Data : le glossaire

Quelques points de repère

Nous vous proposons un petit glossaire sans prétention afin de mieux vous repérer dans le monde des data. Ce glossaire s'enrichira ensuite, sur le site d'ADELI, de vos contributions et références.

Bases de données

Ce terme peut être employé pour désigner n'importe quel ensemble de données, mais devrait être réservé aux ensembles de données interrogeables par le contenu, selon n'importe quel critère.

Définitions :
« Collection de données structurées reliées par des relations, interrogeable et modifiable par des langages de haut niveau » (Georges Gardarin)
Voir : http://georges.gardarin.free.fr/

« Ensemble structuré de données enregistrées sur des supports informatiques pour satisfaire simultanément plusieurs utilisateurs de façon sélective et en temps opportun » (Michel Adiba, Claude Delobel 1982)

Bases de données relationnelles

Les bases de données relationnelles ont occupé le terrain au point de laisser croire qu'il n'y avait pas d'autre modèle.
Les données y sont organisées sous forme de table et sont accessibles via des langages déclaratifs tels que SQL.
Les bases de données réseau (IDS, IDMS,..) et hiérarchiques ou arborescentes (IMS), qui utilisaient des systèmes de pointeur, les ont précédées.
Aujourd'hui le modèle relationnel a trouvé ses limites dans l'interrogation des données non structurées. On parle de bases « non SQL » ou « NoSQL ».

Voir : Serge Abiteboul — Informatique et sciences numériques — Cours du 14 mars 2012. Modèle relationnel
http://www.college-de-france.fr/site/serge-abiteboul/cours-du-14-mars-2012-model__1.htm

NoSQL Solutions

Des solutions NoSQL sont apparues afin de permettre le traitement de grands volumes de données non structurées (phénomène big data). Des performances extrêmes peuvent être obtenues en renonçant à certaines fonctionnalités des systèmes relationnels, par exemple la conformité à un schéma préexistant lorsque l'on enregistre ou que l'on analyse des données qui peuvent être non structurées ou partiellement structurées.
Google a ainsi développé BigTable pour l'indexation du web : il s'agit d'une base de données « orientée colonnes », qui permet l'extensibilité horizontale des tables. Parmi les plus connues, on peut également citer Hadoop.

Voir : http://nosql-database.org
http://www.college-de-france.fr/site/serge-abiteboul/cours-du-21-mars-2012-au-dela-du-modele-relationnel.htm

Big data

Les solutions « Big data » répondent aux besoins engendrés par l'explosion des volumes de données et plus particulièrement de traitement de données non structurées.
Les bases de données noSQL font partie de ces solutions.
Hadoop, souvent cité dans le contexte Big data est une plate-forme (framework) Java Open source de la fondation Apache qui permet de construire des architectures de traitement sur de très gros volumes de données.
Voir : http://hadoop.apache.org

Business Intelligence (BI)

La Business Intelligence ou « informatique décisionnelle » est une grande consommatrice des données non structurées du « Big data ».
Voir : http://www.piloter.org (site de M. Alain Fernandez)

Connaissances

Les données véhiculent de l'information qui porte de la connaissance.
Extraire la connaissance des monceaux de données qui nous submergent est aujourd'hui l'enjeu des technologies de l'information…
La gestion des connaissances (« KM » ou « Knowledge Management ») se préoccupe de la capitalisation et transmission des connaissances, qu'elles soient documentaires ou humaines.

Voir : http://www.apqc.org

Data

Le terme data vient du latin, pluriel de datum, participe passé neutre de dare (« donner ») et a été repris en langue anglaise pour désigner les données.
Les mots circulent d'une langue à l'autre et nous reviennent avec des différences subtiles.
En français, données et data prennent aujourd'hui des significations proches mais légèrement différentes.
La data, c'est l'information brute, non modélisée : texte, voix, image, non rattachée à un modèle de données particulier, mais qui peut potentiellement être intégrée à une multitude de modèles.
Contrairement à la donnée, la data préexiste au modèle de données et n'est pas forcément attachée à un objet du monde réel.

Exemple : un courriel a une existence propre en dehors de tout modèle de donnée.

Datacenter

On parlait autrefois de centre de calcul. On emploie plutôt aujourd'hui l'expression « Datacenter » ou « Centre de traitement de données » qui reflète mieux les préoccupations des DSI, en particulier le besoin d'accès permanent aux données.

Data journalisme

Le data journalisme est une évolution des pratiques du métier de journaliste permise par l'utilisation et la combinaison de data d'origine multiple (en particulier les données publiques issues de l'open data), leur analyse systématique et leur recoupement.

Voir : http://datajournalismhandbook.org

Datamart

Magasin de données : c'est un sous-ensemble spécialisé de l'entrepôt de données, ciblé sur un métier spécifique de l'entreprise.

Data mining

Data mining, ou fouille de données : extraction de connaissances à partir de données. Il s'agit de créer un modèle à partir de données en quantité suffisante. Le marketing utilise en particulier le data mining pour modéliser le comportement des consommateurs.

Voir : http://data.mining.free.fr/

Data warehouse

Entrepôt de données : Définition Wikipedia : « base de données regroupant l'ensemble des données fonctionnelles d'une entreprise.
Il entre dans le cadre de l'informatique décisionnelle ; son but est de fournir un ensemble de données servant de référence unique, utilisée pour la prise de décisions dans l'entreprise par le biais de statistiques et de rapports réalisés via des outils de reporting. ».
L'entrepôt de données est la source dans laquelle les outils de data mining vont puiser l'information.

Document

Dictionnaire de l'Académie française (9ème édition) : « tout objet pouvant apporter un renseignement, établir ou infirmer un fait ». La notion de document tend aujourd'hui à se réduire à celle de contenu.

Voir la Lettre d'ADELI n°57 Gestion de contenu

Donnée

La donnée est une information descriptive attachée à un objet du monde réel dont elle n'est qu'un attribut. La donnée est interprétable via le modèle de données.

Donnée publique

Définition Wikipedia : « La notion de donnée publique couvre l'ensemble des données qui sont ou devraient être (légalement ou volontairement) publiées ou tenues à disposition du public, et qui sont produites ou collectées par un état, une collectivité territoriale, un organe parapublic, dans le cadre de leurs activités de service public ».
Par principe, sauf exception pour raison de sécurité ou de protection des données personnelles, les données produites ou collectées par un état appartiennent à ses citoyens et devraient être publiques.

Données personnelles

Les données personnelles ou plus précisément « données à caractère personnel » sont des données associées ou associables à une personne physique identifiable.
Leur manipulation est réglementée en France par la loi Informatique et Libertés.

Voir : http://www.cnil.fr

Information

Extrait du glossaire ADELI (http://www.adeli.org/glossary/9/letteri) :
Ce qui est transmis : objet de connaissance, de mémoire.

Linked data

En français « Web des données » : notion introduite en 2006 par Tim Berners-Lee pour désigner la mise en relation des données pour constituer un réseau global sur le web, permettant le partage des données entre machines.
Le web des données repose sur un ensemble de standards permettant d'identifier les données (URI) et de décrire leurs associations (liens RDF).

Voir : « Linked Data : Evolving the Web into a Global Data Space »
http://linkeddatabook.com/editions/1.0/

MDM

Master Data Management ou Gestion des données de référence
Ensemble de méthodes, outils et processus permettant de constituer le référentiel des données de l'entreprise.

Voir : http://www.mdmalliancegroup.com/

Metadata

Les metadata ou métadonnées sont des éléments textuels descriptifs associés aux data ou aux données, afin de pouvoir les interroger.
Une donnée, au sens classique du terme, n'a pas besoin de métadonnée, sa place dans le modèle de données fournissant toutes les informations contextuelles.
La data, qui existe par contre hors contexte, a besoin de meta data pour être interrogée : c'est le cas par exemple des photos et des vidéos dont la recherche nécessite des éléments d'informations multiples : date, lieu, auteur, conditions de prise de vue, description du sujet, liste des personnes présentes sur un cliché…

Voir : http://www.w3.org/Metadata/

Modélisation

La modélisation des données permet de séparer la description des données et les langages de manipulation… Manipuler de grands volumes de données non structurées (data) et donc non modélisées a priori est le nouvel enjeu du Big data

Voir la Lettre d'ADELI n°69

Ontologie

Les ontologies sont des vocabulaires contrôlés et formalisés de termes, compréhensibles à la fois par les humains et les machines, couvrant généralement un domaine spécifique et partagés par une communauté d'utilisateurs. Leur intérêt est de permettre une expression logique de requêtes sur des sources de données ainsi que l'intégration de diverses sources de données.

Une ontologie est composée de classes et de relations entre classes organisées de façon hiérarchique en taxonomies.

Voir :
http://www.w3.org/standards/semanticweb/ontology

Open data

Donnée publique ouverte, librement accessible par tous, non soumise à copyright ni à protection d'aucune sorte. Plusieurs critères définissent la « véritable » donnée ouverte (voir article dans la Lettre n°86). Les données doivent être lisibles à la fois par un humain et par une machine, ce qui suppose la présence de métadonnées et d'identifiants.

Voir Lettre d'ADELI n°86

Ressource

Le terme ressource, utilisé dans URL (Uniform Resource Locator plus connu sous l'appellation « adresse Web ») ou URI (Uniform Ressource Identifier), désigne tout objet identifiable pouvant être atteint au travers du Web. Au départ cela désignait essentiellement un document ou un fichier, avant de s'étendre plus généralement à tout type de donnée pouvant être identifiée précisément sur le web.

Voir : http://www.w3.org/DesignIssues/HTTP-URI.html

RDF

« Resource Description Framework » : langage définissant le cadre général de la standardisation des métadonnées des ressources du Web.
RDF modélise les données sous forme de triplets « sujet-prédicat-objet » pouvant être représentés sous forme de graphes.
D'autres langages de description plus complexes ont été développés : RDFS, puis OWL qui étend RDF en permettant d'exprimer des contraintes complémentaires (classes disjointes,…).
SPARQL est un langage de requête pour RDF.

Voir : http://www.w3.org/RDF/

Web des données (Web of data)

Synonyme de « Linked data ».
Pour que le Web des Données soit une réalité, il est nécessaire que la quantité énorme de données présentes sur le Web soit disponible dans un format standard, accessible et gérable par des outils de Web sémantique.
Il est également nécessaire que les relations entre les données soient disponibles.

Voir W3C :
http://www.w3.org/standards/semanticweb/data

Web documentaire

Le « web documentaire », première forme historique du Web, permet la navigation entre documents via des liens hypertextes.

Web sémantique

Le « web sémantique », ou toile des connaissances, associe des annotations sémantiques aux ressources du web, sous forme de metadata.
Il s'appuie sur des standards en évolution permanente avec des syntaxes décrites par des ontologies.

Voir :
Serge Abiteboul – cours du 28 mars 2012
http://www.college-de-france.fr/site/serge-abiteboul/cours-du-28-mars-2012-.htm
et http://www.w3.org/standards/semanticweb/data
▲