



Données Semi Structurées

Chapitre 1: Contexte et problématique

S. Bella
Bella.salyma@gmail.com

L3 – Systèmes Informatiques

2021/2022

Plan

- 01 Problèmes & Nouveaux Besoins
- 02 Type de Données: données structurées, non structurées et semi Structurées



Problèmes & Nouveaux Besoins



3



Explosion des données manipulées par les SI...

- Les données des utilisateurs, produits, historiques d'opérations,...
- Les documents (Word, multimédias,...).
- Les réseaux sociaux (commentaires, opinions,...).
- Les données géospatiales (google Maps).

4

Les problèmes commencent...

- Plus de données → de plusieurs choix, des bonnes décisions, des meilleures expériences,...
- Comment **stocker** toutes ces données ou **documents** ?
- Comment **échanger** ces données avec les différentes applications des SI ?

5

Document, Document Multimédia et Hyper Document (1/3)

Document :

- Designer/Contenir une **information écrite** (texte, programmes,...), **sonore** (audio, musique,...), **visuelle** (photo, dessin, ...) ou **multimédia** (intègre les données de différentes natures).
- Peut être **physique** (papier, livre,...) ou **électronique** (Word, Excel, image,...).
- Document électronique** est, généralement, un **fichier** stocké sur une **mémoire secondaire** (non volatile) (un disque dur, un disque SSD, un CD-ROM, une clé USB, une disquette,...).
- Organiser et stocker par un **système de fichier** (NTFS, FAT,...).
- Fait partie d'une **hiérarchie**: par exemple, dans une répartition d'un lecteur C, on trouve des dossiers et à l'intérieur de celle-ci on trouve des fichiers. Ce dernier est le seul qui contient du contenu.
- Extension** indique le **type du fichier**: .txt, .docx, .xlsx, .jpg, .png, .mp3, .m4a, .mp4, .mov,...

6

Document, Document Multimédia et Hyper document (2/3)

Document multimédia : un document qui intègre sur un même support des **ressources (données) numériques de différentes natures** (texte, son, image, vidéos, présentation,...). Il peut être se présenter sous forme de :

Papier	Diaporama	Montage audio –vidéo	Site web
			
Publication Assistée par Ordinateur Communiquer sur un support papier en y intégrant des textes, images, graphiques, tableaux ...	Présentation Assistée par Ordinateur Présentation animée avec une succession de diapositives intégrant des textes, images, vidéos, son ... Elle soit attirer l'attention de l'auditoire.	Montage vidéo Son usage permet de réaliser des films avec des incrustations d'images, du texte, du son ...	La publication dans Internet Communiquer avec des pages web consultables à distance au contenu varié avec des textes, images, vidéos ...

Le **choix du forme** (mode de représentation) est choisi en fonction de **son usage** et selon des **contraintes** comme le temps, le lieu, l'objectif, le public ciblé, le matériel de l'événement, ...

7

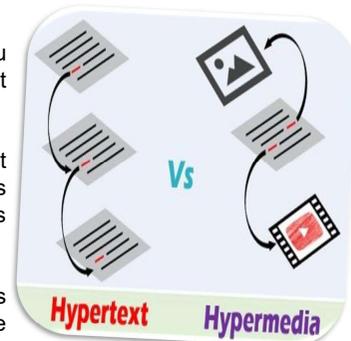
Document, Document Multimédia et Hyper document (3/3)

Hyper documents:

- Il désigne un **hypertexte** ou un **hypermédia**,
- Un produit/document, à créer ou à consulter, dont le contenu est **informatisé** (toutes les informations sont numérisées) et organisé de manière **non linéaire**.

Hypertexte: **système de renvois** permettant de passer directement d'une partie d'un document à une autre partie (informations textuelles) via des **liens ou hyperliens** ou d'un document à d'autres documents.

Hypermédia: il est une **extension de hypertexte** à des **données multimédias**, qui ajoute aux informations de type texte d'autres données multimédias. Il permet d'inclure des hyperliens vers d'autres textes, mais aussi vers des documents multimédias.



8

Stockage de Données

Différentes solutions:

- Les structures volatiles (tableaux, matrices, ...),
- Les fichiers (textes, multimédias, ...),
- Les **bases de données**,
- Les DataWarehouse (BDD relationnelle hébergée sur un serveur dans un Data Center ou dans le Cloud).



9

Base de Données

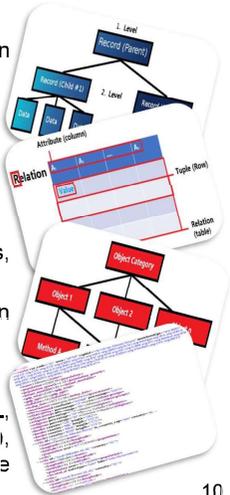
Collection structurée de données bien organisées et mémorisées sur un support permanent.

Pièce centrale (mémoire centralisée) d'un système d'information ou d'un système de base de données.

Utilisation du **SGBD** (Système de Gestion de Base de Données):

- Gestion de la manipulation des données: l'authentification des utilisateurs, insertion/extraction des données, modification/suppression de données,...
- Garantir la qualité, la pérennité et la confidentialité des informations, tout en cachant la complexité des opérations.

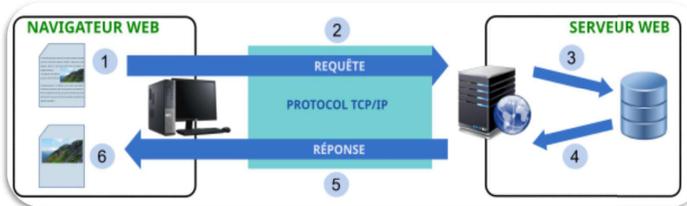
Modèle de base de données: hiérarchique(IMS, DB,...), réseau (CODASYL, UDS,...), relationnel(ORACLE, MySQL,...), orienté-objet (db4o, ObjectStore,...), NoSQL (Hbase, MangoDB,...), orienté texte (fichier plat: fichier texte ou du texte combiné avec un fichier binaire contenant un seul enregistrement par ligne),...



10

Echange des Données du WEB

Exemple: une page web avec du texte et une image qui est affichée dans le navigateur.



1. L'utilisateur clique sur le lien hypertexte de l'image avec l'intention de l'afficher dans sa taille d'origine;
2. La requête **HTTP** (fichier requête) est envoyée par le navigateur (Ex: chrome) sur internet. Ce fichier circule sur internet en utilisant le protocole **TCP/IP**;
3. Le serveur WEB analyse puis localise le fichier demandé;
4. En générale, 2 cas de figures existent : si le fichier demandé est trouvé dans la BDD (ici une image, mais cela peut être un **fichier HTML, une vidéo...**), un fichier de réponse est construit avec une en-tête de réponse valide et l'image. Sinon, un fichier de réponse est construit avec une en-tête de réponse HTTP non valide;
5. Le fichier réponse est envoyé au navigateur avec plusieurs paramètres et bien sûr la réponse elle-même;
6. Le navigateur analyse la réponse afin d'afficher, dans notre exemple, l'image. Si l'image n'est pas trouvée : un message Erreur 404 sera alors affiché.

11

Nouveaux Besoins: Intégrer/Echanger de Données Hétérogènes

Les **nouvelles applications** du Web (marketing digital du e-commerce (réseaux sociaux, chatbots,...), e-business (écrire/vendre des e-books, créer une boutique en ligne de vendre des produits d'autres sites/personnes par dropshipping,...), e-tourisme,...) ont besoin d'**intégrer** des **données hétérogènes**:

- Données des base de données (relationnelles, orienté textes,...),
- Données des documents (Word, HTML, CSV,...),
- Données multimédia (photos, dessins, audio, vidéo, ...),
- Données spatiale.

Mais comment traduire des données de BDD vers HTML, Word,... ?

12

Solution

Nécessité d'un **modèle général et souple**, «sans schéma», avec un **langage de requêtes** associé pour faire le traitement et l'échange des données : **modèles semi-structurés**.

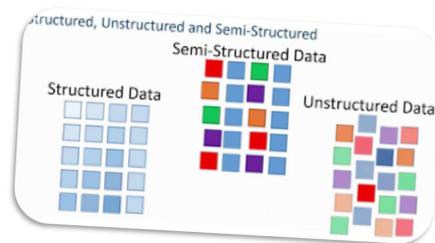
13

Types de Données : Structurées, Non Structurées et Semi Structurées

Types de Données

Trois **types de sources de données** peuvent être considérés pour les systèmes d'information ouverts sur le Web :

- Données **structurées**,
- Données **semi structurées**,
- Données **non structurées**.



15

Données Structurées

- Informations qui ont été formatées et transformées en un **modèle de données bien défini** (structure régulière).
- **Schéma de structure bien précis** qui permet **l'accès et le traitement efficace** des données, et la **facilité** de les analyser.

- **Exemple parfait**: les bases de données relationnelles
→ constituer 20% du total des données des entreprises.

ID	Name	Age
1	Alex	22
2	Bob	24
3	Emily	32

- Les données structurées sont générées à la fois par les humains et les machines (ex: codes-barres).

- **Complexe & difficile à appliquer sur des données de type ou format hétérogène.**



16

