

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, à l’approche des Jeux Olympiques de Paris, d’étudier l’existence ou non de liens entre le succès d’un pays aux Jeux Olympiques et sa richesse. Notre rôle sera de comparer le nombre de médailles remportées par pays et par sport à diverses éditions des Jeux pour cerner la présence ou l’absence d’une logique économique.

Notre hypothèse de départ est la suivante : les pays développés gagnent significativement plus de médailles en raison de leur niveau d’investissement dans le sport. À l’aune de l’examen et du traitement des données, nous avons pour ambition de déterminer si la politique d’investissement dans les infrastructures sportives de ces pays influe – tant positivement que négativement – sur leurs résultats aux différentes épreuves des Jeux Olympiques.

Notre travail consistera, dans un deuxième temps, en une analyse comparative de la politique d’investissement du Royaume-Uni et de la France. Plusieurs points de convergence distinguent ces deux pays : leur nombre d’habitants, leur économie et leur qualité de pays organisateur des Jeux.

Le résultat du traitement des données constituera une base à la réalisation de datavisualisations et d’une application web pouvant servir de base à l’écriture d’articles journalistiques.

2 Jeux de données

2.1 Jeux collectés

En amont du traitement des données, nos jeux étaient au nombre de quatre et présentent les informations suivantes :

- Nombre de médailles remportées par sport et par pays sur 120 ans¹ ;
- PIB par habitant, investissements dans le domaine sportif² ;
- Investissements dans les infrastructures sportives en France³ ;
- Investissements dans les infrastructures sportives au Royaume-Uni⁴ .

Un cinquième jeu a dû faire son entrée au cours du traitement, pour pallier la maigre qualité des données fournies par le Fonds Monétaire International (*confer infra.* p. 7). Ce jeu provient du site de la Banque mondiale.⁵

1. *120 years of Olympic history : athletes and results*. URL : <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results> (visité le 17/01/2024).

2. *Download entire World Economic Outlook database, April 2019*. IMF. URL : <https://www.imf.org/en/Publications/WE0/weo-database/2023/October/download-entire-database> (visité le 17/01/2024).

3. *Data ES - Base de données*. URL : <https://equipements.sports.gouv.fr/explore/dataset/data-es/information/> (visité le 17/01/2024).

4. *Active Places Power*. URL : <https://www.activeplacespower.com/> (visité le 20/01/2024).

5. *World Bank Open Data*. World Bank Open Data. URL : <https://data.worldbank.org> (visité le 20/01/2024).

2.1.1 Jeu de Kaggle

Analyse du jeu à fournir.

2.1.2 Jeu du FMI

Analyse du jeu à fournir.

2.1.3 Jeu du ministère des sports et des jeux olympiques et paralympiques

Analyse du jeu à fournir.

2.1.4 Jeu de Sport England

Analyse du jeu à fournir.

2.1.5 Jeu de la Banque mondiale

Analyse du jeu à fournir.

2.2 Enrichissement Wikidata *via* requête SPARQL

Dans l'optique d'enrichir nos données *via* Wikidata, nous avons mis au point une requête SPARQL à même de renvoyer le nombre d'habitants de l'ensemble des pays du monde sur une période de trente ans (1993 - 2003). En voici la première version :

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

```

?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 l'intégralité des pays a été la première étape de la construction de notre requête. Le premier triplet permet de spécifier la nature – notée `wdt:P31` – de notre variable inconnue `?pays` en la faisant correspondre à l'objet `country` – noté `wd:Q6256` :

```

?pays wdt:P31 wd:Q6256.

```

La deuxième partie de notre requête doit renvoyer le nombre d'habitants par pays en prenant en compte une dimension chronologique. La complexité de cette demande requiert un parcours de graphique en quatre temps.

Pour ce faire, nous avons consulté la liste de préfixes⁶ de Wikidata afin de créer une nouvelle variable, `?populationStatement`. Le deuxième triplet a recours à la classe `population` – notée `p:P1082` – et signifie que la variable `?populationStatement` a pour valeur la population des pays :

```

?pays p:P1082 ?populationStatement.

```

6. *Wikidata prefixes (E49)* - Wikidata. URL : <https://www.wikidata.org/wiki/EntitySchema:E49> (visité le 20/01/2024).

L'obtention du nombre d'habitants a nécessité le recours au préfixe `ps`, lequel permet de récupérer la valeur de la propriété relative à la population :

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

Enfin, nous avons rédigé le dernier triplet sur la base du préfixe `pq` pour attribuer la valeur chronologique à la variable `?date`. Un filtre y a été appliqué pour exprimer les limites extrêmes 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**⁷ et **date**.⁸

Nous avons également pris en exemple la requête *Population in Europe after 1960*.⁹ La consulter a été l'occasion d'un meilleur apprentissage sur les 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 *infra* (page 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". }
```

7. *population*. URL : <https://www.wikidata.org/wiki/Property:P1082> (visité le 20/01/2024).

8. *point in time*. URL : <https://www.wikidata.org/wiki/Property:P585> (visité le 20/01/2024).

9. *Wikidata Query Service*. URL : <https://w.wiki/8uHH> (visité le 20/01/2024).

```
}  
  
ORDER BY ?paysLabel ?date
```

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 code 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 aberrantes¹⁰) à propos des investissements en pourcentage du PIB dans le domaine du sport. Nous avons donc fait le choix d'utiliser les valeurs absolues¹¹ 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.

10. Donner des exemples.

11. ?

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 mondiale.¹² 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.

4 Visualisation des données

4.1 Présentation des visualisations

4.2 Analyse

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

12. *World Bank Open Data.*

5 Sitographie

120 years of Olympic history : athletes and results. URL : <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results> (visité le 17/01/2024).

Active Places Power. URL : <https://www.activeplacespower.com/> (visité le 20/01/2024).

Data ES - Base de données. URL : <https://equipements.sports.gouv.fr/explore/dataset/data-es/information/> (visité le 17/01/2024).

Download entire World Economic Outlook database, April 2019. IMF. URL : <https://www.imf.org/en/Publications/WE0/weo-database/2023/October/download-entire-database> (visité le 17/01/2024).

point in time. URL : <https://www.wikidata.org/wiki/Property:P585> (visité le 20/01/2024).

population. URL : <https://www.wikidata.org/wiki/Property:P1082> (visité le 20/01/2024).

Wikidata prefixes (E49) - Wikidata. URL : <https://www.wikidata.org/wiki/EntitySchema:E49> (visité le 20/01/2024).

Wikidata Query Service. URL : <https://w.wiki/8uHH> (visité le 20/01/2024).

World Bank Open Data. World Bank Open Data. URL : <https://data.worldbank.org> (visité le 20/01/2024).

Table des matières

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 2 | Jeux de données | 2 |
| 2.1 | Jeux collectés | 2 |
| 2.1.1 | Jeu de Kaggle | 3 |
| 2.1.2 | Jeu du FMI | 3 |
| 2.1.3 | Jeu du ministère des sports et des jeux olympiques et paralympiques | 3 |
| 2.1.4 | Jeu de Sport England | 3 |
| 2.1.5 | Jeu de la Banque mondiale | 3 |
| 2.2 | Enrichissement Wikidata <i>via</i> requête SPARQL | 3 |
| 3 | Traitement des données | 7 |
| 3.1 | Objectif du traitement | 7 |
| 3.2 | Chaîne de traitement | 7 |
| 3.2.1 | Semaine du 1 ^{er} janvier 2024 | 7 |
| 4 | Visualisation des données | 8 |
| 4.1 | Présentation des visualisations | 8 |
| 4.2 | Analyse | 8 |
| 5 | Sitographie | 9 |