Avant-propos

Les mots-clés

Avant de commencer la lecture de cet ouvrage, nous avons souhaité rappeler différentes définitions de certains mots-clés utilisés. Bien évidemment nous ne serons pas exhaustifs.

– *Analyse des données* : c'est une famille de méthodes statistiques permettant de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Certaines méthodes aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. D'autres techniques permettent de regrouper les données de façon à faire apparaître clairement ce qui les rend homogènes, et ainsi mieux les connaître.

– *Big Data* : le terme de Big Data est utilisé lorsque la quantité de données qu'une organisation doit gérer, atteint une taille critique qui nécessite de nouvelles approches technologiques pour leur stockage, leur traitement et leur utilisation. Volume, vitesse et variété sont souvent les trois critères qui permettent de qualifier une base de données de Big Data.

– *Cloud Computing* ou « l'informatique en nuage » : désigne un ensemble de processus qui consiste à utiliser la puissance de calcul et/ou de stockage de serveurs informatiques distants à travers un réseau, généralement Internet. C'est un modèle qui permet l'accès au réseau à la demande. Les ressources sont partagées et la puissance de calcul est configurable en fonction des besoins.

– *Data* ou « données » : sont constituées par les faits, les observations, les éléments bruts. Les données en elles-mêmes ont peu de signification si elles ne sont pas traitées.

– *Information* : elle consiste en données interprétées, porteuses de sens. Elle est contenue dans des descriptions, des réponses à des questions comme qui, quoi, quand et combien.

– *Connaissance* : c'est un savoir-faire qui rend possible la transformation d'une information en instructions. La connaissance peut être obtenue soit par le transfert de celui qui la détient à travers des instructions ou par son extraction d'après l'expérience.

– *Data Journalism* ou « journalisme de données » : désigne une nouvelle façon de faire de l'investigation journalistique en se basant sur l'analyse de données et (souvent) la représentation visuelle. Le journaliste utilise des bases de données comme sources et en déduit des connaissances, des corrélations ou des intuitions qui ne seraient pas accessibles par les méthodes traditionnelles de l'enquête journalistique. Même si l'article reste la composante de base, l'illustration des idées par l'image graphique, un schéma, une carte, etc., prend une place plus importante.

– *Data.gouv.fr* : site officiel servant de répertoire pour les données publiques du gouvernement français, qui a été mis en ligne le lundi 5 décembre 2011 par la Mission Etalab. En décembre 2013, data.gouv.fr a subi une profonde transformation, en changeant sa structure et la philosophie de son site. Elle est en effet devenue une plateforme collaborative orientée vers la communauté, au bénéfice d'une meilleure réutilisation des données publiques.

– *Data Mining* : ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Le Data Mining utilise un ensemble d'algorithmes issus de disciplines diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles à l'entreprise. Il regroupe l'ensemble des technologies susceptibles d'analyser les informations d'une base de données pour y trouver des informations utiles et d'éventuelles corrélations signifiantes et utilisables entre les données.

– *Data Visualisation* : aussi nommée « DataViz », il s'agit de technologies, méthodes et outils de visualisation des données. Elle peut se concrétiser par des graphiques, des camemberts, des diagrammes, des cartographies, des chronologies, des infographies, ou même des créations graphiques inédites. La présentation sous une forme illustrée rend les données plus lisibles et compréhensibles.

– *Etalab* : mission proposée dans le rapport Riester de novembre 2010, mise en place en 2011, chargée de mettre en œuvre la politique d'ouverture des données de l'administration française, et de mettre en place un annuaire des données publiques françaises data.gouv.fr.

– *La gouvernance des données* : constitue un cadre de contrôle qualité pour la gestion et la protection de ressources d'information-clé à travers l'entreprise. Sa mission est de veiller à ce que les données soient gérées dans le respect des valeurs et des convictions de l'entreprise, de valider leur qualité et de mettre en œuvre des processus de suivi et de maintien de cette qualité. La gouvernance des données recouvre les processus de gestion des données, les contrôles, les bilans visant à améliorer la qualité, la cohérence, l'intégrité et la sécurité des ressources d'information d'une entreprise.

– *Hadoop* : infrastructure logicielle pour application Big Data qui inclut un système de stockage et un outil d'exécution parallèle d'applications.

– *Informations structurées* : se trouvent, par exemple, dans les bases de données ou encore dans les langages informatiques. Ainsi, on les reconnaît au fait qu'elles sont disposées de façon à être traitées automatiquement et efficacement par un logiciel, mais non nécessairement par un humain. D'après Alain Garnier, l'auteur du livre L'information non structurée dans l'entreprise, « une information est structurée lorsqu'elle est répétable, systématique et calculable ». Il peut s'agir de formulaires, de factures, de fiches de paie, de libellés…

– *Informations non structurées* : par opposition à la catégorie précédente, les informations non structurées représentent l'ensemble des informations pour lesquelles il est impossible de retrouver une structure prédéfinie. Elles sont toujours destinées à des humains et il s'agit donc essentiellement de documents textes et multimédias, comme des lettres, des livres, des rapports, des collections d'images ou de vidéos, des brevets, des images satellites, des offres de services, des CV, des appels d'offres… Et la liste est encore longue.

– *Informations semi-structurées* : il est à noter que la frontière entre informations structurées et informations non structurées demeure assez floue et qu'il n'est pas toujours aisé de classer un document dans l'une ou l'autre des catégories. Dans ce cas précis, vous avez sans doute affaire à de l'information semi-structurée.

– *Innovation* : reconnue comme source de croissance et de compétitivité. Le Manuel d'Oslo définit quatre types d'innovations :

- *l'innovation de produit* : l'introduction d'un bien ou d'un service nouveau. Cette définition inclut les améliorations sensibles des spécifications techniques, des composants et des matières, du logiciel intégré, de la convivialité ou autres caractéristiques fonctionnelles ;

- *l'innovation de procédé* : la mise en œuvre d'une méthode de production ou de distribution nouvelle ou sensiblement améliorée. Cette notion implique des changements significatifs dans les techniques, le matériel et/ou le logiciel ;

- *l'innovation de commercialisation* : la mise en œuvre d'une nouvelle méthode de commercialisation impliquant des changements significatifs de la conception ou du conditionnement, du placement, de la promotion ou de la tarification d'un produit ;

- *l'innovation d'organisation* : la mise en œuvre d'une nouvelle méthode organisationnelle dans les pratiques, l'organisation du lieu de travail ou les relations extérieures de la firme.

– *Innovation ouverte* ou « Open Innovation » : est définie comme l'utilisation accrue, en amont, de sources d'informations et de connaissances externes à l'entreprise, et la multiplication, en aval, des canaux de commercialisation des actifs immatériels de celle-ci dans le but d'accélérer l'innovation.

– *Intelligence économique*, « IE » : est l'ensemble des activités coordonnées de collecte, de traitement et de diffusion de l'information utile aux acteurs économiques, en vue de son exploitation. Selon le rapport Marte, l'intelligence économique peut être définie comme l'ensemble des actions coordonnées de recherche, de traitement et de distribution, en vue de son exploitation, de l'information utile aux acteurs économiques. Ces diverses actions sont menées légalement avec toutes les garanties de protection nécessaires à la préservation du patrimoine de l'entreprise, dans les meilleures conditions de qualité, de délais et de coûts. L'information utile est celle dont ont besoin les différents niveaux de décision de l'entreprise ou de la collectivité, pour élaborer et mettre en œuvre de façon cohérente la stratégie et les tactiques nécessaires à l'atteinte des objectifs définis par l'entreprise dans le but d'améliorer sa position dans son environnement concurrentiel. Ces actions, au sein de l'entreprise, s'ordonnent autour d'un cycle ininterrompu, générateur d'une vision partagée des objectifs de l'entreprise.

– *Interopérabilité* : désigne la capacité que possède un produit ou un système dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs, et ce, sans restriction d'accès ou de mise en œuvre.

– *Jeu de données* ou « Dataset » : collection structurée et documentée de données sur laquelle s'appuient les réutilisateurs.

– *Linked Open Data*, « LOD » : désigne une approche du web poussée par les tenants du « web sémantique » qui décrit toutes les données d'une manière lisible par les ordinateurs, et à les lier entre elles en décrivant leurs relations, ou encore, en facilitant leur mise en relation. Données publiques ouvertes dans un format de type « web sémantique », où les entités ont un identificateur unique et les jeux de données sont liés entre eux par ces identificateurs.

– *Open Knowledge Foundation Network* : association britannique à but non lucratif œuvrant pour l'ouverture des données, elle a notamment développé CKan.

– *Open Data* ou « ouverture des données » : principe selon lequel les données publiques (celles recueillies, maintenues et utilisées par les organismes publics) doivent être disponibles pour accès et réutilisation par les citoyens et les entreprises.

– *Réutilisation des données* : prendre un jeu de données pour le visualiser, le fusionner avec d'autres jeux, l'utiliser dans une application, le modifier, le corriger, le commenter, etc.

– *Science des données* ou « Data Science » : est une nouvelle discipline qui comprend des éléments de mathématiques, de statistiques, d'informatique et de visualisation des données. L'objectif est d'extraire des informations de sources de données. En ce sens la Data Science s'intéresse à l'exploration et à l'analyse de bases de données. Cette discipline a reçu beaucoup d'attention dernièrement grâce à l'intérêt grandissant pour le Big Data.

– *Smart Data* ou « données intelligentes » : le déluge de données auxquelles seront confrontés les citoyens et le monde économique va entraîner des bouleversements de comportements, le développement de nouveaux services et la création de valeur. Ces données doivent être traitées et valorisées pour devenir « Smart Data ». Les données intelligentes représentent le résultat d'un travail d'analyse et d'interprétation des données brutes qui va permettre de retirer effectivement quelque valeur. L'important est de donc de savoir partir des données existantes pour créer de la valeur.

– *Text Mining* : est une technique permettant d'automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière statistique les différents sujets évoqués.

– *Tim Berners-Lee* : co-inventeur du web, inventeur du web sémantique, très actif et impliqué dans data.gov.uk : il a notamment défini la notation en cinq étoiles pour mesurer le niveau d'ouverture du web sémantique pour une mise en ligne de jeu de données.

– *Web 1.0* : est Internet permettant d'accéder à des sites constitués de pages web liées entre elles par des hyperliens. Ce web a été créé au début des années 1990. C'est une relation entre un site éditeur qui publie un contenu ou des services, et des internautes qui le visitent, et qui surfent ainsi de site en site.

– *Web 2.0* : désigne l'ensemble des techniques, des fonctionnalités et des usages du *World Wide Web* qui ont suivi la forme originelle du web. Elle concerne en particulier les interfaces permettant aux internautes ayant peu de connaissances techniques de s'approprier de nouvelles fonctionnalités du web. Les internautes peuvent d'une part contribuer à l'échange d'informations et interagir (partager, échanger, etc.) de façon simple.

– *Web 3.0* : (aussi appelé web sémantique) est un réseau de données qui permet à des machines de comprendre la sémantique c'est-à-dire le sens de l'information publiée sur le web. Il étend le réseau de pages web compréhensibles par l'humain en y ajoutant des métadonnées qui sont compréhensibles par la machine et qui créent des liens entre les contenus des différentes pages, ce qui permet à des agents automatiques d'accéder au web de façon plus intelligente et d'effectuer des tâches à la place des utilisateurs.

– *Web sémantique* : désigne un ensemble de technologies visant à rendre l'ensemble des ressources du web accessibles, intelligibles et utilisables par des programmes et agents logiciels, grâce à un système de métadonnées. Les machines pourront traiter, relier et combiner automatiquement un certain nombre de données. Le web sémantique représente un ensemble de standards développés et promus par le W3C pour permettre la représentation et la manipulation de connaissances par les outils du web (navigateurs, moteurs de recherche ou agent dédiés). Parmi les plus importants, on peut citer :

- *RDF* : modèle conceptuel permettant de décrire toute donnée sous forme de graphe, afin de constituer des bases de connaissances ;

- *RDF Schéma* : langage permettant de créer des vocabulaires : ensemble de termes utilisés pour décrire des choses ;

- *OWL* : langage permettant de créer des ontologies, vocabulaires plus complexes servant de support aux traitements logiques (inférences, classification automatique…) ;

- *SPARQL* : langage de requêtes pour obtenir des informations à partir de graphes RDF.