GTnum IMS DEFI #DEFI – Glossaire de la formation et de l'éducation aux données - Groupes thématiques numériques de la Direction du numérique pour l'éducation (Ministère de l'Éducation nationale et de la Jeunesse) 2021-2024

Camille Capelle, Anne Lehmans, Marie Chagnoux, Aude Seurrat, Sarah Labelle

MINISTÈRE
DE L'ÉDUCATION
NATIONALE
Liberté
Égalité
Fraternité

Direction du numérique
pour l'éducation

THÉMATIQUE
Littératie des données : évolution des compétences, enjeux éducatifs, pédagogiques et éthiques, perspectives pour la formation

Glossaire de la formation et de l'éducation aux données

#GTnum DEFI
2021-2024

GLOSSAIRE DE LA FORMATION ET DE L'EDUCATION AUX DONNEES

Les auteurs / autrices
Camille Capelle, Université Bordeaux – INSPE, laboratoire IMS
Anne Lehmans, Université Bordeaux – INSPE, laboratoire IMS
Marie Chagnoux, Université Paris 8, laboratoire CEMTI
Aude Seurrat, Université Paris Créteil - INSPE, laboratoire CEDITEC, GIS2IF
Sarah Labelle, Université Paul Valéry, LERASS

Projet piloté par
Camille Capelle, Université Bordeaux – INSPE, laboratoire IMS

Ces travaux sont publiés dans le cadre des groupes thématiques numériques soutenus par la Direction du numérique pour l'éducation.

• Eduscol https://eduscol.education.fr/2174/enseigner-et-apprendre-avec-la-recherche-les-groupes-thematiques-numeriques-gtnum
• Carnet Hypothèses « Éducation, numérique et recherche » https://edunumrech.hypotheses.org/

Décembre 2024

SOMMAIRE

Sommaire 3

Avant-Propos 5

Base de données 6

Carnet de bord numérique (notebook) 6

Cartographie de données 6

Culture des données 7

Cycle de vie de la donnée 7

Démocratisation des données 8

Donnée (data) 8

Données massives (Big data, ou mégadonnées) 9

Données ouvertes (Open Data) 9

Données personnelles 9

Données publiques 10

Donnée d'entrée 10

Données d'entraînement 10

Données de l'éducation 11

Données sensibles 11

Education aux données 11

FAIR 12

Fouille de données (data mining) 12

Gouvernance des données 12

Infographie 13

Intelligence artificielle (IA) 13

Interopérabilité des données 13

Jeu de données 14

Littératie des données 14

Médiation des données 14

Métadonnées 15
Métiers de la donnée 15
Enquêtes en sources ouvertes (Open Source Intelligence - OSINT) 16
Physicalisation des données 16
Pilotage par les données (data driven strategy) 16
Protection de la vie privée dès la conception (privacy by design) 17
Règlement Général sur la Protection des données (RGPD) 17
Souveraineté des données 18
Traces numériques 18
Visualisation de données (Datavisualisation ou dataviz) 19
Références 20

AVANT-PROPOS

Ce glossaire recense différents termes et notions avec lesquels nous avons été amenés à travailler dans le cadre du GTnum DEFI (Données pour l'Education, la Formation, l'Innovation). Pour certains, il s'agit de termes employés par les acteurs que nous avons enquêtés sur le terrain (ex : pilotage par les données, souveraineté des données). Pour d'autres, il s'agit de notions que nous avons nous-mêmes choisi d'employer pour caractériser nos actions de formation et de recherche (ex : cartographie de données, gouvernance des données, physicalisation des données, visualisation de données…). Chacune renvoie à des définitions d'abord vulgarisées, puis issues de la littérature scientifique ou professionnelle. Le document comporte une bibliographie de référence associée aux différentes définitions.

BASE DE DONNEES

Une base de données est un ensemble volumineux de données, structuré de façon logique, stocké sur des supports numériques, conçu pour répondre aux besoins d'une ou plusieurs applications, interrogeable et modifiable par un ou plusieurs utilisateurs.

> Ensemble de données organisé en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes. (Enrichissement du vocabulaire de l'informatique, 1981)

CARNET DE BORD NUMERIQUE (NOTEBOOK)

Les calepins électroniques, plus communément désignés sous le terme de *notebooks,* sont des interfaces interactives qui combinent des sections de code à des sections en langage naturel pour documenter ce code et partager les données utilisées. Très répandus en science des données, ils ont transformé la manière dont les scientifiques partagent largement leurs approches en publiant leurs idées combinées à du code exécutable et ils ont aujourd'hui intégré le monde éducatif où plusieurs initiatives visent à la mise en place de notebooks pédagogiques adaptés aux élèves et aux étudiants.

> Les notebooks éducatifs embarquent une activité pédagogique contenant des instructions textuelles guidant les étudiants à travers les différentes tâches à réaliser. Ensuite, l'enseignant cherche à reproduire les résultats des étudiants en suivant un ordre le plus souvent linéaire. (Casseau, 2024)

CARTOGRAPHIE DE DONNEES

Technique qui consiste à représenter graphiquement des données numériques et leurs relations sur une carte de manière à en obtenir une visualisation compréhensible.

> Historiquement, la cartographie est une discipline scientifique et technique exercée par des spécialistes formés pour cela. Les cartes topographiques ont ainsi connu un essor important au cours des xviie et xviiie siècles, à l'image des quatre générations de Cassini qui ont travaillé à l'amélioration du relevé topographique du territoire français, et en ont proposé en 1815 une carte exhaustive. Au xixe siècle, les spécialistes de la cartographie relèvent essentiellement de l'Armée et des corps de géomètre. Les applications civiles en dehors du cadastre demeurent assez faibles. Elles concernent principalement les cartes administratives et les plans de ville. Les usages de la carte se diversifient tout au long du xxe

> siècle avec l'apparition des cartes routières et touristiques à l'attention des automobilistes et les représentations cartographiques en urbanisme qui relèvent d'une activité de mise en forme graphique de l'information plus que de levers de carte sur le terrain. (Joliveau et al., 2013)

> La cartographie, auparavant associée à un savoir-faire propre aux représentations graphiques de portions de l'espace terrestre, se dissout désormais dans la visualisation – ce que ne perçoivent évidemment pas les utilisateurs de logiciels d'itinéraire en ligne, ces gros consommateurs de cartes numériques. (Guichard, 2019)

CULTURE DES DONNEES

Ensemble de connaissances, de compétences et de pratiques liées à la fabrique, la collecte, l'analyse et/ou l'utilisation de données pour s'informer et agir au quotidien de façon critique et en comprenant les enjeux des données. Le terme est parfois considéré comme une traduction française de la notion de "littératie des données".

> C'est véritablement un ensemble de représentations, un univers de sens, une capacité à se projeter, un sentiment d'appartenance qui sont en jeu, et pas seulement la capacité à manipuler des outils informatiques. Cette culture se cristallise et se transforme autour des données, comme élément de base de l'information. (Lehmans, 2017)

> L'expression anglaise « data culture » a émergé dès les années 1960 avec le boom de l'industrie informatique. Tombée en désuétude dans les années 1980 et 1990, elle tend alors à être supplantée par celle d'« information culture ». L'expression a suscité nouvel intérêt dans les années 2010 en même temps que la croissance exponentielle du phénomène de « mise en données » (datafication) de la société, avec en corollaire l'émergence de nouveaux enjeux liés à la « littératie des données ». L'étude des cultures de données avoisine ces autres notions tout en s'en distinguant. (Casemajor, 2024)

> « Assemblage de répertoires collectifs de représentation, d'action et de justification situé dans un contexte spécifique et caractérisé par une sensibilité, un univers de sens et une rationalité construits en relation avec des ensembles de données. » (Casemajor, 2024)

CYCLE DE VIE DE LA DONNEE

Le cycle de vie d'une donnée représente toutes les étapes qu'une donnée traverse depuis sa création jusqu'à sa destruction, en passant par le stockage, la gestion, l'utilisation, le partage, l'archivage.

> Le cycle de vie des données de recherche (Research data lifecycle) décrit le processus d'utilisation des données de leur création à la publication et à leur réutilisation ultérieure. Ce modèle définit les six étapes suivantes :
>
> Création ou collecte des données (creating data) ;
>
> Traitement des données (processing data) ;
>
> Analyse des données (analysing data) ;
>
> Conservation des données (preserving data) ;
>
> Accès aux données (giving access to data / data discovery) ;
>
> Réutilisation des données (reusing data). (Inist, inspiré de UK Data archive)

DEMOCRATISATION DES DONNEES

La démocratisation des données est un processus qui consiste à utiliser des données pour en faire un bien commun.

> La démocratie des données fait référence à la réflexion sur où, comment et avec qui les ingénieurs et autres participants d'une organisation génèrent, utilisent et traduisent les données pour le bien social. (Cutts et al. 2024)

> La démocratisation des données est un processus de transition d'une politique autoritaire à un régime représentatif (Kauffman, 2018)." (...) Dans une organisation, cela fait référence à "un paradigme selon lequel tous les salariés ont accès aux données" (...). "Ces consommateurs de données comprennent des experts en données traditionnels (par exemple, data scientist, gestionnaires de données ou architectes de données), mais aussi tous les employés (parfois appelés des citoyens des données), qui n'ont peut-être même jamais travaillé avec des données auparavant. (Labadie et al., 2020)

DONNEE (DATA)

Représentation d'un fait ou d'une caractéristique d'un objet, d'une personne, d'un lieu ou d'un événement. Elle peut être exprimée sous différentes formes, telles que des chiffres, du texte, des images, des sons ou des vidéos. En informatique, la donnée est une information présentée sous une forme traitable par une machine. En sciences de l'information et de la communication, la donnée représente l'unité de base qui permet de construire une information quand on y ajoute du sens.

> Toute représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement. (Enrichissement du vocabulaire de l'informatique, 1981)

> Élément (fait, chiffre, etc.) qui est une information de base sur laquelle peuvent s'appuyer des décisions, des raisonnements, des recherches et qui est traité par l'humain avec ou sans l'aide de l'informatique. (Office Québécois de la Langue Française, 2024)

> Une information-objet, considérée comme un élément informatif, un contenu formaté, qui peut être traité par une machine. (Buckland, 1991)

> Une donnée numérique est la description élémentaire, représentée sous forme codée, d'une réalité (chose, évènement, mesure, transaction, etc.) en vue d'être collectée, enregistrée, traitée, manipulée, transformée, conservée, archivée, échangée, diffusée, communiquée. (Open Data France)

> La donnée peut être définie comme la plus petite représentation conventionnelle et fondamentale d'une information (fait, notion, objet, nom propre, chiffre, statistique, etc.) sous une forme analogique ou digitale permettant d'en effectuer le traitement manuel ou automatique (informatique). (Rousseau et Couture, 1994)

> Mot, nombre, signal, chaîne de caractères, séquence de bits, morceau de matière ou tout autre élément brut enregistré dans un système d'information où il pourra être corrélé à d'autres objets et interprété pour constituer une information. (Chabin, 2010)

DONNEES MASSIVES (BIG DATA, OU MEGADONNEES)

Grands ensembles des données liés à la numérisation des activités humaines, caractérisés par le grand volume, la vitesse de circulation et la variété, nécessitant des outils, techniques, méthodes visant à stocker, analyser, traiter, éditorialiser les données.

> Les données du big data manquent d'une définition sérieuse. (...) Des chercheurs européens décèlent en arrière plan de ce terme, une alliance de Big Brother (datafication, dataisme, dataveillance) et de Big Business (measure, manipulate, monetize) : les citoyens consommateurs ont cédé une partie de leur intimité en échange de sécurité et de gratuité. Cette alliance est exploitée par des experts, dont on observe qu'ils passent sans difficulté du monde académique, de l'industrie et au renseignement. (Delort, 2018)

> L'accélération des processus de calcul et le déluge de nouvelles données que favorise la digitalisation progressive de toutes les traces de la vie quotidienne se déploient dans tous les domaines : prolifération des sources d'information, digitalisation du dossier médical et des informations personnelles, modèles probabilistes de l'assurance, développement des outils de data mining dans la relation client, capture et suivi des traces de mobilité, de communication ou de navigation. (Cardon, 2012)

DONNEES OUVERTES (OPEN DATA)

Les données ouvertes sont des données numériques dont l'accès et l'utilisation sont libres, accessibles et utilisables, pour tous. Elles peuvent provenir de sources diverses, mais sont souvent produites par des organismes publics (gouvernements, collectivités, etc.). L'ouverture des données nécessite une action politique et une organisation des services chargés de leur collecte et de leur diffusion (gouvernance des données).

> Données librement accessibles, mises à disposition dans un format ouvert et réutilisable par toute personne. (CNIL, 2019)

> Données collectées par des organismes publics ou privés chargés d'un service public, ou les citoyens, et mises à disposition en format numérique sur des plateformes nationales ou locales permettant leur accès et leur réutilisation. (Lehmans, 2018)

DONNEES PERSONNELLES

Une donnée personnelle est toute information qui permet d'identifier directement ou indirectement une personne physique. La protection des données personnelles contre les risques liées à leur stockage et leur utilisation dans des bases de données informatiques fait

l’objet d’une législation depuis la loi Informatique et libertés du 6 janvier 1978, qui a créé la Commission nationale de l’informatique et des libertés (CNIL).

> Une « donnée personnelle » est « toute information se rapportant à une personne physique identifiée ou identifiable ». Une personne peut être identifiée directement (exemple : nom, prénom) ou indirectement (exemple : par un identifiant (n° client), un numéro (de téléphone), une donnée biométrique, plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale, mais aussi la voix ou l’image). (CNIL)

DONNEES PUBLIQUES

Les données publiques désignent l'ensemble des informations collectées, produites ou conservées par les administrations publiques (État, collectivités territoriales, etc.) dans le cadre de leurs missions de service public.

> Les données publiques concernent les informations poursuivant un objectif d’intérêt général. Autrement dit, celles produites par les collectivités territoriales, l’État ou toutes autres institutions publiques. (Open data soft, 2024)

DONNEE D’ENTREE

Une donnée d'entrée est une information qui est fournie à un système, un programme ou un processus pour qu'il puisse effectuer un traitement. La qualité et la pertinence des données d'entrée déterminent en grande partie la qualité des résultats obtenus. Des données erronées, incomplètes ou mal formatées peuvent conduire à des résultats inexacts ou aberrants.

> Dans le domaine de l’intelligence artificielle, une donnée d’entrée est une donnée utilisée pour l’apprentissage automatique ou la prise de décision du système d’IA (en phase de production). (CNIL)

DONNEES D’ENTRAINEMENT

Les données d'entraînement sont l'ensemble des données que l'on fournit à un modèle d'intelligence artificielle pour qu'il apprenne et effectue des tâches spécifiques.

> Jeu de données (texte, sons, images, listes, etc.) utilisé lors de la phase d'entraînement / d’apprentissage : le système s’entraîne sur ces données pour effectuer la tâche attendue de lui. (CNIL)

DONNEES DE L'EDUCATION

Les données d'éducation regroupent l'ensemble des informations numériques collectées et produites par et pour le système éducatif. Elles concernent aussi bien les élèves, les enseignants, les établissements scolaires et leurs activités.

> Les données de l'éducation peuvent être définies comme les données liées à la vie scolaire (depuis l'administration jusqu'à la pédagogie) et concernant les acteurs de la communauté éducative. Elles sont produites, stockées et analysées pour des objectifs spécifiques, exploitées à des fins de suivi pédagogique des élèves, d'organisation et de pilotage du service public éducatif, d'élaboration de ressources pédagogiques ainsi que de statistiques d'évaluation et de recherches (Atal et Froidevaux, 2020). Elles concernent, en premier lieu, les élèves avec des données administratives et de scolarité (inscriptions, emplois du temps, absences, retards, redoublements, notes…) et leurs familles, par exemple avec les données sociales. Il s'agit également des données pédagogiques produites par les enseignants dans les activités d'enseignement (contenus et supports de cours, évaluations…). Il existe également des « données d'interactions » considérées comme fondamentales par le Comité d'éthique pour les données d'éducation : celles-ci sont produites dans les interactions entre les élèves et leurs enseignants à travers des documents ou sur les dispositifs numériques utilisés au cours de travaux collaboratifs entre les élèves, par exemple lors d'échanges sur l'évaluation d'un travail, ou encore dans les messages rédigés par un enseignant à destination d'une famille. (Lehmans et Capelle, 2023)

DONNEES SENSIBLES

Une donnée sensible est une donnée personnelle qui, par sa nature, révèle des informations particulièrement sensibles sur une personne. Ce sont des informations qui, si elles étaient divulguées, pourraient porter atteinte à la vie privée de la personne concernée ou lui causer un préjudice. Elles font l'objet d'une protection renforcée.

> Les données sensibles forment une catégorie particulière des données personnelles. Ce sont des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique. (CNIL, https://www.cnil.fr/fr/definition/donnee-sensible)

EDUCATION AUX DONNEES

L'éducation aux données correspond à une activité pédagogique visant la compréhension critique et éthique des données, qui ne se résume pas à l'enseignement de compétences techniques. Elle repose sur le constat qu'il est crucial de pouvoir identifier les leviers et les biais potentiels dans les usages des données, de comprendre les implications éthiques de leur utilisation et de prendre des décisions éclairées. Elle vise également à rendre les élèves capables d'utiliser voire de produire de façon autonome des données pour leur vie personnelle et professionnelle future.

> L'enseignement traditionnel de la statistique s'est principalement concentré sur des problèmes décontextualisés et des preuves mathématiques, plutôt que sur le raisonnement réel avec les données comme le font les statisticiens. La plupart des ensembles de données utilisés dans ces cours de statistiques sont petits et « propres », souvent construits pour illustrer une notion en mathématique en particulier. L'émergence et la visibilité récentes de la science des données ont permis d'attirer l'attention sur des aspects du travail avec des données qui ont été pour la plupart ignorés dans les cours de statistique. De grandes quantités de données sont désormais accessibles au grand public, ce qui accroît la possibilité – et le besoin – pour les non-statisticiens de s'engager dans un raisonnement basé sur les données. (Rubin, 2020)

> Construction d'une culture de la donnée à travers, d'une part la production de connaissances au moyen de la documentarisation des activités réalisées à partir de données, d'autre part, la compréhension des enjeux éthiques et politiques liés aux usages des données dans les représentations de ce qui est acceptable ou pas. (Lehmans et Liquète, 2022)

FAIR

FAIR (Findable, Accessible, Interoperable, Reusable) désigne un ensemble de principes appliqués dans le domaine de la gestion des données scientifiques et de la science ouverte. La retrouvabilité, l'accessibilité, l'interopérabilité et la réutilisabilité à long terme sont des principes qui permettent de s'assurer de la possibilité d'utiliser les données scientifiques dans plusieurs contextes différents. Ces principes sont également largement utilisés dans le cadre des données ouvertes.

> Les principes FAIR sont un ensemble de principes directeurs élaborés pour gérer les données de la recherche visant à les rendre faciles à trouver (Findable), accessibles (Accessible), interopérables (Interoperable) et réutilisables (Re-usable) par des opérateurs mais aussi par des machines. Ces bonnes pratiques en matière de gestion de données sont suffisamment générales pour être valables dans tous les domaines. (Quimbert et al. 2022)

FOUILLE DE DONNEES (DATA MINING)

La fouille de données, également dénommée exploration de données ou forage de données, est un ensemble de techniques qui visent à extraire des informations pertinentes à partir de grandes quantités de données.

> Processus de recherche et d'analyse qui permet de trouver des corrélations cachées ou des informations nouvelles, ou encore, de dégager certaines tendances. (Office Québécois de la Langue Française, 2024)

GOUVERNANCE DES DONNEES

La gouvernance des données est un ensemble de processus, de structures et de pratiques mis en place pour gérer efficacement les données, depuis leur collecte jusqu'à leur mise à

disposition, en passant par le stockage et l'organisation. Elle s'appuie sur des règles d'organisation politique, administrative et technique des différents services de l'organisation en charge des données.

> Dans le rapport remis par le Secrétariat général pour la modernisation de l'action publique en 2015, la gouvernance de la donnée désigne « l'ensemble de principes et de pratiques qui visent à assurer la meilleure exploitation du potentiel des données. (Lehmans, 2018)
>
> Exercice de l'autorité et du contrôle sur la gestion des données par l'institution d'un système de normes et de procédures (Plotkin, 2013). (Verdi et al. 2024)

INFOGRAPHIE

Souvent utilisé de manière interchangeable avec le terme "datavisualisation", une infographie désigne une modalité de communication graphique d'un message à l'aide de données ou d'informations alors que la datavisualisation se concentre sur la représentation précise des données.

> Un objet de communication construisant un discours et rassemblant en ce but des graphiques, des diagrammes, des textes, des visuels iconiques ou photographiques. Le propos d'une infographie est de rendre intelligibles le plus immédiatement possible des données complexes aussi bien quantitatives que qualitatives en les recontextualisant… des visuels plurimodaux construits par un infographiste qui joue son rôle de transformateur et de passeur d'information. (Pérès, 2016)

INTELLIGENCE ARTIFICIELLE (IA)

L'intelligence artificielle est un domaine de l'informatique qui vise à créer des machines capables de simuler certaines fonctions de l'intelligence humaine. Ces machines sont conçues pour apprendre, raisonner, prendre des décisions et s'adapter à de nouvelles situations, en s'appuyant sur de vastes quantités de données et sur des algorithmes pour les traiter à l'aide de modèles.

> L'intelligence artificielle repose sur l'utilisation d'algorithmes et l'exploitation de corpus de données massives avec l'application de règles d'apprentissage. Elle permet ainsi de réaliser, selon un certain rendement établi en fonction de degrés de précision attendus et du temps d'apprentissage, des traitements dits « intelligents ». (Raulin, 2022)

INTEROPERABILITE DES DONNEES

L'interopérabilité des données désigne la propriété de différents systèmes informatiques d'être susceptibles d'échanger des informations et des données de manière fluide et sans entrave.

Elle repose sur des processus de standardisation des outils, des protocoles de communication et des règles sémantiques.

> Différents systèmes logiciels numériques travaillant tous ensemble automatiquement sans qu'il soit nécessaire de recourir à un codage personnalisé ou à des processus manuels compliqués pour importer des données d'un système à l'autre (Ben Henda, 2019).

JEU DE DONNEES

Un jeu de données est un ensemble organisé de valeurs, souvent présentées sous forme de tableau. Chaque colonne représente une variable (comme l'âge, le sexe, le revenu) et chaque ligne représente une observation (une personne, un produit, un événement). Les jeux de données sont mis à disposition et utilisés pour produire des informations à partir des données. Ils doivent être organisés pour être utilisables le plus efficacement possible.

> Un jeu de données, ou dataset, regroupe plusieurs données ayant un lien cohérent entre elles. Il se présente sous forme de tableau permettant d'analyser chaque donnée qui le compose. Chaque donnée peut être composée de texte, de chiffres, de coordonnées géographiques ou encore d'éléments multimédia (par exemple une image ou une vidéo). (Open Data Soft)

LITTERATIE DES DONNEES

La littératie des données désigne la capacité à trouver, collecter, organiser, analyser, interpréter et communiquer des données de manière efficace et critique.

> Capacité à comprendre les enjeux de la production, de l'organisation et de l'exploitation des données, et à les utiliser efficacement et de manière critique et créative. Elle fait partie des compétences nécessaires pour évaluer et utiliser l'information. (Schield, 2004)

> La littératie des données inclut la capacité à lire, travailler, analyser, et argumenter, avec les données dans le cadre d'un processus plus large d'enquête sur le monde. (D'Ignazio & Bhargava 2016 ; Letouzé et al. 2015, cité par D'ignazio, 2017).

> Capacité d'accéder, d'interpréter, d'évaluer, de gérer, de manipuler et d'utiliser de manière critique et éthique les données (Calzada-Prado et Marzal, 2013) grâce à une connaissance des caractéristiques de la donnée, de son cycle de vie et des différents impacts engendrés par son usage, notamment en termes de sécurité et de protection de la vie privée (Crusoe, 2016). (Verdi, 2023)

MEDIATION DES DONNEES

La médiation des données est un processus d'intervention permettant de mettre en relation les données avec leurs utilisateurs, visant à garantir que les données sont accessibles,

compréhensibles et exploitables. Elle peut être humaine (par l'accompagnement de médiateurs par exemple), technique (par la mise à disposition d'outils facilitant les usages des données) ou documentaire (par l'organisation et l'indexation des données). Les médiateurs de données agissent comme des intermédiaires entre les données brutes et les personnes susceptibles de les utiliser. La médiation peut porter sur des méthodes, des outils (de visualisation, de traitement…) ou des documents.

> Les médiations offrent un cadre qui permet à des publics de s'impliquer, d'accéder à des données, de les traiter. (Labelle, 2023)

METADONNEES

Données sur les données, les métadonnées sont des informations structurées et normalisées qui décrivent, expliquent, localisent ou facilitent l'exploitation et la gestion des ressources d'information. Elles fonctionnent comme des étiquettes (labels) qui donnent des indications sur le contexte de fichiers, photos, vidéos ou tout autre type de contenu numérique. On trouve trois types de métadonnées : descriptives (titre, auteur, sujet), administratives (format de fichier, droits d'accès), structurelles (concernant l'organisation de la base de données). En informatique, elles permettent de structurer le traitement des données dans les systèmes.

> Les métadonnées déterminent le cadre d'usage et d'échange (d'un document numérique) à travers les communautés d'utilisateurs et les réseaux de spécialistes. (...) Elles permettent d'identifier et de décrire les diverses ressources numériques d'une manière lisible et compréhensible à la fois par les machines et les humains. (Ben Henda, 2006)

METIERS DE LA DONNEE

Les métiers de la donnée regroupent un large ensemble de professions liées à toute la chaîne de traitement qui se base sur les données. Cela peut concerner la collecte, le nettoyage et l'enrichissement, l'analyse et la valorisation, mais aussi la mise en oeuvre des solutions expertes pour la prise de décisions et l'optimisation de processus ou encore la responsabilité éthique de leur utilisation.

> Ce que semble souvent oublier le débat sur les nouvelles données numériques, c'est le travail de la donnée et le rôle qu'y jouent les travailleurs de la donnée, codeurs, statisticiens, modélisateurs, designers d'algorithmes et l'ensemble des métiers, dont ceux des sciences sociales, qui se donnent pour tâche d'en extraire de la signification. (Bastard et al., 2014)

ENQUETES EN SOURCES OUVERTES (OPEN SOURCE INTELLIGENCE - OSINT)

L'OSINT est une méthode de collecte et d'analyse d'informations à partir de sources publiques et librement accessibles. Cette méthode est utilisée par les journalistes, mais aussi les professionnels de la cybersécurité informatique et des militants qui ont besoin de mener des enquêtes en dehors des sources d'information strictement officielles ou issues des médias traditionnels.

> Ces enquêtes en sources ouvertes, également appelées OSINT pour « open source intelligence », sont désormais une composante visuelle et médiatique reconnaissable de notre quotidien. À partir d'une collecte d'informations disponibles librement sur Internet (parfois comparées à d'autres tirées d'une expérience « de terrain », sur place), ces éléments (photos et vidéos présentes sur les médias sociaux, cartographies numériques, documents administratifs, données gouvernementales, articles scientifiques, etc.) visent désormais à documenter un large éventail d'événements. (Deneuville et Rasmi, 2024)

> Recueil d'informations à partir de sources ouvertes (...) L'OSINT constitue un nouveau régime de vérité qui repose sur l'étude des traces pour établir des preuves qui viennent étayer une logique démonstrative et explicative, notamment pour répondre à la désinformation. (Le Deuff, 2021)

PHYSICALISATION DES DONNEES

Approche qui consiste à représenter les données sous forme d'objets physiques tangibles plutôt que sur des écrans numériques.

> Mise en scène de données matérialisées par des objets tangibles. (Catoir-Brisson, 2021)

> La physicalisation de données est un champ de recherche à la croisée de l'informatique tangible et de la visualisation d'information qui a pour but de faciliter l'exploration, la compréhension et la communication de données par le biais de leurs représentations physiques et numériques. (Jansen et Dragicevic, 2015)

PILOTAGE PAR LES DONNEES (DATA DRIVEN STRATEGY)

Le pilotage par les données consiste à baser des décisions stratégiques sur l'analyse de données pertinentes. En d'autres termes, plutôt que de se fier uniquement à l'intuition, à

l'expérience ou à l'intelligence humaines, les organisations utilisent des données quantifiables pour éclairer leurs choix et optimiser leurs performances.

> Data driven strategy, soit pilotage par les données, c'est-à-dire que l'organisation et la finalité de l'activité sont redessinées à partir de la donnée, notion qui devient un descripteur universel des objets. (Pène, 2020)

PROTECTION DE LA VIE PRIVEE DES LA CONCEPTION (PRIVACY BY DESIGN)

Le concept de *Privacy by Design*, en français protection de la vie privée dès la conception, désigne une approche proactive de la protection des données personnelles. Il s'agit d'intégrer la protection de la vie privée dès les premières étapes de la conception d'un produit, d'un service ou d'un système.

> Le Privacy by Design, pouvant se traduire en français par l'expression « la prise en compte de la vie privée dès la conception », est problématisé comme le principe techno-juridique selon lequel toute technologie exploitant les données personnelles doit intégrer la protection de la vie privée à partir des premières phases de sa conception, et s'y conformer tout au long de son cycle de vie (Musiani, 2015)

REGLEMENT GENERAL SUR LA PROTECTION DES DONNEES (RGPD)

Le Règlement Général sur la Protection des Données entré en vigueur le 23 mai 2018 a pour objectif de **renforcer la protection des données personnelles** des citoyens européens et d'harmoniser les législations nationales en matière de protection des données à travers l'Union européenne. En France, les données personnelles font l'objet d'une protection spécifique depuis la loi Informatique et libertés du 6 janvier 1978. Le RGPD renforce cette protection et pose des principes qui doivent être respectés pour toute activité concernant des données personnelles : le principe de nécessité (la collecte de données doit être justifiée), le droit à l'information (sur la collecte elle-même et sur les droits de la personne), la mise à disposition de moyens d'exercer ses droits (demandes d'accès, consultation, opposition, rectification, suppression), la limitation de la durée de conservation, la sécurisation, la continuité de la démarche de protection. C'est la CNIL, en France, qui est chargée de la mise en œuvre de ces principes, et les délégués à la protection des données (souvent appelés DPO, data protection officer) dans les organisations.

Le règlement général de protection des données (RGPD) est un texte réglementaire européen qui encadre le traitement des données de manière égalitaire sur tout le territoire de l'Union européenne (UE). Il est entré en application le 25 mai 2018.

Le RGPD s'inscrit dans la continuité de la loi française « Informatique et Libertés » de 1978, modifiée par la loi du 20 juin 2018 relative à la protection des données personnelles, établissant des règles sur la collecte et l'utilisation des données sur le territoire français. Il a été conçu autour de trois objectifs :

renforcer les droits des personnes

responsabiliser les acteurs traitant des données

crédibiliser la régulation grâce à une coopération renforcée entre les autorités de protection des données. (Ministère de l'Economie, des finances, et de la souveraineté industrielle et numérique)

SOUVERAINETE DES DONNEES

Principe qui stipule que les individus, les organisations et les Etats ont le droit de contrôler leurs données numériques. Cela signifie qu'ils ont le pouvoir de décider de la façon dont leurs données sont collectées, stockées, utilisées et partagées, qui a accès à leurs données, des espaces de stockage et de traitement, et des conditions d'utilisation et de transfert de leurs données. Ce principe est surtout politique et concerne le souci des Etats de conserver la maîtrise des données qu'ils produisent. La souveraineté peut aussi concerner le fait que l'État garde le contrôle des données qu'il produit sans en laisser la maîtrise au secteur privé.

La souveraineté numérique est souvent analysée de deux points de vue : le régalien, car les actions des plateformes numériques s'interprètent comme une remise en cause des prérogatives régaliennes ; l'économique car les plateformes, par nature, exploitent des externalités positives de réseau (les effets de réseaux) qui ont des conséquences économiques nombreuses sur les dynamiques concurrentielles (Cremer et al., 2019). (Isaac, 2022)

TRACES NUMERIQUES

Une trace est toute manifestation numérique qui atteste d'une activité, d'un événement ou d'une interaction. C'est un peu comme une empreinte digitale dans le monde numérique. Les traces numériques représentent des données, souvent personnelles, qui, traitées par les algorithmes des logiciels ou des plateformes, permettent de cibler les informations, mais aussi de surveiller les activités des personnes et de les influencer.

Les empreintes que nous laissons sur les réseaux sont au cœur de ce processus qui permet aux récepteurs – destinataires ou non – de réarticuler les contenus selon leur interprétation. Utiles et signifiantes sans être encore des documents, les traces dépendent des opérations d'extraction, d'annotation et de réagencement auxquelles elles sont soumises. (Merzeau, 2009)

VISUALISATION DE DONNEES (DATAVISUALISATION OU DATAVIZ)

La visualisation de données est une modalité, le plus souvent automatisée, de représentation graphique des données visant à les rendre faciles à comprendre et à interpréter. Elle utilise des éléments tels que des graphiques, des diagrammes, des cartes, des tableaux et des infographies pour communiquer des informations de manière claire, concise et engageante. Elle est largement utilisée dans la presse pour permettre au public de visualiser les informations statistiques par exemple. Elle peut poser des problèmes quant aux choix graphiques et à la création de biais d'interprétation possibles.

> La visualisation de données désigne tous types de représentations visuelles qui facilitent l'exploitation, l'analyse et la communication de données. Le design d'information et la visualisation scientifique en constituent des sous-ensembles. (Fredriksson, 2015)

> La datavisualisation sert à présenter rapidement les données et à les rendre lisibles. Elle permet d'analyser et de faciliter la compréhension des données en les retranscrivant efficacement sous forme visuelle. (Lagnel, 2021)

> Ensemble de procédures graphiques destinées à faciliter l'interprétation de jeux de « données » trop complexes ou trop massifs pour pouvoir être rapidement compris par le chercheur ; ces procédures sollicitent de nombreux algorithmes, même si la part humaine reste importante (paramétrages, recodages, choix ultimes). Elles s'inscrivent dans une logique de « preuve graphique » : la production d'un résultat visuel (d'une image organisée par l'enquêteur ou la structure des données, à l'inverse d'une peinture) afin d'infirmer ou de confirmer une hypothèse. (Guichard, 2019)

RÉFÉRENCES

Bastard, I., Cardon, D., Fouetillou, G., Prieur, C. et Raux, S. (2014) . Chapitre 8. Travail et travailleurs de la donnée Les sciences sociales et les données du web dans l'enquête Algopol. Big Data Nouvelles partitions de l'information. Actes du séminaire IST Inria, octobre 2014. (p. 133 -148). De Boeck Supérieur. https://doi.org/10.3917/dbu.caan.2014.02.0133

Ben Henda, M. (2019). TICE: normativité, interopérabilité et pratiques convergentes. In ECOTIDI: Numérique et Formation à Distance.

Ben Henda, M. (2006). Les standards de métadonnées pédagogiques : quelle interopérabilité ?. 23e Congrès de l'AIPU "Innovation, Formation et Recherche en pédagogie universitaire", Association Internationale de Pédagogie Universitaire, May 2006, Monastir, Tunisie. ⟨hal-04554697⟩

Cardon, D (2012). Regarder les données. Multitudes, 2012/2 n° 49. pp. 138-142. https://doi.org/10.3917/mult.049.0138.

Casseau, C. (2024). Accompagnement à l'exécution des notebooks Jupyter en milieu éducatif. Informatique. Thèse de doctorat. Université de Bordeaux.

Casemajor, N. (2024). Cultures de données et publics de données: conceptualisation critique. In Millerand, F., Coutant, A., Latzko- Toth, G., Millette, M. Datafication et publics de données. Penser la mise en donnée de la société, Presses de l'Université de Montréal, A paraître. hal-04254885v2

Catoir-Brisson, M.-J. (2021). Design d'information et physicalisation des données : une expérience pédagogique en Humanités numériques. *Sens public*, 1–23. https://doi.org/10.7202/1089653ar

Chabin, M.A. (2010). Nouveau glossaire de l'archivage. 45 p. URL : http://www.archive17.fr/index.php/l-archivage-pour-les-nuls/nouveau-glossaire-de-larchivage.html)

Cutts, B. B., Osia, U., Bray, L. A., Harris, A. R., Long, H. C., Goins, H., ... & Schnetzer, A. (2024). Shifting power: data democracy in engineering solutions. Environmental Research Letters, 19(10).

Delort, P. (2018). Le big data. Que sais-je.

Deneuville, A., et Rasmi, J. (2024, septembre 19). L'usage des données en accès libre questionne la pratique des journalistes. The Conversation. http://theconversation.com/lusage-des-donnees-en-acces-libre-questionne-la-pratique-des-journalistes-227239

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. Information Design Journal, 23(1), 6-18.

Fredriksson, S. (2015). « Du design d'information à la visualisation de données : un enjeu de transmission de sens auprès de la société civile », I2D - Information, données & documents, vol. 52, no. 2, 2015, pp. 36-36

Guichard, É. (2019). Cartographie et visualisation. Annales des Mines - Responsabilité & environnement, N° 94(2), 38-41. https://doi.org/10.3917/re1.094.0038.

Isaac, H. (2022). Quelle souveraineté numérique européenne?. Revue française de gestion, 305(4), 63-77.

Jansen, Y. et Dragicevic, P. (2015). Les représentations physiques de données. I2D - Information, données & documents, Volume 52(2), 37-37. https://doi.org/10.3917/i2d.152.0037.

Joliveau, T., Noucher, M. et Roche, S. (2013). La cartographie 2.0, vers une approche critique d'un nouveau régime cartographique. L'Information géographique, Vol. 77(4), 29-46. https://doi.org/10.3917/lig.774.0029.

Labadie, C., Eurich, M. & Legner, C. (2020). Data Democratization in Practice: Fostering Data Usage with Data Catalogs (La democratisation des données en pratique : Favoriser l'utilisation des données avec les catalogues de données). Conférence AIM, Marrakech, juin 2020.

Labelle, S. (2023). Pouvoir des médiations dispositives et consécration de l'agir ingénieur. Enquête sur les politiques de données en France. Approches Théoriques en Information-Communication (ATIC) 2023/2 N° 7. pp. 101-117.

Lagnel, J. M. (2021). Manuel de datavisualisation-2e éd.: Méthodes-Cas pratiques. Dunod.

Le Deuff, O (2021). L'Open Source Intelligence (OSINT) : origine, définitions et portée, entre convergence professionnelle et accessibilité à l'information. I2D - Information, données & documents, 2021/1 n° 1. pp. 14-20. https://doi.org/10.3917/i2d.211.0014.

Lehmans, A. (2017). "Données ouvertes et redéfinition de la culture de l'information dans les organisations", Communication et organisation [Online], 51 | 2017. URL: http://journals.openedition.org/communicationorganisation/5495

Lehmans, A. (2018). Les réinventions de la démocratie à l'aune de l'ouverture des données: du discours de la participation aux contraintes de la gouvernance. Les Enjeux de l'information et de la communication, (2), 135-146

Lehmans, A. et Capelle, C. (2023). Le cadre de l'expérience des données en éducation : gouvernance, représentations et intelligibilité des données dans l'éducation nationale. Communication & Organisation, n° 64(2), 33-49. https://doi-org.docelec.u-bordeaux.fr/10.4000/communicationorganisation/12583.

Lehmans, A, & Liquète, V (2022). Des littératies aux approches culturelles du numérique : l'exemple de la culture des données dans le champ de l'éducation aux médias et à l'information (EMI) Approches Théoriques en Information-Communication (ATIC) 2022/2 N° 5. pp. 35-46. https://doi-org.docelec.u-bordeaux.fr/10.3917/atic.005.0035.

Merzeau, L (2009). Du signe à la trace : l'information sur mesure. Hermès, La Revue, 2009/1 n° 53. pp. 21-29. https://doi.org/10.4267/2042/31471.

Musiani, F (2015). Les architectures P2P Une solution européenne originale pour la protection des données personnelles ? Réseaux, 2015/1 n° 189. pp. 47-75. https://doi.org/10.3917/res.189.0047.

Pène, S. (2020). Métiers de la fonction publique : motifs et modèles de transition numérique. Approches Théoriques en Information-Communication (ATIC), 1, 58-80. https://doi-org.docelec.u-bordeaux.fr/10.3917/atic.001.0058

Pérès, M. (2016). Analyser avec des outils descriptifs et statistiques des infographies d'infographies Questionner les processus de visualisation. Les Cahiers du numérique, 2016/4 Vol. 12. pp. 65-92.

Quimbert, E., Fichaut, M., Maudire, G. (2022). Guide principes FAIR. Ref. Principes FAIR dans le contexte du pôle ODATIS. ODATIS. https://doi.org/10.13155/87107

Raulin, A. (2022). L'intelligence artificielle dans la gestion et la valorisation de l'information : clés de repérage (histoire et analyse) I2D - Information, données & documents, n° 1(1), 14-21. https://doi-org.docelec.u-bordeaux.fr/10.3917/i2d.221.0014.

Rousseau, J.-Y., Couture, C. (1994). Les fondements de la discipline archivistique. Sainte-Foy (Québec) : Presses de l'Université du Québec, p. 123.

Rubin, A. (2020). Learning to reason with data: How did we get here and what do we know? Journal of the Learning Sciences, 29, 154-164.

Schield M., (2004). Information Literacy, Statistical Literacy and Data Literacy. IASSIST Quarterly, International Association for Social Science Information Service and Technology 28 (2–3), 7–14.

Verdi, U. (2023). “Quelle(s) réponse(s) à l’enjeu d’acculturation aux données ? Un état de l’art des caractéristiques de la data literacy”, Revue française des sciences de l’information et de la communication [Online], 26 | 2023. URL: http://journals.openedition.org/rfsic/14589

Verdi, U., Pinède, N. et Melançon, G. (2024). La gouvernance des données en contexte universitaire : proposition d'un modèle de maturité. INFORSID 2024, May 2024, Nancy, France. pp.21-36. ⟨hal-04585975⟩

Ce document est rédigé par les équipes de recherche dans le cadre des GTnum du ministère de l'Éducation nationale.

La responsabilité des contenus publiés leur appartient.