Tableur et base de données

Cycle de vie des données, mai 2022

Plan

- (Acquisition de données sur le web)
- Organisation des données
- Nettoyage des données
- Modélisation

- Différences tableur / base de données relationnelles
- Formats

https://github.com/clement-plancq/formation-donnees-msh

“Raw data” is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.

« Raw Data » is an Oxymoron. Lisa Gitelman (dir.), Cambridge, MIT Press, 2013

Décidément, on ne devrait jamais parler de « données » mais d’« obtenues »

Petites leçons de sociologie des sciences, Bruno Latour, 1993

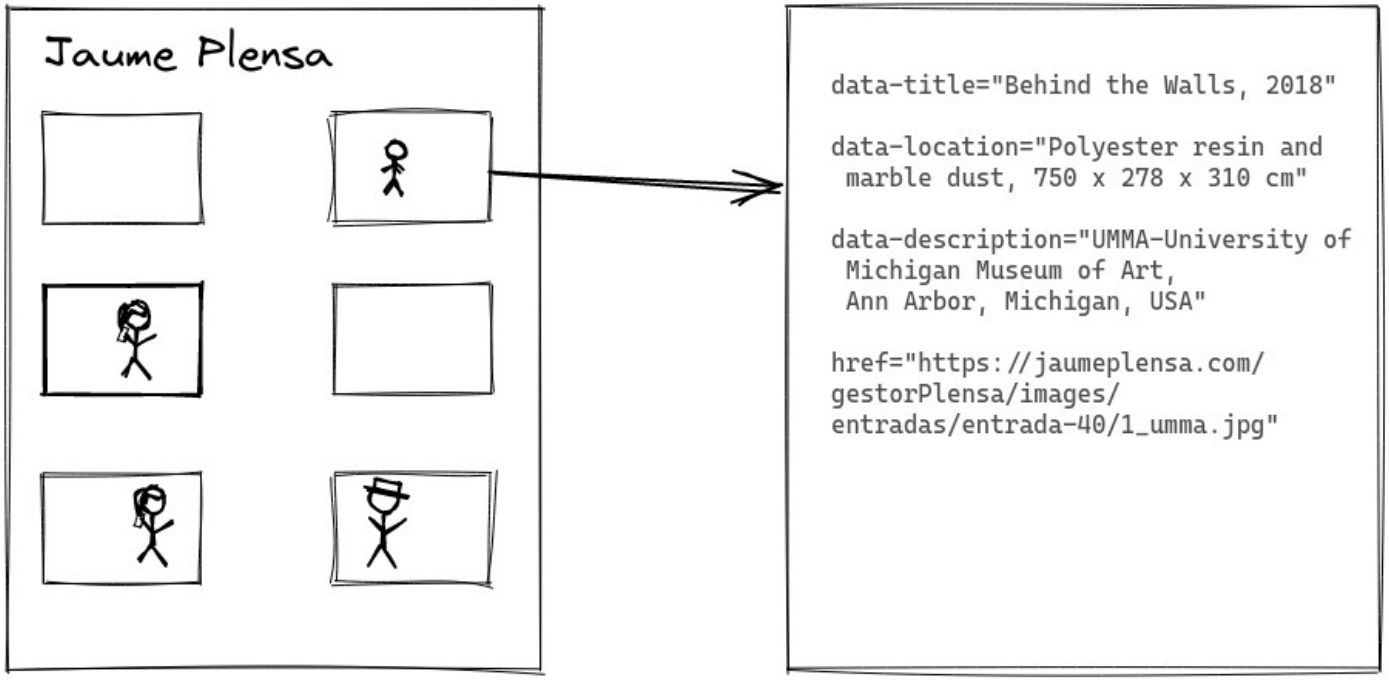
Jaume Plensa

https://jaumeplensa.com/works-and-projects/sculpture

Catalogue, liste, tableau

- Objectifs du catalogue :
  - pouvoir répondre à des questions comme : combien de sculptures entre 2015 et 2020 ? dans combien de pays d'Europe différents ?
  - Pouvoir ajouter des informations comme : sculpture exposée en extérieur / intérieur

- Données déjà organisées en liste et attributs en HTML
  - Comment passer à un tableur ?



Titre	Matériaux	Taille	Lieu	Image
Behind the Walls, 2018	Polyester resin and marble dust	750 x 278 x 310 cm	UMMA-University of Michigan Museum of Art, Ann Arbor, Michigan, USA	https://jaumeplensa.com/gestorPlensa/images/entradas/entrada-40/1_umma.jpg

jaume-plensa-0.csv

Titre	Description	Lieu	Image
Talaia, 2017	Polyester resin and marble dust, 1400 x 478 x 703 cm	Lanai, Hawaii, USA 2021	https://…
Behind the Walls, 2018	Polyester resin and marble dust, 750 x 278 x 310 cm	UMMA-University of Michigan Museum of Art, Ann Arbor, Michigan, USA	https://…
Nuria, 2017	Stainless steel, 350 x 275 x 344 cm	Philadelphia Museum of Art, Philadelphia, PA, USA. Gift of Aileen and Brian Roberts	https://

Organisation, nettoyage des données

Tidy data :

- Every column is a variable
- Every row is an observation
- Every cell is a single value

OpenRefine

- Logiciel de nettoyage et mise en forme des données
- Logiciel libre (Licence BSD)

OpenRefine

- (Détecter des cellules avec des valeurs nulles)
- Détecter des valeurs dupliquées
- Supprimer des espaces en début et/ou fin de cellule
- Séparer des cellules multi-valeurs (‘titre’, ‘description’, ‘lieu’)
- Rassembler des colonnes
- Facets and clustering
- Utiliser des expressions régulières

- OpenRefine garde la trace des modifications
  - On peut revenir en arrière et reproduire le traitement opéré sur les données initiales

Modélisation

Une œuvre peut être exposée à plusieurs endroits, peut être associée à plusieurs images

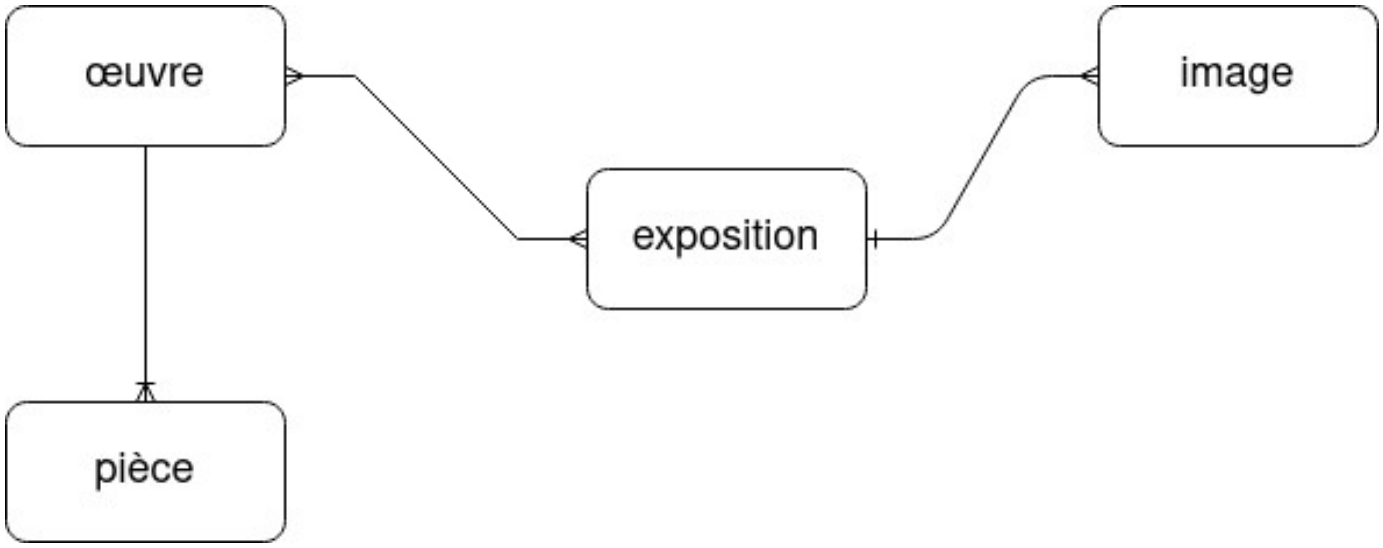
Titre	Expo 1	Expo 2	Image expo 1	Image expo 2
Behind the Walls, 2018	Ann Arbor, Michigan, USA	New York, USA	https://	https://

Modélisation

Une œuvre peut être exposée à plusieurs endroits, peut être associée à plusieurs images

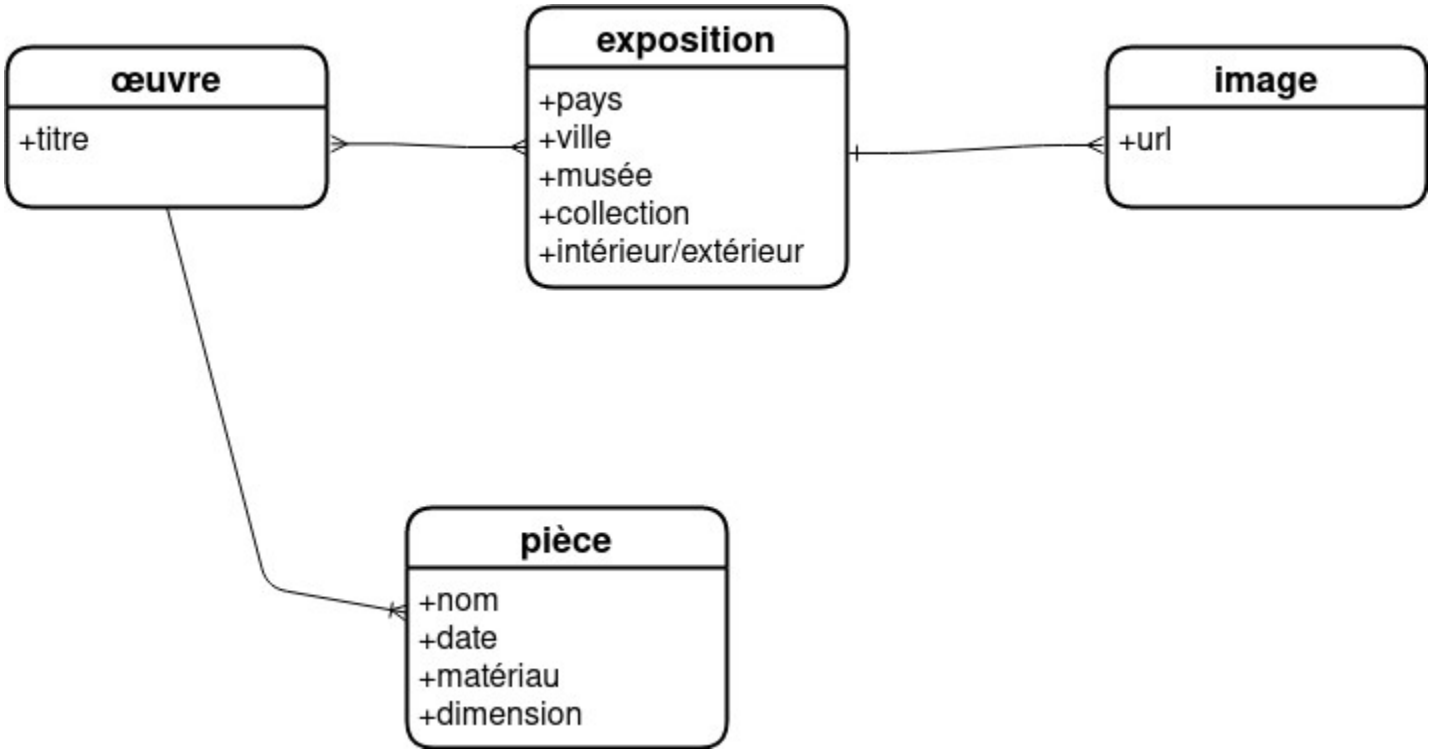
Titre	Expo 1	Image expo	Matériau	Date
Behind the Walls, 2018	Ann Arbor, Michigan, USA	https://	…	…
Behind the Walls, 2018	New York, USA	https://	…	…

Modélisation



Modèle conceptuel de données (MCD) de type entité-association

Modélisation



Modèle logique de données (MLD)

"oeuvre"	
PK	**"id" INTEGER**
	"titre" TEXT

"piece"	
PK	**"id" INTEGER**
	"nom" TEXT
	"date" INTEGER
	"materiau" TEXT
	"dimension" TEXT
	"oeuvre" INTEGER
	FOREIGN KEY("oeuvre") REFERENCES "oeuvre"("id

"exposition"	
PK	**"id" INTEGER**
	"oeuvre" INTEGER
	"pays" TEXT
	"ville" TEXT
	"musee" TEXT
	"collection" TEXT
	"int_ext" TEXT
	FOREIGN KEY("oeuvre") REFERENCES "oeuvre"("id

"image"	
PK	**"url" TEXT**
	"expo" INTEGER
	FOREIGN KEY("expo") REFERENCES "exposition"("

Implémentation dans SQLite

Formats

- OpenDocument (`.ods`), LibreOffice, format ouvert, documenté, fichier binaire
- Office Open XML (`.xlsx`), Microsoft Office, format ± ouvert, documenté, fichier binaire

- csv, tsv, format ouvert, fichier texte. C’est le format d’import/export pour les données tabulaires
- Json, format ouvert, fichier texte. Pour les données structurées

Avantages d’un SGBD

- Contrôle de l'intégrité des données
- Contrôle de la conformité des données saisies (typage)
- Gestion des accès concurrents, plusieurs utilisateurs en même temps
- Gestion de gros volumes de données
- Indexation systématique, rapidité des recherches
- Pas d'information redondante ou invalide

Inconvénients d’un SGBD

- Pas ou peu d’interfaces utilisateur clé en main
- Suppose du développement logiciel
- Pas d’export des données facile à part SQL : prévoir une stratégie d’export des données à plat