Generating a Linked Soccer Dataset

Tanja Bergmann* Stefan Bunk* Johannes Eschrig*
Christian Hentschel** Magnus Knuth** Harald Sack**
Ricarda Schüler*

Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
*firstname.lastname@student.hpi.uni-potsdam.de
**firstname.lastname@hpi.uni-potsdam.de

ABSTRACT

The provision of Linked Data about sporting results enables extensive statistics, while connections to further datasets allow enhanced and sophisticated analyses. Moreover, providing sports data as Linked Open Data may promote new applications, which are currently impossible due to the locked nature of today's proprietary sports databases. Though the sport domain is strongly underrepresented in the Linked Open Data Cloud. In this paper we present a dataset containing information about soccer entities crawled from heterogeneous sources and linked to related entities from the LOD cloud. To enable easy exploration and to illustrate the capabilities of the dataset a web interface is providing a structured overview and extensive statistics.

Keywords

Linked Data, Soccer, Information Extraction, Triplification

1. INTRODUCTION

The Linked Open Data (LOD) Cloud comprises 870 datasets containing more than 62 billion triples [1]. The majority of triples describes governmental (42 %) and geographic data (19 %), whereas Linked Data about sports is strongly underrepresented. Sport competition results are collected by various authorities and other parties, they are connected to events, teams, players, etc. Providing Linked Data about sports and sporting results enables extensive statistics, while connections to further datasets allow enhanced and sophisticated analyses. Moreover, providing sports data as Linked Open Data may promote new applications, which are currently impossible due to the locked nature of today's proprietary sports databases. By enabling linkage to additional resources such as geographical, weather, or social network data, interesting statistics for the sport enthusiast can be easily derived and provide further information that would be hidden otherwise.

In this paper we describe an extensive RDF dataset of soccer data providing soccer matches, teams, and player information, collected from heterogeneous sources and linked to LOD datasets like the DBpedia. The raw data was collected via APIs and by crawling authorities' websites, like UEFA.com or Fussballdaten.de, and is linked to further web resources for supportive information, such as Twitter postings for most recent information, Youtube videos for multimedia support, and weather information.

Based on this aggregated new dataset we have implemented an interactive interface to explore this data. The interface provides various statistics that have been made possible solely by the aggregation of diverse sources.

2. RELATED WORK

The BBC Future Media and Technology department applies semantic technologies according to their Dynamic Semantic Publishing (DSP) strategy [3] to automate the publication, aggregation, and re-purposing of inter-related content objects. The first launch using DSP was the BBC Sport FIFA World Cup 2010 website[1] featuring more than 700 team, group and player pages. However, the data internally exploited by the system is not published as Linked Data.

An extensive dataset of soccer data is aggregated by footytube. According to their website[2] the data is crawled from various sources and connected by semantic technologies. The recipes, however, are not described in detail. Footytube's data include statistics about soccer matches and teams, as well as related media content, such as videos, news, podcasts, and blogs. The data is accessible via the openfooty API but is subject to restrictions that forbid the republication as Linked Data.

Generally, it is hard to find open data about sport results, since exploitation rights are possessed by responsible administrative body organizations. An approach to liberate sport

[1] `http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/default.stm`
[2] `http://www.footytube.com/aboutus/search-technology.php`

ISEM '13 September 04 – 06 2013, Graz, Austria
 ACM 978-1-4503-1972-0/13/09...$15.00.
http://dx.doi.org/10.1145/2506182.2506192

results are community-based efforts, such as OpenLigaDB[3], which collect sports data for public use. As to the authors' best knowledge, the presented dataset provides the first extensive soccer dataset published as Linked Data, consisting of more than 9 million triples.

3. LINKED SOCCER DATASET

Our intention was to create a dataset including reliable information about soccer events covering as much historical data as available including recent competition results. A cross domain dataset such as DBpedia is not sufficient for this purpose, since soccer data in DBpedia is incomplete and unreliable.

The dataset is aggregated from raw data originating from multiple sources, namely Fussballdaten.de[4], Uefa.com[5], DBpedia[6], the Twitter feed of the Kicker magazine[7], the Sky Sport HD Youtube Channel[8], and weather information from Deutscher Wetterdienst[9]. Fussballdaten.de and Uefa.com provide match results and player information. By analyzing the Twitter feed, live match data (Kicker updates its feed with live results) is extracted and free text tweets are used to find latest news about players or teams. Our data collection ranges from the 1960s until today and is updated constantly. Updates are scheduled every matchday, while the Twitter feeds are analyzed every 30 seconds during running matches. Currently, the dataset contains information about the German 1. and 2. Bundesliga, UEFA Champions League, European and World Championships. Additional leagues can easily be included by setting up new data sources.

In order to be able to describe detailed information about soccer data we have created a dedicated vocabulary *Soccer Voc*[10], which reuses basic parts of the *BBC Sport Ontology* [2] but includes also soccer specific classes and properties to describe in-game-events (e. g. goal, substitution, booking), various persona classes (soccer player, referee, soccer manager), soccer teams and clubs, stadiums, seasons, as well as relations between them.

The data from the named sources is converted to and persistently stored as RDF triples. Each entity is referenced by a unique URI, which unites all facts about the entity, regardless of the originating data source. As an example, Listing 1 shows a selection from a total of 1,064 triples about the player *Mehmet Scholl*.

Listing 1: Triples applying the player entity *Mehmet Scholl*

```
1  smm:Mehmet_Scholl_1970-10-16 rdf:type smm:
       SoccerPlayer ;
2    rdfs:label  "Mehmet Scholl"@en ;
3    smm:name "Mehmet Scholl" ;
4    smm:birthDate "1970-10-16" ;
5    smm:nationality "D" ;
6    smm:height "177" ;
7    smm:weight "70" ;
8    smm:website <http://www.mehmet-scholl.com> ;
9    smm:image <http://fussballdaten.de/bilder/vereine/
         bayernmuenchen/2000-2001/schollmehmet.jpg> ;
10   smm:fussballdatenUrl <http://www.fussballdaten.de/
         spieler/schollmehmet/> .
11 smm:Booking_Mehmet_Scholl_1970-10-16_YellowCard
       _Match_Real_Madrid_CF_FC_Bayern_Muenchen_2000
       -05-03
12   smm:cautionedPlayer smm:Mehmet_Scholl_1970-10-16 .
13 smm:Goal_Mehmet_Scholl_1970-10-16
       _Match_Deutschland_Rumaenien_2000-06-12_28
14   smm:scorer smm:Mehmet_Scholl_1970-10-16 .
15 smm:Match_Real_Madrid_CF_FC_Bayern_Muenchen_2000
       -05-03
16   smm:startPlayer smm:Mehmet_Scholl_1970-10-16 .
```

Due to the legal restrictions on the crawled sources it is not possible to make the aggregated dataset as a whole publicly available. Therefore we publish an unrestricted subset, which comprises merely the list of all entities with the respective label, resource type information, and mappings to DBpedia and other web resources. This dataset is available as an RDF dump at `http://mediaglobe.yovisto.com/SoccerLD/dump/public.ttl.gz`.

3.1 Matching

The representation of an entity by a source can vary strongly. Therefore, it is necessary to recognize, which representations incorporate the same entity and to combine the information of multiple representations into a single entity. The algorithm for matching a player entity from Fussballdaten.de and UEFA.com to DBpedia can be outlined as follows:

1. *Querying DBpedia for potential entities matching a query name:* The DBpedia SPARQL endpoint does not support fuzzy search functionality. In order to cover variations, the player's name is split into substrings at whitespace characters and DBpedia is queried separately for each part of the name. Each part needs to match one of the following attributes: `rdfs:label`, `dbprop:name`, or `foaf:name`. The check is done using the Virtuoso full text matching function `bif:contains`. Doing so, a list of possible candidates is returned.
2. *Filtering results from previous step:* The name listed for each of the returned DBpedia entities is compared to the player's name using the Levenshtein distance. If the distance is below a certain threshold (we apply a value of 2 here), the DBpedia entity is considered as a potential match.
3. *Validating the DBpedia entity:* The entity is regarded as correct if it is a person, the date of birth (if present) matches that of the potential match and the person is listed as a football player.

Although it might seem plausible to start the process by matching on the date of birth first in order to narrow down search space, we observed that often player entities in DBpedia do not come with a valid date of birth. Therefore, matching on this attribute in first place would strongly affect the number of potential entity candidates.

[3] `http://www.openligadb.de/`
[4] `http://www.fussballdaten.de/`
[5] `http://www.uefa.com/`
[6] the original `http://dbpedia.org/` as well as German DBpedia `http://de.dbpedia.org/` have been applied for matching
[7] `http://twitter.com/kicker_bl_li`
[8] `http://www.youtube.com/user/SkySportHD`
[9] `http://www.dwd.de/`
[10] `http://purl.org/hpi/soccer-voc/`

Local entities are connected to their respective DBpedia entity using `owl:sameAs`.

3.2 Matching Twitter

Analysis of Kicker tweets is performed for two purposes: (1) obtaining live data about running matches for presentation on the website and (2) finding news articles with background information about specific teams. Using the Kicker twitter streams[11] it is possible to find both, as these feeds offer live data, as well as free text tweets teasing articles of the Kicker website. Game result tweets have a common, strict pattern, which renders them simple to parse. It can be assumed that these tweets are generated automatically, as there are no pattern violations. One example tweet could be

Borussia Dortmund - Bayern München 1:2 Tor: Robben (89., Ribery) #BVBFCB http://bit.ly/14O58cp

which reveals the pattern

[home team] - [guest team] [current result] Tor: [scorer] ([minute], [assist]) [hashtag, containing of three letters for each team] [link to detail page]

The parts can be extracted using regular expressions. In order to map these elements to entities in our data, we firstly match the team names using the Levenshtein distance. As we have only 18 teams to consider for a Bundesliga match, this works in 100 % of all cases. For matching both players (scorer and assist) we consider all players who played for the goal-scoring team in the last two years. Matching is done by last name and works in 98 % of all cases.

In order to find links to background information about a team, free text tweets are parsed. An example for such a tweet is:

Bayern's triumph - All about the magnificent final in Wembley #BL http://bit.ly/Z8KpyE"

Since team names are usually abbreviated or paraphrased within Kicker tweets (e. g. 'Bayern' or 'FCB' instead of 'FC Bayern Munich' a list of common short names for each team has been assembled automatically. This list is generated from parts of the team name, employed hashtags, and DBpedia nicknames. In the case of Bayern Munich this list contains names like 'Bayern', 'Munich', 'Rekordmeister', 'FCB' and more. By using this list we are able to match team entities within a tweet.

If one or more of these names are part of a tweet, it is considered to refer to the specific team.

3.3 Statistics

At the time of this writing, the dataset comprises descriptions of 56,537 soccer players, 1,402 clubs, 1,514 teams, 38,098 matches, 98,794 goals, 1,686 referees, 1,870 managers, 717 stadiums, and 207 seasons or competition series. In total, 9 million triples have been generated up to now, 3.35 million of which originate from raw data of Fussballdaten.de and 2.10 million triples from the UEFA.com website.

3.4 Evaluation

In order to evaluate the quality of the matching, a percentage of matched entities has been reviewed. The correctness of these matches has been confirmed by manually judging the results. For Bundesliga, all teams (54) and about 78 % of all players (6,790) have been matched successfully to DBpedia entities. Missing matches were mostly due to missing player entities in DBpedia.

The algorithm for matching the results of soccer matches from the Twitter feed was capable of matching 99.60 % results correctly, which was confirmed by a evaluation on a random data set of 200 tweets. The algorithm for matching team names in tweets containing unstructured text resulted in matching 1,574 of a total of 1,780 Tweets. The unmatched tweets contained no team names, so that a match was not possible. An evaluation of a random data set of 150 Tweets showed that all had been correctly matched.

3.5 Application

The dataset can be accessed via a demonstrator website[12], where each entity is presented on its own page with relevant information, statistics, and links to related entities. Additionally, a variety of possible complex queries are demonstrated, such as "Which player scored most often on himself?" or "From which foreign country do most players in the last Bundesliga season come from?". These show the possibilities that come with providing soccer data as Linked Data. In Figure 1, two different views of the website are shown.

4. CONCLUSION AND OUTLOOK

In this paper we presented a rich soccer dataset, which, to our best knowledge, is the first comprehensive linked soccer dataset. We published non-restricted portions of the dataset, since the publication of the dataset as a whole is prevented by legal rights belonging to the respective authorities. An application based on this data not only allows to browse the dataset and to provide various statistics about players, teams and matches but also exploits the advantages of Linked Data principles in order to provide additional information currently not considered by available soccer datasets.

Possible additions could include advanced and more detailed data such as the number of ball contacts, played passes, or the distance covered by a player during a match. Currently, the developed extraction framework covers data mainly from the Bundesliga. Integration of leagues from further countries as well as lower regional leagues, however, is possible since a lot of this data is published by enthusiasts on the web. An interface for data editing in order to provide users of our dataset with the ability to submit own data, as well as to edit and correct present data is likewise imaginable. Future extensions will also aim to include rich media content, such

[11] e. g. `https://twitter.com/kicker_bl_li`

[12] `http://mediaglobe.yovisto.com/SoccerLD/`

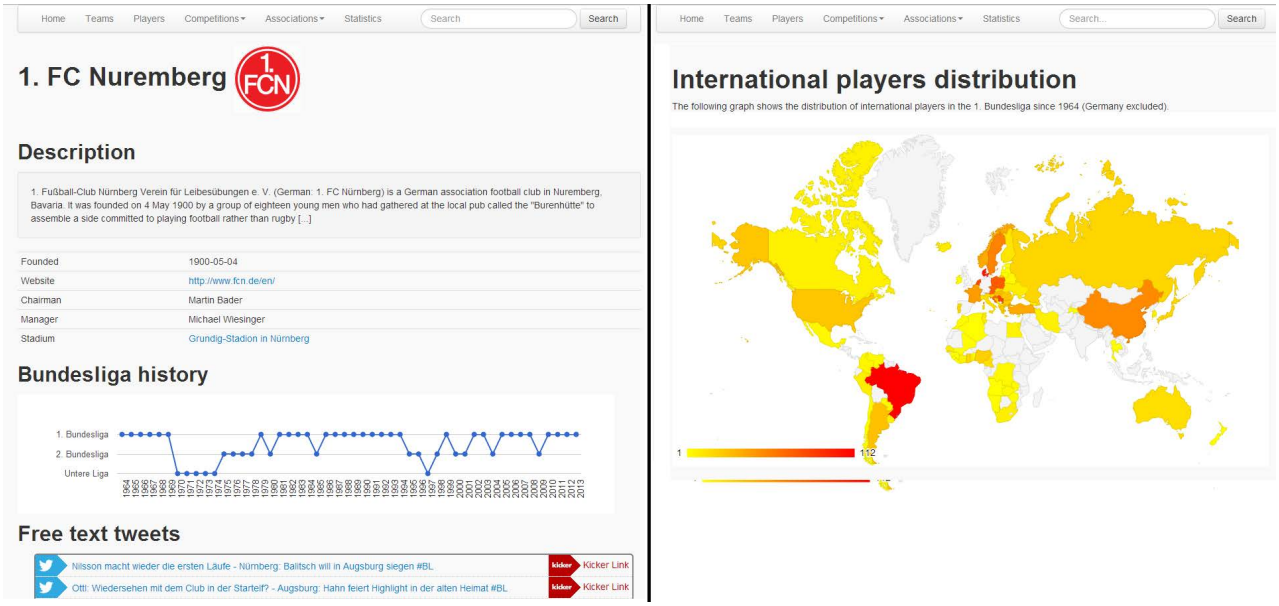

Figure 1: *Left*: Information about a German soccer club, among others a graph showing promotions and relegations (generated from match data) and free text tweets belonging to this club, *Right*: Map visualization about the distribution of international players in the Bundesliga since 1963, generated from player data.

as articles from sport magazines, interviews, team presentations, or background stories of players.

5. REFERENCES

[1] J. Demter, S. Auer, M. Martin, and J. Lehmann. Lodstats – an extensible framework for high-performance dataset analytics. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.

[2] S. Oliver. Enhancing the bbc's world cup coverage with an ontology driven information architecture. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[3] J. Rayfield. *Semantic Technologies in Content Management Systems*, chapter Dynamic Semantic Publishing, pages 49–64. Springer Berlin, Heidelberg, 2012.