

Sciences du Numérique et Technologie



Les données numériques

TD n° 1 Données structurées, histoire et formats



Objectif de la séance :

- Lire et repérer des informations sur un support numérique
- Définir une donnée personnelle.
- Identifier les principaux formats et représentations de données



Mots-clés :

Cloud computing : utilisation de ressources situées dans des serveurs informatiques distants.

Descripteur : élément servant à décrire une donnée.

Données personnelles : informations identifiant une personne.

Format : type d'un fichier numérique.

Identifiant : code permettant d'identifier une personne.

Métadonnées : informations relatives à un fichier image, son, vidéo, etc.

Objet : élément d'une table de données.

Requête : texte composé de mots-clés permettant une action sur des données.

Synchronisation : copie des données stockées en local sur un serveur.

Table de données : données organisées en tableau.

Traitement de données : ensemble d'actions permettant d'extraire de l'information.

Valeur d'un descripteur : valeur d'une caractéristique d'un objet.



A la fin de cette séance j'aurai acquis ...

Des connaissances	Des compétences
- Données structurées/non structurées - Format de données structurées - Table de données - Les noms de champ ou descripteurs.	- Lire et repérer des informations sur un support numérique - Définir une donnée personnelle - Identifier les principaux formats et représentations de données - Utiliser les fonctions simples d'un traitement de texte

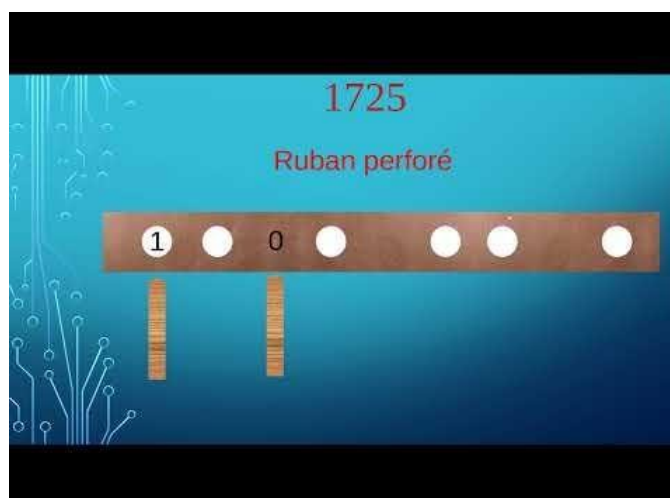
Vous écrierez les réponses aux questions sur le document

I. Les données numériques :

Données, comment les manipuler ? : Visionner la vidéo en cliquant sur l'image ou consulter la vidéo sur le site : Données, comment les manipuler ?



Histoire des données structurées : Visionner la vidéo en cliquant sur l'image ou consulter la vidéo sur le site présentant une histoire rapide des données structurées.



Exercice 1 :

A l'aide de la vidéo précédente, répondre aux questions suivantes :

- 1) Quelle invention datant du XVIIIe siècle a été mise à profit par les informaticiens pour stocker des données ?
- 2) En quelle année a été créé le premier disque dur ?
- 3) En quelle année a été développé le premier modèle permettant de structurer et d'indexer des bases de données ?
- 4) Comment s'appelle le premier tableur mis à la disposition du grand public ?
- 5) En quelle année l'**Open Government Initiative** a-t-elle été mise en place ?
- 6) Quel est le nom du texte de loi européen qui protège les citoyens européens quant à leurs données personnelles ?

II. Les données personnelles et les métadonnées

Définition 1 :

Une donnée est une valeur numérique ou textuelle représentant une information.

Pour pouvoir utiliser cette donnée d'un point de vue informatique, il va être utile de la « structurer ».

Exercice 2 :

DOC 1 Qu'est-ce qu'une donnée personnelle ?

Une **donnée** est **personnelle** si elle se rapporte à une personne identifiable. Elle peut être de différente nature, textuelle ou non. Elle doit permettre d'identifier une personne directement, grâce à un **identifiant**, ou indirectement, par recoupement de plusieurs informations.

Définition et nature des données personnelles

Toute information relative à un particulier identifié ou identifiable, directement ou indirectement, grâce à un identifiant ou à un ou plusieurs éléments propres à son identité

Par exemple :

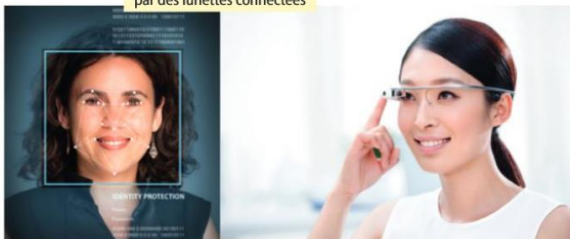


Source : Règlement général sur la protection des données (RGPD) du 27 avril 2016

DOC 2 La collecte des données

Les données peuvent être renseignées par un humain, comme lorsque nous nous inscrivons sur un site qui nous incite à renseigner nos données personnelles pour créer un compte. Les données peuvent aussi être capturées et enregistrées par un dispositif matériel, comme les caméras de surveillance associées à un logiciel de reconnaissance faciale ou encore des lunettes connectées.

Reconnaissance de visage par des lunettes connectées



Créer votre compte client

Civilité *	
☒ Mme ☐ M.	
Prénom *	Nom *
Date de naissance *	Téléphone mobile
E-mail *	Confirmation e-mail *
Mot de passe * (entre 8 et 25 caractères)	Confirmation du mot de passe *

Formulaire de création de compte

DOC 3 Retrouver les métadonnées d'un fichier personnel

Une **métadonnée** est une donnée particulière qui donne des informations sur la donnée principale, comme le titre d'une vidéo. Les données personnelles sont souvent associées à de nombreuses métadonnées qui donnent davantage d'informations sur le fichier principal et sur leur propriétaire. Ces métadonnées varient selon le type de fichier qu'elles décrivent.



A l'aide des documents, répondre aux questions suivantes :

- 1) **Donner des exemples de données personnelles et préciser dans chaque cas si l'identification est directe ou indirecte.**

Un nom permet d'identifier une personne directement alors qu'un numéro de compte en banque permet d'identifier une personne indirectement grâce à la base de données de la banque.

- 2) **Les données personnelles sont-elles toujours données consciemment par l'utilisateur ?**

Les données personnelles sont parfois fournies consciemment par l'utilisateur dans des formulaires, mais elles peuvent aussi être fournies par d'autres utilisateurs ou captées par un dispositif comme les lunettes connectées, des caméras de surveillance, le téléphone

- 3) **Préciser quelle est la donnée principale et quelles sont les métadonnées dans le Document 3. Comment peut-on retrouver les métadonnées d'un fichier vidéo ?**

Les métadonnées fournies ici sont le titre, le nom de l'artiste, le nom de l'album, le genre, la date, la langue,

III. Format de données structurées :

Définition 2 :

- Une donnée structurée est organisée sous forme d'un tableau **appelé table de données**.
- La première ligne décrit la forme des lignes suivantes ; les éléments de cette ligne donnant la signification portent plusieurs noms possibles : **ce sont les noms de champ ou descripteurs**.
- Les lignes suivantes sont appelées des **entrées ou enregistrements** : elles sont la liste des valeurs de chacun des descripteurs.

Exercice 3 :

Document 1 : Les caractéristiques des données structurées

Pour retrouver et traiter des données facilement, on les organise sous la forme d'un tableau appelé « **table de données** ». Une collection regroupe des objets partageant les mêmes **descripteurs**. Un **objet** est donné par la liste des valeurs de tous ses descripteurs. Les données sont alors dites « structurées ».

Collection					
Descripteurs	→	Nom	Capitale	Hymne	Superficie (km²)
Une valeur du descripteur « Nom »	→	France	Paris	La Marseillaise	632 734
		Chine	Pékin	La Marche des volontaires	9 596 961
		États-Unis	Washington	The Star-Spangled Banner	9 833 517
Un objet	→	Argentine	Buenos Aires	Himno Nacional Argentino	2 791 810

Document 2 : La représentation des données

Le choix de la structure de données est essentiel pour retrouver les informations. Ainsi, selon les éléments que l'on souhaite mettre en avant, on choisira certains descripteurs plutôt que d'autres. Une même donnée peut donc être représentée dans différentes tables et de différentes manières.



Dans l'exemple ci-contre, l'image ci-dessus est un objet dans deux collections, avec des descripteurs différents.

Collection 1			Collection 2		
Nom	Pays	Date	Nom	Paysage	Date
Vacances papy	France	19/12/1960	Vacances papy	Montagne	19/12/1960
New-York	États-Unis	17/05/1901	New-York	Ville	17/05/1901
Ski 2019	France	01/01/2019	Ski 2019	Montagne	01/01/2019
Vacances 2018	France	05/06/2018	Vacances 2018	Plage	05/06/2018

Document 3 : Les formats des données

Selon les besoins, les données peuvent être représentées sous différentes formes. Les formats CSV, JSON et XML sont très utilisés. Le format CSV se présente sous forme de table. Les valeurs des descripteurs sont séparées, ici, par des points-virgules. Le format JSON associe les données avec une étiquette (descripteur) sous forme d'une liste. Le format XML utilise des balises (mot entre "<" et ">") pour organiser les informations en sous-éléments.

Le CSV et le JSON sont simples à écrire et à lire, contrairement au XML qui nécessite de longues lignes de code. Le CSV est rigide (tous les descripteurs doivent être renseignés), contrairement au XML et au JSON. Le XML est le format le plus rapide à traiter par la machine.

Données au format CSV

```
1 prenom;nom;classe
2 Antoine;Ledoux;seconde
3 Pauline;Darcis;premiere
```

Données au format JSON

```
1 {
2   {
3     "prenom": "Antoine",
4     "nom": "Ledoux",
5     "classe": "seconde"
6   }
7   {
8     "prenom": "Pauline",
9     "nom": "Darcis",
10    "classe": "premiere"
11  }
12 }
```

Données au format XML

```
1 <eleve>
2   <prenom>Antoine</prenom>
3   <nom>Ledoux</nom>
4   <classe>seconde</classe>
5 </eleve>
6 <eleve>
7   <prenom>Pauline</prenom>
8   <nom>Darcis</nom>
9   <classe>premiere</classe>
10 </eleve>
```

Définition 3 :

Pour pouvoir stocker, transmettre et exploiter ce type de tableau de valeurs, il peut être représenté sous différentes formes appelées format.

Les principaux formats utilisés pour représenter un ensemble de données sont le CSV, le JSON et le XML.

- L'acronyme CSV signifie « Comma Separated Values », c'est-à-dire « valeurs séparées par une virgule ».
- L'acronyme JSON signifie « JavaScript Object Notation », c'est-à-dire « notation d'objets en Javascript ».
- L'acronyme XML signifie « eXtensible Markup Language », c'est-à-dire « langage de balisage extensible ».

En CSV, les données sont stockées dans un fichier texte :

- Où la première ligne donne la liste des descripteurs,
- Où chaque autre ligne du tableau correspond à un objet,
- Où les valeurs (d'un objet) sont séparées par une virgule.

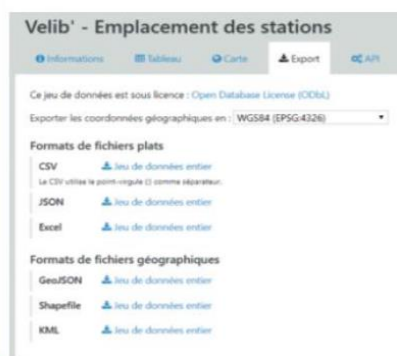
Document 4 : Les données ouvertes

Les données ouvertes, ou *Open Data*, sont des données totalement publiques et libres de droit. De nombreux sites offrent l'accès à ces données. Il ne doit y avoir aucun obstacle technique et aucune restriction à l'utilisation des données, même commerciale. Sur le site opendata.paris.fr, il est par exemple

possible de retrouver l'emplacement exact de toutes les stations de vélos en libre service et d'utiliser ces informations pour produire des cartes ou d'autres documents.



Récupération de données sur l'emplacement des stations de Velib sur le site opendata.paris.fr



	Code de la station	Nom de la station	Etat des stations	Etat du Totem	Nombres
1	6021	Beaux-Arts - Bonaparte	Operative	yes	20
2	9020	Toudouze - Clauzel	Operative	yes	21
3	12109	Mairie du 12ème	Operative	yes	30
4	17026	Jouffroy d'Abbans - Wagram	Operative	yes	0
5	9104	Caumartin - Provence	Operative	yes	22
6	6108	Saint-Romain - Cherche-Midi	Operative	yes	17
7	41301	Bois de Vincennes - Gare	Operative	yes	51
8	15047	Morillons - Dantzig	Operative	yes	52
9	31024	Romainville - Vaillant-Couturier	Operative	yes	38

Remarques :

- Le format XML avec son système de balises est plus long à écrire mais permet une bonne interopérabilité entre des systèmes d'informations hétérogènes. Il est aussi plus rapide à traiter par la machine.
- Attention ! La virgule est le séparateur standard pour les données anglo-saxonnes, mais pas pour les données aux normes françaises puisque, en français, la virgule est le séparateur des chiffres décimaux (= "nombres à virgule"). C'est pourquoi on utilise un autre séparateur : le point-virgule (;)
- Dans certains cas, le fait d'avoir comme séparateur une virgule ou un point-virgule peut engendrer quelques problèmes. Vous devez donc rester vigilants sur le type de séparateur utilisés. Un tableur tel que "Calc" (de Libre Office) est capable de lire les fichiers au format CSV mais d'autres tableurs (comme Excel) gèrent moins bien le séparateur CSV "point-virgule".

A l'aide des documents et des vidéos, répondre aux questions suivantes :

1) Donner les descripteurs et une valeur d'un descripteur.

Descripteurs : « Nom », « Capitale », « Hymne » et « Superficie ». Par exemple, « Pekin » est une valeur du descripteur « Capitale ».

2) Retrouver la date de l'image. Quelle collection permet de retrouver cette information le plus rapidement ?

La deuxième collection permet de déterminer la date (05/06/2018) car on peut constater qu'une seule photo a une plage pour paysage. Avec la première collection, on peut le déduire également, mais plus difficilement. La date élimine les deux premiers objets et le nom laisse penser qu'on peut éliminer le troisième objet.

3) Combien y a-t-il d'objets dans la collection représentée sous forme CSV, XML et JSON ? Quels en sont les descripteurs et comment sont-ils placés dans les différents formats ?

On distingue deux objets et trois descripteurs : « prénom », « nom » et « classe ».

4) Peut-on utiliser librement les informations sur les vélos en libre-service pour faire une application payante qui indique la localisation des stations ?

Sur des sites de données ouvertes à différents formats (CSV, JSON...), tout le monde peut utiliser librement les informations pour faire ce qu'il souhaite, même pour une application commerciale.

5) Conclusion : Quels sont les points communs aux différents formats de données ?

On retrouve le système objet/descripteurs/valeur dans tous les formats (et dans tous les documents).

Collection			Objets
Descripteurs →	Prénom	Nom	
Une valeur du descripteur Prénom →	Antoine	Ledoux	
	Pauline	Darcis	
	Karim	Chanhoun	
Un objet →	Sarah	Goldberg	

