Université
Nice
Sophia Antipolis

Semantic Web

La Liga Prediction

Date: 7 - February - 2017

GALDO SEARA Luis
GONZÁLEZ HUESCA Juan Manuel

1. Introduction

While looking for different ideas where the semantic web could find a good and novel use we came across the Spanish Football League “La Liga” [1] and how we could predict a match result based on historical data. We noticed there was almost nothing about semantic web on the field, just some information related to sports, but not in the way we were looking for; so we dig into it to create a new ontology that could be used as the foundation of the online gambling vocabulary for different sports.

We found many challenges along the way with the system design, on how to represent in the best way our information, with the tools to create the OWL and the RDF files; but all of them were sorted out by researching and trying different approaches and, at the end, choosing the best one, the more suitable.

The most difficult part of the project was the final integration, once we had the ontology, the RDF files and the SPARQL queries already defined, we had to integrate it all in an HTML website with JavaScript, this was the most time consuming part. It was hard because we had to make sure everything was in the right format, not only SPARQL for instance, and we had to make sure there were no conflicts with all the queries in the same place; we also had to work on the website design and the presentation format, but at the end we manage to create a fully functional website for Spanish Football predictions.

It was a very interesting project, we introduced a new ontology into the online gambling field and we consolidated the knowledge acquired during the lectures and made sure it just doesn’t stay in the classroom but in a real application, which is the final purpose when we learn something in the University.

The report will start with some background of the solution, then we will move to the analysis and design where the tool will be outlined, next the implementation will be explained. After that we will move to the testing, results and to finalize: conclusion, evaluation and further work will be detailed.

2. Background

La Liga Prediction is an online tool to forecast the result of a football match between two teams of the Spanish Football League “La Liga” based on real statistics from the last 5 years and betting odds from three leading online gambling companies.

Nowadays there are many companies that design models and make tools for statistical football prediction which are sold to individuals and/or bookmakers so they can have a starting point to define their own odds for the football matches (or any event). These statistical tools are very sophisticated and have a relatively good accuracy but they don't use semantic web, they have their own algorithms and they don't use W3C standards, which is where La Liga Prediction come to fill this gap, with common data formats and exchange protocols on the web.

As it's well known, the web is constantly evolving and this new phase, web 3.0, it's about getting the web more connected, more open and more intelligent; to have the data linked. So La Liga Prediction is adapting the statistical models to predict football matches results with semantic web standards and protocols to make sure it's aligned with the new web wave and it can be shared and used by different parties without restrictions. The goal is to introduce the semantic web to the online gambling world and make our ontology the new standard to be shared to all the companies and individuals in the area. This will simplify the process and will allow companies to share data from multiple sources in a seamless manner, so companies can focus on the design and implementation of more sophisticated algorithms which potentially have other applications in other fields.

3. Analysis and design

At the beginning of the project we considered to build our own ontology but we also planned that most of our queries would be pointing to the DBPedia endpoint. But we discovered they didn't have the data required for our project and further, we couldn't find anywhere an SPARQL endpoint with football statistics and bookmakers odds. This scenario leaded us to download the databases of the last five seasons of La Liga from "Football Data" [2] and from there we were able to make the RDF files based on our OWL, with some queries being made directly to DBPedia.

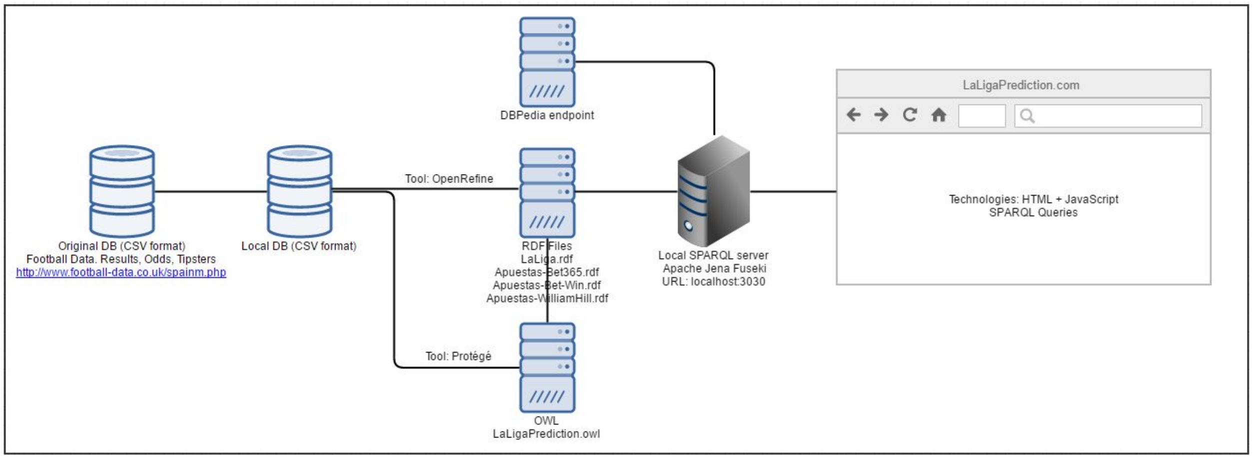


Figure 1. Current system design

The final system design is based on the SPARQL server Apache Jena Fuseki [3] which let us do our queries to this localhost endpoint, as shown in figure 1. The website was designed with HTML and javascript (the SPARQL is embedded in the code). OWL and RDF files were loaded to the server. In this configuration, we need to constantly take data from the original DB (Football Data in this case), make a local copy with the required information and then transform it into RDF following our ontology, once there we can proceed to make the SPARQL queries using our local endpoint. But the purpose for the future is to reduce this architecture like in figure 2, when the online gambling companies share the information in RDF format using our ontology, that would be the point where our website can be fully automated and every entity in this field can “profit” out of the shareable information.

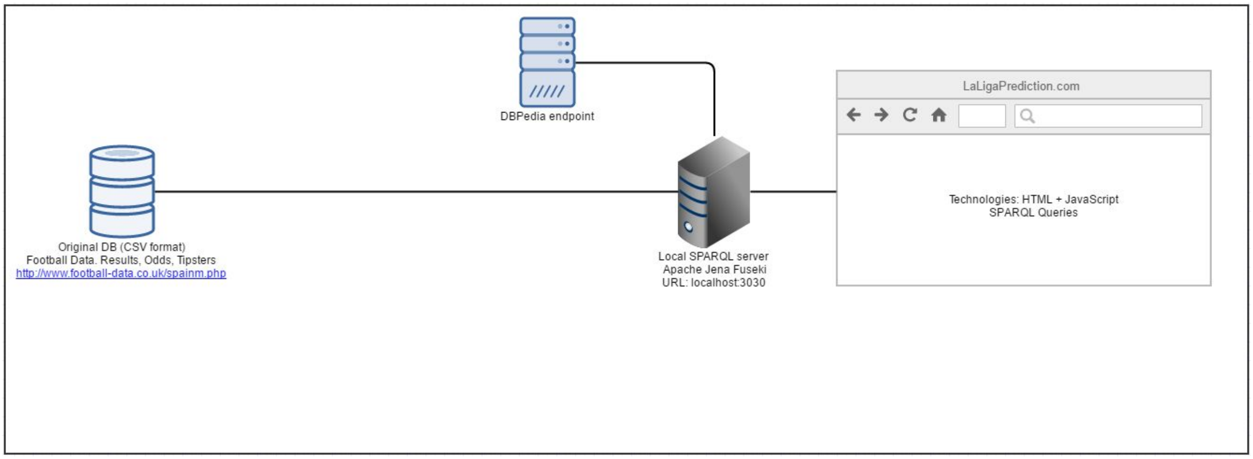


Figure 2. Future system design (when our ontology is used)

4. Implementation

The data in csv format was taken from the database of statistics of football records mentioned before (of the last five years) were: Season, Date, Home Team, Away Team, Home Team Goals, Away Team Goals, Home Team Shots, Away Team Shots, Home Team Shots on Target, Away Team Shots on Target, Home Team Fouls Committed, Away Team Fouls Committed, Home Team Corners, Away Team Corners, Home Team Yellow Cards, Away Team Yellow Cards, Home Team Red Cards, Away Team Red Cards. Besides the previous records we also took Home Win Odds, Draw Odds and Away Win Odds from three main online gambling companies to help us in our predictions, although we had more companies we just took the more representative as the others didn’t provide much different data. All those records were chosen as they are the most representative ones in a football match and with them we can have a better big picture of how the matches of two particular teams could be based on recent statistical data. Once we had the table with the data mentioned before, we gathered other data from DBPedia (Team Link, Stadium Link, Stadium Capacity, Stadium Longitude, Stadium Latitude) and from our own vocabulary to construct the RDF files based on our ontology. It’s worthy to mention there are some columns in the data base we didn’t consider such as companies odds for over/under 2.5 goals per match, Asian handicap betting odds, Closing odds (last odds before match starts), and many others as they didn’t provide meaningful information for our purpose.

The ontology OWL was built on Protégé [4] (A free, open-source ontology editor and framework for building intelligent systems) and it can be analyzed in figure 3. We have nine classes: City, Stadium, Team, BettingCompany, Season, Match, BetSet, Bet, Statistics; and each one of them has their properties related to their nature (i.e. The class Stadium has properties: type, hasTeam, belongsToCity, hasLink, hasCapacity, hasLatitude, hasLongitude, hasName).

The RDF files were built using OpenRefine [5] (A free, open source, powerful tool for working with messy data) but due to the nature of the csv file we were not able to create just one single RDF but four of them: one for the football matches statistics and three for the odds of the online gambling companies (one each). Each match was identified by a record and has all the properties mentioned in the ontology as can be seen on figure 4.

In order to show the three available algorithms (described below), the user will need to chose a match between two teams from the Spanish Football League “La Liga” (Home Team vs Away Team).

Available algorithms:

- Individual Comparison: From the matches on the last five seasons with a particular home team vs away team it’s presented the score, goals scored and received, average shots per team, average shots on target per team, average fouls committed per team, average corner kicks per team and average yellow and red cards per team.
- Group Comparison: The statistics here are not related to this particular match but with the selected teams and their individual performance over the last five seasons, showing total number of matches, number matches win-draw-lost, points and number of goals scored/received.
- Prediction: Based on the odd provided by three main online gambling companies we are able to compute the probabilities that home team wins, draw and away team wins. The probabilities represent an absolute percentage and it the average of all three companies odds for the last 5 seasons. The predicted score depends on the average score with the same teams on the last matches from the last five seasons.

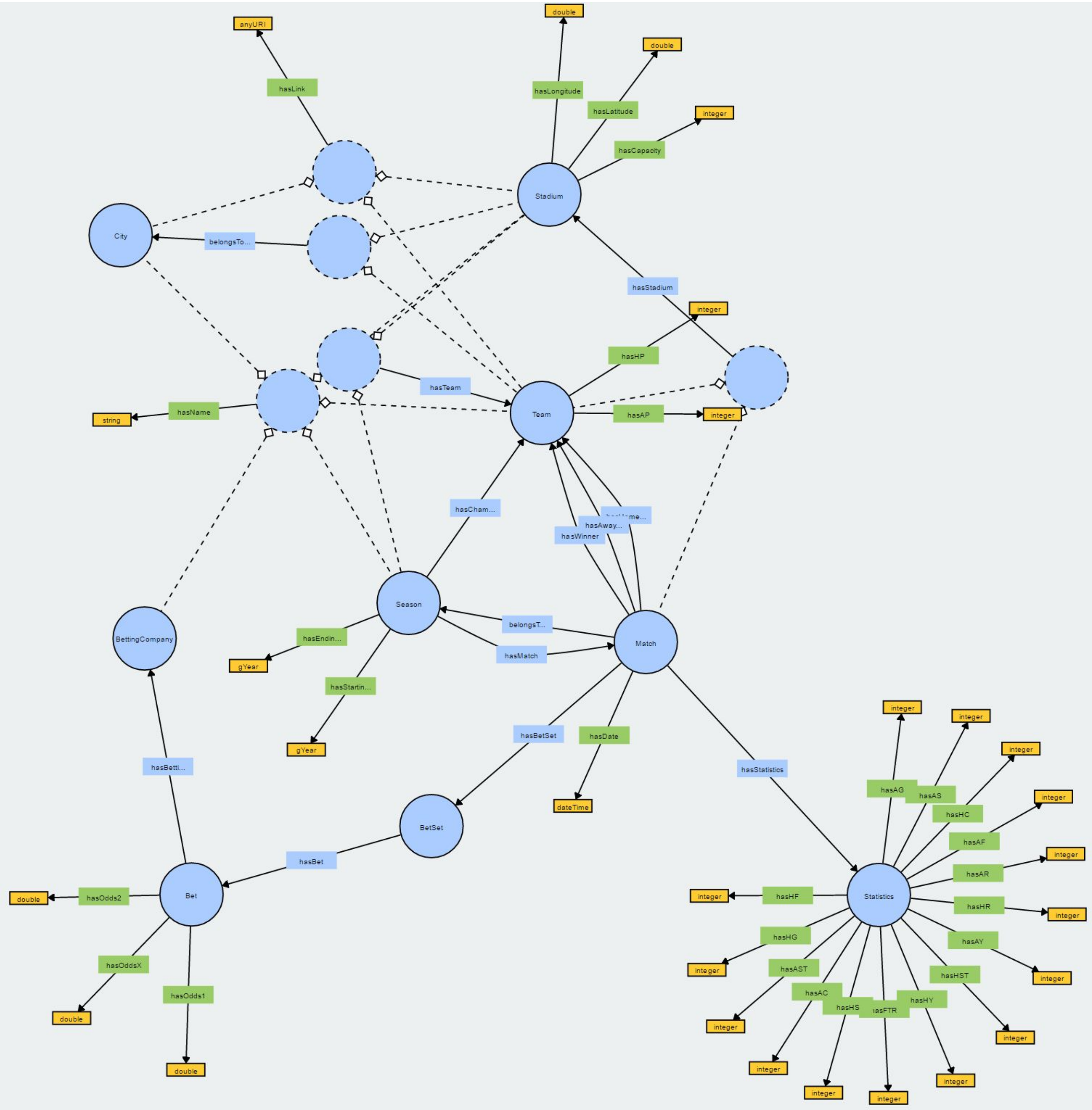
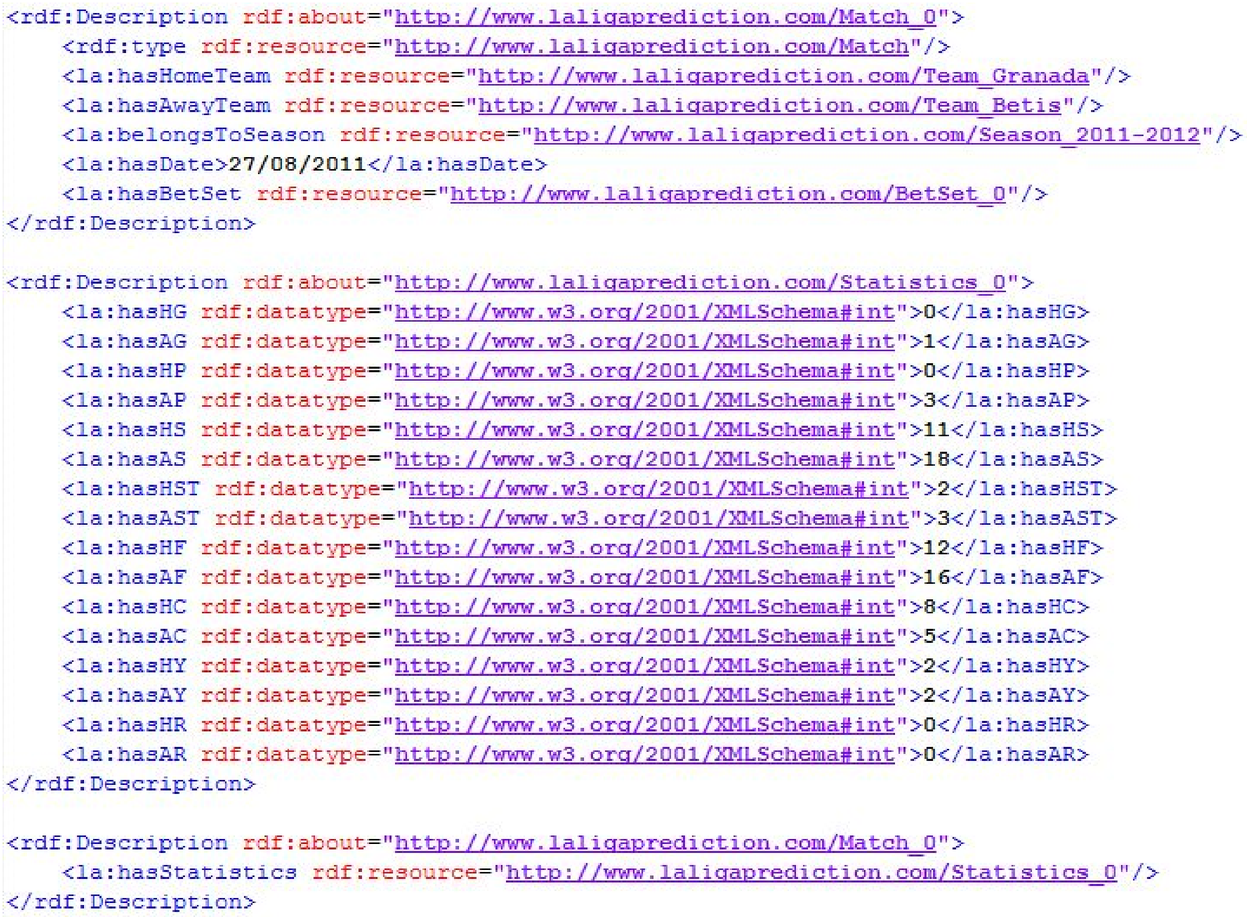


Figure 3. Ontology



```
<rdf:Description rdf:about="http://www.laligaprediction.com/Match_0">
    <rdf:type rdf:resource="http://www.laligaprediction.com/Match"/>
    <la:hasHomeTeam rdf:resource="http://www.laligaprediction.com/Team_Granada"/>
    <la:hasAwayTeam rdf:resource="http://www.laligaprediction.com/Team_Betis"/>
    <la:belongsToSeason rdf:resource="http://www.laligaprediction.com/Season_2011-2012"/>
    <la:hasDate>27/08/2011</la:hasDate>
    <la:hasBetSet rdf:resource="http://www.laligaprediction.com/BetSet_0"/>
</rdf:Description>

<rdf:Description rdf:about="http://www.laligaprediction.com/Statistics_0">
    <la:hasHG rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</la:hasHG>
    <la:hasAG rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</la:hasAG>
    <la:hasHP rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</la:hasHP>
    <la:hasAP rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</la:hasAP>
    <la:hasHS rdf:datatype="http://www.w3.org/2001/XMLSchema#int">11</la:hasHS>
    <la:hasAS rdf:datatype="http://www.w3.org/2001/XMLSchema#int">18</la:hasAS>
    <la:hasHST rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2</la:hasHST>
    <la:hasAST rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</la:hasAST>
    <la:hasHF rdf:datatype="http://www.w3.org/2001/XMLSchema#int">12</la:hasHF>
    <la:hasAF rdf:datatype="http://www.w3.org/2001/XMLSchema#int">16</la:hasAF>
    <la:hasHC rdf:datatype="http://www.w3.org/2001/XMLSchema#int">8</la:hasHC>
    <la:hasAC rdf:datatype="http://www.w3.org/2001/XMLSchema#int">5</la:hasAC>
    <la:hasHY rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2</la:hasHY>
    <la:hasAY rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2</la:hasAY>
    <la:hasHR rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</la:hasHR>
    <la:hasAR rdf:datatype="http://www.w3.org/2001/XMLSchema#int">0</la:hasAR>
</rdf:Description>

<rdf:Description rdf:about="http://www.laligaprediction.com/Match_0">
    <la:hasStatistics rdf:resource="http://www.laligaprediction.com/Statistics_0"/>
</rdf:Description>
```

Figure 4. Section of the RDF file related to a particular match

We also constructed some inferences based on the data we have, for instance: a) the stadium where the match will be played based on who is the home team; b) all the matches from a season based on the match property belongsToSeason; c) the city of a football team based on the location of their stadium; and d) the winner of a season based on the number of points gained in all the respective season matches.

Finally, it was all consolidated in an HTML website using JavaScript to execute the SPARQL queries organized on different functions based on the algorithm the user want to use (Individual Comparison, Group Comparison, Prediction).

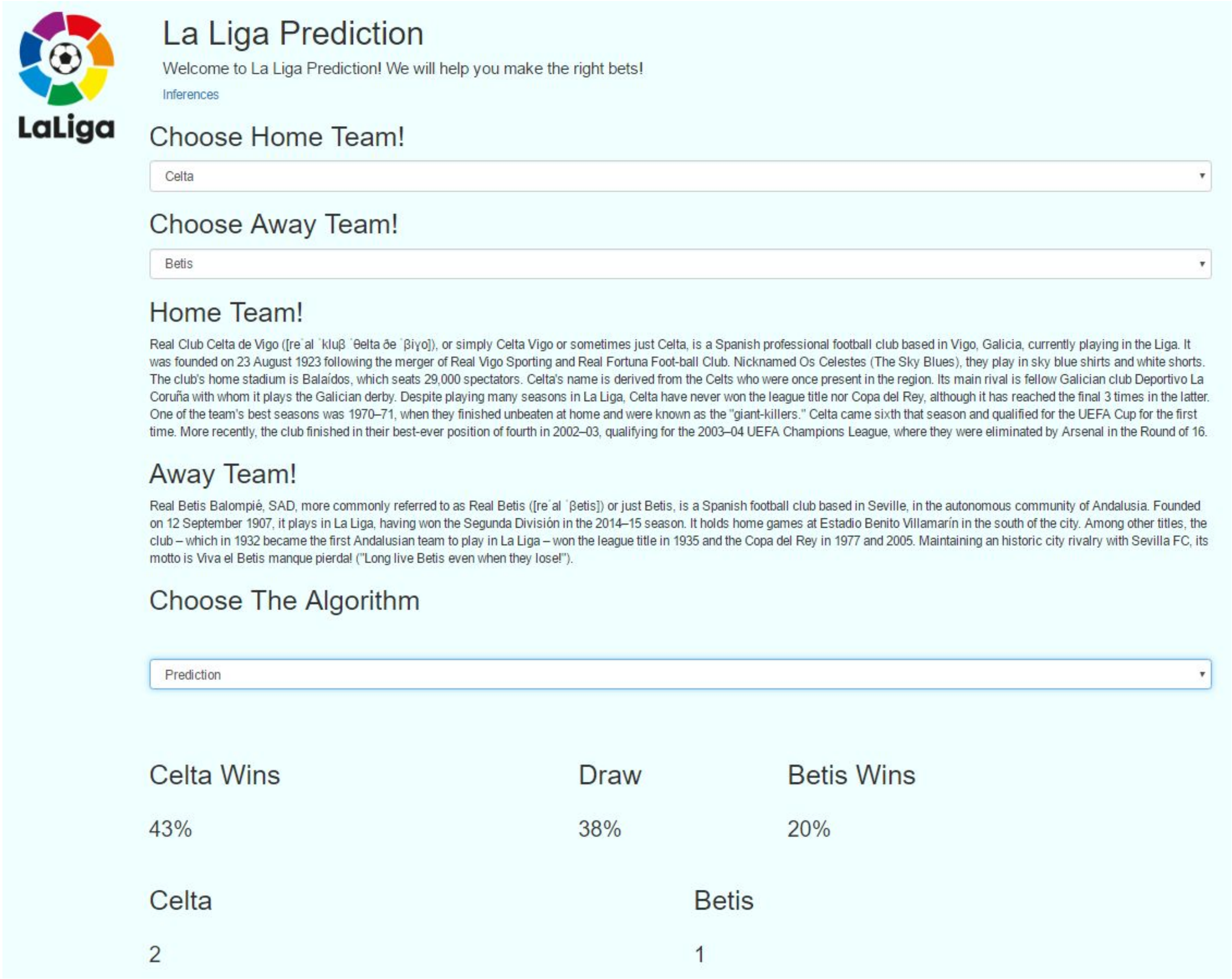
5. Testing

The tests were performed choosing different teams as home team vs away team, and selecting in each one of the cases a different algorithm (Individual Comparison, Group Comparison, Prediction) and then the result was compared with real statistics from the games over the last five seasons; for the prediction we used empirical knowledge to measure the accuracy of the algorithm. In the case of the inferences, we validated all the results based on the current real data.

All the tests were successful and can be verified on the demo of the website.

6. Results

In order to demonstrate the performance of La Liga Prediction, we will test our main algorithm, the prediction. In this case we select “Celta” as home team and “Betis” as away team. The abstract of each team is extracted from DBPedia using the property “dbo:abstract”, a query is done to the DBPedia endpoint and it’s shown to the user so they can have an small summary of the information for the teams. Then we choose the “Prediction” algorithm (figure 5) and the probability that the team “Celta” wins is displayed, together with the draw and the “Betis” probability wins. Also, we can see a predicted score based on the last season data.



La Liga Prediction

Welcome to La Liga Prediction! We will help you make the right bets!

Inferences

Choose Home Team!

Celta ▾

Choose Away Team!

Betis ▾

Home Team!

Real Club Celta de Vigo ([reˈal ˈkluβ ˈθelta ðe ˈβiɣo]), or simply Celta Vigo or sometimes just Celta, is a Spanish professional football club based in Vigo, Galicia, currently playing in the Liga. It was founded on 23 August 1923 following the merger of Real Vigo Sporting and Real Fortuna Foot-ball Club. Nicknamed Os Celestes (The Sky Blues), they play in sky blue shirts and white shorts. The club's home stadium is Balaídos, which seats 29,000 spectators. Celta's name is derived from the Celts who were once present in the region. Its main rival is fellow Galician club Deportivo La Coruña with whom it plays the Galician derby. Despite playing many seasons in La Liga, Celta have never won the league title nor Copa del Rey, although it has reached the final 3 times in the latter. One of the team's best seasons was 1970–71, when they finished unbeaten at home and were known as the "giant-killers." Celta came sixth that season and qualified for the UEFA Cup for the first time. More recently, the club finished in their best-ever position of fourth in 2002–03, qualifying for the 2003–04 UEFA Champions League, where they were eliminated by Arsenal in the Round of 16.

Away Team!

Real Betis Balompié, SAD, more commonly referred to as Real Betis ([reˈal ˈβetis]) or just Betis, is a Spanish football club based in Seville, in the autonomous community of Andalusia. Founded on 12 September 1907, it plays in La Liga, having won the Segunda División in the 2014–15 season. It holds home games at Estadio Benito Villamarín in the south of the city. Among other titles, the club – which in 1932 became the first Andalusian team to play in La Liga – won the league title in 1935 and the Copa del Rey in 1977 and 2005. Maintaining an historic city rivalry with Sevilla FC, its motto is Viva el Betis manque pierda! ("Long live Betis even when they lose!").

Choose The Algorithm

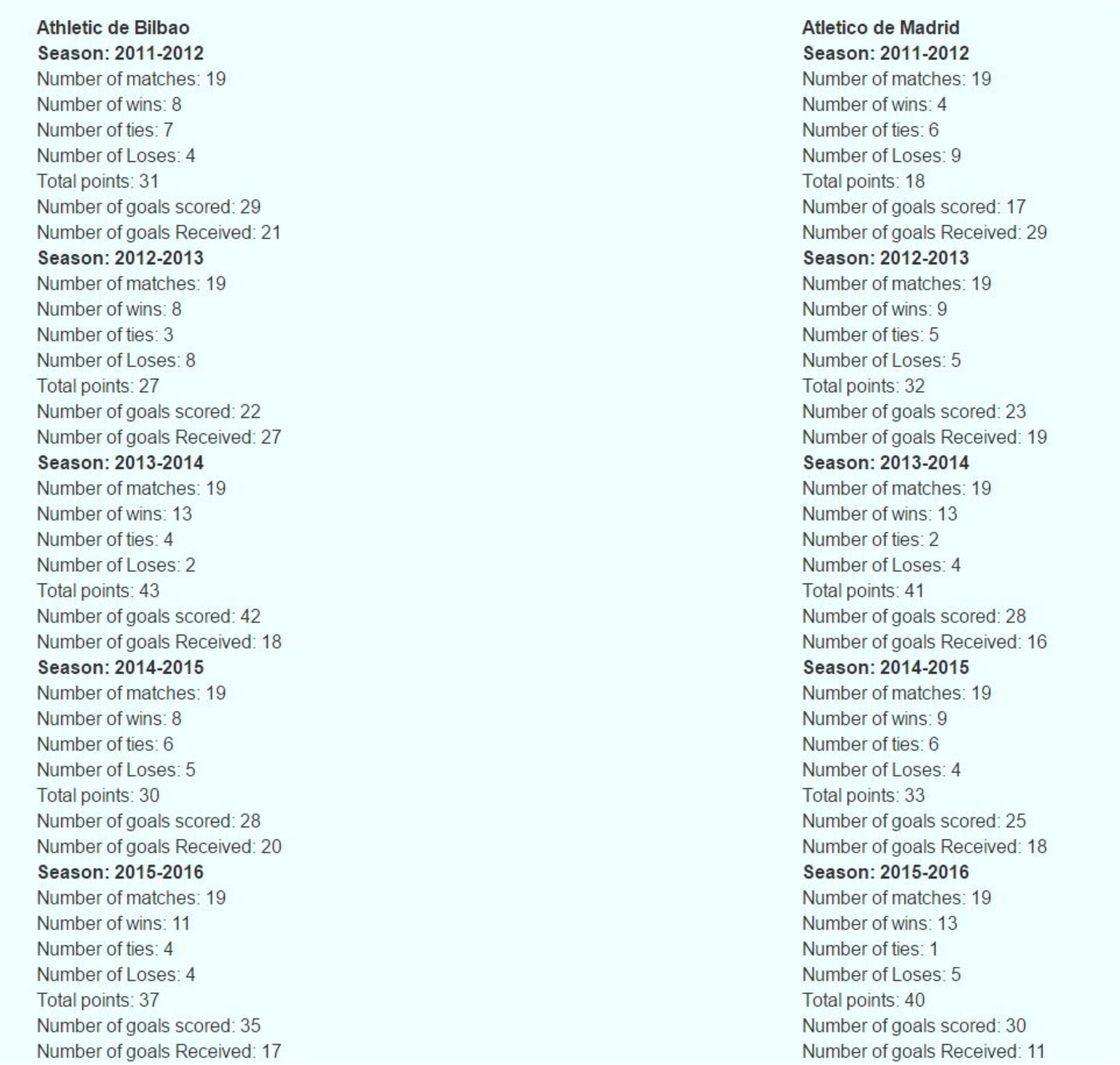
Prediction ▾

Celta Wins	Draw	Betis Wins
43%	38%	20%

Celta	Betis
2	1

Figure 5. Performance of the “Prediction” algorithm with Celta vs Betis

We also tested the algorithm for Group Comparison with different teams to check the performance and validate the results as we can see in figure 6.

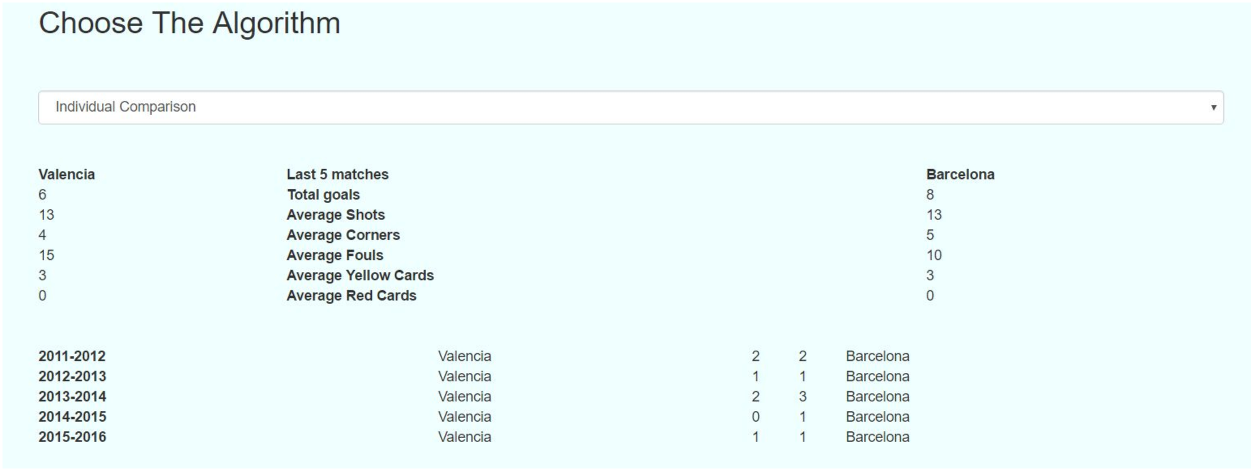


Athletic de Bilbao
Season: 2011-2012
Number of matches: 19
Number of wins: 8
Number of ties: 7
Number of Loses: 4
Total points: 31
Number of goals scored: 29
Number of goals Received: 21
Season: 2012-2013
Number of matches: 19
Number of wins: 8
Number of ties: 3
Number of Loses: 8
Total points: 27
Number of goals scored: 22
Number of goals Received: 27
Season: 2013-2014
Number of matches: 19
Number of wins: 13
Number of ties: 4
Number of Loses: 2
Total points: 43
Number of goals scored: 42
Number of goals Received: 18
Season: 2014-2015
Number of matches: 19
Number of wins: 8
Number of ties: 6
Number of Loses: 5
Total points: 30
Number of goals scored: 28
Number of goals Received: 20
Season: 2015-2016
Number of matches: 19
Number of wins: 11
Number of ties: 4
Number of Loses: 4
Total points: 37
Number of goals scored: 35
Number of goals Received: 17

Atletico de Madrid
Season: 2011-2012
Number of matches: 19
Number of wins: 4
Number of ties: 6
Number of Loses: 9
Total points: 18
Number of goals scored: 17
Number of goals Received: 29
Season: 2012-2013
Number of matches: 19
Number of wins: 9
Number of ties: 5
Number of Loses: 5
Total points: 32
Number of goals scored: 23
Number of goals Received: 19
Season: 2013-2014
Number of matches: 19
Number of wins: 13
Number of ties: 2
Number of Loses: 4
Total points: 41
Number of goals scored: 28
Number of goals Received: 16
Season: 2014-2015
Number of matches: 19
Number of wins: 9
Number of ties: 6
Number of Loses: 4
Total points: 33
Number of goals scored: 25
Number of goals Received: 18
Season: 2015-2016
Number of matches: 19
Number of wins: 13
Number of ties: 1
Number of Loses: 5
Total points: 40
Number of goals scored: 30
Number of goals Received: 11

Figure 6. Performance of the “Group Comparison” algorithm with Athletic de Bilbao vs Atletico de Madrid.

Finally, to complete the tests with the whole algorithm set, the Individual Comparison option was chosen and the results can be found on figure 7.



Choose The Algorithm

Individual Comparison ▾

Valencia	Last 5 matches	Barcelona
6	Total goals	8
13	Average Shots	13
4	Average Corners	5
15	Average Fouls	10
3	Average Yellow Cards	3
0	Average Red Cards	0

2011-2012	Valencia	2	2	Barcelona
2012-2013	Valencia	1	1	Barcelona
2013-2014	Valencia	2	3	Barcelona
2014-2015	Valencia	0	1	Barcelona
2015-2016	Valencia	1	1	Barcelona

Figure 7. Performance of the “Individual Comparison” algorithm with Valencia vs Barcelona

7. Conclusion, Evaluation and Further Work

Along the project we finally materialized (code in GitHub [7]) all what we learned in web of data and semantic web in a real application, which is first initiative to get the online gambling world into the web 3.0, the semantic web. We have developed a simple and concise ontology which can be the base and the foundation for further development, but starting from now it can be used and shared among all the football gambling community and bookmakers so the data can be reutilized by different entities.

The application is a solid tool to know the statistics from different teams and from different matches, and also, most importantly, to make future predictions based on historical real data. Even though the algorithm can be improved in the future, we have a good start here.

The next steps are on different directions, on one side there is the need to make a more robust prediction algorithm based on more parameters such as players in the match, weather conditions, couch, etc., and on the other side the ontology could be made broader for all the different sports where gambling companies define the odds.

This is a good start, and as many companies and individuals use our ontology as better it will be because by having everybody sharing their information in RDF format, it will make life easier to everybody. Our systems can be automatically updated with the newest data by simply automating the SPARQL queries so the effort long term will be drastically reduce and it can be put on some other opportunity areas.

Bibliography

[1] La Liga
http://www.laliga.es/en
[2] Football Data. Results, Odds, Tipsters
http://www.football-data.co.uk/spainm.php
[3] Apache jena fuseki
http://jena.apache.org/documentation/serving_data/index.html
[4] Protégé
http://protege.stanford.edu/
[5] OpenRefine
http://openrefine.org/
[6] Linked Open Vocabularies (LOV)
https://lov.okfn.org/dataset/lov/
[7] Github La Liga Prediction
https://github.com/luisgaldo/LaLigaPrediction