Glossaire de la donnée

Rencontre organisée par le Sieeen
« Que pouvez-vous faire de vos données ? »
19/11/2019 à la ferme du Marault à Magny-Cours

Glossaire des données

Table des matières

AGD Administrateur général des données. 3
API (application programming interface) 3
Algorithme 3
Base de données 3
Big Data ou grandes données 3
CSV (Comma Separated Values) 3
Crowdsourcing ou production participative 4
Donnée ou Donnée numérique 4
Donnée ouverte 4
Données pivot ou Donnée clé 4
Données liées / Web sémantique (Linked Data) 4
Données de référence 5
GAFAM 5
Interopérables 5
Jeu de données (dataset) 5
Licence 5
Data mining ou fouille de données 5
Datavisualisation ou “Dataviz” 6
Droit de communication 6
Gouvernance de la donnée 6
Machine learning 6
Métadonnée 6
Mission de service public 6
Pseudonymisation 7
Quantified self (mesure de soi) 7
Service public de la donnée 7
CRÉDITS : 8
Vos interlocuteurs open data : 8

AGD Administrateur général des données.

En France, la fonction a été créée par décret du Premier ministre le 16 septembre 2014. L'AGD coordonne l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données par les administrations. (AGD).

API (application programming interface)

ou interface de programmation, ou web service C'est une interface de dialogue, technique et normalisée, qui permet d'échanger des informations et des services entre machines. Une API permet à un service de fournir des données de façon standardisée. Une API peut renvoyer les coordonnées GPS d'une adresse postale donnée par exemple la Base d'Adresse Nationale Ouverte.

Algorithme

Un algorithme est une méthode de traitement des données. Ce traitement est automatisé : la machine l'effectue à partir de données et en produit d'autres (composante technique). Ce traitement est produit par une personne ou un service en fonction d'un objectif (composante sociale). Par exemple, le calcul du jour de la semaine pour une date donnée. La loi pour une République numérique impose la mention de la finalité d'un traitement algorithmique.

Anonymisation Le terme d'anonymisation est réservé aux opérations irréversibles. On utilise le terme de pseudonymisation lorsque l'opération est réversible. Une anonymisation irréversible consiste à supprimer tout caractère identifiant à un ensemble de données. Concrètement, cela signifie que toutes les informations directement ou indirectement identifiantes sont supprimées ou modifiées, rendant impossible toute ré-identification des personnes. Voir Pseudonymisation. (CNIL)

Base de données

C'est un ensemble de données organisé dans le but de faciliter leur usage. Une base de données contient un ensemble d'informations structurées permettant de les stocker pour opérer des traitements et fournir des services.

Big Data ou grandes données

Cette expression désigne un ensemble de données très volumineux qui doit être traité par des outils spécifiques.

CSV (Comma Separated Values)

Le csv est un nom d'un format de fichier contenant des données textuelles. Les données sont lisibles par un très grand nombre d'outils : les tableurs, les éditeurs de texte... L'organisation du fichier étant répandue et simple, elle peut être plus facilement traitée par un programme.

Crowdsourcing ou production participative

C'est l'utilisation de la créativité, de l'intelligence et du savoir-faire d'un grand nombre de personnes, en sous-traitance, pour réaliser certaines tâches traditionnellement effectuées par un employé ou un entrepreneur
(Wikipédia juillet 2017 : https://fr.wikipedia.org/wiki/Production_participative).
Par exemple, les contenus de Wikipédia et d'Open Street Map sont réalisés en crowdsourcing.

Donnée ou Donnée numérique

Une donnée numérique est la description élémentaire de nature numérique, représentée sous forme codée, d'une réalité (chose, événement, mesure, transaction, etc.). (AGD).

Donnée ouverte

L'open data ou donnée ouverte est une donnée numérique caractérisée, à minima, par plusieurs propriétés :

- elle est librement accessible,
- elle est compréhensible,
- elle est dans un format suffisamment répandu pour être exploitable
- par une machine,
- elle est réutilisable par tous, ses conditions de réutilisation sont
- précisées dans une licence.

Source : liberTIC
Une donnée ouverte n'est pas forcément une donnée publique. Associations, citoyens ou entreprises peuvent mettre des données à disposition, sans que celles- ci soient produites dans le cadre d'une mission de service public. L'ouverture des données se réalise par la publication des données sur des sites web, des portails ou via des API (Interface pour l'accès Programmé aux Applications). Les données sont mises à disposition sous différents formats de fichiers qui permettent la manipulation et le traitement des données.

Données pivot ou Donnée clé

Une donnée pivot est une donnée qui permet de relier plusieurs jeux de données, comme, par exemple, le numéro SIRET d'une entreprise. (AGD).

Données liées / Web sémantique (Linked Data)

On appelle web sémantique l'extension du web traditionnel pour permettre à toute donnée d'être publiée et documentée de façon standard. Les données liées, c'est la possibilité d'attribuer une adresse à un objet, une URI, et de pouvoir pointer vers elle de façon fixe. Il s'agit de créer une clef qui permet d'aller chercher l'objet et d'y faire référence. Par exemple, Wikidata permet de donner accès à des données de base (dates de naissance, capitale d'un pays...).

Données de référence

Dans le cadre du Service Public de la Donnée, les données de référence sont précisées par l'article 14 de la Loi Pour une République Numérique. Les données de référence sont des informations publiques qui satisfont les conditions suivantes :

- Elles constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes ;
- Elles sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient ;
- Leur réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité.

Un décret dresse la liste des données de références, ainsi que l'administration responsable de leurs conditions de production et de publication. La Base Adresse Nationale, la base Siren, le Répertoire Opérationnel des Métiers et des Emplois (code ROME produit par Pôle Emploi) sont des données de référence.

GAFAM

C'est l'abréviation de Google, Amazon, Facebook, Apple, Microsoft : il s'agit des entreprises les plus puissantes de l'internet et accessoirement celles qui détiennent et/ou manipulent le plus de données.

Interopérables

Le terme interopérable désigne le fait que deux systèmes techniques peuvent s'échanger aisément des données. Plus les systèmes respectent les normes et les standards ouverts, plus ils sont interopérables.

Jeu de données (dataset)

Un jeu de données est un ensemble de données qui forme un tout. Par exemple, la liste de présence des conseillers municipaux lors des assemblées en 2012, est un jeu de données.

Licence

Une licence est un contrat qui précise les conditions de réutilisation d'un jeu de données. Par exemple, des données sous licence ODBL ou Licence Ouverte sont réputées en open data.

Data mining ou fouille de données

La fouille de données consiste en l'exploration de masse de données issues de documents ou base de données pour les analyser à partir de méthodes comme la statistique, un traitement automatisé/algorithmique, une intelligence artificielle. L'objectif de cette analyse est de comprendre, résoudre ou encore prévoir des actions.

Datavisualisation ou “Dataviz”

Il s’agit de représentation graphique de données. Quelques formes simples et connues de visualisation de données sont le “camembert”, l’histogramme, le nuage de points. La visualisation de données peut s’appuyer sur différentes sources de données. Elle a pour objectif de rendre les données plus lisibles et compréhensibles.

Droit de communication

Les administrations sont tenues de publier en ligne ou de communiquer les documents administratifs qu’elles détiennent aux personnes qui en font la demande. Cela ne s’applique qu’à des documents achevés et ne concerne pas les documents préparatoires à une décision administrative tant qu’elle est en cours d’élaboration. Dans le cas où la demande permet de bénéficier d’une décision individuelle créatrice de droits, les documents sont communicables à l’auteur de cette demande dès leur envoi à l’autorité compétente pour statuer sur la demande.
Le droit de communication perdure, même si le document est déposé aux archives. Le droit de communication cesse lorsque les documents font l’objet d’une diffusion publique.

Gouvernance de la donnée

Ensemble de principes et de pratiques qui visent à assurer la meilleure exploitation du potentiel des données. (AGD).

Machine learning

Le machine learning est une forme d’intelligence artificielle, cela peut se traduire par l’apprentissage automatique. Le machine learning est un ensemble de techniques où les algorithmes sont dits apprenants. C’est-à-dire que les algorithmes se perfectionnent et s’améliorent d’eux-mêmes en traitant des données renseignées par les humains. (AGD).

Métadonnée

Une métadonnée est une information descriptive liée à une donnée. Par exemple, la date de production de la donnée, son producteur, son format, sa licence constituent des métadonnées. Pour qu’un jeu de données soit facilement accessible et réutilisable, la qualité des métadonnées joue un rôle déterminant.

Mission de service public

Une mission de service public est une action menée par une administration pour satisfaire l’intérêt général. Il peut s’agir de service public administratif ou d’un service public industriel et commercial. La jurisprudence a établi un faisceau d’indices permettant de déterminer si l’on est en présence d’un service public :

- une activité d’intérêt général,
- la présence directe ou indirecte d’une administration,
- la présence de prérogatives de puissance publique,
- les modalités de financement public.

Pour approfondir la notion, le wiki du CNFPT : https://frama.link/Def-servicepublic-cnfpt.

Pseudonymisation

La pseudonymisation est une technique qui consiste à remplacer un identifiant (ou plus généralement des données à caractère personnel) par un pseudonyme. Cette technique permet la ré-identification ou l'étude de corrélations en cas de besoin particulier. Lors d'une pseudonymisation, il faut être vigilant dans la mesure où une ré-identification peut intervenir à partir d'informations partielles (par exemple, la combinaison des données ville et date de naissance peut être suffisante). Voir Anonymisation (CNIL).

Quantified self (mesure de soi)

C'est un mouvement qui regroupe les outils, les principes et les méthodes permettant à chacun de mesurer ses données personnelles, de les analyser et éventuellement de les partager. Les outils du quantified self peuvent être des objets connectés, des applications mobiles ou des applications Web (Wikipédia juillet 2017 : https://fr.wikipedia.org/wiki/Quantified_self). Le fait de mesurer le nombre de pas effectués par jour constitue une pratique de quantified self.

Self data Le self data désigne la production, l'exploitation et le partage de données personnelles par les individus, sous leur contrôle et à leurs propres fins : pour mieux se connaître, prendre de meilleures décisions, se faciliter la vie, etc.

Service public de la donnée

Le service public de la donnée créé par l'Article 14 de la loi pour une République numérique vise à mettre à disposition, en vue de faciliter leur réutilisation, les jeux de données de référence qui présentent le plus fort impact économique et social. Il s'adresse principalement aux entreprises et aux administrations pour qui la disponibilité d'une donnée de qualité est critique. Les producteurs et les diffuseurs prennent des engagements auprès de ces utilisateurs. La mission Etalab est chargée de la mise en œuvre et de la gouvernance de ce nouveau service public. Elle référence l'ensemble des données concernées.

(Extrait : https://www.data.gouv.fr/fr/reference).

CRÉDITS :

Pour le glossaire de la donnée, est une compilation de plusieurs sources :

- la documentation produite par OpenData France dans le cadre de la démarche OpenDataLocale. http://opendatalocale.net/ressources/. Elle-même est une mise à jour de “Les mots de l’Infolab” de la Fing http://infolabs.io/mots-infolab

- le glossaire dans le rapport 2016-2017 de l’Administrateur général des données https://www.etalab.gouv.fr/wp-content/uploads/2018/04/RapportAGD_2016-2017_web.pdf. Dans ce cas, le terme est précisé par l’annotation (AGD)

- la documentation de la Cnil en particulier Le pack de conformité Logement Social (https://www.cnil.fr/sites/default/files/typo/document/FICHE10_PackConf_LOGEMENT_SOCIAL_web.pdf). Dans ce cas, le terme est précisé par l’annotation (CNIL).

Document mis à jour par La Reine Merlin contact@lareinemerlin.org 06 27 80 41 41 www.lareinemerlin.org @ArmelleGilliard dans le cadre de la rencontre Que pouvez-vous faire de vos données ? du 19/11/2019 à ferme du Marault à Magny-Cours.

Vos interlocuteurs open data :

- **Philippe JEANNET**, Chef de projet qualité, organisation, dématérialisation
  - (03) 86 59 76 90 Poste N° 141
  - 06.84.54.91.67