Linked Soccer Data

Tanja Bergmann[1], Stefan Bunk[1], Johannes Eschrig[1], Christian Hentschel[2], Magnus Knuth[2], Harald Sack[2], and Ricarda Schüler[1]

[1]firstname.lastname@student.hpi.uni-potsdam.de
[2]firstname.lastname@hpi.uni-potsdam.de

Hasso Plattner Institute for Software Systems Engineering, Potsdam, Germany

Abstract. The sport domain is strongly under-represented in the Linked Open Data Cloud, whereas sport competition results can be linked to already existing entities, such as events, teams, players, and more. 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. We present a dataset containing information about soccer matches, teams, players and so forth crawled from from heterogeneous sources and linked to related entities from the LOD cloud. To enable exploration and to illustrate the capabilities of the dataset a web interface is introduced providing a structured overview and extensive statistics.

Keywords: Linked Data, Soccer, Information Extraction, Triplification

1 Introduction

The Linked Open Data (LOD) Cloud includes 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 under-represented. Sport competition results are collected by various authorities and other parties, they are connected to events, teams, players, etc. Providing also 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

[1] http://stats.lod2.eu/

sources and linked to LOD datasets like the DBpedia. The raw data was collected via APIs and crawling from 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.

2 Related Work

The BBC Future Media and Technology department applies semantic technologies according to their Dynamic Semantic Publishing (DSP) strategy [2] 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[2] featuring more than 700 team, group and player pages. But, the data used by the system internally is not published as Linked Data.

An extensive dataset of soccer data is aggregated by footytube. According to their website[3] the data is crawled from various sources and connected by semantic technologies, though the recipes are not described in detail. Footytube's data include soccer 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 interdict the re-publishing 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 results are community-based efforts, such as OpenLigaDB[4], which collect sport data for public use. Van Oorschot aims to extract in-game events from Twitter [3]. 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 many historical data as available including recent competition results. For this purpose DBpedia as cross domain dataset is not sufficient, since soccer data in DBpedia is incomplete and unreliable.

The dataset is aggregated from raw data originating from Fussballdaten.de[5], Uefa.com[6], DBpedia[7], the Twitter feed of the Kicker magazine[8], the Sky Sport

[2] `http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/default.stm`
[3] `http://www.footytube.com/aboutus/search-technology.php`
[4] `http://www.openligadb.de/`
[5] `http://www.fussballdaten.de/`
[6] `http://www.uefa.com/`
[7] the original `http://dbpedia.org/` and German DBpedia `http://de.dbpedia.org/` have been applied for matching
[8] `http://twitter.com/kicker_bl_li`

HD Youtube Channel[9], and weather information from Deutscher Wetterdienst[10]. Fussballdaten.de, Uefa.com, and Kicker.de offer match results and player information. The Twitter feed is used both for parsing live match data (Kicker updates its feed with live results) and to analyse free text tweets for latest news about players or teams. The time frame of our data collection ranges from the 1960s until today and is updated constantly. Updates are scheduled every matchday, while the Twitter feeds are refreshed every 30 seconds during running games. Additional leagues can be included by setting up new crawlers, or by providing an interface for manual submission. Currently, the dataset contains information about 1. and 2. Bundesliga, the Champions League, European and World Championships.

The data from these sources is converted and persistently stored as RDF triples describing resources such as soccer player, soccer teams, matches, associations, different types of in-game-events, and seasons. Each entity is referenced by a unique URI, which unites all facts, from whatever source they originate, about the entity.

For describing the information about soccer data we have created a vocabulary *Soccer Voc*[11], which extends the *BBC Sport Ontology*[1] with soccer specific classes and properties.

The dataset comprises descriptions of about 57,000 soccer players, 1,500 teams, 1,400 clubs, 1,500 referees, 1,800 managers, 700 stadiums, 38,000 matches, 97,000 goals, and 207 seasons or competition series. In total 9 million triples have been generated up to now. About 3.35 million triples originate from raw data from Fussballdaten.de and 2.10 million triples from the UEFA.com website.

In order to evaluate the quality of the matching, a percentage of matched entities has been reviewed. The correctness of these matches was confirmed by manually comparing the results to a data sample. 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.

4 Application

The soccer dataset comprises a diverse amount of information, both historic and present data. As the data set contains data about every match played, it is possible to create queries for all types of entities in a soccer match, e. g. all games of a particular referee, or all games played in a specific stadium. By querying the data, the user can find interesting statistics about the world of soccer, or find information about his or hers favorite club.

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

[9] `http://www.youtube.com/user/SkySportHD`
[10] `http://www.dwd.de/`
[11] `http://purl.org/hpi/soccer-voc/`
[12] `http://mediaglobe.yovisto.com/SoccerLD/`

related entities. Additionally, a variety of possible complex queries are demonstrated, such as "Which player is most important for his team?", "From which foreign country do most players in the last Bundesliga season come from?", or "Which team performs best in rainy weather?". In Figure 1, two different views of the website are shown.

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

5 Conclusion and Outlook

We presented a rich soccer dataset, which is to our best knowledge the first comprehensive linked soccer dataset. We published non-restricted parts of the dataset, the publication of the dataset as a whole is prevented by legal rights belonging to the respective authorities. Applications based on this data not only allow for typical statistical information about players and matches but also exploit the advantages of Linked Data principles in order to provide additional information currently not considered by available soccer datasets. We have developed and deployed a website in order to conveniently browse the dataset and provide various statistics that exemplify the advantage of aggregating multiple resources as Linked Data.

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. Integrating such data even more sophisticated queries could be answered. Further extensions of the dataset include also articles from sport magazines like interviews, team presentations, or background stories of players.

References

1. S. Oliver. Enhancing the BBC's world cup coverage with an ontology driven information architecture. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *9th International Semantic Web Conference (ISWC2010)*, November 2010.
2. J. Rayfield. Dynamic semantic publishing. In W. Maass and T. Kowatsch, editors, *Semantic Technologies in Content Management Systems*, pages 49–64. Springer Berlin Heidelberg, 2012.
3. G. van Oorschot, M. van Erp, and C. Dijkshoorn. Automatic extraction of soccer game events from twitter. In M. van Erp, L. Hollink, W. R. van Hage, R. Troncy, and D. A. Shamma, editors, *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, volume 902, pages 21–30, Boston, USA, 11 2012. CEUR.