Les mots de l'infolab

API (*application programming interface*) ou interface de programmation, est une interface technique normalisée au travers de laquelle un logiciel offre des services à d'autres logiciels. Une API permet typiquement à un service de fournir des données de façon standardisée. Une API peut par exemple renvoyer les coordonnées GPS d'une adresse postale donnée.

Algorithme : une suite d'opérations ou instructions [informatiques] permettant de résoudre un problème. (Wikipedia.) La plupart du temps, un algorithme consomme des données pour en produire d'autres. Ex. : le calcul du jour de la semaine pour une date donnée.

Base de données ou *database*, est un ensemble de données organisé dans le but de faciliter leur usage.

Big Data : désigne un ensemble de données très volumineux qui doit être traité par des outils spécifiques.

CSV : fichier contenant des données dans un format textuel. Les données sont ainsi lisibles par un très grand nombre d'outils -- les tableurs, mais aussi de simples éditeurs de texte -- et plus faciles à traiter par des programmes.

Corrélation : montre la similitude de deux phénomènes. Deux données corrélées sont deux données qui évoluent de la même manière. Pour autant, "corrélation ne vaut pas causalité" : la corrélation peut être causée par un facteur tiers. L'augmentation corrélée de la vente des lunettes noires et des glaces ne signifie pas que les consommateurs associent ces deux produits : le soleil de l'été l'explique mieux !

Crowdsourcing ou production participative, 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. (Source Wikipedia.)

Datajournalisme : littéralement journalisme de données : travaux d'analyse ou de communication journalistiques fondés sur les données. Ex. : http://www.letemps.ch/data/

Data mining (ou fouille de données) : exploration des données pour en extraire -- par le biais de la statistique, l'intelligence artificielle, etc. -- des méthodes pour comprendre, résoudre ou encore prévoir des actions.

Data scientist✣ : "Job le plus sexy du 21° siècle" selon la Harvard Review. Un data scientist possède des compétences en statistiques, traitement de données et en programmation informatique. Il met en oeuvre un ensemble de techniques d'analyse, dont les algorithmes et le machine learning. A remplacé le terme "analyste de données".

Dataset : voir jeu de données.

Datavisualisation ou "**Dataviz**" : représentation graphique de données, par exemple sous la forme de camemberts, histogrammes, nuages de points, etc. Une dataviz rend les données plus lisibles et compréhensibles.

Donnée (*data*) : une description élémentaire d'une réalité. C'est par exemple une observation ou une mesure. (Source Wikipédia.)

Donnée brute : l'expression "donnée brute" a été popularisée pour indiquer une donnée n'ayant subit aucun traitement ; pour autant, parce qu'elle est construite par l'homme, une donnée est forcément le résultat d'un traitement, d'une action. Il est donc toujours important d'interroger le mode de production de chaque donnée.

Donnée froide : donnée dont la valeur ne change pas dans le temps. Par exemple : une liste électorale est une donnée froide. La circulation à un instant T dans une ville est, par opposition, une donnée chaude.

✣ source : Datanomics de L.-D. Benayer et S. Chignard.

Donnée personnelle : Données attachées et produites par une personne identifiée et identifiable.

Donnée pivot ou parfois donnée de référence **:** donnée ayant valeur de référence pour plusieurs métiers/écosystèmes, permettant ainsi de croiser ou relier des données. Le code postal est une donnée pivot.

Géolocalisé : se dit d'un objet localisé dans l'espace sous la forme de coordonnées X et Y. Un POI est géolocalisé. Votre téléphone vous géolocalise grâce au GPS.

Hacker : le fait de résoudre, modifier, reproduire, transformer, bricoler ou pirater un système, souvent pour le détourner et l'améliorer. On peut hacker un programme mais aussi un lieu, un événement, un objet, etc.

Fichier plat : fichier texte représentant une base de données rudimentaire, contenant généralement un seul enregistrement par ligne. Typiquement, un fichier CSV.

Fouille de données : voir Data mining.

GAFA : abréviation de Google, Amazon, Facebook, Apple, soit les entreprises les plus puissantes de l'internet (et accessoirement celles qui détiennent et/ou manipulent le plus de données sur nous).

Infolab (parfois *datalab*) : un espace collaboratif dédié à la compréhension, la manipulation et l'exploration de données.

Interopérables : se dit de deux systèmes techniques qui peuvent s'échanger naturellement des données. Souvent employé à tord hors de tout contexte d'usage : "il faut plus d'interropérabilité".

Jeu de données (*dataset*) : 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 : conditions juridiques d'usage. Par exemple des données sous licence ODBL ou Licence Ouverte sont réputées en Open Data.

Métadonnée : une information liée à une donnée. Par exemple, l'auteur d'un livre est une métadonnée du dit livre.

ODBL (*Open DataBase Licence*) : voir licence.

Open Data : données réutilisables par tous, gratuitement, sans restriction technique ou juridique. L'open data désigne plus largement le mouvement qui promeut l'ouverture des données en libre accès ou réutilisation, pour des objectifs d'innovation économique et sociale.

POI (*Point of interest*) ou point d'intérêt, représente un site utile, un point digne d'intérêt. Ce terme est utilisé par différents logiciels de cartographie et appareils de navigation GPS.

Proxy✣ : En statistique, le proxy est utilisé lorsque la variable d'intérêt est manquante. Par exemple, dans une série temporelle longue sur le prix du blé, il se peut que certaines données soient manquantes. On pourra utiliser le prix d'une autre matière première si celui-ci est fortement corrélé au prix du blé.

Quantified self (mesure de soi) est un mouvement qui regroupe les outils, les principes et les méthodes permettant à chacun de mesurer ses données personnelles, de les analyser et de les partager. Les outils du quantified self peuvent être des objets connectés, des applications mobiles ou des applications Web. (Wikipedia).

Repository (ou "repo" ou "entrepôt de données") : lieu de stockage et d'accès aux données.

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 meilleurs décisions, se faciliter la vie, etc.

SHAPE : format de fichier technique très utilisé dans le monde de la géomatique, pour représenter des modèles spatiaux.

Web des données (ou web sémantique) : extension du web traditionnel pour permettre à toute donnée d'être publiée sur le web de façon standard. La capitale de la France devient ainsi une URL que des programmes peuvent interroger.

✣ source : Datanomics de L.-D. Benayer et S. Chignard.