Représentation et échange de données

3ème année du Cycle Ingénieur en Informatique par Apprentissage

Aurélien Max

Année 2020-21

Projet de groupe #1

Production d'une ressource en données ouvertes et FAIR

(v2 du 13.01.2021)

L'ouverture des données (*reprise*)

- Besoins forts de transparence, partage, accès universel
  - émergence des initiatives **données ouvertes** (**open data**)
    - conception de l'information publique comme un bien commun
    - diffusion de données structurées selon une licence ouverte garantissant un libre accès et une réutilisation sans restrictions
    - nombreuses initiatives, ex. `https://www.data.gouv.fr`
  - applications en **science ouverte**
    - partage et documentation des données issues de la recherche
    - réponse nécessaire aux problèmes de **reproductibilité**

- Besoins de garanties sur les données
  - ex. principes des données **FAIR** (*Findable, Accessible, Interoperable, Reusable*)
    - Findable : description par métadonnées riches, utilisation d'identificateurs uniques et persistants
    - Accessible : utilisation de licences claires (pas nécessairement données ouvertes), métadonnées toujours accessibles
    - Interoperable : formats fondés sur des standards établis
    - Reusable : métadonnées de provenance, données vérifiées et de qualité

Rôle du projet

- Sensibilisation aux principes et intentions des **données ouvertes** et des **données FAIR** (*Findable, Accessible, Interoperable, Reusable*)
- Découverte et exploration de **répertoires** et **sources de données**
- Familiarisation avec des **jeux de données** dans un domaine d'intérêt choisi (ex. vie urbaine, télécommunication, agriculture, alimentation, géographie)
- **Ingestion de jeux de données** pour la **production d'un jeu de données original**
- Analyse critique des principes suivis et de la ressource produite et compte-rendu

Cahier des charges

- Ingestion d'au moins **3 jeux de données distincts** (idéalement dans des formats différents)
- Le jeu de données produit doit correspondre à un **ensemble de données original**
  - celui-ci pourrait être obtenu par morceaux via un service web mais sera rendu disponible sous forme de fichiers
  - certaines données exposées pourront être obtenues par calcul (ex. distances entre points géographiques)
  - les formats pour décrire le jeu de données et son schéma sont à proposer/défendre (ex. `JSON`, `XML`)
- Aucune exploitation des données n'est attendue
  - (*en option, possibilité de proposer des visualisations*)
- Liberté totale sur les moyens techniques (langages de programmation)
- Travail uniquement sur des **données ouvertes**
  - éviter tout problème de licence et d'accès
- Respect *autant que possible* des principes FAIR
  - (*en limitant l'effort : projet pédagogique*)

Aspects pratiques

- Liste des projets par groupe
  1. Fréquentation des salles de cinéma
  2. Indices de qualité des jeux vidéo
  3. Modes de déplacements doux pour l'accès aux loisirs
  4. Covid-19 et capitalisations boursières

- Échéance du rendu : date de rendu globale pour le module
- Composition du rendu
  - base de code (lien ou archive)
  - lien vers données source utilisées
  - données produites et schéma
  - court rapport du projet sous forme de présentation (avec des transparents) au format PDF (cf. transparent suivant)

Présentation/rapport de projet

1. Description critique des caractéristiques du projet
   - sources de données utilisées
   - questions liées à l'ingestion des données
   - format produit
   - métadonnées de provenance des données regroupées ou calculées
   - pérennité de la stratégie automatique de la construction de la ressource
   - questions liées à la qualité/fiabilité de la ressource produite

2. Discussion (principalement théorique) de l'applicabilité des principes FAIR au projet
   - objectifs poursuivis par la mise à disposition des données
   - plan de gestion de données (*ce qu'il faudrait faire pour être FAIR*)
   - faisabilité à discuter des possibilités suivantes :
     - mise à jour (périodique) des données, mise à jour du schéma des données, invalidation/correction de données, contribution de données par des tiers