

Sciences humaines, sciences sociales à l'ère numérique : la mise en données et méthodes de modélisation des connaissances

Stéphane Pouyllau

Ingénieur de recherche hors classe CNRS

Professeur attaché à l'université d'Evry-Paris-Saclay

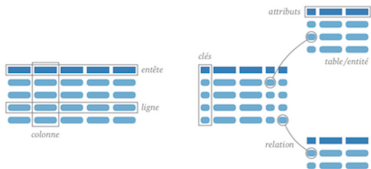
Orcid : 0000-0002-9619-1002

Séance du 18 novembre 2024

Introduction

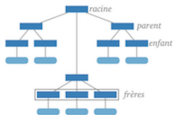
- ▶ Objectifs du cours
- ▶ Importance de la structuration des données
- ▶ Aperçu des formats CSV, SGBDR, XML (TEI), et RDF

Introduction



Données tabulaires

Les données d'un livre sont structurées sur une ligne, composée d'un ensemble de valeurs venant de différents colonnes. Chaque colonne contient le même type de données et l'en-tête indique leur nom.



Langages de balisage

Les documents XML ont une structure hiérarchique, comme un arbre inversé. Chaque élément peut avoir un ou plusieurs enfants mais il y a exactement un seul « root élément » ou racine.

Modèle relationnel

Les données sont structurées à l'aide de tables, qui disposent chacune d'un ensemble d'attributs. Les données d'une table peuvent être liées à une autre table en utilisant les clés primaires.



RDF

Chaque information est exprimée sous la forme d'un triplet, qui connecte un sujet à un objet grâce à un prédicat. Cela donne un graphe qui peut prendre n'importe quelle forme.

Figure: Seth van Hooland & al., 2016.

Contexte et applications

- ▶ Utilisations courantes de chaque format
- ▶ Exemples d'industries et de cas d'utilisation

Introduction à CSV

- ▶ Définition et historique
- ▶ Structure de base d'un fichier CSV

Syntaxe CSV

- ▶ Séparateurs de colonnes et de lignes
- ▶ Gestion des valeurs manquantes

Avantages et inconvénients de CSV

- ▶ Simplicité et lisibilité
- ▶ Limitations en termes de structure et de types de données

Exercice pratique : CSV

- ▶ Créer un fichier CSV
- ▶ Importer et manipuler des données CSV avec un outil de tableur (Excel, Google Sheets)

Introduction au modèle relationnel

- ▶ Définition et historique
- ▶ Concepts de base : tables, colonnes, lignes, clés primaires et étrangères

Syntaxe SQL

- ▶ Création de tables
- ▶ Insertion de données
- ▶ Requêtes de sélection

Avantages et inconvénients du modèle relationnel

- ▶ Intégrité des données et transactions
- ▶ Complexité de gestion des relations

Exercice pratique : SGBDR

- ▶ Créer un petit schéma relationnel
- ▶ Exécuter des requêtes SQL pour manipuler les données

Introduction à XML

- ▶ Définition et historique
- ▶ Structure de base d'un document XML

Syntaxe XML (TEI)

- ▶ Éléments et attributs
- ▶ Balises et nœuds

Avantages et inconvénients de XML

- ▶ Flexibilité et extensibilité
- ▶ Complexité et verbosité

Exercice pratique : XML (TEI)

- ▶ Créer un petit document XML TEI
- ▶ Valider le document XML avec un outil en ligne

Introduction à RDF

- ▶ Définition et historique
- ▶ Concepts de base : triples, sujets, prédicats, objets

Syntaxe RDF

- ▶ Représentation en XML, Turtle, JSON-LD

Avantages et inconvénients de RDF

- ▶ Flexibilité et interopérabilité
- ▶ Complexité de gestion des données

Exercice pratique : RDF

- ▶ Créer un petit document RDF
- ▶ Valider le document RDF avec un outil en ligne

Résumé des points clés

- ▶ Récapitulatif des formats CSV, SGBDR, XML (TEI), et RDF
- ▶ Comparaison des avantages et inconvénients