FootyPedia: a semantic web approach to football data

L. Miguel Pinto
up201806206@up.pt

Luís Rafael Afonso
up201406189@up.pt

Nuno Oliveira
up201806525@up.pt

December 13, 2022

Abstract

The work described in this paper aims to provide a clean interface for users to search for football information. In order to achieve that, we performed some web scraping to gather the relevant information and then created an ontology that describes the interactions between the entities we wished to present. Finally, we created a simple website to display the information gathered. The usage of the semantic web enables us to create an expandable and machine-readable platform to gather information regarding football.

1 Introduction

Football is an ever-growing sport, and so is the desire to access information about players, competitions, and clubs. Although there are solutions for that problem, they are either behind a paywall or cluttered with visual pollution, making access to information difficult and unpleasant. Moreover, access to information tends to be human-centered, not offering that information in a structured and standardized format to enable easier use by machines.

With this paper, we analyze existing solutions and propose to create one that thrives on the power of linked data and offers this information in a pleasant way to users. Our solution also provides ways for the data to be accessible by machines through linked JSON files.

2 Motivation

2.1 Where to get football information nowadays?

The way users consume their football information can be mainly divided into two categories: Either sports-oriented news outlets or highly specialized tools that summarize a wide variety of statistics and data regarding the game, the teams, and its players.

2.1.1 News websites

News outlets tend to only provide the basic information regarding teams and games, giving greater focus to news article pieces, leading to a very cluttered platform for data visualization. For example, one of the most complete Portuguese websites, Zerozero[1], is primarily a news platform and as such, most of the information present is linked to news instead of relevant football statistics or data.

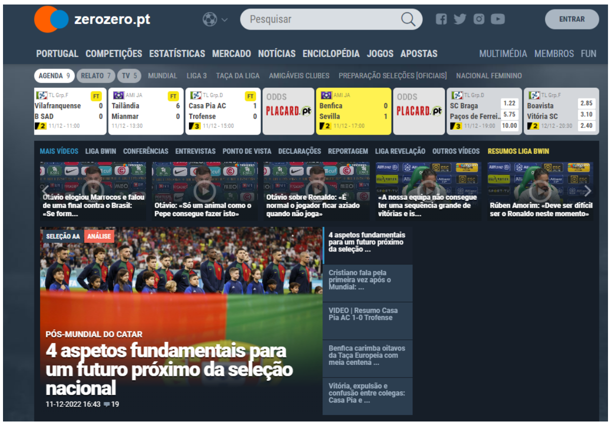

zerozero.pt
Pesquisar
ENTRAR
PORTUGAL COMPETIÇÕES ESTATÍSTICAS MERCADO NOTÍCIAS ENCICLOPÉDIA JOGOS APOSTAS
MULTIMÉDIA MEMBROS FUN
AGENDA 9 RELATO 7 TV 5 MUNDIAL LIGA 3 TAÇA DA LIGA AMIGÁVEIS CLUBES PREPARAÇÃO SELEÇÕES [OFICIAIS] NACIONAL FEMININO
TL Grp.F FT Vilafranquense 0 B SAD 0 11/12 - 11:00
AMI JA FT Tailândia 6 Mianmar 0 11/12 - 13:30
TL Grp.D FT Casa Pia AC 1 Trofense 0 11/12 - 15:00
ODDS PLACARD.pt
AMI JA Benfica 0 Sevilla 1 11/12 - 17:00
ODDS PLACARD.pt
TL Grp.D SC Braga 1.22 Paços de Ferrei.. 5.75 11/12 - 19:00 10.00
TL Grp.F Boavista 2.85 Vitória SC 3.10 12/12 - 20:30 2.40
MAIS VÍDEOS LIGA BWIN CONFERÊNCIAS ENTREVISTAS PONTO DE VISTA DECLARAÇÕES REPORTAGEM LIGA REVELAÇÃO OUTROS VÍDEOS RESUMOS LIGA BWIN
Otávio elogiou Marrocos e falou de uma final contra o Brasil: «Se form...
Otávio: «Só um animal como o Pepe consegue fazer isto»
Otávio sobre Ronaldo: «É normal o jogador ficar aziado quando não joga»
«A nossa equipa não consegue ter uma sequência grande de vitórias e is...
Rúben Amorim: «Deve ser difícil ser o Ronaldo neste momento»
SELEÇÃO AA ANÁLISE
PÓS-MUNDIAL DO CATAR
4 aspetos fundamentais para um futuro próximo da seleção nacional
11-12-2022 16:43 19
4 aspetos fundamentais para um futuro próximo da seleção ...
Cristiano fala pela primeira vez após o Mundial: ...
VIDEO | Resumo Casa Pia AC 1-0 Trofense
Benfica carimba oitavos da Taça Europeia com meia centena ...
Vitória, expulsão e confusão entre colegas: Casa Pia e ...

Figure 1: *ZeroZero's Homepage*

2.1.2 Specialized websites

With the increase of online betting on sports events, there has been a surge in websites dedicated to collecting the most detailed information possible in order to offer betting tips to players. As expected, these types of websites usually charge users for this information like Soccerment[2] and FootyStats[3]. Alternatively, there are also some platforms with vast amounts of information that are displayed in a very complex nature with very low usability.

2.2 Proposed solution

A common thing in existing solutions is the complexity of data, be it visual cluttering or the amount of data presented. With our solution, we first want to remove the complexity and create an application with a clean and "easy on the eyes" interface.

Furthermore, an ever-increasing complexity of data to handle (in variety and detail) comes with complex databases, which constitute difficulties in information retrieval and knowledge representation. To solve that problem we intend to use linked data and ontologies, which, surprisingly, none of the existing solutions use, at least the ones we found. With this approach, there is no defined structure. Instead, classes have attributes and they connect with each other through relations. As a consequence, it is more prepared to handle complex queries on complex databases. Consequently, information retrieval is much simpler when compared to a common relational or non-relational database [4].

3 FootyPedia

The development of this platform is anchored on three pillars:

- **Data Collection** — Collection of information;
- **Ontology Creation** — Creation of a dedicated ontology;
- **Website Creation** — Creation of a clean interface.

3.1 Data Collection

The first stage of the project involved determining which information we wanted to choose and defining our knowledge sources. The datasets that our search yielded were lackluster and insufficient to meet our standards, as most datasets only contained information regarding games and lacked information regarding the players and the teams. Mainly which players played in each team for varying seasons. We also observed that many of these datasets were very betting oriented, which was not our intended objective. As such we decided to use game history collected from Football-Data[5], and we decided to use web scraping to collect the missing information, namely the teams and their players. For this, we decided to scrape some information from ZeroZero[1], which appeared to be the most reliable Portuguese source.

We encountered some difficulties to achieve this task as the website proved to have varying countermeasures regarding bots even though the website's terms of service failed to provide any information regarding that this practice was not welcomed.

As such, our initial naive approach quickly encountered the bot detection capabilities of the website, which quickly gave us a timeout barring us from accessing the website. We then turned to Selenium[6] to aid us in retrieving the desired information. It enabled us to emulate a browser and simulate clicks instead of simply issuing requests to the website. This tactic coupled with a varying delay between each request enabled us to collect all of the information that we had initially planned to acquire.

After retrieving the information for each web page requested we used Beautiful Soup[7] in order to parse the HTML code retrieved to turn the process of collecting and transforming this information into RDF triples faster. To allow for the insertion and manipulation of these triples we opted to use Owlready2[8], a package for ontology-oriented programming in Python

3.2 Creation of Ontology

The first solution considered while approaching the ontology needed was to search the existing literature and previous work for an ontology that could be used as a base for our case. However, the existing solutions were either too complex for our problem or too specific to use as is or to morph into our needs. The solution was to create our own ontology from scratch and better structure the system to our needs.

Next, we needed to model the ontology, and for that, we chose *protégé*, which allowed us to create the ontology with a more visual approach than a plain XML editor. Three different types of entities were created, Classes that represent the central ideas of the problem, object-properties that represent connections between classes, and data-properties that represent attributes of said classes.

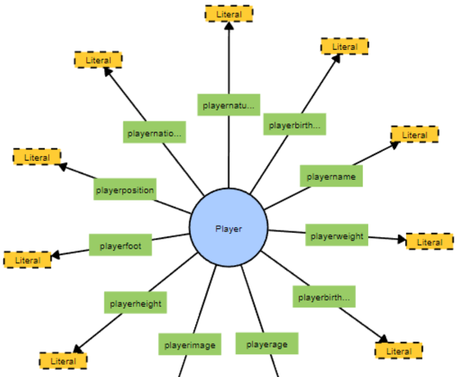


Figure 2: *Player class and its properties*

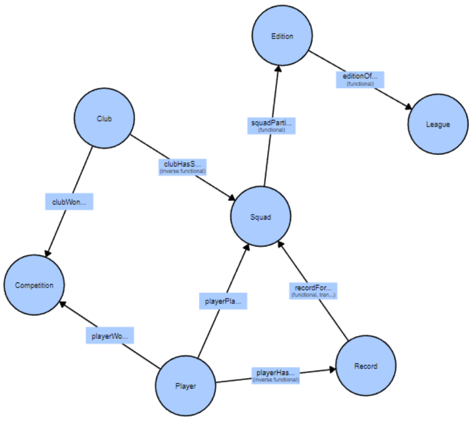


Figure 3: *Class diagram*

3.3 Website creation

Finally, all that was left was making all our data available for real users and machines as intended.

3.3.1 Frontend

First, we tackled the creation of the website. Since we were already using Python for scraping and ontology manipulation, it made sense also to use it for the website to keep the project simple. We decided to use the web framework Flask [9], which allows for the fast creation of a simple web application. As aforementioned, our focus was making a simple user interface without much visual clutter, but none of us is particularly keen on web design. For that reason, we resorted to MDB, a Bootstrap based UI toolkit for "easy theming and customization".

This stage was hindered due to the complications found in the project's earlier phases, which took time off this one. We initially planned to have multiple leagues and seasons in our dataset but ultimately decided against it since it would be time-consuming and would not add value to our proposal. Therefore, due to time complications, we determined it would be best to keep the implementation to a single season of the Portuguese league and flesh out the power of our ontology.

3.3.2 API

To provide easy access to our dataset, we also opened an endpoint in */api?q=<query>* where a user can provide a SPARQL query in a URL-encoded format. The response is in JSON-LD format, a list of lists with resources that match the said query. In that response, we can find literals and links to entities pages. That page returns a map in JSON.-LD format containing all relations that the entity appears in any side of the relation. We chose this format so the data could be easily accessible and usable by machines and also fairly readable for humans.

3.4 Open Data principles Ranking

In accordance with the Linked Open Data principles, we believe to have fulfilled all requirements for this project to have a ranking of five stars. All Data is available in a non-proprietary machine-readable format, mostly JSON-LD. All entities depicted in this work are properly identified as resources in accordance with the open standards from the W3C, and finally, we try to offer outgoing links to relevant information available on the internet.

4 Conclusion

4.1 Final Remarks

This work encompassed three main stages in order to produce a semantic web approach to a website regarding football data. Firstly the information-gathering phase allowed us to observe the lack of machine-readable data in this field, which demonstrates interest in further pursuing this work. Secondly, we realized that the available ontologies regarding football were very incomplete and were hard to adapt to the information we wanted to offer. Finally, the creation of the website enabled us to offer an endpoint capable of receiving SPARQL queries and delivering the results in a machine-readable manner.

4.2 Future Work

In order to improve the work done, we could increase the amount of data that is available, such as more leagues and more league seasons. Also, the website is only a proof-of-concept. It does not yet gather all the information we have available through it that can be accessed through the API.

References

[1] zerozero.pt :: Porque todos os jogos começam assim... https://www.zerozero.pt/. Access in December, 2022.

[2] Football Stats | Player Comparisons | Team Analyses | Soccerment Analytics. https://analytics.soccerment.com/. Access in December, 2022.

[3] Football Stats, Tables & Results | Soccer Stats - FootyStats. https://footystats.org/. Access in December, 2022.

[4] Kamran Munir e M. Sheraz Anjum. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14:116–126, 7 2018.

[5] Football Betting | Football Results | Free Bets | Betting Odds. https://football-data.co.uk/. Access in December, 2022.

[6] Selenium. https://www.selenium.dev/. Access in December, 2022.

[7] Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. https://beautiful-soup-4.readthedocs.io/en/latest/. Access in December, 2022.

[8] pwin. Owlready2. https://github.com/AureClai/stream-python. Access in December, 2022.

[9] Welcome to flask — flask documentation (2.2.x). https://flask.palletsprojects.com/en/2.2.x/. Access in December, 2022.

Appendices

A Ontology

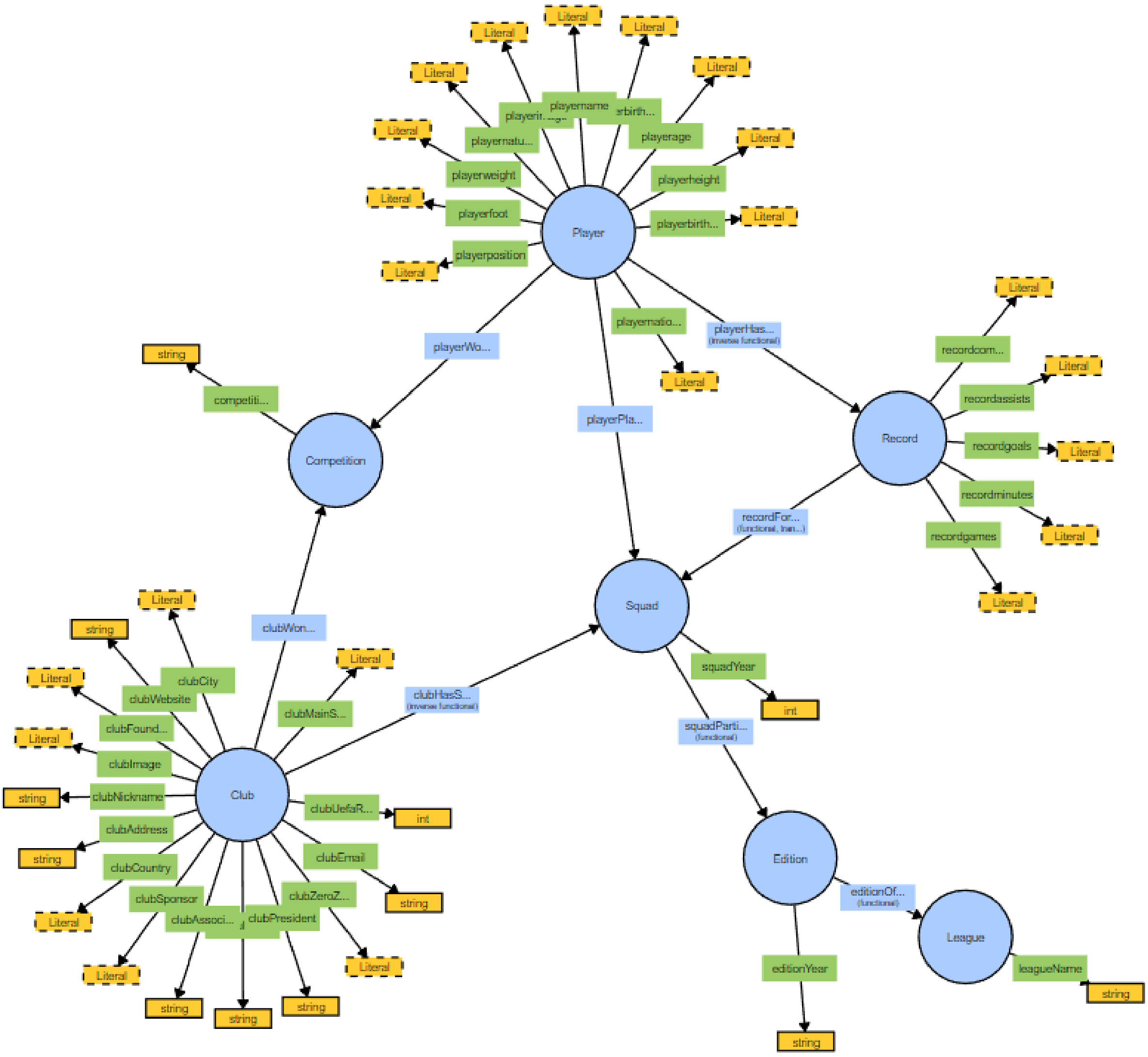


Figure 4: *Complete ontology diagram*

B Frontend

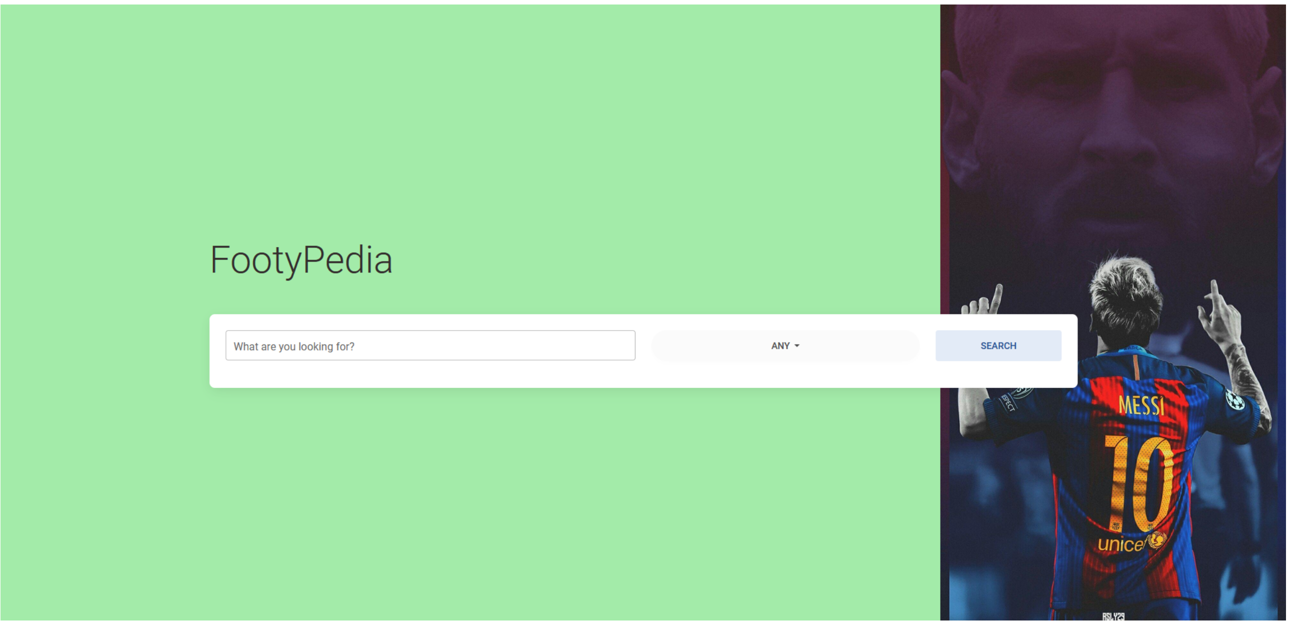
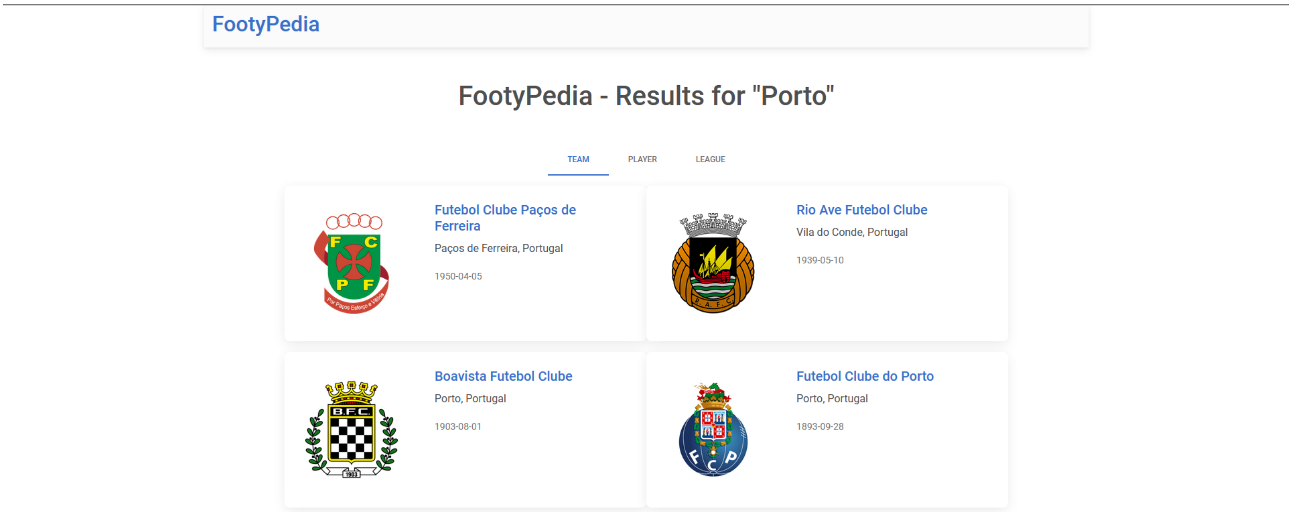


Figure 5: *Search Page*



FootyPedia - Results for "Porto"

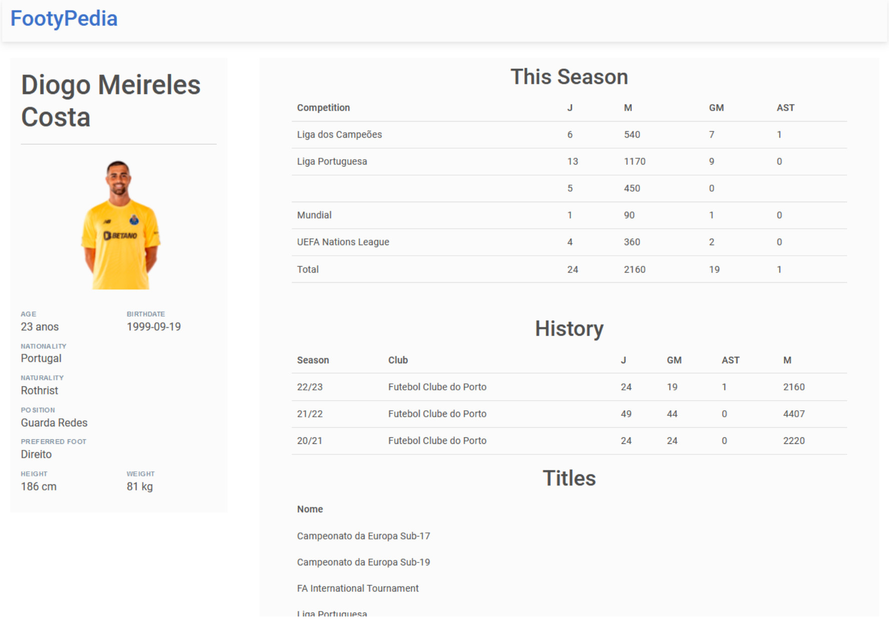
Futebol Clube Paços de Ferreira
Paços de Ferreira, Portugal
1950-04-05

Rio Ave Futebol Clube
Vila do Conde, Portugal
1939-05-10

Boavista Futebol Clube
Porto, Portugal
1903-08-01

Futebol Clube do Porto
Porto, Portugal
1893-09-28

Figure 6: *Search Results*



FootyPedia

Diogo Meireles Costa

AGE
23 anos

BIRTHDATE
1999-09-19

NATIONALITY
Portugal

NATURALITY
Rothrist

POSITION
Guarda Redes

PREFERRED FOOT
Direito

HEIGHT
186 cm

WEIGHT
81 kg

This Season

Competition	J	M	GM	AST
Liga dos Campeões	6	540	7	1
Liga Portuguesa	13	1170	9	0
	5	450	0	
Mundial	1	90	1	0
UEFA Nations League	4	360	2	0
Total	24	2160	19	1

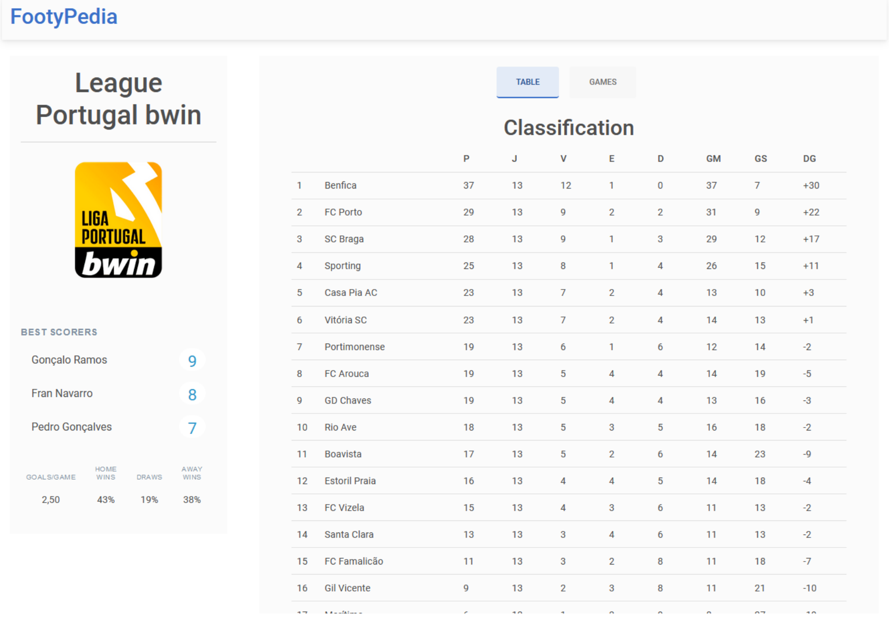
History

Season	Club	J	GM	AST	M
22/23	Futebol Clube do Porto	24	19	1	2160
21/22	Futebol Clube do Porto	49	44	0	4407
20/21	Futebol Clube do Porto	24	24	0	2220

Titles

Nome
Campeonato da Europa Sub-17
Campeonato da Europa Sub-19
FA International Tournament
Liga Portuguesa

Figure 7: *Player Page*



FootyPedia

League Portugal bwin

BEST SCORERS

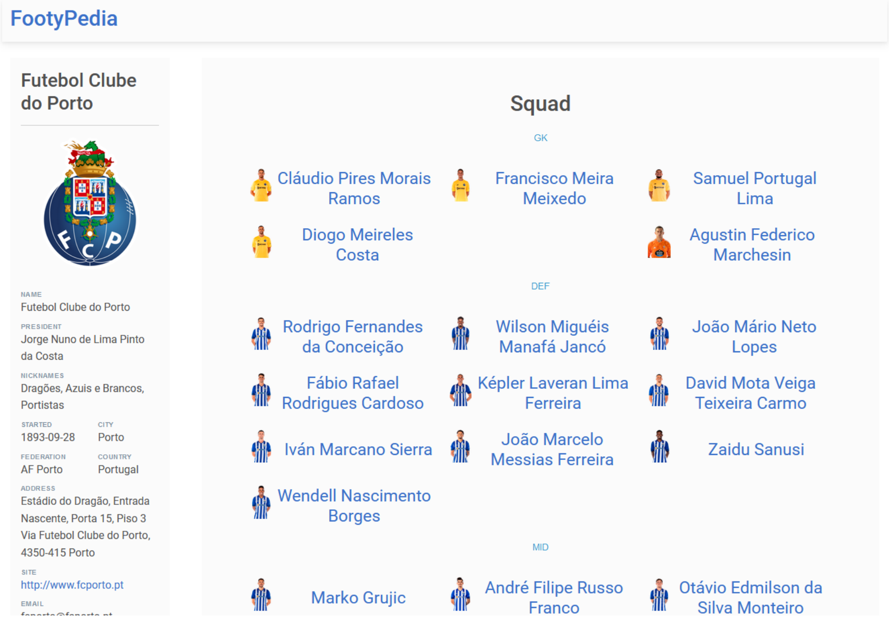
Gonçalo Ramos	9
Fran Navarro	8
Pedro Gonçalves	7

GOALS/GAME	HOME WINS	DRAWS	AWAY WINS
2,50	43%	19%	38%

Classification

		P	J	V	E	D	GM	GS	DG
1	Benfica	37	13	12	1	0	37	7	+30
2	FC Porto	29	13	9	2	2	31	9	+22
3	SC Braga	28	13	9	1	3	29	12	+17
4	Sporting	25	13	8	1	4	26	15	+11
5	Casa Pia AC	23	13	7	2	4	13	10	+3
6	Vitória SC	23	13	7	2	4	14	13	+1
7	Portimonense	19	13	6	1	6	12	14	-2
8	FC Arouca	19	13	5	4	4	14	19	-5
9	GD Chaves	19	13	5	4	4	13	16	-3
10	Rio Ave	18	13	5	3	5	16	18	-2
11	Boavista	17	13	5	2	6	14	23	-9
12	Estoril Praia	16	13	4	4	5	14	18	-4
13	FC Vizela	15	13	4	3	6	11	13	-2
14	Santa Clara	13	13	3	4	6	11	13	-2
15	FC Famalicão	11	13	3	2	8	11	18	-7
16	Gil Vicente	9	13	2	3	8	11	21	-10

Figure 9: *League Page*



Futebol Clube do Porto

NAME
Futebol Clube do Porto

PRESIDENT
Jorge Nuno de Lima Pinto da Costa

NICKNAMES
Dragões, Azuis e Brancos, Portistas

STARTED 1893-09-28 CITY Porto

FEDERATION AF Porto COUNTRY Portugal

ADDRESS
Estádio do Dragão, Entrada Nascente, Porta 15, Piso 3 Via Futebol Clube do Porto, 4350-415 Porto

SITE
http://www.fcporto.pt

EMAIL

Squad

GK

Cláudio Pires Morais Ramos | Francisco Meira Meixedo | Samuel Portugal Lima

Diogo Meireles Costa | Agustin Federico Marchesin

DEF

Rodrigo Fernandes da Conceição | Wilson Miguéis Manafá Jancó | João Mário Neto Lopes

Fábio Rafael Rodrigues Cardoso | Képler Laveran Lima Ferreira | David Mota Veiga Teixeira Carmo

Iván Marcano Sierra | João Marcelo Messias Ferreira | Zaidu Sanusi

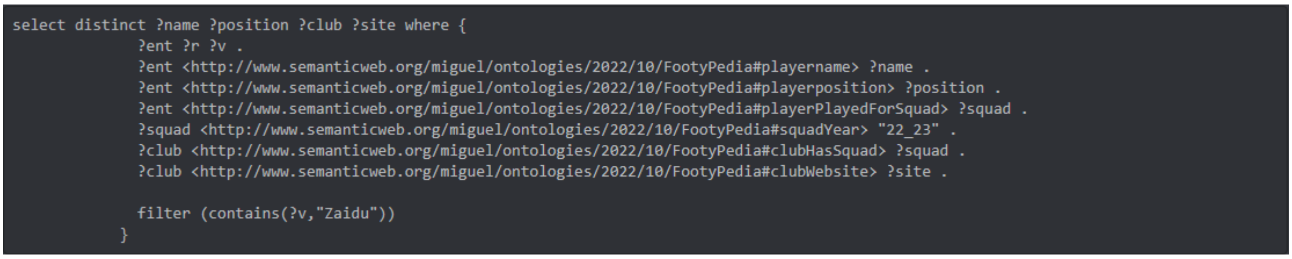
Wendell Nascimento Borges

MID

Marko Grujic | André Filipe Russo Franco | Otávio Edmilson da Silva Monteiro

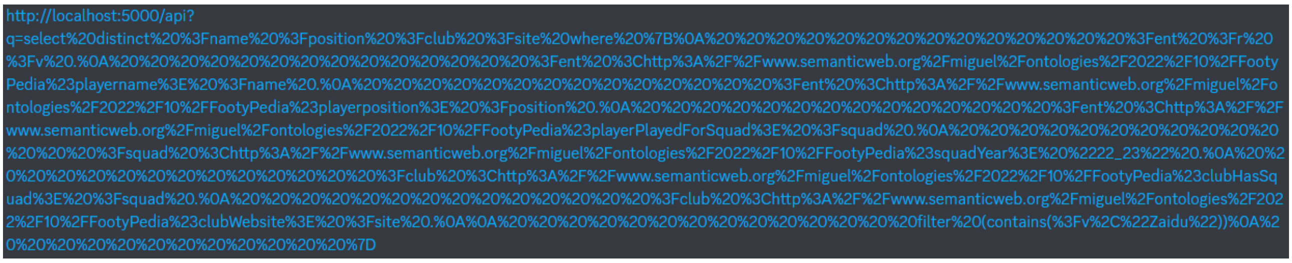
Figure 8: *Club Page*

C API



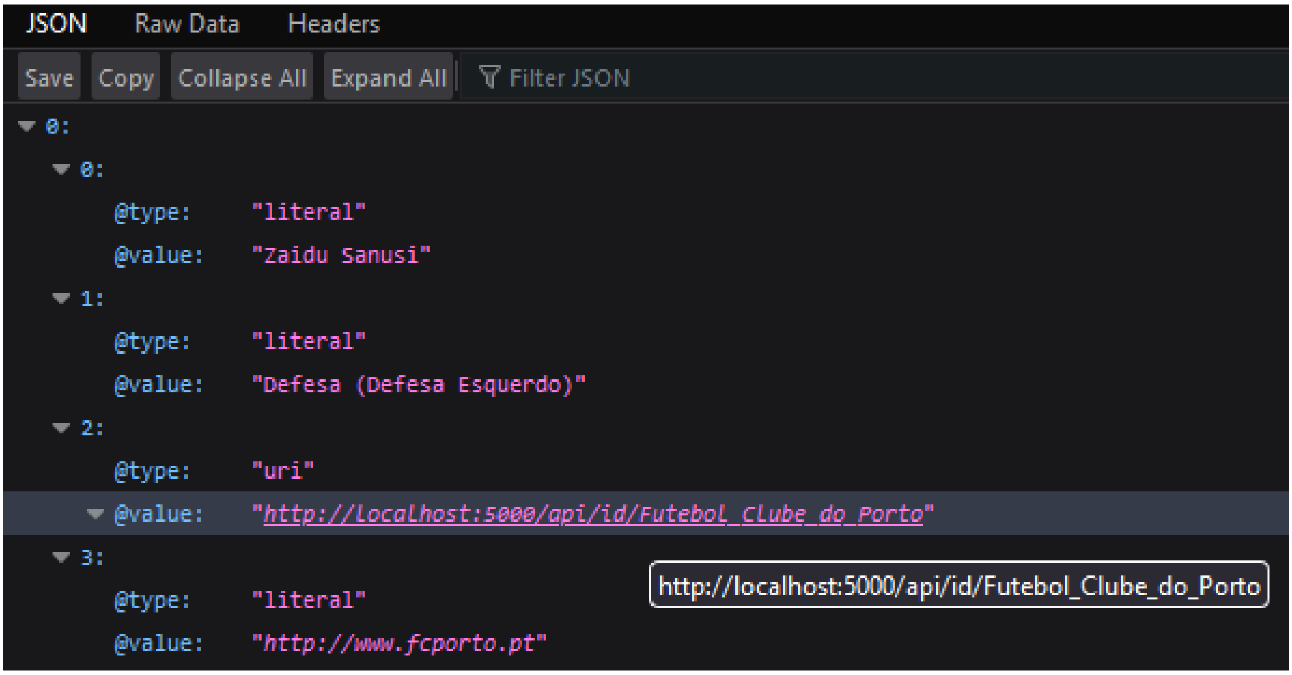
```
select distinct ?name ?position ?club ?site where {
              ?ent ?r ?v .
              ?ent <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#playername> ?name .
              ?ent <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#playerposition> ?position .
              ?ent <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#playerPlayedForSquad> ?squad .
              ?squad <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#squadYear> "22_23" .
              ?club <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#clubHasSquad> ?squad .
              ?club <http://www.semanticweb.org/miguel/ontologies/2022/10/FootyPedia#clubWebsite> ?site .

              filter (contains(?v,"Zaidu"))
            }
```

Figure 10: *SPARQL query*



```
http://localhost:5000/api?
q=select%20distinct%20%3Fname%20%3Fposition%20%3Fclub%20%3Fsite%20where%20%7B%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%3Fent%20%3Fr%20
%3Fv%20.%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%20%3Fent%20%3Chttp%3A%2F%2Fwww.semanticweb.org%2Fmiguel%2Fontologies%2F2022%2F10%2FFooty
Pedia%23playername%3E%20%3Fname%20.%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%20%3Fent%20%3Chttp%3A%2F%2Fwww.semanticweb.org%2Fmiguel%2Fo
ntologies%2F2022%2F10%2FFootyPedia%23playerposition%3E%20%3Fposition%20.%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%20%3Fent%20%3Chttp%3A%2F%2F
www.semanticweb.org%2Fmiguel%2Fontologies%2F2022%2F10%2FFootyPedia%23playerPlayedForSquad%3E%20%3Fsquad%20.%0A%20%20%20%20%20%20%20%20%20%20%20
%20%20%20%3Fsquad%20%3Chttp%3A%2F%2Fwww.semanticweb.org%2Fmiguel%2Fontologies%2F2022%2F10%2FFootyPedia%23squadYear%3E%20%2222_23%22%20.%0A%20%2
0%20%20%20%20%20%20%20%20%20%20%20%20%3Fclub%20%3Chttp%3A%2F%2Fwww.semanticweb.org%2Fmiguel%2Fontologies%2F2022%2F10%2FFootyPedia%23clubHasSq
uad%3E%20%3Fsquad%20.%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%20%3Fclub%20%3Chttp%3A%2F%2Fwww.semanticweb.org%2Fmiguel%2Fontologies%2F202
2%2F10%2FFootyPedia%23clubWebsite%3E%20%3Fsite%20.%0A%0A%20%20%20%20%20%20%20%20%20%20%20%20%20%20filter%20(contains(%3Fv%2C%22Zaidu%22))%0A%2
0%20%20%20%20%20%20%20%20%20%20%20%7D
```

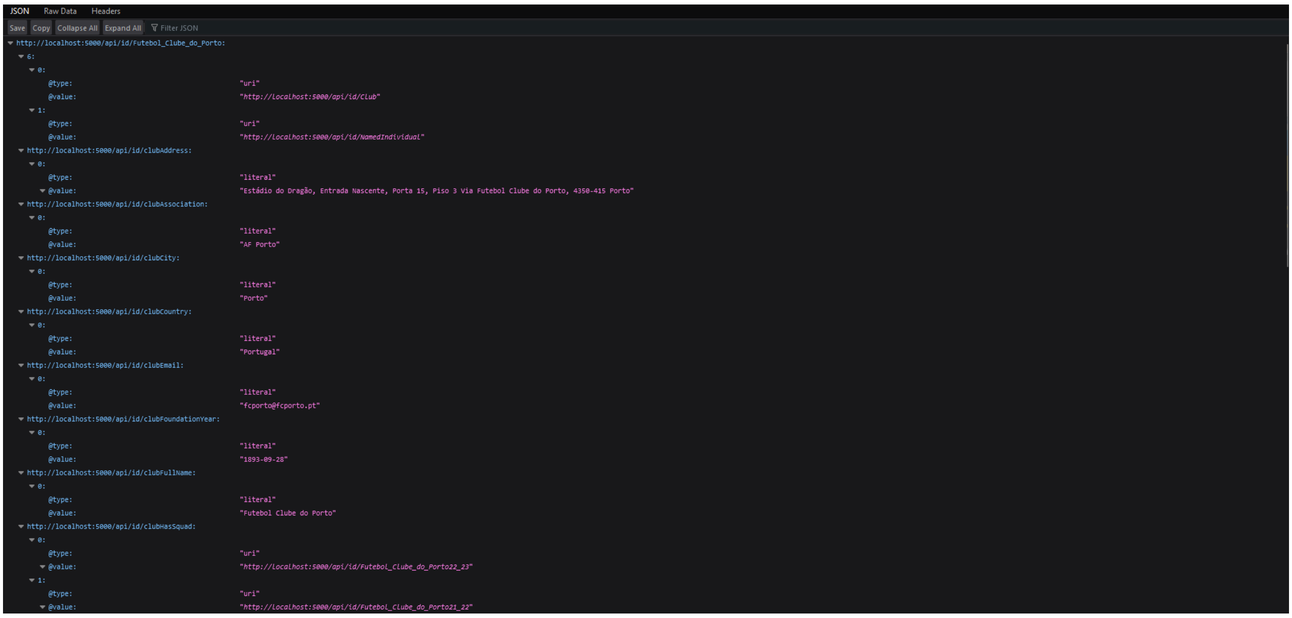
Figure 11: *Corresponding URL-encoded query for fig.5*



```
▼ 0:
  ▼ 0:
      @type:     "literal"
      @value:    "Zaidu Sanusi"
  ▼ 1:
      @type:     "literal"
      @value:    "Defesa (Defesa Esquerdo)"
  ▼ 2:
      @type:     "uri"
    ▼ @value:    "http://localhost:5000/api/id/Futebol_Clube_do_Porto"
  ▼ 3:
      @type:     "literal"
      @value:    "http://www.fcporto.pt"
```

http://localhost:5000/api/id/Futebol_Clube_do_Porto

Figure 12: *Answer of fig.6*



```
▼ http://localhost:5000/api/id/Futebol_Clube_do_Porto:
  ▼ 6:
    ▼ 0:
        @type:                      "uri"
        @value:                     "http://localhost:5000/api/id/Club"
    ▼ 1:
        @type:                      "uri"
        @value:                     "http://localhost:5000/api/id/NamedIndividual"
  ▼ http://localhost:5000/api/id/clubAddress:
    ▼ 0:
        @type:                      "literal"
      ▼ @value:                     "Estádio do Dragão, Entrada Nascente, Porta 15, Piso 3 Via Futebol Clube do Porto, 4350-415 Porto"
  ▼ http://localhost:5000/api/id/clubAssociation:
    ▼ 0:
        @type:                      "literal"
        @value:                     "AF Porto"
  ▼ http://localhost:5000/api/id/clubCity:
    ▼ 0:
        @type:                      "literal"
        @value:                     "Porto"
  ▼ http://localhost:5000/api/id/clubCountry:
    ▼ 0:
        @type:                      "literal"
        @value:                     "Portugal"
  ▼ http://localhost:5000/api/id/clubEmail:
    ▼ 0:
        @type:                      "literal"
        @value:                     "fcporto@fcporto.pt"
  ▼ http://localhost:5000/api/id/clubFoundationYear:
    ▼ 0:
        @type:                      "literal"
        @value:                     "1893-09-28"
  ▼ http://localhost:5000/api/id/clubFullName:
    ▼ 0:
        @type:                      "literal"
        @value:                     "Futebol Clube do Porto"
  ▼ http://localhost:5000/api/id/clubHasSquad:
    ▼ 0:
        @type:                      "uri"
      ▼ @value:                     "http://localhost:5000/api/id/Futebol_Clube_do_Porto22_23"
    ▼ 1:
        @type:                      "uri"
      ▼ @value:                     "http://localhost:5000/api/id/Futebol_Clube_do_Porto21_22"
```

Figure 13: *API page of FC Porto club entity*