut à mille milliards de gigaoctets.

* Prévission en date de mars 2021.

Sources : IDC, Seagate, Statista



Mais le big data est aussi lié à l'hyper-connexion des systèmes et à notre consentement...

► Que sont les GAFAM et les BATX ?

De la nature des données

Quels niveaux de détail pour quels usages ?

Une donnée brute contient des informations élémentaires, détaillées, trop ou pas assez...

Une donnée enrichie possède des informations ajoutées ou calculées

Une donnée agrégée est le résultat d'une opération : somme, statistique, nettoyage

Une donnée anonymisée ne contient plus d'information à caractère personnel

Donnée chaude, la donnée froide

Froide : qui change peu (le cadastre, le résultat d'une élection, la population)

Chaude : qui change souvent (la fréquentation d'un parking, la circulation, la météo)

Donnée fraîche

c'est une donnée récemment actualisée

▶ un mémo : "un plat chaud peut être réalisé avec des produits frais" (*donc froid est différent de frais*)

Donnée géographique

Donnée structurée contenant des informations spatiales (surface, ligne) ou géolocalisées (point)

Les données de référence

Données de référence et service public de la donnée

Le Code des Relation Public Administration définit les données de référence comme des informations publiques qui constituent une référence commune pour **nommer ou identifier** des produits, des services, des territoires ou des personnes.

Donnée souveraine

Aux termes de la lettre de mission du Premier ministre, la souveraineté d'une donnée qui sert de **support direct aux décisions de la puissance publique**.

Une donnée est donc souveraine si sa disponibilité conditionne la possibilité même de la décision publique, c'est-à-dire qu'elle présente pour la puissance publique une véritable criticité.

Donnée d'intérêt général

Les données d'intérêt général sont des données produites et détenues par des entreprises dont certaines modalités de réalisation impliquant un organisme public, ou la portée publique des informations, amène une mise à disposition dans le bien public.


La forme des données

Fichiers plats

Suite organisée de données, généralement un tableau

Jour	Entrée	Plat
Lundi	Tomates	Sauté de dinde
Mardi	Taboulé	Chausson bolognaise

```
<ville nom="Ambronay">
  <an2005>
    <population>2247</population>
    <inscrits>35</inscrits>
  </an2005>
  <an2008>
    <population>2351</population>
    <inscrits>571</inscrits>
  </an2008>
</ville>
```

Lundi	Mardi
Tomate mozzarella au pesto	Taboulé
Sauté de dinde au curry	Chausson bolognaise
Blé	Duo de haricots
Fromage blanc Bio 	Fromage
Fruit	Pêche au sirop

Base de données relationnelle

Ce sont des entrepôts de données structurés

Elles sont constituées de plusieurs tables liées entre-elles par des relations, du type "appartient à", "est composé de"

API ou interface de programmation

Accès à l'information en ligne (internet) et en temps-réel

Mode d'accès réservé aux "machines"



Les standards de données

Les **standards** de données (ou **schémas** de données) définissent la façon dont les données doivent être structurées.

Le standard se réfère aux **champs** de la donnée. Il en donne le nom, le type (numérique, alpha-numérique, date,...), la description et les contraintes (liste, min/max, codification ex. tél)

Voici des exemples de sites proposant des standards :

- [Schema.data.gouv.fr](https://schema.data.gouv.fr)
- [Standards du Conseil National de l'Information Géographique](#)
- ISO ou [Schema.org](https://schema.org) (international)

Champ	Type	Longueur	Description
Nom	alpha-numérique	255 (maximale)	Nom de la commune
Code INSEE	alpha-numérique	5	Code de la commune, au sens de l'INSEE
Département	alpha-numérique	255 (maximale)	Nom du département
Population	numérique	par exemple, 6	Nombre d'habitants de la commune

N° Siret : 9 chiffres (SIREN) + 5 chiffres (établissement)

Montant : en euro ou Keuro ?

Date : commençant par le jour ou l'année ?
(10-12-21 = 10 déc 2021 ou 21 décembre 2010 ?)

Géolocalisation : quelle précision ?

- 1 décimale : 10 000 m (en France métropolitaine)
- 2 décimales : 1 000 m
- 3 décimales : 100 m
- 4 décimales : 10 m
- 5 décimales : 1 m

48.84969, 2.38126 c'est mon adresse à 1 m près

► Quels avantages des producteurs de données ont-ils à convenir d'un standard de données commun ?

Les licences

Elles fixent les droits et devoirs des producteurs et des réutilisateurs.
C'est l'équivalent des "Conditions Générales d'Utilisation".

Il existe à ce jour deux licences homologuées par l'état :

La licence ouverte

La licence qui n'impose au réutilisateur que de mentionner la **source** et le **millésime** de l'information publique. L'utilisateur est libre de réutiliser les données, y compris à titre **commercial**. La licence Etalab est une licence ouverte.



La licence ODbL

La licence ODbL impose en plus à l'utilisateur de **repartager à l'identique** les données dérivées, notamment les enrichissements (biens communs)
(La licence ODbL est comparable à la licence CC-BY-SA : Creative Commons, Paternité (BY), Share Alike)



Nous vous remercions de votre attention

mac@marieannechabin.fr

jmbourgogne@opendatafrance.email @opendata_fr

Faites connaître le #MoisdelaData

Avec l'aimable autorisation du Cerema / Mathieu Rajérison pour son concours à la présentation.

Version : 1.0 date : 1 déc. 2021

Licence : CC-BY-SA

