05 Données et formats

Durée
~2h

Thématique

DONNEES STRUCTUREES

Activité Branchée ou Débranchée

Description de l'activité

Dans cette activité, on présente le vocabulaire lié aux données structurées et on l'illustre par différents exemples.

Objectifs pédagogiques ou compétences

Objectifs généraux	Objectifs intermédiaires	Compétences
Notions de cours	- Connaître le vocabulaire lié aux données structurées - Connaître quelques formats liés aux données structurées - Comprendre les bases des enjeux autour de la collecte et de l'usage des données	- Manipuler un court code (csv, xml, json) - Faire des recherches sur Internet - Réfléchir sur les données auxquelles on donne accès

Matériel et outils

- Fiche élève à imprimer
- Eventuellement, 1 poste par élève ou binôme pour effectuer des recherches

Tags

#données structurées ; #données ; #descripteurs ; #informations ; #collections ; #métadonnées ; #bases de données ; #format ; #données numériques ouvertes ;

Déroulé de l'activité

Introduction : *(~20 minutes)*

- **Présenter les objectifs de la séance (contenu théorique et productions attendues)** *(2-3 minutes)*
- **Introduire la thématique :** *(~10 minutes)*

Pour lancer la thématique, on propose un quiz, auquel les élèves répondent seuls ou en binômes. L'enseignant.e apporte des compléments d'information lorsque besoin.

Q1&2 : Une donnée est une valeur décrivant un objet, une personne, un événement digne d'intérêt pour celui qui choisit de la conserver, etc. Cette donnée est la valeur prise par un descripteur qui précise son sens. Plusieurs descripteurs sont utiles pour décrire un même objet.

- On peut demander aux élèves de donner des exemples de descripteurs et de données, pourquoi pas en amenant les jeunes à parler de ce qu'ils aiment (artiste / streameur ou streameuse préféré.e, dernier manga lu, jeu vidéo détesté, sports pratiqués, allergies …). Dans ce cas, en prévision de la Q3, on peut déjà les structurer en « table » pour avoir un exemple concret.

Q3 : Une collection regroupe des objets partageant les mêmes descripteurs. Les collections sont en général représentées par une table : les objets sont disposés en ligne, les descripteurs en colonne, et les données à l'intersection. Lorsque des collections sont représentées sous la forme d'une table, on parle de « données structurées ».

Q4 : Une base de données (data base en anglais) regroupe plusieurs collections de données reliées entre elles. Elle permet de stocker et de retrouver l'intégralité des données en rapport avec un thème ou une activité. La base de données d'une bibliothèque regroupe la collection sur les livres, les abonnés et les emprunts effectués.

Q5 : Une métadonnée est une donnée servant à définir ou à décrire une autre donnée, quel que soit son support (papier ou électronique). C'est un mot composé du préfixe grec "méta," indiquant l'auto-référence ; le mot signifie donc proprement "donnée de/à propos de donnée.

- On peut en profiter pour évoquer d'autres mots construits sur la même racine, comme :
  - **Métacognition :** La capacité de réfléchir sur sa propre pensée, de surveiller et de réguler sa propre compréhension et son processus de résolution de problèmes.
  - **Métalangage :** Un langage utilisé pour décrire ou discuter d'un autre langage. Il est souvent utilisé dans le contexte de la linguistique pour analyser et discuter des langues.
  - **Métaphysique :** Une branche de la philosophie qui traite de la nature fondamentale de la réalité, de l'existence, de la causalité et d'autres concepts abstraits.
  - **Métagame :** Un concept utilisé dans les jeux vidéo pour décrire les stratégies et les tactiques qui émergent au-delà des règles formelles du jeu.

Q6 : En informatique, une extension de fichier est un suffixe du nom de fichier destiné à identifier son format. Il est séparé du nom par un point. C'est grâce à cette extension que le système d'exploitation sait quel logiciel il doit lancer pour lire ou exécuter le fichier. En général, les extensions correspondent aux formats. Par exemple, un fichier au format CSV a pour extension .csv.

- On peut ici demander aux élèves quels formats de fichiers ils connaissent.

- **Des données … Pour faire quoi ? :** *(~10 minutes)*

En binômes ou petits groupes, on leur demande de lister les applications qu'ils utilisent, et pour chacune, les différents types de données qu'ils collectent, et quel usage peut en être fait.

Le plus important ici est de sensibiliser à l'omniprésence des données, mais aussi à la notion de confidentialité ainsi que des multiples usages, plus ou moins acceptables (voire légaux), qui peut être fait de ces données

Déroulement – Exercices multiples *(~20 minutes)*

Les élèves suivent la fiche d'activité, seuls ou en binômes. L'enseignant.e corrige régulièrement et apporte des informations complémentaires.

Conclusion *(15 minutes)*

- **Bilan de la séance :** *(5 minutes)*

Pour clôturer la séance, on peut revenir sur les principales difficultés rencontrées pendant l'activité, et faire un bilan sur les informations à retenir. Voici quelques pistes :

- Méthodes pour collecter et stocker les données, sélection de données pertinentes
- La sécurité des données
- L'intérêt de conserver les métadonnées
- La question de la correction des données erronées et de leur mise à jour
- La question de l'intégrité et de la conformité aux règlements, notamment pour les entreprises et structures internationales
- …

Éventuellement, il est possible de finir sur un court échange autour :

- **Les données et nous** *(5 minutes)*

On peut profiter de cette activité pour rappeler aux élèves des aspects négatifs liés à la collecte et au stockage des données, notamment en se basant sur ce qui a été dit au cours de la séance.

Pistes de discussion :

- Aspects négatifs :
  - **Violation de la vie privée :** La collecte excessive de données personnelles peut violer la vie privée des individus en divulguant des informations sensibles sans leur consentement. Cela peut entraîner une surveillance indésirable et une exploitation de la vie privée.
  - **Risque de vol d'identité :** Les données mal gérées peuvent être piratées ou volées, exposant les usagers au risque de vol d'identité, de fraude et d'autres activités criminelles.

- **Ciblage publicitaire excessif :** Les entreprises utilisent souvent les données pour cibler les publicités de manière agressive, ce qui peut créer une expérience en ligne intrusive et dérangeante pour les usagers.
- **Discrimination et profilage :** L'utilisation de données pour établir des profils peut conduire à la discrimination, car les décisions basées sur ces profils peuvent être injustes, notamment dans le domaine de l'emploi, de l'assurance et du crédit.
- **Manque de transparence :** Lorsque les usagers ne sont pas informés de la manière dont leurs données sont collectées et utilisées, cela crée un manque de confiance dans les services en ligne.

- Moyens de réglementer :
  - **Lois sur la protection des données :** Les gouvernements peuvent mettre en place des lois telles que le RGPD en Europe ou le CCPA en Californie, qui définissent des normes strictes pour la collecte et l'utilisation des données personnelles.
  - **Consentement éclairé :** En expliquant clairement quelles données seront collectées et comment elles seront utilisées, les usagers auront une meilleure compréhension de ce à quoi ils consentent.
  - **Droit à l'effacement :** Les utilisateurs devraient avoir le droit de demander la suppression de leurs données personnelles des bases de données des entreprises, et surtout, le processus devrait être facile d'accès et rapide pour éviter l'abandon des processus.
  - **Transparence :** Les entreprises doivent être transparentes sur leurs pratiques de collecte et d'utilisation des données, en fournissant aux utilisateurs un accès facile aux informations sur la politique de confidentialité et les pratiques de données.
  - **Audits et sanctions :** Les régulateurs doivent avoir le pouvoir de réaliser des audits et d'imposer des sanctions aux entreprises qui enfreignent les lois sur la protection des données.
  - **Éducation et sensibilisation :** Informer les utilisateurs sur la manière de protéger leur vie privée en ligne et sur les risques liés à la collecte de données est également essentiel pour une réglementation efficace.

- **Les métiers en lien** *(5 minutes)*

Cette activité a permis de sensibiliser les jeunes aux usages des données, mais aussi aux méthodes de traitement et de stockage. On peut donc évoquer les principaux métiers en lien avec ces domaines pour les rendre plus concrets :

- **Domaine de l'Analyse de Données :**
  - **Data Analyst :** Analyse des données pour fournir des informations exploitables.
  - **Data Scientist :** Utilisation de techniques avancées pour extraire des informations à partir de grandes quantités de données.
  - **Business Analyst :** Analyse des données pour aider les entreprises à prendre des décisions stratégiques.

Domaine de la Gestion des Données :

- **Data Manager :** Gestion de la collecte, du stockage et de la maintenance des données.
- **Database Administrator (DBA) :** Gestion et maintenance des bases de données.
- **Data Engineer :** Conception et gestion de l'infrastructure nécessaire au stockage et à la manipulation des données.

Domaine du Marketing et de la Publicité :

- **Digital Marketer :** Utilisation des données pour orienter les stratégies marketing en ligne.
- **Ad Operations Specialist :** Gestion des données publicitaires pour les campagnes en ligne.

Domaine de la Sécurité des Données :

- **Data Security Analyst :** Protection des données contre les menaces de sécurité.
- **Chief Information Security Officer (CISO) :** Responsable de la sécurité des données et de l'information dans une organisation.

Données et formats

Fiche activité - *Correction*

Introduction – Que savez-vous sur la « data » ?

- **Définissons les concepts …**

Trouvez la bonne réponse !

- Question 1 : Que sont les données ?
  - Des informations sur la météo
  - Des faits, des chiffres, ou des informations brutes
  - Des images et des vidéos
  - Des fichiers texte

- Question 2 : Que sont les descripteurs dans le contexte des données ?
  - Des informations permettant de décrire quelque chose (forme, couleur, texture, …) ou quelqu'un (physique, caractère, …)
  - Des informations personnelles
  - Une information décrivant le sens ou la valeur d'une donnée (« âge » désigne 19, « date de naissance » désigne 10/10/2010)
  - Des images en haute résolution

- Question 3 : Que sont les collections de données ?
  - Des ensembles de données qui n'ont aucun lien entre elles
  - Des groupes de données partageant un point commun
  - Des informations importantes
  - Des fichiers audios ou vidéo

- Question 4 : Qu'est-ce qu'une base de données ?
  - Un ensemble de données qui n'a pas été sauvegardé correctement
  - Un programme informatique contenant les sauvegardes ou les données liés à un compte dans un jeu vidéo
  - Un système organisé pour structurer, stocker, gérer et récupérer des données
  - Un format de fichier contenant des données récupérées de manière illégale

- Question 5 : Que sont les métadonnées ?
  - Des données qui ne sont pas importantes
  - Des données sensibles
  - Des données qui décrivent d'autres données
  - Des données financières

- Question 6 : Que sont les formats de données ?
  - Des ordinateurs spécifiques utilisés pour stocker des données
  - Des types de fichiers qui déterminent comment les données sont stockées et organisées
  - Des collections de données
  - Des personnes spécialisées dans l'analyse de données

- **Des données … Mais pour faire quoi ?**

En groupe, listez quelques applications et sites que vous utilisez au quotidien. Pour chacun, listez les données que vous pensez être collectées sur vous, et essayez de trouver quels usages peuvent en être faits.

Propositions :

Réseaux sociaux (par exemple, Instagram, Snapchat, Tiktok, …) :

- **Données collectées :** Photos, messages, emplacement, âge, centres d'intérêt.
- **Usages potentiels :** Personnalisation de la publicité, recommandation de contenu, Collecte de données sensibles sans consentement, …

Applications de messagerie (par exemple, WhatsApp, Telegram, Messenger) :

- **Données collectées :** Messages, contacts, numéros de téléphone.
- **Usages potentiels :** Analyse de conversations pour cibler des publicités, Espionnage ou vol de données personnelles, Usage de messages privés pour entraîner des IA, …

Jeux en ligne (par exemple, Fortnite, Minecraft, Roblox) :

- **Données collectées :** Identifiant de joueur, performances en jeu.
- **Usages potentiels :** Personnalisation de l'expérience de jeu, Incitation à des achats in-app, Publicité ciblée pour d'autres jeux ou contenus, …

Services de streaming vidéo (par exemple, Netflix, YouTube) :

- **Données collectées :** Historique et comportements de visionnage, préférences de contenu
- **Usages potentiels :** Recommandations de vidéos, publicité ciblée, …

Étape 1 – Données, descripteurs, informations et collections

- **Données et descripteurs**

Le tableau suivant est tire d'un fichier de contacts :

Nom	Adresse	Numéro de téléphone
Martin Jean	11 rue de la paix, Paris	06 11 23 45 66
Martin Audrey	11 rue de la paix, Paris	06 15 75 88 54
Dupont Julie		06 44 33 22 11

  - Quels sont les descripteurs utilisés pour caractériser un contact ?

Nom, Adresse, et Numéro de téléphone.

  - Quelle est la donnée correspondante au descripteur "Nom" du premier contact ?

Martin Jean.

  - Quelle est la donnée correspondante au descripteur "Adresse" du troisième contact ?

Aucune.

- **Informations**

Les données sont factuelles et neutres, mais l'information qui en découle en général ne l'est pas. Pour devenir une information qui a du sens, les données doivent être traitées, manipulées, transformées, ou encore croisées, et nécessitent une interprétation.

  - Quelles sont les données correspondantes aux descripteurs "Nom" et "Adresse" des deux premiers contacts ?
  - Quelle information en découle-t-il ?

Même nom de famille et même adresse : il y a sûrement un lien de famille (frère/sœur, mari/femme, père/fille, etc.) ou peut-être pas (il faudrait croiser avec d'autres données).

- **Collections**

Voici une table, c'est-à-dire un ensemble de données structurées représentant les sorties des films Naruto au Japon et en France :

Ordre de parution	Nom du film	Sortie japonaise	Sortie française
1	Naruto et la Princesse des neiges	21 août 2004	5 avril 2009
2	La Légende de la pierre de Guelel	6 août 2005	3 mai 2009
3	Mission spéciale au pays de la Lune	5 août 2006	7 juin 2009
4	Naruto Shippuden : Un funeste présage	4 août 2007	16 mai 2010
5	Naruto Shippuden : Les Liens	2 août 2008	23 mai 2010
6	Naruto Shippuden : La Flamme de la volonté	1er août 2009	? septembre 2010
7	Naruto Shippuden: The Lost Tower	31 juillet 2010	25 mai 2011
8	Naruto Shippuden: Blood Prison	30 juillet 2011	14 novembre 2012
9	Naruto Shippuden: Road to Ninja	28 juillet 2012	6 juin 2013
10	Naruto the Last, le film	6 décembre 2014	13 mai 2015
11	Boruto : Naruto, le film	7 août 2015	16 septembre 2015

- Quels autres descripteurs pourraient être intéressants à voir ?

Le nombre d'entrées en salles et/ou de recettes pour chaque film, en France et au Japon, bénéfices dans les deux pays, éventuellement ventes de produits dérivés ou complémentaires pour analyser si la sortie des films a eu un impact, etc.

Nom	Adresse	Numéro de téléphone
Martin Jean	11 rue de la paix, Paris	06 11 23 45 66
Martin Audrey	11 rue de la paix, Paris	06 15 75 88 54
Dupont Julie		06 44 33 22 11

- Reprenons le tableau précédent. Ce tableau est-il un ensemble de données structurées ? Si oui, que représente la collection qui y est décrite ?

Oui, il s'agit d'un fichier de contacts.

- Les données présentes sont-elles bien structurées ?

Non, il faudrait ajouter au moins le descripteur "Prénom".

- Aujourd'hui, quels autres descripteurs sont présents dans l'application « Contacts » de nos téléphones ?

Photo de profil, Statut de favori (pour trier) ou inscription dans un groupe, derniers appels et messages, autres connexions via applications, …

Étape 2 – Les métadonnées

- Quels exemples de métadonnées connaissez-vous ?

Associer à une donnée la date à laquelle elle a été produite ou enregistrée.
Associer à une photo les coordonnées GPS du lieu où elle a été prise.
Associer à un fichier audio la durée totale de la piste.
Associer à un document le nom de l'auteur et la date de création.
Associer à une vidéo le format de compression utilisé.
Associer à un e-mail l'expéditeur, le destinataire, la date d'envoi et l'objet.
Associer à un fichier PDF le titre du document et les mots-clés associés.
Associer à une image la résolution en pixels.
Associer à un tweet le nombre de retweets et de likes.
Associer à un fichier musical le genre musical et les informations de l'album.
…

Sous le système d'exploitation Windows, on peut accéder aux métadonnées d'un fichier en effectuant successivement les actions suivantes :

- Effectuer un clic droit sur l'icône du fichier dont on veut connaître les métadonnées.
- Choisir Propriétés.
- Cliquer sur l'onglet Détails.

- Accédez aux métadonnées d'une image et d'un fichier texte stockés sur votre ordinateur. Quelles informations avez-vous eu grâce aux métadonnées ?

- Quelle est l'utilité des métadonnées ? Vous pouvez effectuer des recherches si besoin.

Les métadonnées améliorent l'efficacité de la gestion des données et leur utilisation. Elles sont essentielles car elles ajoutent des informations contextuelles aux données, facilitant leur organisation, leur recherche et leur compréhension. Elles permettent de décrire des détails tels que la date de création, l'auteur, le format, ou l'emplacement géographique, ce qui est précieux pour la gestion de l'information, la recherche de données spécifiques, et l'identification rapide des ressources numériques.

Étape 3 – Les bases de données

- Que signifie l'acronyme SGBDR. À quoi est-ce que cela fait référence ?

L'acronyme SGBDR signifie "Système de Gestion de Base de Données Relationnel". Il fait référence à un type de système de gestion de base de données qui est conçu pour stocker, organiser et gérer des données de manière structurée, en utilisant des tables et en établissant des relations entre ces tables.

- Quel est le SGBDR le plus connu ?

MySQL, mais il existe également Oracle Database, PostgreSQL ou encore Microsoft SQL Server.

- Pourquoi utilise-t-on le terme « relationnel » ?

Le terme "relationnel" est utilisé car les bases de données relationnelles organisent les données sous forme de tables dans lesquelles les données sont stockées en lignes et en colonnes. Ces tables peuvent avoir des relations entre elles, ce qui permet de créer une structure de données complexe et de gérer des ensembles de données interconnectés. Le modèle relationnel a été introduit par Edgar F. Codd en 1970 et est devenu la base de la plupart des systèmes de gestion de base de données utilisés aujourd'hui.

Prenons pour exemple la base de données d'une bibliothèque, qui regroupe plusieurs collections, par exemple la collection de livres, les informations sur les abonnés et les enregistrements des emprunts effectués. Le descripteur du numéro (unique) de l'abonné peut apparaître dans les trois tables afin d'établir des liens directs entre celles-ci.

- Quel est l'intérêt de procéder ainsi ?

Par exemple, cela permet d'obtenir directement le nom de la personne qui a emprunté tel ou tel livre sans avoir besoin d'ajouter des données redondantes dans la base de données (comme réécrire systématiquement le nom à chaque emprunt). Ceci permet notamment de prévenir certaines erreurs, telles que la suppression involontaire d'éléments.

Étape 4 – Les formats des données

- **Les différents types de formats**

Pour assurer la persistance des données, celles-ci sont stockées dans des fichiers**. Il existe deux types de formats de fichiers** :

- **Les fichiers de type "texte"**, **c'est-à-dire ceux lisibles par des logiciels de traitement de texte** comme Notepad++, et qui **ne comportent que des caractères alphanumériques**. Les principaux formats de fichiers texte sont les formats .txt, CSV, XML, JSON et vCard. **Les caractères sont codés en mémoire en respectant une norme d'encodage.**
- **Les fichiers de type "binaire"**, **c'est-à-dire ceux non lisibles par des logiciels de traitement de texte.** Il s'agit principalement des formats de tableurs (ODS, XLS, XLSX, etc.), des SGBDR (Systèmes de Gestion de Bases de Données Relationnelles) et de certaines images (JPG, PNG, TIFF, etc.).

Le format CSV (Comma Separated Values)**, qui est à privilégier car il est le plus universel et le plus simple.**

- **Bien qu'il soit à la base un fichier texte, il peut être facilement importé dans un tableur.** De plus, un fichier créé dans un tableur peut être facilement exporté au format CSV.
- **Il est possible de remplacer la virgule par un point-virgule ou une tabulation comme séparateur de données.** Ceci est particulièrement intéressant dans le cas où la virgule est déjà utilisée dans une notation spécifique des données, par exemple, dans l'écriture des nombres décimaux en français."

Remarque : Les extensions de fichiers étant masquées par défaut, on procède de la façon suivante pour pouvoir les afficher (sous Windows 10) :

- Lancer l'explorateur de fichiers.
- Cliquer sur l'onglet **« Affichage »**.
- Cocher la case **« Extensions de noms de fichiers »**.

  - Ouvrez un fichier PDF avec le logiciel Notepad++. Qu'obtient-on ? Était-ce prévisible ?

Le résultat est illisible. C'était prévisible car les fichiers PDF sont des fichiers binaires.

- **Les formats textes CSV, XML et JSON**

On considère une même table codée sous les trois formats textes : CSV, XML et JSON.

CSV	XML	JSON

Nom, Prénom, Décès
Hugo, Victor, 1885
Camus, Albert, 1960
	<Auteur>
	<Nom>Hugo</Nom>
	<Prénom>Victor</Prénom>
	<Décès>1885</Décès>
	</Auteur>
	<Auteur>
	<Nom>Camus</Nom>
	<Prénom>Albert</Prénom>
	<Décès>1960</Décès>
	</Auteur>
		[
		{
		"Nom" :"Hugo",
		"Prénom" :"Victor",
		"Décès" :1885
		},
		{
		"Nom" :"Camus",
		"Prénom" :"Albert",
		"Décès" :1960
		}
		]

  - Quels sont les descripteurs et les données de la table ?

Descripteurs : Nom, Prénom, Décès ; **Données :** Hugo, Victor, 1885, Camus, Albert, 1960

  - Sous quelle forme est codée la table suivant le format utilisé ?

En CSV, la première ligne correspond aux descripteurs et les suivantes aux données ; les éléments sont séparés par une virgule.

En XML, chaque enregistrement est écrit sous la forme <descripteur>donnée</descripteur> ; le format XML est un langage descriptif utilisant des balises (comme le HTML).

En JSON, chaque enregistrement est écrit sous la forme descripteur donnée et séparé par des virgules ; un individu de la collection est délimité par des accolades et les individus sont séparés par des virgules. Ce format est assez similaire au XML.

On désire ajouter le descripteur "Nationalité" ainsi que l'auteur Jane Austen. Modifier les trois codes ci-dessus en conséquence.

CSV	XML	JSON

Nom, Prénom, Nationalité, Décès
Hugo, Victor, Française, 1885
Camus, Albert, Française, 1960
Austen, Jane, Anglaise, 1817
	<Auteur>
	<Nom>Hugo</Nom>
	<Prénom>Victor</Prénom>
	<Nationalité>Française</Nationalité>
	<Décès>1885</Décès>
	</Auteur>
	<Auteur>
	<Nom>Camus</Nom>
	<Prénom>Albert</Prénom>
	<Nationalité>Française</Nationalité>
	<Décès>1960</Décès>
	</Auteur>
	<Auteur>
	<Nom>Austen</Nom>
	<Prénom>Jane</Prénom>
	<Nationalité>Anglaise</Nationalité>
	<Décès>1817</Décès>
	</Auteur>
		[
		{
		"Nom" :"Hugo",
		"Prénom" :"Victor",
		"Nationalité" :"Française",
		"Décès" :1885
		},
		{
		"Nom" :"Camus",
		"Prénom" :"Albert",
		"Nationalité" :"Française",
		"Décès" :1960
		}
		{
		"Nom" :"Austen",
		"Prénom" :"Jane",
		"Nationalité" :"Anglaise",
		"Décès" :1810
		}
		]

Étape 5 – Les données numériques ouvertes

Remarque : Les données ouvertes (Open Data) **sont des informations accessibles librement et gratuitement, sous la forme de fichiers respectant un format spécifique. La finalité de telles données est de donner la possibilité à tout citoyen, toute entreprise ou association de les utiliser à ses propres fins d'analyse pour en extraire l'information désirée.**

Les données ouvertes peuvent être d'origine publique (c'est-à-dire émanant de services publics, de collectivités, de communes, etc.)**, mais également d'origine privée** (c'est-à-dire provenant d'entreprises et d'institutions dont les données contribuent à des projets d'utilité publique, comme la SNCF, la RATP, etc.)**.**

- Quels sont les principes auxquels les données publiques ouvertes doivent répondre ?

Les données publiques sont considérées comme ouvertes si elles répondent aux huit principes suivants (2007, Open Government Data, USA) :

- Complètes : toutes les données doivent être rendues disponibles, sauf les données pouvant porter atteinte à la vie privée des citoyens ou à la sécurité.
- Primaires : les données doivent être brutes, telles qu'elles ont été collectées à la source, non agrégées, non modifiées.
- Récentes et actualisées : elles doivent être rendues disponibles aussi vite que possible afin de préserver leur valeur.
- Accessibles : les données sont disponibles pour le plus large spectre d'utilisateurs.
- Exploitables : elles doivent être structurées et documentées afin de permettre un traitement informatisé.
- Accès non discriminatoire : elles sont disponibles pour tout le monde de manière anonyme et ne nécessitent pas d'enregistrement.
- Format non-propriétaire : elles doivent être rendues disponibles au moins dans un format sur lequel aucune entité ne détient le monopole (ex : non PDF, non Excel).
- Libre de droits : les données ne doivent pas être l'objet de droits d'auteurs, marques déposées, brevets, etc.

- Qui coordonne la mise à disposition des données ouvertes en France ?

La mission gouvernementale Etalab.

- Sur quel site le gouvernement français met-il à disposition des données ouvertes ?

data.gouv.fr.

Données et formats

Fiche activité - *Correction*

Introduction – Que savez-vous sur la « data » ?

- **Définissons les concepts …**

Trouvez la bonne réponse !

- Question 1 : Que sont les données ?
  - Des informations sur la météo
  - Des faits, des chiffres, ou des informations brutes
  - Des images et des vidéos
  - Des fichiers texte

- Question 2 : Que sont les descripteurs dans le contexte des données ?
  - Des informations permettant de décrire quelque chose (forme, couleur, texture, …) ou quelqu'un (physique, caractère, …)
  - Des informations personnelles
  - Une information décrivant le sens ou la valeur d'une donnée (« âge » désigne 19, « date de naissance » désigne 10/10/2010)
  - Des images en haute résolution

- Question 3 : Que sont les collections de données ?
  - Des ensembles de données qui n'ont aucun lien entre elles
  - Des groupes de données partageant un point commun
  - Des informations importantes
  - Des fichiers audios ou vidéo

- Question 4 : Qu'est-ce qu'une base de données ?
  - Un ensemble de données qui n'a pas été sauvegardé correctement
  - Un programme informatique contenant les sauvegardes ou les données liés à un compte dans un jeu vidéo
  - Un système organisé pour structurer, stocker, gérer et récupérer des données
  - Un format de fichier contenant des données récupérées de manière illégale

- Question 5 : Que sont les métadonnées ?
  - Des données qui ne sont pas importantes
  - Des données sensibles
  - Des données qui décrivent d'autres données
  - Des données financières

- Question 6 : Que sont les formats de données ?
  - Des ordinateurs spécifiques utilisés pour stocker des données
  - Des types de fichiers qui déterminent comment les données sont stockées et organisées
  - Des collections de données
  - Des personnes spécialisées dans l'analyse de données

- **Des données … Mais pour faire quoi ?**

En groupe, listez quelques applications et sites que vous utilisez au quotidien. Pour chacun, listez les données que vous pensez être collectées sur vous, et essayez de trouver quels usages peuvent en être faits.

……

……

……

……

……

……

……

……

……

……

……

……

……

Étape 1 – Données, descripteurs, informations et collections

- **Données et descripteurs**

Le tableau suivant est tire d'un fichier de contacts :

Nom	Adresse	Numéro de téléphone
Martin Jean	11 rue de la paix, Paris	06 11 23 45 66
Martin Audrey	11 rue de la paix, Paris	06 15 75 88 54
Dupont Julie		06 44 33 22 11

- Quels sont les descripteurs utilisés pour caractériser un contact ?

.....

.....

- Quelle est la donnée correspondante au descripteur "Nom" du premier contact ?

.....

.....

- Quelle est la donnée correspondante au descripteur "Adresse" du troisième contact ?

.....

.....

- **Informations**

Les données sont factuelles et neutres, mais l'information qui en découle en général ne l'est pas. Pour devenir une information qui a du sens, les données doivent être traitées, manipulées, transformées, ou encore croisées, et nécessitent une interprétation.

- Quelles sont les données correspondantes aux descripteurs "Nom" et "Adresse" des deux premiers contacts ?

.....

.....

- Quelle information en découle-t-il ?

.....

.....

- **Collections**

Voici une table, c'est-à-dire un ensemble de données structurées représentant les sorties des films Naruto au Japon et en France :

Ordre de parution	Nom du film	Sortie japonaise	Sortie française
1	Naruto et la Princesse des neiges	21 août 2004	5 avril 2009
2	La Légende de la pierre de Guelel	6 août 2005	3 mai 2009
3	Mission spéciale au pays de la Lune	5 août 2006	7 juin 2009
4	Naruto Shippuden : Un funeste présage	4 août 2007	16 mai 2010
5	Naruto Shippuden : Les Liens	2 août 2008	23 mai 2010
6	Naruto Shippuden : La Flamme de la volonté	1er août 2009	? septembre 2010
7	Naruto Shippuden: The Lost Tower	31 juillet 2010	25 mai 2011
8	Naruto Shippuden: Blood Prison	30 juillet 2011	14 novembre 2012
9	Naruto Shippuden: Road to Ninja	28 juillet 2012	6 juin 2013
10	Naruto the Last, le film	6 décembre 2014	13 mai 2015
11	Boruto : Naruto, le film	7 août 2015	16 septembre 2015

- Quels autres descripteurs pourraient être intéressants à voir ?

.....

.....

.....

.....

Nom	Adresse	Numéro de téléphone
Martin Jean	11 rue de la paix, Paris	06 11 23 45 66
Martin Audrey	11 rue de la paix, Paris	06 15 75 88 54
Dupont Julie		06 44 33 22 11

- Reprenons le tableau précédent. Ce tableau est-il un ensemble de données structurées ? Si oui, que représente la collection qui y est décrite ?

.....

.....

- Les données présentes sont-elles bien structurées ?

.....

.....

- Aujourd'hui, quels autres descripteurs sont présents dans l'application « Contacts » de nos téléphones ?

.....

.....

.....

Étape 2 – Les métadonnées

- Quels exemples de métadonnées connaissez-vous ?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Sous le système d'exploitation Windows, on peut accéder aux métadonnées d'un fichier en effectuant successivement les actions suivantes :

- Effectuer un clic droit sur l'icône du fichier dont on veut connaître les métadonnées.
- Choisir Propriétés.
- Cliquer sur l'onglet Détails.

    - Accédez aux métadonnées d'une image et d'un fichier texte stockés sur votre ordinateur. Quelles informations avez-vous eu grâce aux métadonnées ?

.....

.....

.....

.....

.....

    - Quelle est l'utilité des métadonnées ? Vous pouvez effectuer des recherches si besoin.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Étape 3 – Les bases de données

- Que signifie l'acronyme SGBDR. À quoi est-ce que cela fait référence ?

- Quel est le SGBDR le plus connu ?

- Pourquoi utilise-t-on le terme « relationnel » ?

Prenons pour exemple la base de données d'une bibliothèque, qui regroupe plusieurs collections, par exemple la collection de livres, les informations sur les abonnés et les enregistrements des emprunts effectués. Le descripteur du numéro (unique) de l'abonné peut apparaître dans les trois tables afin d'établir des liens directs entre celles-ci.

- Quel est l'intérêt de procéder ainsi ?

.....

.....

.....

.....

.....

.....

Étape 4 – Les formats des données

- **Les différents types de formats**

Pour assurer la persistance des données, celles-ci sont stockées dans des fichiers**. Il existe deux types de formats de fichiers** :

- **Les fichiers de type "texte"**, **c'est-à-dire ceux lisibles par des logiciels de traitement de texte** comme Notepad++, et qui **ne comportent que des caractères alphanumériques**. Les principaux formats de fichiers texte sont les formats .txt, CSV, XML, JSON et vCard. **Les caractères sont codés en mémoire en respectant une norme d'encodage.**
- **Les fichiers de type "binaire"**, **c'est-à-dire ceux non lisibles par des logiciels de traitement de texte.** Il s'agit principalement des formats de tableurs (ODS, XLS, XLSX, etc.), des SGBDR (Systèmes de Gestion de Bases de Données Relationnelles) et de certaines images (JPG, PNG, TIFF, etc.).

  Le format CSV (Comma Separated Values)**, qui est à privilégier car il est le plus universel et le plus simple.**

- **Bien qu'il soit à la base un fichier texte, il peut être facilement importé dans un tableur.** De plus, un fichier créé dans un tableur peut être facilement exporté au format CSV.
- **Il est possible de remplacer la virgule par un point-virgule ou une tabulation comme séparateur de données.** Ceci est particulièrement intéressant dans le cas où la virgule est déjà utilisée dans une notation spécifique des données, par exemple, dans l'écriture des nombres décimaux en français."

Remarque : Les extensions de fichiers étant masquées par défaut, on procède de la façon suivante pour pouvoir les afficher (sous Windows 10) :

- Lancer l'explorateur de fichiers.
- Cliquer sur l'onglet **« Affichage »**.
- Cocher la case **« Extensions de noms de fichiers »**.

    - Ouvrez un fichier PDF avec le logiciel Notepad++. Qu'obtient-on ? Était-ce prévisible ?

.....

.....

.....

- **Les formats textes CSV, XML et JSON**

On considère une même table codée sous les trois formats textes : CSV, XML et JSON.

CSV	XML	JSON

Nom, Prénom, Décès
Hugo, Victor, 1885
Camus, Albert, 1960
	<Auteur>
	<Nom>Hugo</Nom>
	<Prénom>Victor</Prénom>
	<Décès>1885</Décès>
	</Auteur>
	<Auteur>
	<Nom>Camus</Nom>
	<Prénom>Albert</Prénom>
	<Décès>1960</Décès>
	</Auteur>
		[
		{
		"Nom" :"Hugo",
		"Prénom" :"Victor",
		"Décès" :1885
		},
		{
		"Nom" :"Camus",
		"Prénom" :"Albert",
		"Décès" :1960
		}
		]

- Quels sont les descripteurs et les données de la table ?

.....

.....

- Sous quelle forme est codée la table suivant le format utilisé ?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

On désire ajouter le descripteur "Nationalité" ainsi que l'auteur Jane Austen. Modifier les trois codes ci-dessus en conséquence.

CSV	XML	JSON

Étape 5 – Les données numériques ouvertes

Remarque : Les données ouvertes (Open Data) **sont des informations accessibles librement et gratuitement, sous la forme de fichiers respectant un format spécifique. La finalité de telles données est de donner la possibilité à tout citoyen, toute entreprise ou association de les utiliser à ses propres fins d'analyse pour en extraire l'information désirée.**

Les données ouvertes peuvent être d'origine publique (c'est-à-dire émanant de services publics, de collectivités, de communes, etc.)**, mais également d'origine privée** (c'est-à-dire provenant d'entreprises et d'institutions dont les données contribuent à des projets d'utilité publique, comme la SNCF, la RATP, etc.)**.**

- Quels sont les principes auxquels les données publiques ouvertes doivent répondre ?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

- Qui coordonne la mise à disposition des données ouvertes en France ?

.....

- Sur quel site le gouvernement français met-il à disposition des données ouvertes ?

.....