An Ontology for the Common Data Format on Football Match Data

Fajar J. Ekaputra[1,2], Gregor Käfer[1] and Matthias Kempe[3]

[1]Institute of Data, Process, and Knowledge Management, Department of Information Systems and Operation Management, Vienna University of Economics and Business (WU), Vienna, Austria

[2]Data Science Research Unit, Institute of Information Systems Engineering, TU Wien, Vienna, Austria

[3]Department of Biomechanics, Kinesiology and Computer Science in Sport, Centre for Sport Science and University Sports, Vienna, Austria

Abstract

Artificial intelligence (AI) applications in sports, particularly for football (soccer), have been growing in recent years, e.g., for player recruitment, performance monitoring, and selection. To support such applications, the availability of an integrated, high-quality dataset is crucial to ensure accurate results. This aspect is especially vital due to the heterogeneity in data acquired by various stakeholders, e.g., companies and football clubs. Catering to such demand, a recent work proposed a common data format (CDF) schema for football match data to ensure the provided data is precise, sufficiently contextualized, and complete to enable typical downstream analysis tasks. This paper reports on an initial effort to create the Football Common Data Format (FCDF) ontology as a schema for the RDF serialization of the CDF core concepts, focusing on streamlining concepts, properties, and attributes. The FCDF ontology aims to provide a formal, shared conceptualisation of CDF to promote using ontology and KGs for AI applications in football.

Keywords

Football, Common Data Format, Ontology, Data Analytics

1. Introduction

The application of Artificial Intelligence (AI) in sports—particularly football (soccer)—has witnessed significant growth in recent years. AI-driven approaches are increasingly being used for various tasks such as player recruitment, performance monitoring, and team selection. This rising adoption can be attributed to two key developments [1]: (a) the substantial increase in both the quantity and quality of data available, driven by advances in technology providers and associated financial incentives, and (b) widely publicised success stories of data-driven decision-making in sports, which was recently summarized in the book "How to win the premier league" by the former head of analytics of FC Liverpool [2].

Despite this progress, practical barriers often hinder the effective use of football data for AI applications. These include ambiguities and inconsistencies in defining key events across different data vendors, challenges in the distribution formats and data structures, aggregation over temporal data, and semantic inconsistencies such as differing units of measurement or component definitions [3, 4]. Prior approaches for standardizing data exchange for sports data, such as SportsML[1], are developed for general sports and do not specifically aim for Football data, and therefore do not fully address the specific needs of the Football community. Other efforts, such as Soccer Player Action Description Language (SPADL) [5, 6], are more focused on specific aspects of match data and do not provide a comprehensive view of the football match.

To address these issues, the Football Common Data Format (CDF) has recently been proposed as a community-driven effort [4]. The CDF initiative seeks to standardise and synchronise football data representations to enable interoperability and reduce heterogeneity. It involves collaboration between academic researchers, professional and football team practitioners, and regulatory bodies. The CDF

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

fajar.ekaputra@wu.ac.at (F. J. Ekaputra); gregor.kaefer@wu.ac.at (G. Käfer); matthias.kempe@univie.ac.at (M. Kempe)

[1]https://iptc.org/standards/sportsml-g2/

was also evaluated within a Delphi study by 200 experts in the field, who accepted the proposed format. The current implementation of the Football CDF is available as a JSON schema, with data represented in either JSON (for non-streamed data) or JSON Lines (for streamed or video-derived data). These formats were chosen for their compact memory usage, extensibility, and broad tool support. The Football CDF aims to provide precise, contextualised data (e.g., with well-defined provenance) and complete, enabling common downstream tasks typically handled by sub-symbolic AI methods.

Despite the growing adoption of AI technologies in the Football domain, however, most existing approaches in football analytics emphasise sub-symbolic AI [7, 8] and there has been relatively limited exploration of symbolic or neurosymbolic AI techniques. One of the key reasons for this gap is the absence of a standard, widely accepted symbolic representation that effectively bridges symbolic reasoning with practical data formats used in the field. We argue that ontologies and knowledge graphs (KG) have the potential to become the bridge to enable such approaches. Ontologies on football could become a key component that supports researchers and practitioners in developing analytic tools. Several early efforts in the Semantic Web community have introduced ontologies for sports data[2] [9], while others specifically aim to represent football data [10, 11, 12]. However, these ontologies often lack practical alignment with the existing data exchange standards and needs of domain practitioners and data analysts, resulting in limited adoption. The usefulness of having such ontologies, if adopted by a large community, was already shown in other fields such as transportation [13, 14].

This work aims to bridge this gap by introducing the Football Common Data Format (FCDF) Ontology, an RDF-based serialisation of the Football CDF. Our approach seeks to enable symbolic and neurosymbolic AI applications in football by offering a formal, machine-interpretable representation compliant with the football CDF's community-driven exchange data format. The FCDF ontology will open up the possibility of utilising ontological reasoning for football data analysis. It would allow the utilization of various tools and methods developed within the Semantic Web community in the last 25 years, including the SPARQL querying protocol, RDFS/OWL reasoners, and SHACL for data validation.

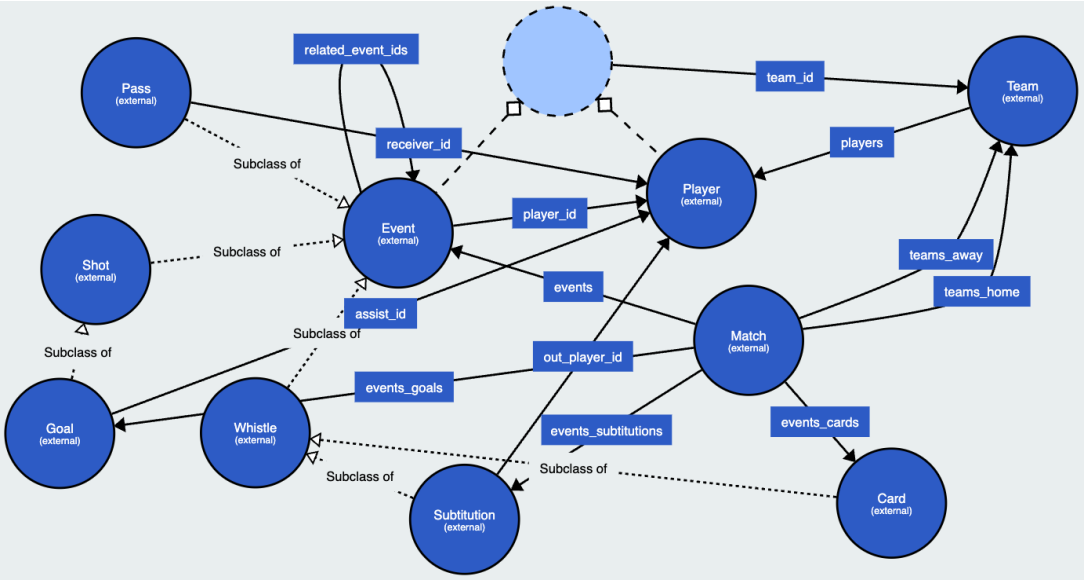

Figure 1: The core classes and object properties of the Football-CDF Ontology

2. The FCDF Ontology

The main goal of the FCDF ontology is to provide a fully compatible representation of the CDF format in RDF. To this end, we aim to allow for bi-directional transformations between JSON and RDF

[2]https://www.bbc.co.uk/ontologies/sport

representations of the Football CDF, which contain the following types of football match data: (a) Match Sheet, (b) Event Data, (c) Match Metadata, (d) Video Footage, and (e) Body Tracking [4].

In this paper, we focus on representing the first three components—Match Sheet, Events, and Match Metadata in FCDF Ontology as they are the most used forms of data [8] —while leaving the Video Footage and Body Tracking for future work. Figure 1 illustrates the core classes and object properties of the current Football-CDF ontology[3], using the WebVOWL notation [15].

Table 1: Mappings between CDF concepts (MatchSheet, Event, MatchMetadata) and FCDF Classes

#	CDF table	CDF JSON root	Description	FCDF class	FCDF superclass
1	MatchSheet	match	Football match.	Match	-
2	MatchSheet	match/status	The status of a football match, including venue neutrality, extra time, or a shootout.	MatchStatus	-
3	MatchSheet	match/result	The result of a football match, including the score and period of the result.	MatchResult	-
4	MatchSheet	teams/{home|away}	The home or away team.	Team	-
5	MatchSheet	teams/home|away/players/{i}	Football player.	Player	-
6	MatchSheet	referees/{i}	Referee of the football match.	Referee	-
7	MatchSheet	events/goals/{i}	Goal event.	Goal	Shot
8	MatchSheet	events/substitutions/i	Substitution event.	Substitution	Whistle
9	MatchSheet	events/cards/{i}	Card event.	Card	Whistle
10	MatchSheet	meta	Metadata of a football match.	Meta	-
11	MatchSheet	meta/vendor	Match sheet data vendor name.	Vendor	-
12	Event	match	(see #1).	Match	-
13	Event	event/{type="referee"}	Referee event (e.g., final whistle, foul, caution).	Whistle	Event
14	Event	event/{type="shot"}	Shot event.	Shot	Event
15	Event	event/{type="pass"}	Pass event.	Pass	Event
16	Event	event/{type="misc"}	Misc event (e.g., tackle, chance without shot).	Misc	Event
17	MatchMetadata	competition	The football competition.	Competition	-
18	MatchMetadata	season	The football season (of the competition, if any).	Season	-
19	MatchMetadata	match	(see #1).	Match	-
20	MatchMetadata	match/periods/{i}	Time period within a football match (e.g., first half, shootout).	MatchPeriod	-
21	MatchMetadata	match/whistles/{i}	(see #13).	Whistle	Event
22	MatchMetadata	teams/{home|away}	(see #4).	Team	-
23	MatchMetadata	teams/{home|away}/players/{i}	(see #5).	Player	-
24	MatchMetadata	stadium	The football stadium where the match is held.	Stadium	-
25	MatchMetadata	meta	(see #10).	Meta	-

Since the original Football-CDF is based on a JSON schema, CDF lacks an explicit class hierarchy or relationship definitions. To support semantic reasoning—particularly for the *Event* data—we enriched the ontology by introducing class hierarchies and formal relationships. Specifically, the Match Sheet and Event components implicitly define several subclasses of the *Event* class. For instance, the Match Sheet defines events such as *Goal*, *Substitution*, and *Card*, while the Event component defines additional types like *Shot*, *Pass*, *Whistle*, and *Miscellaneous*. These event types share several common data properties (e.g., `time`, `period`) but also include properties unique to specific subclasses (e.g., `receiverId` is particular to a *Pass* event). Our ontology formally structures these classes and properties using subsumption hierarchies and specifying domain and range constraints. Table 1 detailed the mapping between concepts described in the CDF data format and the FCDF ontology classes.

[3] http://w3id.org/football-cdf/core

3. Feasibility Evaluation

To demonstrate the capabilities of the ontology, we have developed two small Python-based scripts to populate and generate a knowledge graph according to the FCDF ontology. The scripts are available through our GitHub repository[4] and briefly explained in the following:

- **CDF JSON generator**. As the Football-CDF specification is relatively new, there is currently a limited supply of example data in the format. To address this, we created a tool that transforms open data from the StatsBomb project[5] into the Football-CDF JSON format. The StatsBomb dataset contains a rich collection of football match data, ranging from the FIFA World Cup 1962 to the recently concluded UEFA Women's Euro 2025. The tool takes the StatsBomb lineup, event, and matches files either for a single game or in batch for an entire competition, and for each match outputs three Football CDF JSON formats [4]. Several StatsBomb fields already conform to the CDF format and can be retrieved as is, while others require a key renaming or minor value adjustment. Still, certain information must be derived from the StatsBomb event data (e.g., assist). This generator creates the data representation conforming to the CDF if implemented by the vendors. It therefore provides the basis for the proof-of-concept of the proposed ontology.
- **CDF-JSON to FCDF KG converter**. This tool converts Football-CDF JSON data into RDF format via a JSON-LD representation. It merges the three files produced by the Football-CDF JSON generator: match-sheet, event, and match-meta, and generates a single JSON-LD representation file per match that follows the Football-CDF ontology. We implement the transformation using RDFLib[6]. Like the Football-CDF JSON generator tool, the tool supports both single-match and batch mode. We plan to enhance this tool with JSON-LD context-based transformation to support more flexible and semantically rich data integration in the future.

The UEFA Women's Euro 2025 Knowledge Graph. We executed our scripts on the StatsBomb dataset from the UEFA Women's Euro 2025, which records all 31 matches, including the final match between Spain and the eventual winner, England. The resulting Knowledge Graphs contain more than 1.2 million triples, including information on various match events, such as fouls, goals, cards, and substitutions. We hosted the Knowledge Graph in a triplestore[7].

Through the populated knowledge graph, we can demonstrate the capability of FCDF to answer questions with various complexities. We outline an example SPARQL query on the average goals per team in the first half compared to the full game (without penalty shootout) in Figure 2, and provide both a SPARQL query playground and several predefined SPARQL queries in a simplified user interface[8].

4. Conclusion and Future Work

In this work, we have extended the CDF initiative by introducing the Football Common Data Format (FCDF) Ontology—an RDF-based formalisation that complements the original JSON-based CDF schema. The FCDF Ontology lays a solid foundation for more transparent, explainable, and interoperable AI applications in football analytics. Most importantly, it also allows for speeding up queries and sub-analyses on the data compared to previous implementations. As the community evolves, collaboration between data providers, researchers, and practitioners will be key to refining and adopting such representations. This work makes football data more accessible and actionable for many AI-driven innovations. Furthermore, analogous to the development of the SPADL data format in the past, which gave rise to advanced player evaluation metrics such as Valuing Actions by Estimating Probabilities

[4] https://github.com/wu-semsys/statsbomb-to-football-cdf/
[5] https://github.com/statsbomb/open-data/
[6] https://github.com/RDFLib/rdflib
[7] https://github.com/wu-semsys/statsbomb-to-football-cdf/tree/main/example_output
[8] https://semsys-staging.ai.wu.ac.at/graphdb/

```
PREFIX fcdf: <https://w3id.org/football-cdf/core#>
select ?team_id (?fh_goals/?mc as ?avg_fh_goals)
    (?all_goals/?mc as ?avg_goals) where {
    {   select ?team_id (count(?goal) as ?fh_goals)
        where { ?goal a fcdf:Goal ;
                fcdf:team_id ?team_id ;
                fcdf:event_period "first_half" }
        group by ?team_id }
    {   select ?team_id (count(?goal) as ?all_goals)
        where { ?goal a fcdf:Goal .
                ?goal fcdf:team_id ?team_id .
                ?goal fcdf:event_period ?period .
                FILTER (?period != "shootout") }
        group by ?team_id }
    {   select (count(?match) as ?mc)
        where { ?match a fcdf:Match } }
} order by DESC(?avg_goals)
```

	team_id	avg_fh_goals	avg_goals
1	fcdf:team/863	"0.25806451612903225806451 6"^^xsd:decimal	"0.58064516129032258064516 1"^^xsd:decimal
2	fcdf:team/865	"0.19354838709677419354838 7"^^xsd:decimal	"0.51612903225806451612903 2"^^xsd:decimal
3	fcdf:team/861	"0.19354838709677419354838 7"^^xsd:decimal	"0.38709677419354838709677 4"^^xsd:decimal
4	fcdf:team/858	"0.19354838709677419354838 7"^^xsd:decimal	"0.32258064516129032258064 5"^^xsd:decimal
5	fcdf:team/852	"0.06451612903225806451612 9"^^xsd:decimal	"0.22580645161290322580645 2"^^xsd:decimal
6	fcdf:team/855	"0.09677419354838709677419 4"^^xsd:decimal	"0.19354838709677419354838 7"^^xsd:decimal
7	fcdf:team/857	"0.06451612903225806451612 9"^^xsd:decimal	"0.19354838709677419354838 7"^^xsd:decimal

Figure 2: An example SPARQL query (left) and query results (right) on average goals per match in the tournament (first half vs full time), which suggest the tendencies of some teams to attack stronger in earlier in the game.

(VAEP) and expected threat (xT) [5, 6], we anticipate that the FCDF ontology will similarly stimulate further advancements in football analysis research.

Building upon this work, we identified several promising future exploration and development directions. First, we aim to continue alignment and comparative analysis with the original CDF (Common Data Format) to ensure compatibility while extending its expressiveness and utility. Such alignments will include iterative refinements guided by feedback from real-world deployments. Second, we are planning to integrate retrieval-augmented generation (RAG) techniques for query answering over the FCDF knowledge graph (KG) [16, 17]. Adapting RAG models to query and generate insights would facilitate downstream tasks such as question answering and summarisation. Additionally, neurosymbolic approaches merit deeper investigation, particularly in KG embeddings and graph neural networks (GNNs), e.g., through knowledge graph injection techniques [18]. Furthermore, having a semantic representation of the football data can potentially help address several issues with developing Machine Learning Models, e.g., data bias and contextual errors [19]. To this end, exploring the possible linking with external knowledge graphs, e.g., to explore the use and linking of existing large-scale knowledge bases like Wikidata [20], as well as construct a dedicated, domain-specific KG tailored to the nuances and granularity of football data.

Finally, we will investigate enabling declarative, bi-directional transformations between JSON CDF and RDF representations. These transformation mechanisms include the development of SHACL-based validation for validating input data against given semantic constraints. Such developments would promote interoperability and simplify integration with diverse systems utilising different data serialisation formats.

Acknowledgement

This work was supported by the Austrian Science Fund (FWF) Bilateral AI projects (Grant Nr. 10.55776/COE12) and the Austrian Research Promotion Agency (FFG) FAIR-AI project (Grant Nr. FO999904624).

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Olthof, J. Davis, Perspectives on data analytics for gaining a competitive advantage in football: computational approaches to tactics, Science and Medicine in Football (2025) 1–13. doi:10.1080/24733938.2025.2533784.

[2] I. Graham, How to Win the Premier League: The Inside Story of Football's Data Revolution, Century, 2024.

[3] F. Goes, L. Meerhoff, M. Bueno, D. Rodrigues, F. Moura, M. Brink, M. Elferink-Gemser, A. Knobbe, S. Cunha, R. Torres, et al., Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review, European journal of sport science 21 (2021) 481–496. doi:10.1080/17461391.2020.1747552.

[4] G. Anzer, K. Arnsmeyer, P. Bauer, J. Bekkers, U. Brefeld, J. Davis, N. Evans, M. Kempe, S. J. Robertson, J. W. Smith, J. V. Haaren, Common Data Format (CDF): A Standardized Format for Match-Data in Football (Soccer), Technical Report arXiv:2505.15820, 2025. URL: http://arxiv.org/abs/2505.15820. doi:10.48550/arXiv.2505.15820. arXiv:2505.15820 [cs].

[5] M. Van Roy, P. Robberechts, T. Decroos, J. Davis, Valuing on-the-ball actions in soccer: A critical comparison of xt and vaep, in: Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports, AITS, AI in Team Sports Organising Committee, 2020.

[6] T. Decroos, L. Bransen, J. Van Haaren, J. Davis, Actions speak louder than goals: Valuing player actions in soccer, in: Proceedings of the 25th ACM SIGKDD, ACM, 2019, pp. 1851–1861. URL: https://dl.acm.org/doi/10.1145/3292500.3330758. doi:10.1145/3292500.3330758.

[7] M. Manafifard, H. Ebadi, H. Abrishami Moghaddam, A survey on player tracking in soccer videos, Computer Vision and Image Understanding 159 (2017) 19–46. URL: https://www.sciencedirect.com/science/article/pii/S1077314217300309. doi:10.1016/j.cviu.2017.02.002.

[8] L. Lolli, P. Bauer, C. Irving, D. Bonanno, O. Höner, W. Gregson, V. Di Salvo, Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world, Science and Medicine in Football 9 (2025) 189–198. doi:10.1080/24733938.2024.2341837.

[9] N. Tsolakis, N. Vryzas, C. Dimoulas, C. Maga-Nteve, G. Meditskos, S. Vrochidis, An ontology-based framework for sports media data interpretation, in: 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), IEEE, 2023, pp. 1–8.

[10] N. Bouayad-Agha, G. Casamayor, L. Wanner, F. Díez, S. López Hernández, Footbowl: Using a generic ontology of football competition for planning match summaries, in: Extended Semantic Web Conference, Springer, 2011, pp. 230–244. doi:10.1007/978-3-642-21034-1_16.

[11] A.-Z. Adel, S. Zebari, K. Jacksi, Football ontology construction using oriented programming, Journal of Applied Science and Technology Trends 1 (2020) 24–30. doi:10.38094/jastt1113.

[12] A. Bacaj, Footology: A Football Ontology through Linked Open Terms Approach, Technical Report, Vienna University of Economics and Business, 2024.

[13] M. Katsumi, M. Fox, Ontologies for transportation research: A survey, Transportation Research Part C: Emerging Technologies 89 (2018) 53–82. doi:10.1016/j.trc.2018.01.023.

[14] J. Toledo, D. Doña, E. Ruckhaus, O. Corcho, M. Aguado, D. Patru, G. Atemezing, P. Vasilopoulou, Using Semantic Technologies in the Railway Domain: The Register of Infrastructure (RINF) System, in: International Semantic Web Conference (to appear), Springer, 2025.

[15] S. Lohmann, S. Negru, F. Haag, T. Ertl, Visualizing ontologies with VOWL, Semantic Web 7 (2016) 399–419. doi:10.3233/SW-15020.

[16] J. Ongris, F. Tjitrahardja, E.and Darari, F. Ekaputra, FrOG: Framework of Open GraphRAG, in: 4th International Workshop on LLM-Integrated Knowledge Graph Generation from Text (Text2KG) Co-located with the Extended Semantic Web Conference (ESWC 2025 - to appear), 2025.

[17] A. Gashkov, A. Perevalov, M. Eltsova, A. Both, Sparql query generation with llms: Measuring the impact of training data memorization and knowledge injection, in: Proceedings of 25th International Conference on Web Engineering (to appear), 2025.

[18] M. Llugiqi, F. J. Ekaputra, M. Sabou, Semantic-based data augmentation for machine learning

prediction enhancement, Neurosymbolic Artificial Intelligence 1 (2025) 29498732251340160. doi:10.1177/29498732251340160.

[19] J. Davis, L. Bransen, L. Devos, A. Jaspers, W. Meert, P. Robberechts, J. Van Haaren, M. Van Roy, Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned 113 (2024) 6977–7010. doi:10.1007/s10994-024-06585-0.

[20] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85. doi:10.1145/26294.