Mini-projet - Données semi structurées et sémantique

1. Données

J'utilise pour ce projet les données <u>Formula 1 World Championship</u> provenant de Kaggle. Ce jeu de données regroupe des données sur les circuits, les pilotes, les constructeurs, les courses, les résultats des courses, les temps aux tours, les temps des pit stops, et d'autres données sur la Formule 1. Les données sont collectées à partir du site <u>Ergast</u>, qui est un service permettant d'accéder aux données historiques de la Formule 1. Les données sont stockées au format CSV et possèdent des clés pour lier les différents objets. De plus, de nombreux liens wikipedia sont liés à certains objets.

Je fais une sous-sélection des fichiers à utiliser : "races.csv", "results.csv", "drivers.csv", "circuits.csv", et "status.csv".

2. Modélisation en RDF

"races.csv" → Contient les informations sur les courses.

Type: dbo:GrandPrix

Clé: "race_" + identifiant unique de la course

Utilisation de noeud blanc pour représenter les différents temps de la course.

x >-:fp1_date→	✓ x L: fp1_date Add object	x >-:fp1→	▼ 🗴 B: Blank 🗶 :DateTime Add type	▼ x >-:date-> x >-:time->	▼ x L: fp1_date Add object ▼ x L: fp1_time Add object
x >-:fp1_time→ x >-:fp2_date→	▼ x L: fp1_time Add object ▼ x L: fp2_date	x >-:fp2→	Add object Add object B: Blank C: DateTime	Add property ▼ × >-:date->	▼ x L: fp2_date Add object
x >-:fp2_time→	Add object ▼ L: fp2_time Add object		Add type	→:time→ Add property	▼ & L: fp2_time Add object
x >-:fp3_date→	▼ L: fp3_date Add object	x >-:fp3→	▼ 🗶 B: Blank ★ :DateTime Add type	▼ 	▼ x L: fp3_date Add object ▼ x L: fp3_time Add object
x >-:fp3_time→ x >-:quali_date→	✓ x L: fp3_time Add object ✓ x L: quali_date	x >−:qualification→	Add object ▼ B: Blank	Add property ▼ ×>-:date→	▼ 🗴 L: quali_date
x >-:quali_time→	Add object L: quali_time Add object		★ :DateTime Add type	➤-:time→ Add property	Add object L: quali_time Add object
x>-:sprint_date->	▼ x L: sprint_date Add object	x >-:sprint→	Add object ▼ 🗴 B: Blank ★ :DateTime	▼ ※ >-:date>	▼ 🗴 L: sprint_date Add object
➤-:sprint_time→	▼ x L: sprint_time Add object		Add type	→:time→ Add property	▼ x L: sprint_time Add object

"results.csv": Résultat des courses par pilote. On a comme information: le numéro du pilote, sa position sur la grille de départ, sa position en fin de course, le nombre de point gagné, le nombre de tours réalisé, le temps réalisé au total en milliseconde, le numéro du meilleur tours, le range au classement pilote, le temps du meilleurs tours, et le "status" (type de fin de course, finis, panne moteur, accident, ...)

Type: :Result

Clé: "result_" + identifiant unique du résultat d'un pilote par course

"circuits.csv": On retrouve comme information le nom, le pays, la position GPS et l'altitude des circuits ainsi qu'un lien wikipedia vers la page du circuit.

Type: :Circuit

Clé : "circuit_" + identifiant unique d'un circuit

J'ai transformé la colonne "country" en IRI pointant vers le prefix **dbr:** qui permet de lier la page dbpedia d'un pays. La propriété "country" attend donc une valeur de type **dbo:Country**.

De la même manière pour "Location" indiquant une ville, j'ai lié la valeur par l'IRI avec dbpedia en définissant le type sur **dbo:Place**.

"status.csv": Permet de faire le type de fin de course d'un pilote lors d'une course.

Type::Status

Clé: "status_" + identifiant du statut

"drivers.csv": Contient le nom, prénom, nom de code (si il en a un), date de naissance, nationalité, et un lien wikipédia sur un pilote.

Type: :Driver

Clé: "driver_" + identifiant unique d'un pilote

Utilisation de noeud blanc pour le nom du pilote :

```
    ★ B: Blank
    ★ Name
    Add type...

Add property

** L: forename Add object...

** L: surname Add object...
```

"constructors.csv": Contient le nom, nationalité, et un lien wikipédia d'un constructeur automobile.

Type: :Constructor

Clé: "constructor" + identifiant unique d'un constructeur

Une fois les données exportées, j'utilise le service <u>rdfvalidator.mybluemix.net</u> pour m'assurer que les données RDF soient exportées correctement. Je concatène par la suite tous les fichiers et je combine les préfixes pour ne pas qu'il soit en double.

3. Schéma Turtle et extrait du fichier obtenu

Ayant cinq fichiers contenant différent type de donnée, le schéma Turtle en est assez long à inclure dans ce rapport. Celui-ci est consultable dans le git : Processed/schema.ttl

Voici des exemples de chaque classe provenant des différents types d'objets de l'ontologie.

Exemple d'un circuit

```
:circuit_79 a :Circuit;
:country dbr:USA;
:lat "25.9581"^^xsd:float;
:lng "-80.2389"^^xsd:float;
:location "Miami";
:name "Miami International Autodrome";
:wikipedia <a href="http://en.wikipedia.org/wiki/Miami_International_Autodrome">http://en.wikipedia.org/wiki/Miami_International_Autodrome</a>.

dbr:USA a dbo:Country.
```

Exemple d'un Grand Prix

```
:race_1078 a
                      dbo:GrandPrix;
     :circuit
                 :circuit_79;
     :dateTime
                    [ a
                            :DateTime;
                 :date "2022-05-08"^^xsd:date;
                 :time "19:30:00"^^xsd:time
               ];
     :fp1
                 [ a
                        :DateTime ;
                 :date "2022-05-06"^^xsd:date;
                 :time "18:30:00"^^xsd:time
               ];
     :fp2
                 [ a
                        :DateTime;
                 :date "2022-05-06"^^xsd:date;
                 :time "21:30:00"^^xsd:time
               ];
     :fp3
                        :DateTime;
                 [ a
                 :date "2022-05-07"^^xsd:date;
                 :time "17:00:00"^^xsd:time
               ];
                   "Miami Grand Prix";
     :name
     :qualification [ a
                           :DateTime;
                 :date "2022-05-07"^^xsd:date;
                 :time "20:00:00"^^xsd:time
               ];
                   "5"^^xsd:int;
     :round
                   <a href="http://en.wikipedia.org/wiki/2022_Miami_Grand_Prix">http://en.wikipedia.org/wiki/2022_Miami_Grand_Prix</a>;
     :wikipedia
                  "2022"^^xsd:int.
     :year
```

Exemple du résultat du pilote Max Verstappen à ce Grand Prix

```
:result 25486 a
                      :Result;
     :constructor
                    :constructor_9;
     :driver
                  :driver_830;
    :driverNumber
                     "1"^^xsd:int :
    :fastestLap
                    "54"^^xsd:int;
    :fastestLapSpeed "213.255"^^xsd:float;
     :fastestLapTime "1:31.361";
    :grid
                 "3"^^xsd:int :
    :laps
                 "57"^^xsd:int;
    :milliseconds
                   "5664258"^^xsd:int;
                  "26"^^xsd:float;
     :points
    :positionOrder "1"^^xsd:int;
    :race
                  :race 1078;
                  "1"^^xsd:int;
    :rank
    :status
                  :status_1.
```

Exemple du pilote Max Verstappen

Exemple du constructeur Redbull

```
:constructor_9 a :Constructor;
    :name     "Red Bull";
    :nationality "Austrian";
    :wikipedia <a href="http://en.wikipedia.org/wiki/Red_Bull_Racing">http://en.wikipedia.org/wiki/Red_Bull_Racing</a>.
```

Exemple d'un statut de résultat de course

```
:status_1 a :Status;
:status "Finished"@en, "Terminé"@fr;
:type "ok".
```

J'ai traduit les statuts en français pour les afficher correctement dans mon application, d'où le fait que chaque statut est une version écrite en anglais et français.

4. Point d'accès SPARQL

Après téléchargement et lancement de Apache Jena Fuseki, je créé une base de donnée nommé "formula-1" et j'y importe mon fichier concatené qui contient toutes les données de Formule 1

Apache Jena Fuseki



Import du fichier:

+ select files 2 upload all								
name \$	size \$	speed	status	actions				
data.ttl	14.07mb	2.06mb/s	100.00 Triples uploaded: 368755	1 upload now				
				• remove				

5. Requêtes SPARQL

Mon application contient au total dix requêtes différentes pour afficher les graphiques ou les listes. Les requêtes sont visible dans le fichier <u>front-end/src/sparql.js</u>

Requête : Récupérer le nombre de grand prix par année

```
PREFIX: <a href="http://lar.o.o.1:3333/">http://lar.o.o.1:3333/</a>
PREFIX dbo: <a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/</a>
SELECT
?year
(COUNT(?uri) AS ?value)
WHERE {
?uri a dbo:GrandPrix;
:year ?year.
}
GROUP BY(?year)
ORDER BY ASC(?year)
```

Je récupère tous les Grand Prix, puis je les regroupe par année. Je récupère ensuite l'année avec le nombre de valeur contenu dans le groupe. J'ordonne le résultat avec l'année des Grand Prix.

```
Requête : Récupérer les données d'un pilote
```

```
PREFIX: <a href="http://127.0.0.1:3333/">http://127.0.0.1:3333/>
PREFIX foaf: <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/>
PREFIX dbo: <a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/>
SELECT*
WHERE {
   :${driver_iri} a :Driver ;
      :name [ a :DriverName; :forename ?forename; :surname ?surname ].
   OPTIONAL { :${driver_iri} :dateOfBirth ?dateOfBirth }
   OPTIONAL {:${driver_iri} :code ?code }
   OPTIONAL {:${driver iri}:nationality?nationality}
   OPTIONAL {:${driver_iri} :wikipedia ?wikipedia }
   OPTIONAL {
      SERVICE <a href="http://dbpedia.org/sparql">http://dbpedia.org/sparql</a> {
         ?wikipedia foaf:primaryTopic ?page dbo.
         OPTIONAL{
           ?page dbo dbo:abstract ?abstract.
           FILTER(LANG(?abstract) = "fr")
         }
         OPTIONAL{ ?page_dbo dbo:thumbnail ?thumbnail. }
     }
  }
}
```

\${driver_iri} est remplacé par l'identifiant unique du pilote lors de la requête. Le manque d'information sur certains pilotes me force à utiliser des champs OPTIONAL sur toute les valeurs car je peux requêter des valeurs inexistantes. Pour la description du pilote, je requête via DBpedia si possible l'abstract du pilote et si possible une image de celui-ci.

Requête : Récupérer les pilotes d'une équipe par année PREFIX: http://127.0.0.1:3333/> PREFIX dbo: http://dbpedia.org/ontology/> **SELECT** ?driver uri (sample(? driver forename) AS ?driver forename) (sample(? driver surname) AS ?driver surname) (sample(?_driver_number) AS ?driver_number) (count(?gp race) AS ?driver nbr gp) ?gp_year WHERE { ?result a :Result: :constructor:\${constructor_iri}; :driver ?driver uri ; :driverNumber ? driver number ; :race ?gp_race. ?gp race a dbo:GrandPrix; :year ?gp_year . ?driver uri a :Driver ; :name [:forename? driver forename;:surname? driver surname]. } GROUP BY ?gp year ?driver uri

\${constructor_iri} est remplacé par l'identifiant unique d'un constructeur lors de la requête.

ORDER BY DESC(?gp_year)

Je récupère les pilotes à partir des résultats car si un pilote a concouru pour un Grand Prix, il fait alors partie de l'équipe pour laquelle il a concouru. Je groupe ensuite par l'année des Grand Prix ainsi que l'identifiant unique du pilote, pour récupérer les valeurs du nom prénom, et de son numéro associé lors de l'année, j'utilise la fonction *sample*, et pour récupérer le nombre de Grand Prix effectué, je compte le nombre de valeur dans le groupe. J'ordonne enfin le résultat en ayant les derniers Grand Prix en premier.

```
Requête : Récupérer le score d'un pilote par année
     PREFIX: <a href="http://127.0.0.1:3333/">http://127.0.0.1:3333/>
     PREFIX dbo: <a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/>
     SELECT
        ?gp_year
        ?status_type
        (count(?status_type) as ?count)
        (sum(?pt) as ?points)
     WHERE {
        ?result_uri a :Result ;
           :driver :${driver iri};
           :race ?gp uri;
           :points ?pt;
           :status ?status uri .
        ?gp_uri a dbo:GrandPrix;
           :year ?gp year .
        ?status_uri a :Status ;
           :type ?status_type.
     GROUP BY ?status_type ?gp_year
     ORDER BY ?gp_year
```

\${driver iri} est remplacé par l'identifiant unique du pilote lors de la requête.

Pour le calcul du score, je le décompose en trois groupes : ok, lap, et out, qui correspondent aux types de statut possible. Cette requête et cette décomposition du score me permet d'afficher les deux graphiques sur la page d'un pilote. En fonction du type de groupe du statut, je calcule la somme des points obtenues par année. Le tout est ordonné par l'année du Grand Prix avec les plus anciens en premier.

```
Requête: Récupérer les données d'un constructeur
PREFIX: <a href="http://127.0.0.1:3333/">http://127.0.0.1:3333/">http://127.0.0.1:3333/</a>
PREFIX foaf: <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>
PREFIX dbo: <a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/</a>

SELECT

?name
?name
?nationality
?wikipedia
?abstract
?thumbnail
WHERE {
:${constructor_iri} a : Constructor ;
:name ?name ;
:nationality ?nationality ;
```

```
:wikipedia?wikipedia.
  OPTIONAL {
     SERVICE <a href="http://dbpedia.org/sparql">http://dbpedia.org/sparql</a> {
          ?page dbo foaf:isPrimaryTopicOf ?wikipedia
        }
        UNION
          ?page dbo 2 foaf:isPrimaryTopicOf ?wikipedia;
                  dbo:wikiPageRedirects ?page_dbo.
        }
        OPTIONAL {
          ?page dbo dbo:abstract ?abstract.
             FILTER(LANG(?abstract) = "fr")
        OPTIONAL { ?page dbo dbo:thumbnail ?thumbnail. }
     }
  }
}
```

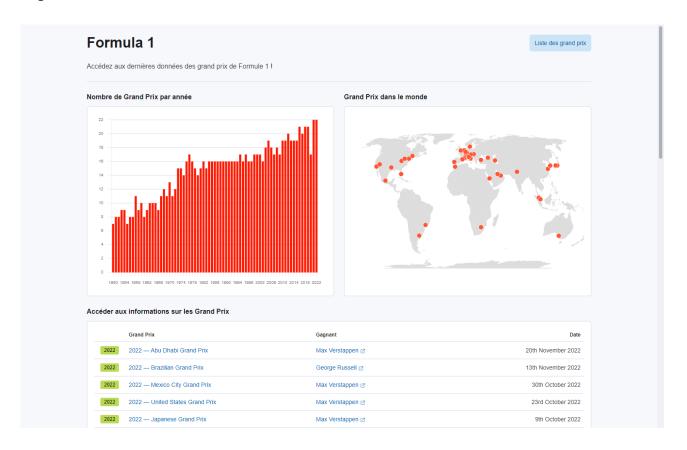
\${constructor_iri} est remplacé par l'identifiant unique d'un constructeur lors de la requête.

Cette requête permet simplement de retourner les données de l'ontologie correspondant à un constructeur via son identifiant unique. Pour récupérer une description, je fais appel à DBpedia. Ici, j'utilise l'opération UNION car dans certains cas, l'adresse wikipedia contenu dans mes données qui est une valeur de *foaf:isPrimaryTopicOf* fait alors rebondir sur une redirection. L'union me permet donc de m'assurer que je récupère la page DBpedia dans les deux cas (redirection ou non). Je peux ensuite récupérer un abstract si disponible, et une image si disponible.

Cinq autres requêtes sont disponibles dans le fichier <u>front-end/src/sparql.js</u> que j'ai jugé moins intéressantes car récupère juste des données via DBpedia ou retournent simplement les données de l'ontologie.

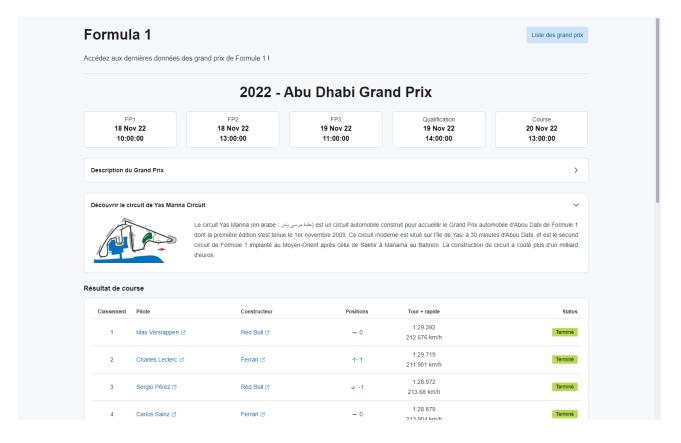
6. Capture d'écran de l'application

Page d'accueil



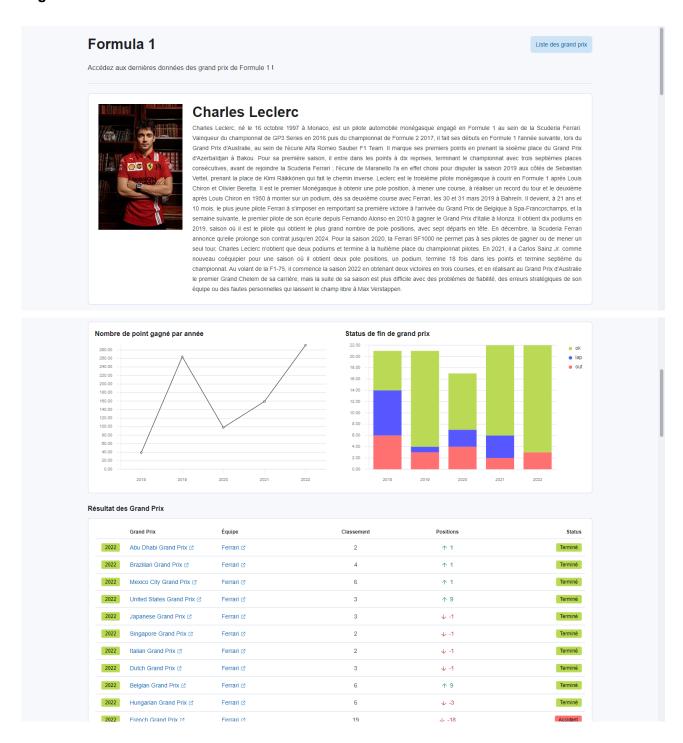
Sur cette page on peut trouver un graphique présentant le nombre de Grand Prix par année, une carte du monde avec un point sur chaque circuit ou un Grand à eu lieu, et un accès aux résultats des derniers Grand Prix avec le nom, l'année, le vainqueur et la date de la course.

Page Grand Prix



Sur la page d'un Grand Prix, on peut trouver la date et heures des différents évènements (FP1, FP2, FP3, Sprint, Qualification, et la Course), une description du grand prix récupérée sur DBpedia, une description avec un image du circuit utilisé pour le Grand Prix, ainsi que les résultats de la course. Les résultats sont composés du classement, du nom du pilote, de l'équipe constructeur du pilote, du nombre de positions gagnées ou perdues par rapport à sa position sur la grille de départ, le temps et la vitesse moyenne du tour le plus rapide, ainsi que le statut. Un statut vert indique que le pilote a fini dans les temps, un statut bleu indique au moins un tour de plus par rapport au vainqueur, et un statut rouge indique une fin de Grand Prix pour le pilote (avec la raison expliquée : accident, disqualification, ...)

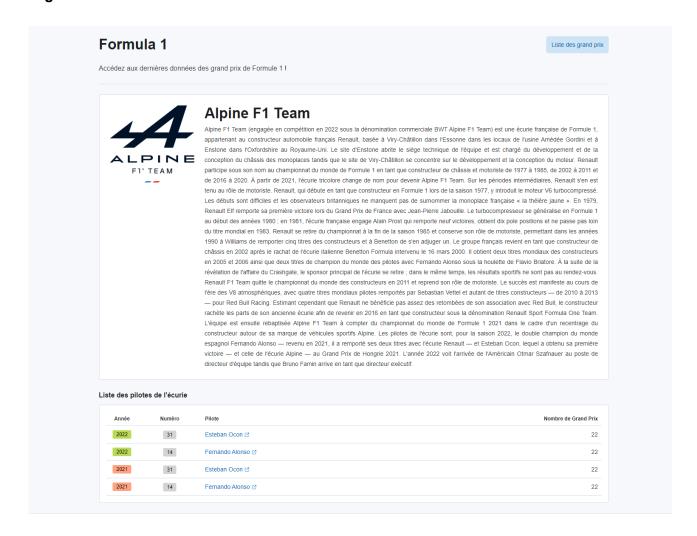
Page Pilote



Cette page permet d'accéder aux détails d'un pilote. La première partie du site est consacrée à des informations plus générales, comme une description et une photo de celui-ci. Puis la deuxième partie est consacrée à son expérience en formule 1. Dans un premier temps, deux graphiques sont affichés. La première à gauche, le graphique indique le nombre de points gagnés par année concouru, et le second à droite, indique le statut de fin de course par année. Ces deux graphiques permettent de voir l'évolution du pilote sur le long terme.

Enfin, une liste des résultats du pilote pour chaque grand prix concouru est affichée avec comme information l'année et le nom du Grand Prix, l'équipe constructeur du pilote, son classement, le nombre de positions gagnées ou perdues, et le statut de sa course.

Page constructeur



Cette page est composée d'une première partie avec le logo de l'équipe ainsi qu'une description de celle-ci. La deuxième partie est composée des pilotes ayant concouru pour cette équipe par année avec le nombre de Grand Prix auxquels ils ont participé.