

Projet Jeux Olympiques - Journal de bord

T. Burnel, N. Grim, M. Griveau, M. Mechentel

Janvier 2024

1 Introduction

Nous sommes une équipe de datajournalistes chargée d'étudier l'existence ou non de liens entre le succès d'un pays aux Jeux Olympiques et sa richesse.

2 Jeux de données

3 Traitement des données

3.1 Objectif du traitement

3.2 Chaîne de traitement

3.2.1 Semaine du 1^{er} janvier 2024

Au cours du traitement, nous avons réalisé que le jeu de données de Kaggle ne représentait pas les pays grâce aux codes basés sur la norme ISO 3166 mais aux codes du CIO (Comité International Olympique). Nous avons

donc adapté notre requête SPARQL pour les obtenir pour être en mesure de réaliser une jointure en bout de chaîne de traitement.

Également, le jeu de données du FMI est trop avare en données fiables (données manquantes, valeurs aberrantes¹) à propos des investissements en pourcentage du PIB dans le domaine du sport. Nous avons donc fait le choix d'utiliser les valeurs absolues² que nous croiserons avec le PIB par habitant de chaque pays. En somme, nous pallions les données fautives du FMI.

Pour ce faire, nous avons de prime abord songé à rédiger une autre requête SPARQL. Nonobstant, le résultat ne concernait que les années 2021 et 2022 – les années précédentes sont vraisemblablement absentes de Wikidata.

Nous avons donc dû trouver un jeu de données contenant le PIB par habitant. Parmi les données sources de Wikidata se trouvait un jeu hébergé sur le site de la Banque mondiale³. Les données correspondent à une profondeur temporelle suffisante pour le cadre de notre projet. Cela nous a permis de résoudre le problème de mauvaise qualité du jeu du FMI.

3.2.2 Requête SPARQL

Pour enrichir nos données *via* Wikidata, notre objectif était de mettre au point une requête SPARQL retournant le nombre d'habitants de tous les pays du monde sur trente ans (1993 - 2003).

```
SELECT ?paysLabel ?population ?date  
WHERE
```

¹Donner des exemples.

²???

³https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_year_desc=false&view=map

```

{
?pays wdt:P31 wd:Q6256.
?pays p:P1082 ?populationStatement.
?populationStatement ps:P1082 ?population.
?populationStatement pq:P585 ?date.
FILTER(YEAR(?date) >= (YEAR(NOW()) - 30)).
SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],fr". }
}
ORDER BY ?paysLabel ?date

```

Être en mesure de requêter la liste de l'ensemble des pays a été la première étape de la construction de notre requête. Le premier triplet utilise la variable inconnue **?pays** dont l'objet est **country** (wd :Q6256) :

```
?pays wdt:P31 wd:Q6256.
```

La deuxième partie de la requête doit retourner la liste du nombre d'habitants de chaque pays en prenant en compte une dimension chronologique. La complexité de cette demande correspond à un parcours de graph en quatre temps – et non pas en trois.

Nous avons donc créé une nouvelle variable, **?populationStatement**, contenant toutes les propriétés utilisées dans la classe **population** grâce au préfixe **p**⁴. Pour obtenir le nombre d'habitants, nous avons utilisé le préfixe **ps** permettant d'obtenir la valeur de la propriété relative à la population :

⁴Pour ce faire, nous avons consulté la liste de préfixes Wikidata : <https://www.wikidata.org/wiki/EntitySchema:E49>.

```
?pays p:P1082 ?populationStatement.  
?populationStatement ps:P1082 ?population.
```

Enfin, nous avons utilisé le préfixe **pq** pour récupérer la valeur chronologique dans la variable **?date**. Un filtre a été appliqué pour exprimer les limites de notre période, soit **NOW** pour 2023 et **-30** pour 1993 :

```
?populationStatement pq:P585 ?date.  
FILTER(YEAR(?date) >= (YEAR(NOW()) - 30)).
```

La toute dernière ligne de la requête – couplée à la toute première – permet de contraindre l’affichage des résultats. Lorsque la requête s’exécute, le service Wikidata renvoie le nom de chaque pays (**?paysLabel**) par ordre alphabétique, le nombre d’habitants (**?population**) et l’année correspondante (**?date**) par ordre croissant :

```
SELECT ?paysLabel ?population ?date  
ORDER BY ?paysLabel ?date
```

Nous nous sommes heurtés à plusieurs erreurs avant de rendre la requête fonctionnelle. Nous avons dû nous éloigner du parcours de graphique à trois étapes vu en cours et comprendre la nécessité de recourir à la variable **?paysStatement**. La documentation a été d’un grand secours pour nous aider à saisir l’utilisation des propriétés **population**⁵ et **date**⁶.

Nous avons également pris en exemple la requête *Population in Europe after 1960*. La consulter a été l’occasion d’un meilleur apprentissage sur les

⁵ *Confer.* <https://www.wikidata.org/wiki/Property:P1082>.

⁶ *Confer.* <https://www.wikidata.org/wiki/Property:P585>.

propriétés Wikidata, une porte ouverte à la lecture de la documentation et à l'appréhension des requêtes en deux temps : d'abord la collecte des propriétés d'une classe puis le requêtage de chacune d'elles en fonction de notre besoin.

Tel que mentionné *supra*⁷, nous avons adapté notre requête pour récupérer les codes CIO des pays, et ce grâce à l'ajout d'une unique ligne :

```
?pays wdt:P984 ?cio.
```

Voici, en somme, notre requête SPARQL :

```
SELECT ?paysLabel ?population ?date ?cio
WHERE
{
  ?pays wdt:P31 wd:Q6256.
  ?pays wdt:P984 ?cio.
  ?pays p:P1082 ?populationStatement.
  ?populationStatement ps:P1082 ?population.
  ?populationStatement pq:P585 ?date.
  FILTER(YEAR(?date) >= (YEAR(NOW()) - 30)).
  SERVICE wikibase:label { bd:serviceParam wikibase:language
    "[AUTO_LANGUAGE],fr". }
}
ORDER BY ?paysLabel ?date
```

⁷Mettre le renvoi

4 Visualisation des données

4.1 Présentation des visualisations

4.2 Analyse

Qu'apprend-on en regardant les visualisations ? Quels sont les biais ?

Table des matières

1	Introduction	1
2	Jeux de données	1
3	Traitement des données	1
3.1	Objectif du traitement	1
3.2	Chaîne de traitement	1
3.2.1	Semaine du 1 ^{er} janvier 2024	1
3.2.2	Requête SPARQL	2
4	Visualisation des données	6
4.1	Présentation des visualisations	6
4.2	Analyse	6