Creating and Using Sports Linked Data: Applications and Analytics

Panagiotis-Marios Philippides
OKF Greece
Thessaloniki, Greece
filippidis.okfgr@gmail.com

Charalampos Bratsas
Mathematics Department
Aristotle University of Thessaloniki
OKF Greece
Thessaloniki, Greece
charalampos.bratsas@okfn.com

Andreas Veglis
School of Journalism & Mass Communications,
Aristotle University of Thessaloniki
veglis@jour.auth.gr

Evangelos Chondrokostas
Mathematics Department
Aristotle University ofThessaloniki
echondrok@gmail.com

Dimitra Tsigari
Mathematics Department
Aristotle University of Thessaloniki
dimitra.tsi@gmail.com

Ioannis Antoniou
Mathematics Department
Aristotle University of Thessaloniki
iantonio@math.auth.gr

ABSTRACT

Linked data have made significant progress over the last few years and many kinds of datasets are transformed into this format at a highly increasing rate, contributing to the openness, connectivity and re-use of web data. However, this progress is not the case for a popular sport like basketball, at least as far as the raw statistics is concerned. This kind of data contains valuable information that can be used by fans, teams and coaches, statisticians and other scientists. In this work, statistical data from Euroleague are transformed into linked data, thereby filling the relevant gap in the LOD Cloud, while ways of exploitation of them are presented, from fascinating applications for the fans, like the Euroleague Timeline, to cases of complex processing, analysis and visualization of data through software like R.

Categories and Subject Descriptors

[**World Wide Web**]: Web data description languages – *Resource Description Framework (RDF)*

[**Information Retrieval**]: Document representation – *Document structure, Ontologies*

[Probability and Statistics] : Statistical paradigms – *Statistical Graphics, Exploratory Data Analysis*

General Terms

Measurement, Design, Experimentation.

Keywords

Sports Open Data, Linked Data, Analytics, Data Visualizations

1. INTRODUCTION

Sports data can be valuable not only to anyone related to sports, but also to the scientific community, because the statistical information they include can widely describe what has happened in a sports game. Processing, modeling and visualization of these data can benefit areas such as sports analysis, either from the perspective of the players and their performance, or from the perspective of coaches and their tactics and can draw inference about the finding of the best players and teams, the detection of the sport's important elements, or the prediction of game situations, performance and results in ways and methods that can potentially then be used in other scientific fields.

A typical example is basketball, a sport full of statistics that can largely be represented through the boxscore, a table containing the performance in every statistical category for every player and team of a game. This amount of statistical data is sufficiently large so that very precise and detailed analytics about the sport of basketball can be made. However, basketball data are not usually available in large quantities, at least in their raw form and this leads the related scientific researches to devote a very large part of their time in searching for these data, that is not easy afterwards to share, or to link with similar data. That generates the need to transform such data in linked data form and this is the primary subject of this work, using Euroleague statistics, the top European basketball competition for clubs.

The benefits of the semantic enrichment of these data is more than obvious, since the openness of large volumes of structured data is valuable not only to the coaches, statisticians and other scientists of basketball, who could have an easy and direct access to data relevant to their job, but also to a large audience such as basketball fans, who could make in depth analyses of their favourite sport on their own. The statistical nature of these data increases the value of their openness, since they can be processed in many ways, from anyone interested, to lead to further inference about the sport. The linked data technologies themselves include means by which such information can be easily used and processed.

Besides the statistics of boxscores, additional data relating to the games of the competition such as the court and the date they took place have been transformed into linked data form too. This kind of information is essential and can link basketball data to other LOD datasets in many ways, so further information about games, teams or players can be reached and retrieved. This connection complements the statistical information and enriches the provided knowledge, while, a common way of modeling such data can benefit the comparison of similar data and increases their processing, analysis and visualization capabilities in favor of every stakeholder of the sport. Additionally, it provides a complete informational source for creating fascinating applications for basketball fans and encourages in turn initiatives to create and make use of linked data, thus benefits the LOD cloud itself, with further data enrichment and linkage. One such

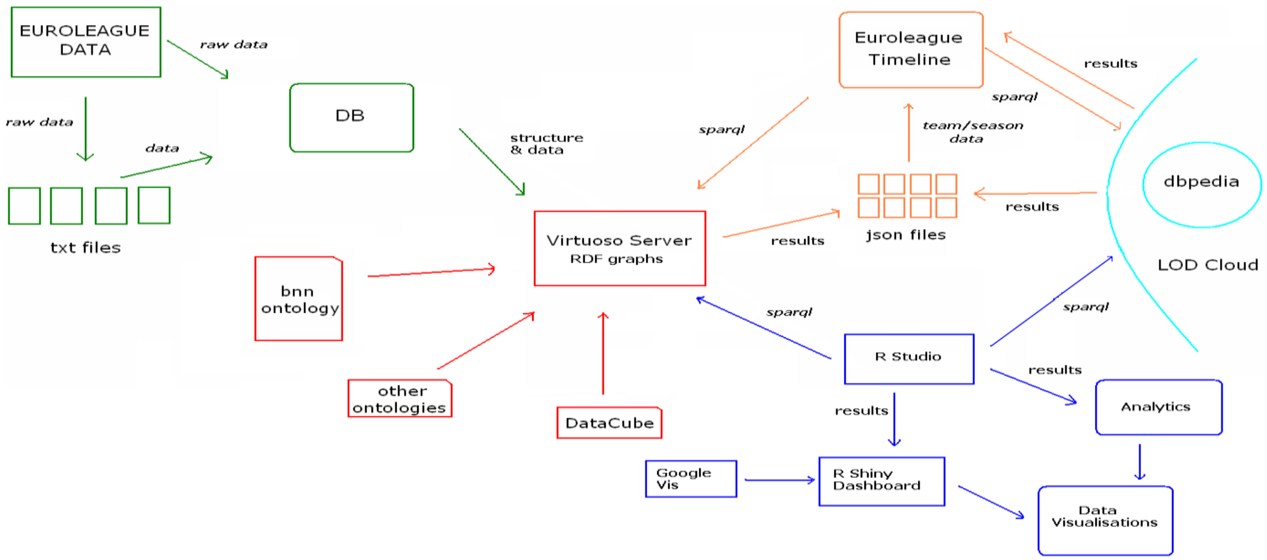


Figure 1. System Architecture

application is the Euroleague Timeline, as presented below, while some examples of data analytics and visualizations of basketball data are introduced, as the second subject of this work.

The entire system architecture is shown in Figure 1. The first stage is about data retrieval and processing, while the second stage involves the creation of linked data. Then, these data can be used in many ways, like the Euroleague Timeline or analytics and visualizations through software like R, while they are also linked with the LOD Cloud and especially DBpedia.

2. EUROLEAGUE LINKED DATA

The whole procedure of creating Euroleague linked data from raw data statistics is presented in this section.

2.1 Creating the RDF Graphs

Initially, there was a handcrafted extraction of data from the official website of Euroleague. Basic data of the games stored directly in databases, while boxscores statistics initially saved in text files to undergo the necessary processing. The cleaning and filtering process of the extracted boxscores involved tasks such as insertion of delimiter characters between the statistics, renaming "team" value with the corresponding team name, filling of empty values with zero values and separation of each shooting column (eg 10/11 2FG means 10 made two pointers and 11 attempted two pointers). After that, the statistics data of text files stored in databases too. For each season of the competition a unique database has been created. The main tables of a database are the boxscores table, containing the statistical information, so that each row of the table is a statline of a player or a team in a Euroleague game and the schedule table, containing the basic data of every game, such as time and date. Additional key tables have been created about the teams, the players, the phases of the competition, the groups of teams and the courts. The structure and data of all databases then stored in a Virtuoso Server.

The first step in transforming Euroleague data in linked data was the creation of an ontology, under which the mapping of relational data to RDF would take place. The basic classes created in the ontology are conceptually related with the main database tables, like the Statline class, whose properties are similar to the columns of the boxscores table, namely statistics such as the points of the player or team in a game. Statline class has additional properties, such as the game and the week the statline has been recorded. The same holds for the Game class, containing properties such as the teams of a game, the final score, the date and time, the court and the week it took place. Other basic classes is the Phase class, containing the name of the phase and its starting and ending week, the Group class, containing teams and the phase of the competition, the Team class, with the names of teams, their players and their courts, the Player class, which contains the names of players and the teams they are part of and the Court class, including the name and the geo-coordinates of the court. The whole ontology schema is illustrated in Figure 2.

Based mainly on this ontology, but also using additional ontologies such as foaf[1], skos[2], event[3] and timeline[4], the mapping of relational data to RDF was made, leveraging the quad map patterns of Virtuoso. A unique RDF graph for each season of the competition has been created.

Updating data with new games and statistics needs to perform almost the whole procedure, only for the specific amount of data, namely, the data retrieval and processing tasks, the insertion to the database, although different php files have to be executed in order to update the database and finally, the recreation of the year's rdf graph.

2.2 Datacube Integration

The next step is to integrate the Datacube vocabulary[5], the most appropriate ontology on statistical data. This process is a work in progress, the structure of the cube however has been quite defined. Most of the dimensions of the cube concern mainly the games data (date, time, teams, court, etc), but there is an extra dimension, the player dimension, referring to a subject of the statistics recorded in the game which is identified by the other dimensions. Since the statistics of the boxscore (points, rebounds, turnovers etc) have been defined as the measures of the cube, each observation is a statline of a player or a team, defined by the game it's been recorded. There are additional attributes in the cube that provide supplementary information, such as the player's team in a game, the number of his jersey and the unit of measure that is defined separately for every statistical measure.

[1] http://xmlns.com/foaf/0.1/

[2] http://www.w3.org/2004/02/skos/core#

[3] http://purl.org/NET/c4dm/event.owl#

[4] http://purl.org/NET/c4dm/timeline.owl#

[5] http://purl.org/linked-data/cube#

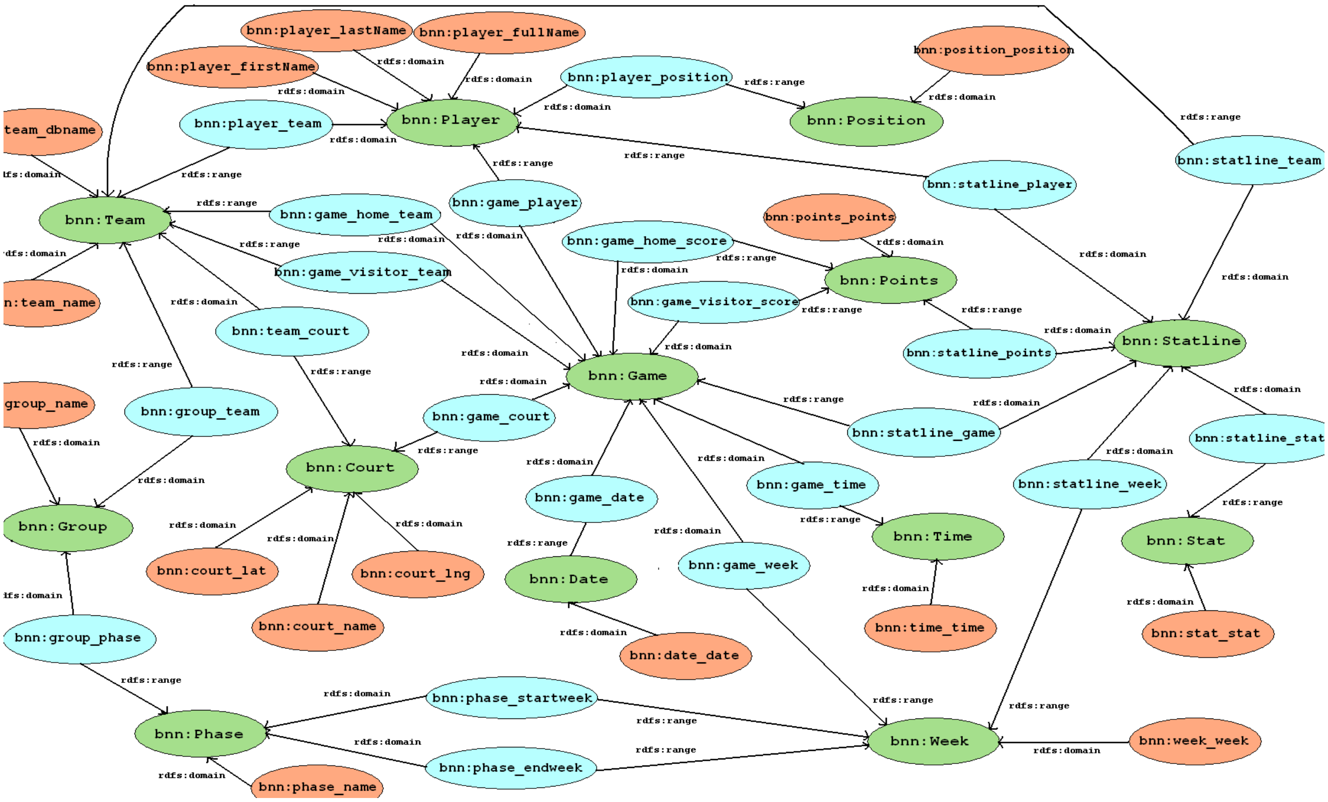


Figure 2. Ontology Schema

The dimensions of the cube may have as range the classes of the ontology that has already been created and thus contain, in this way, its properties, apart from the cube, while some code lists have been created for specific dimensions such as the season, the week, the phase, the group and the time dimension. Meanwhile, some slices that is likely to be used refer to players, teams, weeks and games, with each slice containing the relevant information and finally, all the above concepts have been defined as skos concepts, according to the Datacube specifications.

3. APPLICATIONS AND ANALYTICS

The Euroleague linked data that have been created can be exploited in many ways, such as applications and analytics, as presented below.

3.1 Euroleague Timeline

An application that makes the most of this work and all the advantages of linked data is the Euroleague Timeline, which is a timeline of the results of Euroleague games, containing information both on the basic data of the games and on their boxscores. The Euroleague Timeline involves two types of timelines, the team timeline, including all the games of a team in the competition and the season timeline, containing all the games of a season. In any case, games are displayed in a time series and after the user selection of a season or a team, he or she can navigate through the games either consecutively, or by selection, via the special time bar featured by the Timeline, which contains all the games of the season or the team.

The Euroleague Timeline is based on the TimelineJS, which loads json files to get and display the information, so that was the file format needed to extract data from Virtuoso. A separate json file has been created for every team and every season, while it is possible to update them with new games data. The information stored in each json file and appearing in the timeline is retrieved through a series of queries, both on the sparql endpoint of virtuoso containing the Euroleague data and on the DBpedia endpoint, to extract further information on players and teams. The final result contains all the relevant information (basic data, statistics, additional data from DBpedia) and is shown in Figure 3. The application is online at wiki.el.dbpedia.org/apps/Euroleague.

3.2 Data Visualizations

The plethora of statistical data that has been transformed into linked data is suitable for data processing and analytics and this task was carried through R Studio, which can make sparql queries to any endpoint through its packages. The retrieved data may then undergo any mathematical processing and visualization. Extra visualization capabilities are enabled through the R Shiny package, via its widgets, while the R Shiny Dashboard allows handling many visualizations that interact, simultaneously, thus serving as a complete information visualization framework, that can utilize other applications as well to display information such as Google Vis.

Using this technology, some visualizations exploiting Euroleague linked data have been generated, providing useful information and insights for basketball fans or even coaches such as:

- Table of rosters of teams for each season with the average statistics of the players
- Points and shots distributions of teams along with their shooting percentages
- Relations between turnovers and points and fouls and points for the teams of a game, for all games of a season
- Graph of the results of teams in the competition
- Players comparison via diagrams, based on their average statistics
- Map containing the teams of each season and each phase of the competition, with additional information from dbpedia

3.3 Further Analytics

Besides these visualizations, Euroleague statistics are used for further mathematical processing and analysis in order to examine measures, ratings and relations that could yield useful results. Some examples already done on these data by this work are:

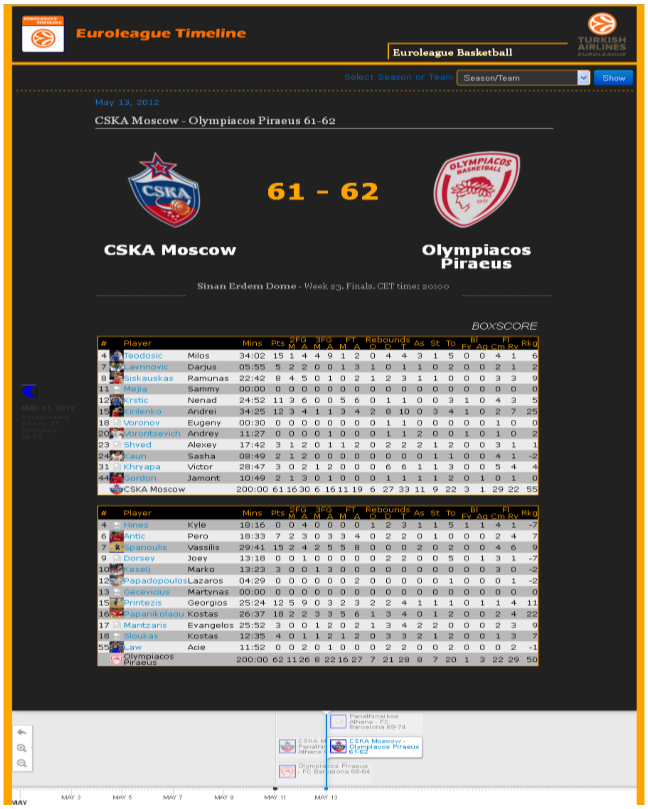
- research on the teamwork of teams and its relation to their success, relying on categories like assists, points and turnovers
- research on individual defensive actions and on relations between the steals, the blocks and the fouls, along with predicting the number of steals and blocks of a player under the fouls and the court he plays at (home/away)
- Evaluating the best players per position on the basis of normalized equations of their statistics
- Creating and analyzing a network of players who have played in Euroleague
- research on the correlation of the shooting percentages of the two teams of a game with the final result and their points difference

4. FUTURE WORK

A large volume of basketball data has been transformed into linked data, however it could be further enriched, especially with the play-by-plays of games, which contain all the actions of the players that are statistically recorded, in order of time. This would increase significantly the information processing, analysis and visualization capabilities. The examples that have been made so far in this work is only the beginning and there are still countless topics on these data that can be explored, as well as many other ways of analysis. Their combination is the step forward and can lead to applications and results that will reveal and provide additional knowledge on basketball, which would be readily accessible to every fan, through tools that leverage linked data.

5. CONCLUSION

Basketball linked data offer a variety of possibilities in sports, statistical and scientific field, because of their large volume of statistics. This work transforms Euroleague basketball data into linked data to enrich the LOD Cloud with valuable sports statistics and to utilize these data in various ways, such as creating fascinating applications, like the Euroleague Timeline, or by processing and analyzing these statistics through R, to draw useful inference and display the corresponding diagrams, thus demonstrating the enormous range of capabilities offered by linked data in a sport that is full of statistical information.



May 13, 2012

CSKA Moscow - Olympiacos Piraeus 61-62

Sinan Erdem Dome - Week 23. Finals. CET time: 20:00

#	Player		Mins	Pts	2FG		3FG		FT		Rebounds			As	St	To	Bl		Fl		Rkg
					M	A	M	A	M	A	O	D	T				Fv	Ag	Cm	Rv	
4	Teodosic	Milos	34:02	15	1	4	4	9	1	2	0	4	4	3	1	5	0	0	4	1	6
7	Lavrinovic	Darjus	05:55	5	2	2	0	0	1	3	1	0	1	1	0	2	0	0	2	1	2
8	Siskauskas	Ramunas	22:42	8	4	5	0	1	0	2	1	2	3	1	1	0	0	0	3	3	9
11	Mejia	Sammy	00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Krstic	Nenad	24:52	11	3	6	0	0	5	6	0	1	1	0	0	3	1	0	4	3	5
15	Kirilenko	Andrei	34:25	12	3	4	1	1	3	4	2	8	10	0	3	4	1	0	2	7	25
18	Voronov	Eugeny	00:30	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0
20	Vorontsevich	Andrey	11:27	0	0	0	0	1	0	0	0	1	1	2	0	0	1	0	1	0	2
23	Shved	Alexey	17:42	3	1	2	0	1	1	2	0	2	2	2	1	2	0	0	3	1	1
24	Kaun	Sasha	08:49	2	1	2	0	0	0	0	0	0	0	0	1	1	0	0	4	1	-2
31	Khryapa	Victor	28:47	3	0	2	1	2	0	0	0	6	6	1	1	3	0	0	5	4	4
44	Gordon	Jamont	10:49	2	1	3	0	1	0	0	0	1	1	1	1	2	0	1	0	1	0
	CSKA Moscow		200:00	61	16	30	6	16	11	19	6	27	33	11	9	22	3	1	29	22	55

#	Player		Mins	Pts	2FG		3FG		FT		Rebounds			As	St	To	Bl		Fl		Rkg
					M	A	M	A	M	A	O	D	T				Fv	Ag	Cm	Rv	
4	Hines	Kyle	18:16	0	0	4	0	0	0	0	1	2	3	1	1	5	1	1	4	1	-7
6	Antic	Pero	18:33	7	2	3	0	3	3	4	0	2	2	0	1	0	0	0	2	4	7
7	Spanoulis	Vassilis	29:41	15	2	4	2	5	5	8	0	0	0	2	0	2	0	0	4	6	9
9	Dorsey	Joey	13:18	0	0	1	0	0	0	0	0	2	2	0	0	5	0	1	3	1	-7
10	Keselj	Marko	13:23	3	0	0	1	3	0	0	0	0	0	0	0	0	0	0	3	0	-2
12	Papadopoulos	Lazaros	04:29	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	1	-2
13	Gecevicius	Martynas	00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Printezis	Georgios	25:24	12	5	9	0	3	2	3	2	2	4	1	1	1	0	1	1	4	11
16	Papanikolaou	Kostas	26:37	18	2	2	3	3	5	6	1	3	4	0	1	2	0	0	2	4	22
17	Mantzaris	Evangelos	25:52	3	0	0	1	2	0	2	1	3	4	2	2	0	0	0	2	3	9
18	Sloukas	Kostas	12:35	4	0	1	1	2	1	2	0	3	3	2	1	2	0	0	1	3	7
55	Law	Acie	11:52	0	0	2	0	1	0	0	0	2	2	0	0	2	0	0	0	2	-1
	Olympiacos Piraeus		200:00	62	11	26	8	22	16	27	7	21	28	8	7	20	1	3	22	29	50

Figure 3. A game in the Euroleague Timeline

6. REFERENCES

[1] Bizer C., Heath T., and Berners-Lee T. 2009. *Linked Data - the story so far*. Int. J. Semantic Web Inf. Syst, 5(3):1-22

[2] Klyne G. and Carroll J. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax.

[3] Lehmann J., Bizer C., Kobilarov G., Auer S., Becker C., Cyganiak R., and Hellmann S. 2009. DBpedia - a crystallization point for the web of data. Journal of Web Semantics, 7(3): 154-16