

Section 3 : Données, données... quelles données ? Les différents types de données

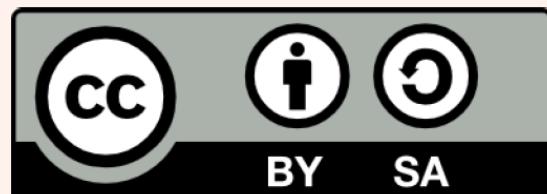
Culture générale des données

Dataactivist, 2018-2019

Ces slides en ligne : <http://dataactivist.coop/SPoSGL/sections/section3.html>

Sources : <https://github.com/dataactivist/SPoSGL/>

Les productions de Dataactivist sont librement réutilisables selon les termes de la licence [Creative Commons 4.0 BY-SA](#).



Plan du cours

1- Données qualitatives, quantitatives, structurées, échappées...

Ecoutez le podcast Dataactivist "La diversité des données - quels liens entre Etat et système statistique"

Recommandé : Lire l'article "GOOGLE MAPS'S MOAT - How far ahead of Apple Maps is Google Maps?"

2- Les données crowdsourcées

Annotez quelques contributions au grand débat sur la [GrandeAnnotation.fr](#)

3- Petit glossaire autour des données

Bibliographie

Quiz section 3

1. Données, données... quelles données ?

En guise d'introduction...

Tous les pays possèdent-ils les mêmes données publiques ?

Ecoutez le podcast "La diversité des données - quels liens entre Etat et système statistique" par Samuel Goëta, Dataactivist



Données quantitatives

Différents types de variables :

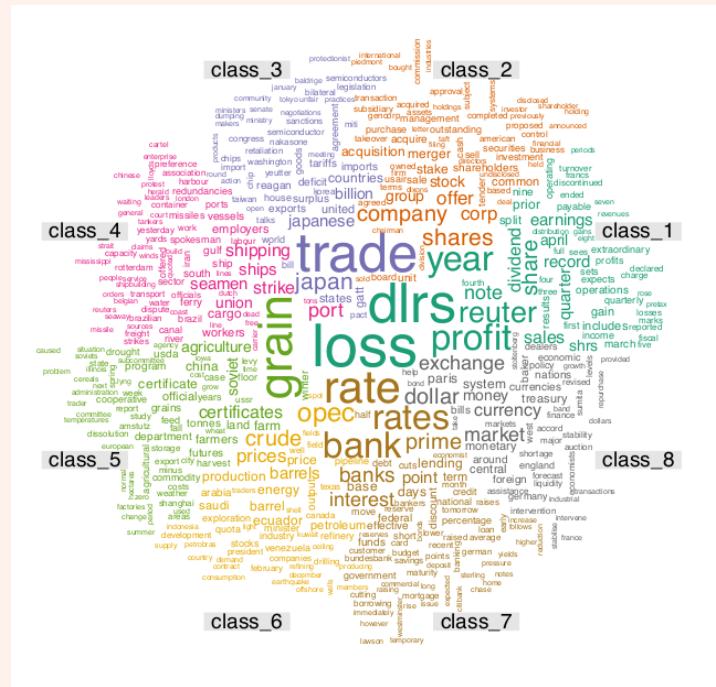
- **Nominale** : des catégories que l'on nomme avec un nom (marié/célibataire/divorcé/veuf)
- **Ordinal** : échelle de mesure dotant chaque élément d'une valeur qui permet leur classement par ordre de grandeur (faible, moyen, fort)
- **Intervalles** : l'intervalle entre deux catégories a toujours la même valeur (12-16°C / 16-20°C / 20-24°C)

6650	623	8960	-588	3759	3648	48
455	545	5511	552	5542	851	54
56	321	4598	564	648	321	381
90	9023	8734	780	283	472	6
2	322	1112	322	322	322	322
388	886	686	686	686	686	686
662	666	666	666	666	666	666
5665	5665	5665	5665	5665	5665	5665

Données qualitatives

Ce sont des données non numériques, par exemple du texte, des images, de la vidéo, du son...

- Ces données peuvent être converties en données quantitatives
- Mais on risque de perdre la richesse des données originales
- Il est possible de réaliser une analyse qualitative de ces données



Exemple : les annotations en text mining

L'annotation (ou l'étiquetage) est une tâche plus spécifiquement linguistique que les précédentes, au sens où elle ne s'applique pas, aux données tabulaires et ne relève donc pas de la fouille de données (data mining)

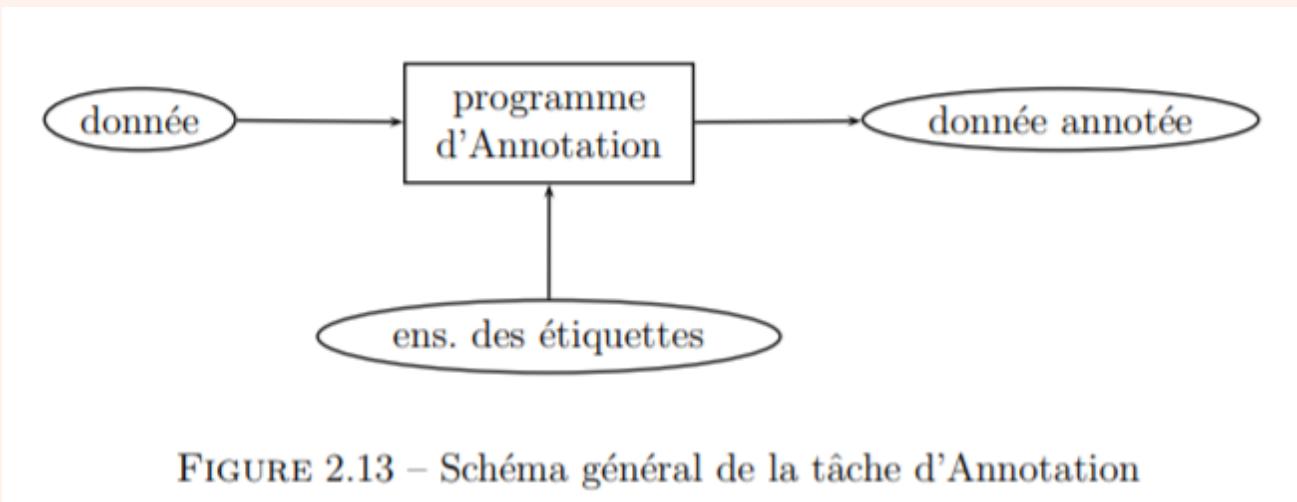


FIGURE 2.13 – Schéma général de la tâche d'Annotation

Source : [Introduction à la fouille de textes université de Paris 3 - Sorbonne Nouvelle](#)

Exemple : les annotations en text mining

La donnée est exclusivement un texte brut ou un document semi-structuré non transformé en tableau : elle est donc composée d'unités respectant au moins une relation d'ordre.

L'ensemble des étiquettes possibles est fini et connu à l'avance au moment où le programme est appelé. Le résultat est la donnée initiale dans laquelle chaque unité est associée à une étiquette prise dans l'ensemble des étiquettes possibles

L'annotation peut aussi s'appliquer à d'autres données structurées que les textes : **on peut ainsi annoter des séquences audio ou vidéo**, ou des bases de données XML par exemple. On parlera d'annotation quand la structure de la donnée d'origine se trouve "reproduite" sur les étiquettes ajoutées par le programme.

Source : [Introduction à la fouille de textes université de Paris 3 - Sorbonne Nouvelle](#)

Exemple : les annotations en text mining

30
Appelant-demandeur

32 75011 PARIS née le 09 Janvier 1990 à Bagnolet (93170)

34 Avocat
Représentés par Me Serge BEYNET de la SELEURL SERGE BEYNET, avocat au barreau

35 Avocat
de PARIS, toque : C0482

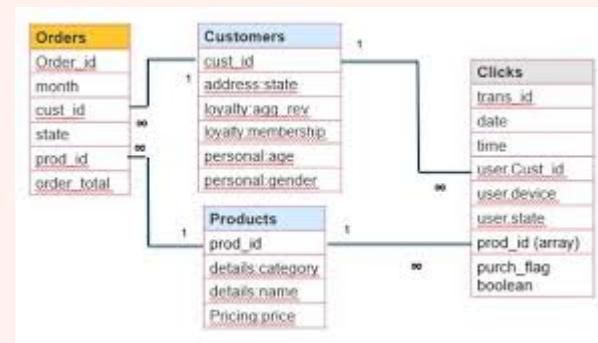
37 Intimé-défendeur
INTIMÉS

39 Intimé-défendeur
Monsieur DANIEL Y...

Données structurées

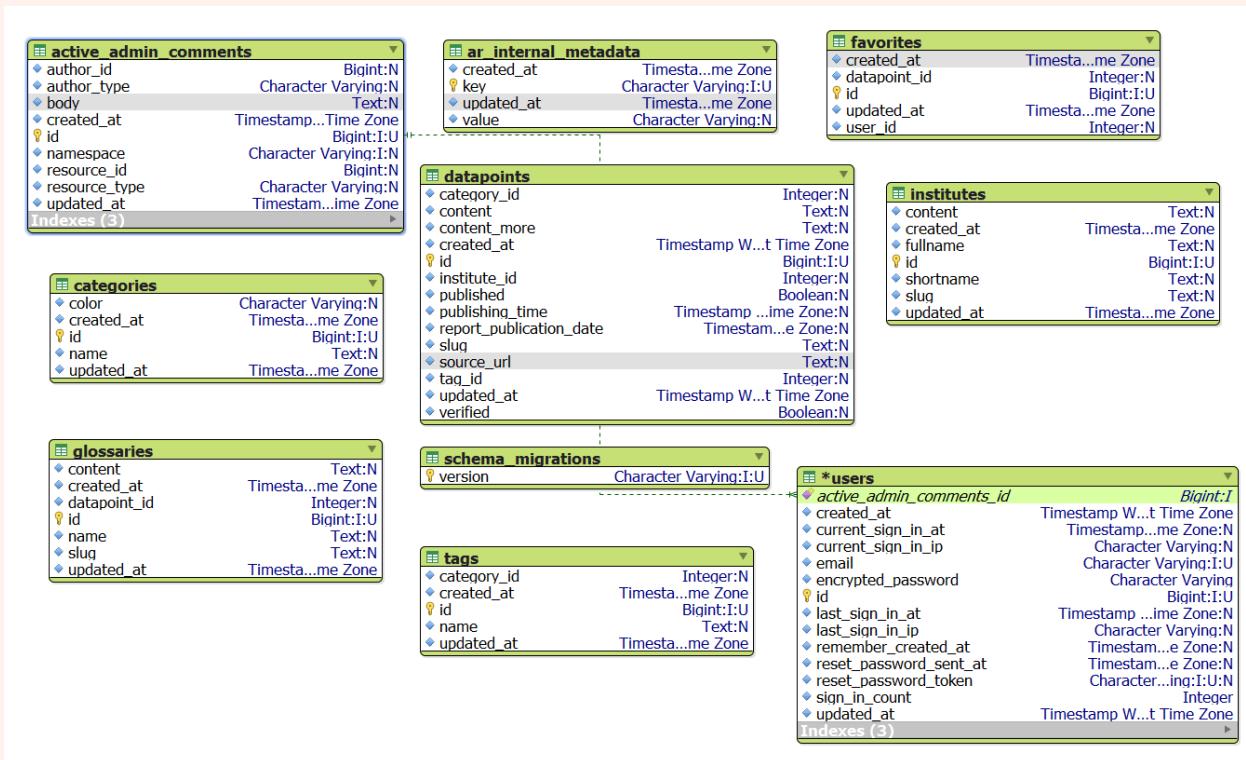
Des données dotées d'un modèle qui définit les relations entre les composantes de la base de données

- Ex : base de données relationnelle SQL
- Lisibles machine
- Faciles à analyser, manipuler, visualiser...



Données structurées

Un exemple concret : le schéma de la base de données relationnelle derrière [la plateforme Datagora](#)



Données semi-structurées

Pas de modèle prédéfini : structure irrégulière, implicite... mais données organisées néanmoins, ensemble raisonnable de champs

Exemple : XML, JSON

Possible de trier, ordonner et structurer les données

```
{  
    "ocid": "56810-2006",  
    "releaseID": "F-Lille: other electrical installation work - 1219084",  
    "releaseDate": "2006-01-12T00:00:00-0500",  
    "releaseTag": "awardNotice",  
    "language": "en",  
    "buyer": {  
        "id": {  
            "name": "RÉGION NORD-PAS DE CALAIS",  
            "uid": "6504162",  
            "uri": "http://www.dgmarket.com/tenders/adminShowBuyer.do?buyerId=6504162"  
        },  
        "address": {  
            "country-name": "France"  
        }  
    }  
}
```

Données non structurées

Pas de structure commune identifiable
Exemple : BDD NoSQL

Généralement qualitatives

Difficilement combinées ou analysées quantitativement

Les données non structurées croisraient 15x plus que les données structurées

Le machine learning est de plus en plus capable d'analyser ces données.

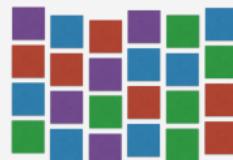
Voir sections 9 et 10

Structured Data



What you find in a DB (typically)

Unstructured Data



What you find in the 'wild' (text, images, audio, video)

R

Données capturées, échappées, transitoires

Données capturées

Données issues d'observations, d'enquêtes, d'expérimentations, de prise de notes, de senseurs... => il y a eu l'**intention de générer des données**

Données échappées

Sous-produit d'un engin ou d'un système dont la fonction première est autre
Avez-vous des exemples de données échappées ?

Données capturées, échappées, transitoires

Données capturées

Données issues d'observations, d'enquêtes, d'expérimentations, de prise de notes, de senseurs... => il y a eu l'**intention de générer des données**

Données échappées

Sous-produit d'un engin ou d'un système dont la fonction première est autre
Avez-vous des exemples de données échappées ?

Parking, borne d'accès... => Des données sur les horaires d'accès, le nombre d'ouvertures, fermetures, la fréquentation

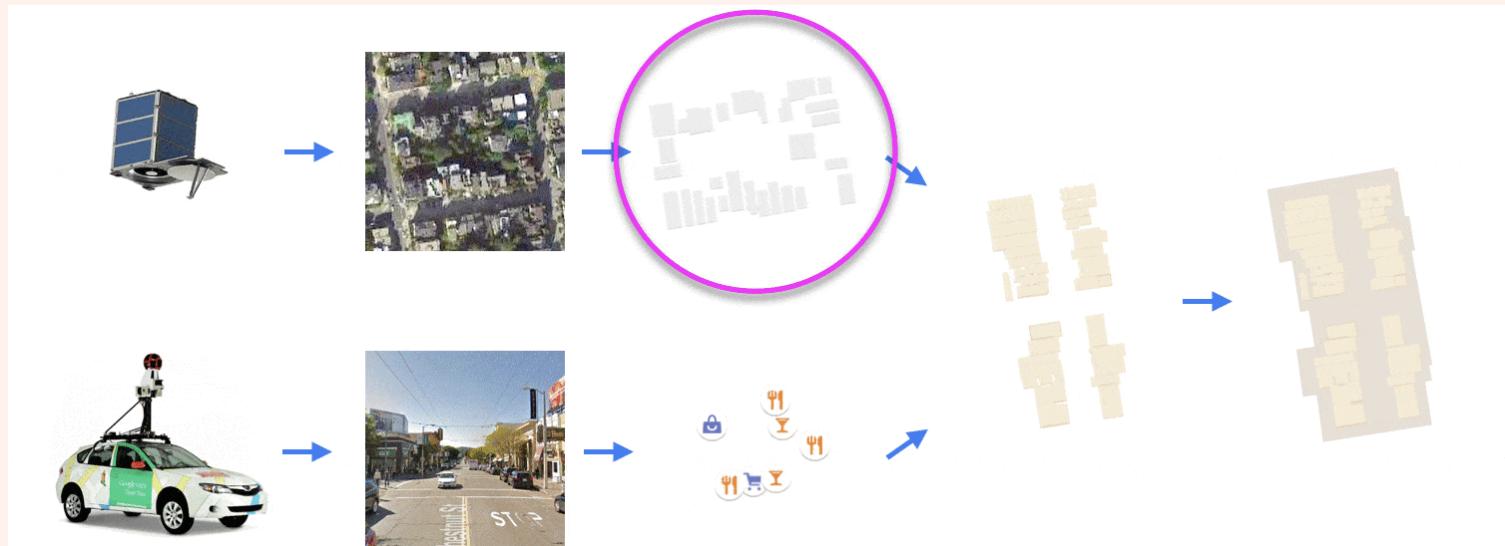
Données transitoires

Ce sont des données échappées qui ne sont jamais examinées, transformées ou analysées

Données dérivées

Résultat d'un traitement ou d'une analyse supplémentaire de données capturées.

Exemple avec les **données de Google Maps** :



Recommandé : Lire l'article "["GOOGLE MAPS'S MOAT - How far ahead of Apple Maps is Google Maps?"](#)"

2. Données, données... quelles données ?

Les données crowdsourcées

Des données produites par des citoyens, des communs partagés et gouvernés par leurs producteurs.

Concrètement, les données sont issues du travail collaboratif de divers acteurs, bénévoles, dans la récolte sur le terrain.

Connaissez-vous un site ou une application fonctionnant via des données crowdsourcées ?

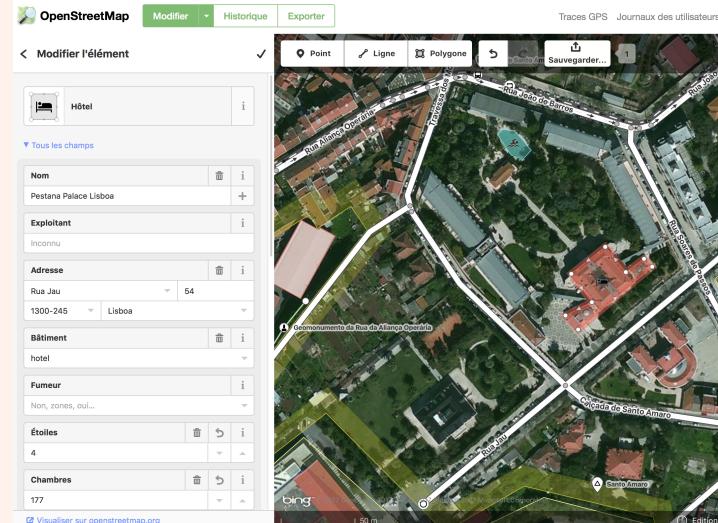
Les données crowdsourcées

Des données produites par des citoyens, des communs partagés et gouvernés par leurs producteurs.

Concrètement, les données sont issues du travail collaboratif de divers acteurs, bénévoles, dans la récolte sur le terrain.

Connaissez-vous un site ou une application fonctionnant via des données crowdsourcées ?

Exemple : OpenStreetMap, le wiki de la carte



Exemple 1 : OpenStreetMap

Pourquoi faites-vous OpenStreetMap ? Les données géographiques (géo-données) ne sont pas libres dans nombre de régions du monde, par exemple en France, en Belgique, au Canada. En général, ces régions ont confié la tâche de cartographie à diverses agences gouvernementales, qui en retour font de l'argent en revendant les données à des gens comme vous et moi. Si vous vivez dans un de ces pays, alors vos impôts servent à payer le travail de cartographie

En France certaines données du ministère des finances (données cadastrales pour l'identification des parcelles) peuvent être réutilisées comme référence, mais avec des conditions qui ne permettent pas une exploitation massive permettant d'obtenir une carte complète (leur précision ne permet pas nécessairement d'identifier tous les chemins, rues et routes qui traversent une même parcelle ; de plus elles ne sont souvent plus à jour).

Source : [La FAQ d'OpenStreetMap](#)

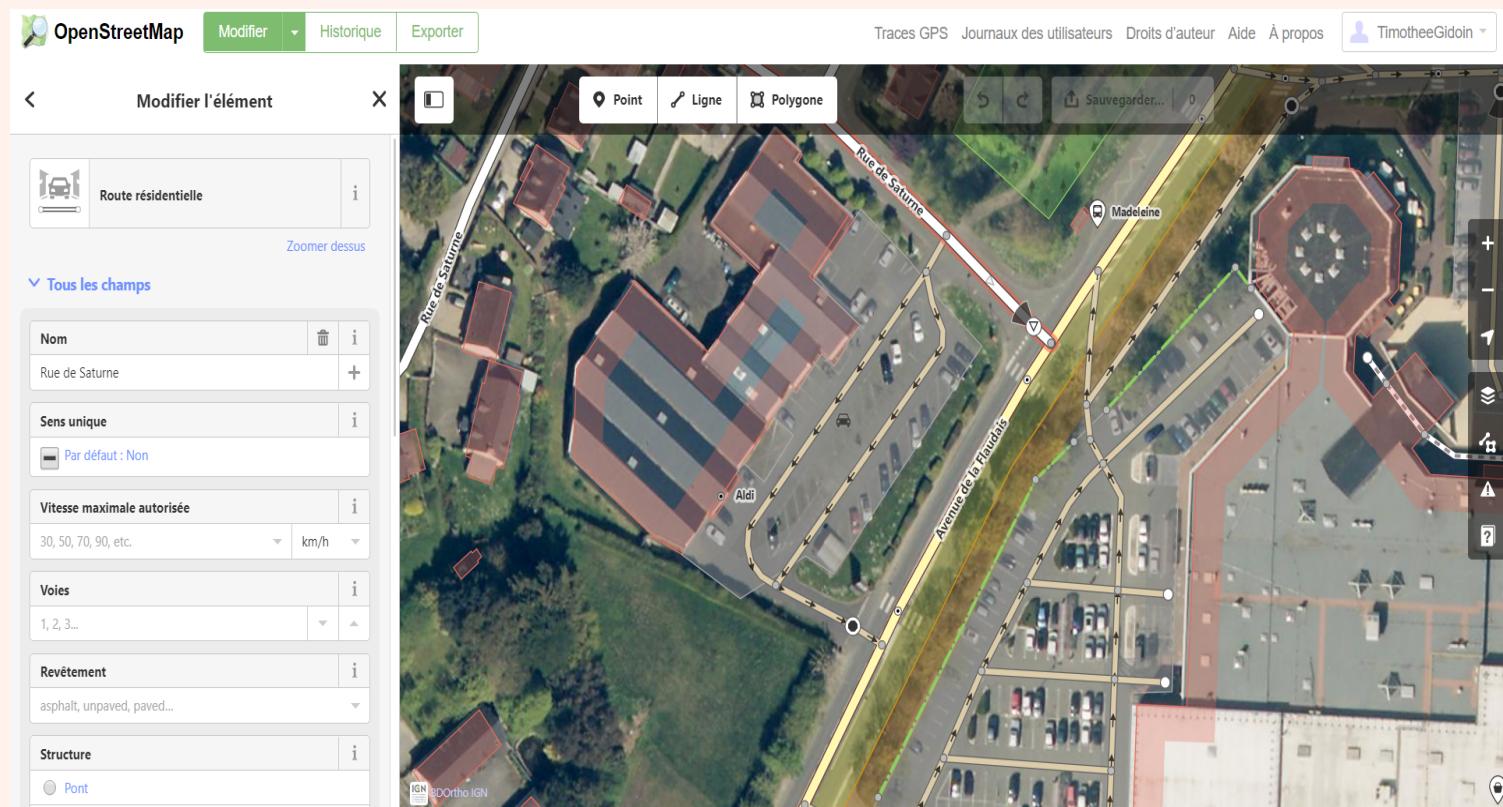
Exemple 1 : OpenStreetMap

En résumé :

- OpenStreetMap (OSM) est un projet de cartographie qui a pour but de constituer une base de données géographiques libre du monde (permettant par exemple de créer des cartes sous licence libre), en utilisant le système GPS et d'autres données libres.
- À la manière de Wikipédia, **tous les internautes naviguant sur le web peuvent contribuer à la création et à la numérisation de cartes**. Des éditeurs permettent de réaliser en ligne des cartes en se basant sur un fond d'image satellitaire. Cependant, ces images satellitaires ne couvrent pas toujours en haute résolution l'ensemble du globe. C'est pourquoi il est possible d'introduire des données provenant de récepteurs GPS. Il suffit pour cela de réaliser un itinéraire et de positionner le récepteur en mode enregistrement, puis de le restituer sur le serveur de données d'OpenStreetMap

Exemple 1 : OpenStreetMap

Vous aussi vous pouvez contribuer à OpenStreetMap !



Exemple 2 : OpenFoodFacts



Open Food Facts est une base de données sur les produits alimentaires faite par tout le monde, pour tout le monde. Elle vous permet de faire des choix plus informés, et comme les données sont ouvertes (open data), tout le monde peut les utiliser pour tout usage.

Open Food Facts est un projet citoyen à but non lucratif créé par des milliers de volontaires à travers le monde. Vous pouvez commencer à contribuer en ajoutant un produit de votre cuisine, et nous avons plein de projets enthousiasmants auxquels vous pouvez participer de beaucoup de façons différentes.

Exemple 2 : OpenFoodFacts

Vous aussi vous pouvez contribuer à OpenFoodFacts !

Chacun peut contribuer

Participez à notre base libre et ouverte sur les produits alimentaires du monde entier !

Open Food Facts est un projet à but non lucratif développé entièrement par des volontaires, nous avons vraiment besoin de vous.

Ajoutez des produits

Utilisez notre app [Android](#), [iPhone](#) ou [Windows Phone](#) pour scanner le code barre des produits que vous possédez ou de vos magasins préférés et envoyer des photos de leur étiquette.

Pas de smartphone ? Pas de problème : vous pouvez tout aussi bien utiliser un appareil photo pour ajouter des produits directement sur le site web.

Sur le site web, vous pourrez également remplir les informations des produits que vous ajoutez ou que d'autres ont ajoutés.

Faites connaître le projet

Parlez d'Open Food Facts autour de vous !

Vous pouvez présenter le projet à votre famille et à vos amis, leur montrer comment installer l'app et contribuer, écrire un billet de blog, partager le site sur les réseaux sociaux etc.

Et vous pouvez faire une présentation en live ! Nous avons présenté Open Food Facts devant des audiences très différentes, dans des lieux très différents, et à chaque fois les retours sont très enthousiastes quand nous présentons tout ce qu'on peut faire lorsque les données des produits sont ouvertes. Nous avons des [présentations](#) (slides etc.) que vous pouvez adapter ou présenter tels quels.

Adaptez-le pour votre pays

Vous pouvez nous aider à [traduire](#) le site et l'application mobile dans votre langue, ainsi que les présentations, annonces etc.

Nous pouvons travailler ensemble pour ajouter des logos pour les labels de votre pays, pour décoder les codes emballeurs afin de les cartographier etc.

Commencez ou rejoignez une communauté locale de contributeurs : ajoutez des produits locaux, recrutez des amis, présentez le projet dans des rencontres et conférences locales etc.

Exemple 2 : OpenFoodFacts

Et vous connaissez très probablement une application qui utilise les données d'OpenFoodFacts pour "évaluer" la composition des produits alimentaires...

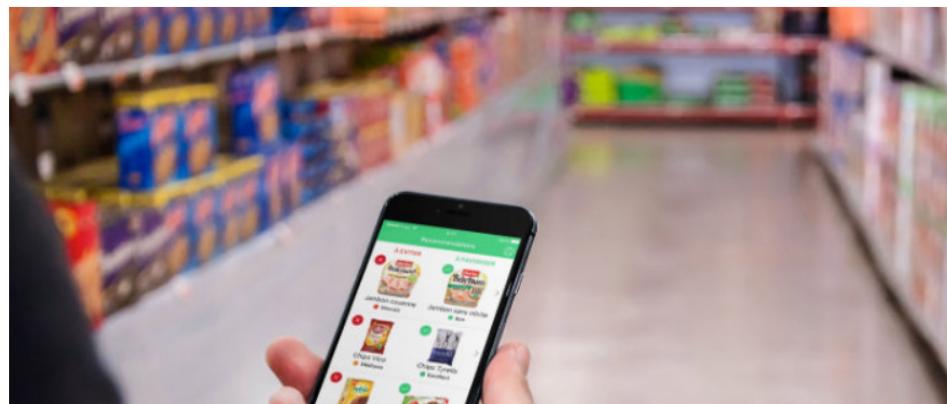
Exemple 2 : OpenFoodFacts

Et vous connaissez très probablement une application qui utilise les données d'OpenFoodFacts pour "évaluer" la composition des produits alimentaires...

Yuka ! Avec plus de 8 millions d'utilisateurs en février 2019, elle a désormais un impact non négligeable, y compris dans la stratégie des grandes marques de l'agroalimentaire..

ALIMENTATION : INQUIETS DU SUCCÈS DE L'APPLI YUKA, INDUSTRIELS ET DISTRIBUTEURS CONTRE-ATTAQUENT

L'appli star Yuka, qui note les aliments en fonction de leurs qualités nutritionnelles, revendique désormais 7,5 millions d'utilisateurs. Une influence grandissante qui a poussé le secteur agroalimentaire à changer la composition de certains produits. Mais son système de notation est critiqué par des industriels qui passent à l'offensive.



La Grande Annotation

L'objectif ? Faire en sorte que les contributions au grand débat puissent être lues et comprises. Tout un chacun peut, sur le site grandeannotation.fr lire ces textes, classés par thème et par question, et les annoter pour en révéler le sens.

Donnons du sens au grand débat

Plus de **250 000 personnes** ont rédigé des réponses aux questions de granddebat.fr

Mais **aucune technologie** n'est aujourd'hui capable de **comprendre leur sens**.

Rejoignez 1 106 humains pour **nous aider à le faire** en les lisant !

Plutôt que d'obtenir une synthèse des contributions au grand débat qui soit non collaborative, opaque (car réalisée par quelques sociétés) et en partie traitée par de l'intelligence artificielle, la Grande Annotation veut construire une synthèse collective, transparente et fondée sur l'intelligence humaine.

La Grande Annotation

Vous aussi vous pouvez contribuer en annotant les réponses au grand débat ! Ce faisant, vous créez de fait de nouvelles données qui viennent qualifier et enrichir les données initiales.

Démocratie et citoyenneté Transition écologique Fiscalité et dépense publique Organisation de l'État et des services publics

Que pensez-vous de la situation de l'immigration en France aujourd'hui et de la politique migratoire ? Quelles sont, selon vous, les critères à mettre en place pour définir la politique migratoire ?

- La population ne se renouvelle pas sans l'immigration - Elle devrait être choisie et chaque immigré légal ou non doit être traité dignement - Nous avons besoin de l'immigration d'un point de vue économique, culturel et solidaire - Mieux répartir sur le territoire les arrivées des personnes

- 23 février, Auterrive

Annoter cette réponse Lire une autre réponse Changer de question

Selon vous, quelles idées clefs résument le mieux ce qu'a voulu dire le répondant ?

Accueillir mieux Aider les pays d'origine Aider les réfugiés, demandeurs d'asile Politique trop laxiste (trop de migrants)

Politique non satisfaisante Politique satisfaisante Contrôler, renvoyer Politique européenne commune

Exiger des efforts d'intégration Expliquer, clarifier, informer Immigration choisie, sélective Instaurer des quotas

Priorité aux Français, SDF Répartir géographiquement les arrivées Respect de la culture française Rester un pays ouvert

Autres + Créer

3. Petit lexique autour des données

Index

Index : Des données permettent l'identification et la mise en relation.

Essentielles pour enrichir les données. Exemple : le numéro de SIRET dans la base Sirene (informations concernant les entreprises et les établissements immatriculés au répertoire interadministratif Sirene depuis sa création en 1973), gérée par l'Insee

Base Sirene v1

The screenshot shows a table with 12 rows of data. The columns are: SIRET, Libellé de la région de l'établissement, Libellé du département de l'établissement, Département et commune de localisation, IRIS de l'établissement, and Libellé de la taille de l'établissement. The data includes various regions like Auvergne-Rhône-Alpes, Occitanie, Ile-de-France, and Nouvelle Aquitaine, along with their respective departments and communes. The IRIS column shows identifiers like 74010, 74143, 46269, etc. The last column indicates the size of the establishment, such as 'Unité urbaine de 100' or 'Agglomération de Paris'.

SIRET	Libellé de la région de l'établissement	Libellé du département de l'établissement	Département et commune de localisation	IRIS de l'établissement	Libellé de la taille de l'établissement
1	83409038300011	Auvergne-Rhône-Alpes	HAUTE-SAVOIE	74010	Unité urbaine de 100
2	83409040400014	Auvergne-Rhône-Alpes	HAUTE-SAVOIE	74143	Unité urbaine de 100
3	83409056500013	Occitanie	LOT	46269	Établissement localisé
4	83409059900012	Auvergne-Rhône-Alpes	PUY-DE-DOME	63143	Établissement localisé
5	83409071400017	Ile-de-France	SEINE-ET-MARNE	77181	Agglomération de Paris
6	83409084700015	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
7	83409097900016	Nouvelle Aquitaine	GIRONDE	33063	Unité urbaine de 200
8	83409103500016	Provence-Alpes-Côte d'Azur	BOUCHES-DU-RHONE	13096	Unité urbaine de moins de 100
9	83409105000015	Provence-Alpes-Côte d'Azur	VAR	83126	Unité urbaine de 200
10	83409119100017	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
11	83409121700010	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
12	83409123300017	Ile-de-France	VAL-D'OISE	95572	Agglomération de Paris

Attributs

Attributs : Des données représentent les aspects d'un phénomène, mais ne sont pas des index (pas identifiants uniques). Exemples avec la base Sirene : région de l'établissement, département de l'établissement, Iris de l'établissement...

Base Sirene v1

The screenshot shows the 'Base Sirene v1' dashboard. At the top, there are navigation links: Informations (selected), Tableau, Carte, Analyse, Développez vos services, Export, and API. To the right are social media sharing icons for Twitter, Facebook, LinkedIn, and Email.

The main content is a table with the following columns: SIRET, Libellé de la région de l'établissement, Libellé du département de l'établissement, Département et commune de localisation, IRIS de l'établissement, and Libellé de la taille de l'établissement.

SIRET	Libellé de la région de l'établissement	Libellé du département de l'établissement	Département et commune de localisation	IRIS de l'établissement	Libellé de la taille de l'établissement
1	83409038300011	Auvergne-Rhône-Alpes	HAUTE-SAVOIE	74010	74010
2	83409040900014	Auvergne-Rhône-Alpes	HAUTE-SAVOIE	74143	Unité urbaine de 10 000 à 20 000 habitants
3	83409056500013	Occitanie	LOT	46269	Établissement localisé dans une zone rurale
4	83409059900012	Auvergne-Rhône-Alpes	PUY-DE-DOME	63143	Établissement localisé dans une zone rurale
5	83409071400017	Ile-de-France	SEINE-ET-MARNE	77181	Agglomération de Paris
6	83409084700015	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
7	83409097900016	Nouvelle Aquitaine	GIRONDE	33063	Unité urbaine de 200 à 500 habitants
8	83409103500016	Provence-Alpes-Côte d'Azur	BOUCHES-DU-RHONE	13096	Unité urbaine de moins de 100 habitants
9	83409105000015	Provence-Alpes-Côte d'Azur	VAR	83126	Unité urbaine de 200 à 500 habitants
10	83409111900017	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
11	83409121700010	Ile-de-France	VAL-D'OISE	95585	Agglomération de Paris
12	83409123300017	Ile-de-France	VAL-D'OISE	95572	Agglomération de Paris

Métadonnées

Métadonnées : Des données sur les données. Peuvent être descriptives, structurelles ou administratives. Exemple de standard : le **Dublin Core**.

Pour notre exemple précédent, les métadonnées (date de modification, producteur, nombre de téléchargements, thématiques...) sont les suivantes :

Avertissement

La base Sirene contenant des données à caractère personnel, l'Insee attire votre attention sur les obligations légales qui en découlent:

- Le traitement de ces données relève des obligations de déclaration de la Loi 78-17 du 6 janvier 1978 modifiée, dite Loi CNIL : <https://www.cnil.fr/fr/loi-78-17-du-6-janvier-1978-modifiee>
- Selon votre usage du jeu de données, il est de votre responsabilité de tenir compte du statut de diffusion le plus récent de chaque personne physique. En effet, l'article A123-96 du code de commerce dispose que : "Toute personne physique peut demander soit directement lors de ses formalités de création ou de modification, soit par lettre adressée au directeur général de l'Institut national de la statistique et des études économiques, que les informations du répertoire la concernant ne puissent être utilisées par des tiers autres que les organismes habilités au titre de l'article R. 123-224 ou les administrations, à des fins de prospection, notamment commerciale."

Le site de l'Insee www.sirene.fr fournit des informations sur le contenu de ces bases, ainsi que la [documentation](#) associée et une [Foire Aux Questions](#) pour vous aider, comprenant plus de 50 questions-réponses.

Identifiant du jeu de données [sirene](#)

Téléchargements 47 998

Thèmes Administration, Gouvernement, Finances publiques, Citoyenneté, Économie, Business, PME, Développement économique, Emploi

Mots clés SIREN, SIRENE, SIRET, entreprises, INSEE

Licence [Licence Ouverte \(Etalab\)](#)

Langue Français

Modifié 11 mai 2018 20:25

Producteur INSEE

Référence <http://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/>

Dernier traitement 27 mars 2019 12:07 (métadonnées)

27 mars 2019 11:08 (données)

Bibliographie

- Libération, Checknews, "Yuka est-elle une appli publicitaire déguisée ?", mis en ligne le 18 mai 2018
- Jérémie Valentin, « Les données environnementales : un cas particulier dans la mise en place des données publiques en Open Data ? », Netcom, 27-1/2 | 2013, 254-263
- Justin O'Beirne "[GOOGLE MAPS'S MOAT - How far ahead of Apple Maps is Google Maps?](#)"
- Alain Desrosières, "[L'État, le marché et les statistiques : Cinq façons d'agir sur l'économie](#)"
- Isabelle Tellier "[Introduction à la fouille de textes](#)" Université de Paris 3 - Sorbonne Nouvelle

Quiz section 3 : rdv sur votre espace e-campus !

Merci !

Contact : timothee@dataactivi.st