FAIR DATA ENGINEERING – 2025-2026

VNL 2023 MEN'S VOLLEYBALL PLAYER DATA AND STATISTICS EXPANSION

Enriching and FAIR-ifying a dataset through the merging with a more extensive one.

AUTHOR(S)

ARDA AKYAZI & BERRY DOMINGUEZ ADILOVA

DATE

1ST OF NOVEMBER 2025

UNIVERSITY OF TWENTE.

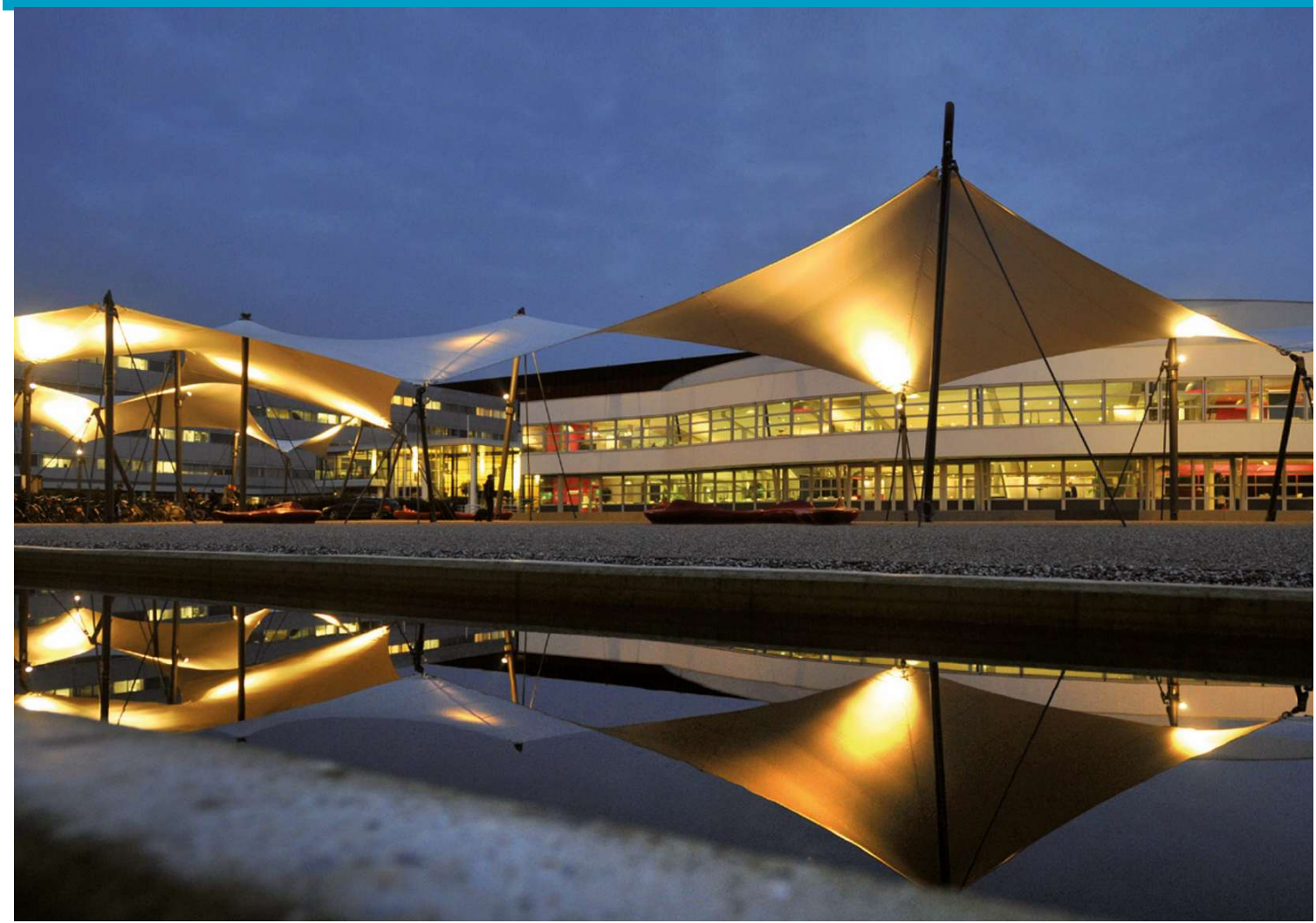


TABLE OF CONTENT

1. Introduction 3
2. FAIRification planning 4
   2.1 FAIRification goals 4
   2.2 Target resource(s) 4
   2.3 Target stakeholders 4
   2.4 Reuse stakeholders 4
   2.5 FAIRification requirements 5
   2.6 Semantic types 5
   2.7 Competency questions 5
3. FAIRification process 5
   3.1 initial fairness assessment 5
   3.2 Semantic data model 6
   3.3 Data triplification 7
   3.4 Semantic metadata model 8
   3.5 Metadata and data publication 8
   3.6 Final FAIRness assessment 8
4. discussions and Conclusions 9
5. annexes ..... **Error! Bookmark not defined.**

1. INTRODUCTION

The main goal of our FAIR-ification project is to create a database centered around volleyball player that adheres to the FAIR principles through the use of a main dataset on player that partook in the 2023 edition of the VNL tournament, and expand it with a vastly larger dataset that contains heaps of volleyball related information. The latter dataset has records not only players, but clubs, transfers, stadiums and much more. The objective would be to extract all the information directly relevant to players from the larger dataset, and use it to enrich the VNL player dataset, whilst FAIR-ifying it in the process.

VNL Men 2023 [Source – kaggle]

The dataset consists of columns that provide name, country, age, performance statistics and position (what role they play in the team). It is important to note that this contains information for players only for 2023.

Column Type	Number of columns	Column name
Decimal	6	Performance statistics columns
String	3	Name, Country, Position
Integer	1	Age

The license for this dataset is claimed as "Other". There is no further mention or clarification in the description. There is a total of 131 unique records in this dataset.

World Men Volleyball Data: Players,Teams, 13 more [Source – kaggle]

The license is Attribution-NonCommercial 4.0 International, meaning that remixing and sharing is allowed whilst crediting and non-profiting from the dataset.

This dataset contains a vast amount of information, mostly centring around players (awards, transfers, player details) and teams (presidents, managers, stadiums, matches, etc). A lot of the information is irrelevant to the ends intended in this project, thus only the following tables will be used:

Table	Relevant columns	Entries
Players	Name, birth date, ranking, height, weight, and position	46.525
Awards	Award name, player name, award date, and league	4.267
Teams	Team name, country, and town	8.582
Transfers	Player name, date, old team, and new team	2.260

2. FAIRIFICATION PLANNING

2.1 FAIRIFICATION GOALS

Findable	Accessible
F2. Describe the data with rich metadata	A2. Metadata has extended longevity
F3. Include the ID of the data in metadata	
Interoperable	**Reusable**
I1. Broadly applicable knowledge representation	R1. Richly described with accurate and relevant attributes
I2. FAIR principle abiding vocabulary use	R1.1 Clear data usage license
I3. Include references to other data	R1.2 Provide clear data provenance

2.2 TARGET RESOURCE(S)

Resource	Source
"VNL Men 2023"	Kaggle
"World Men Volleyball Data…"	Kaggle

2.3 TARGET STAKEHOLDERS

Stakeholder	Relevance
Volleyball personnel (coaches, players, team analysts)	The party that stands to gain the most from such a dataset are the relevant parties themselves, the players on whom the dataset is built, and their immediate surrounding in terms of the sport. To analysts, the data could make evident strengths and weakness. To coaches, it could it provide valuable insight into which players perform statistically better in which situations. To players, it could allow them to keep track of personal feats.

2.4 REUSE STAKEHOLDERS

Reuse Stakeholder	Relevance
Journalists, media, content creators	Journalists and reporters have to document and report on the happenings of their corresponding sport. Usually, they are insightful enough to draw conclusion from the game itself, however, having access to a dataset that provides them with aggregate statistics as well as facts about players, could allow them to do their job more effectively and, more importantly, more factually.
Data Miners	Data mining can have an almost limitless extent, as different contexts can be applied to different datasets. For instance, data miners could use this dataset as a foundation to find trends relative to data acquired in the future.

2.5 FAIRIFICATION REQUIREMENTS

FAIR Principle	Requirements
F2	Both datasets have limited metadata. Expand on it by providing rich and relevant metadata on the combined FAIRified data object.
F3	Mention the ID of the new data object within the metada.
A2	Ensure metadata is stored in a different service that offers higher longevity that the data object itself.
I1	Move from plain language to representing knowledge through links to established ontologies, which also abide by the FAIR principles. Expand the linked data as much as possible within needed relevance.
I2	
I3	
R1	Define and provide clear attributes to the (meta)data
R1.1	Include data usage license
R1.2	Provide as much detail as possible on the origin of the data and the steps it went through.

2.6 SEMANTIC TYPES

Semantic Type	Relevance
vball:Player	Represents a single volleyball player. The focal point of the dataset.
vball:Award	Any award won by a specific player.

2.7 COMPETENCY QUESTIONS

Player Performance & History

- Which players who participated in the 2023 VNL tournament have won a specific type of Award (e.g., MVP, Best Blocker) in their careers?
- What are the age range and position distribution of players who played in the 2023 VNL?

Data Enrichment & Interoperability

- Can the dataset provide the height and weight for players identified in the VNL 2023 dataset? (Leveraging the 'Players' table in the secondary dataset)

3. FAIRIFICATION PROCESS

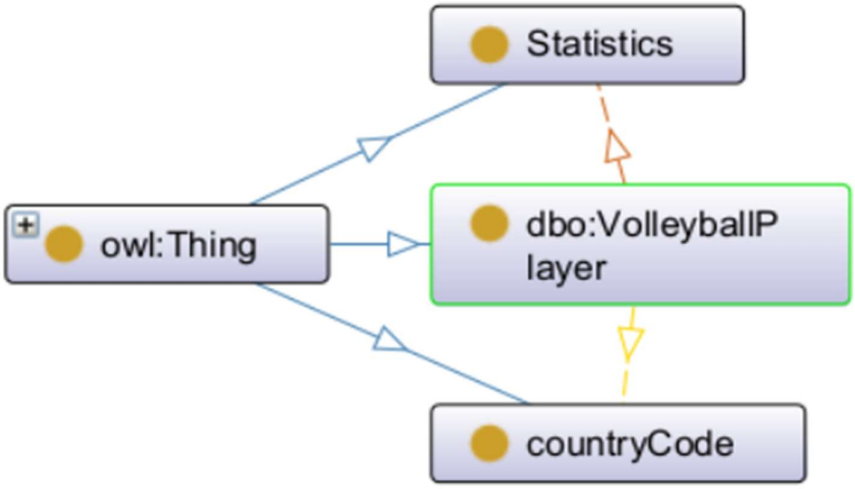
3.1 INITIAL FAIRNESS ASSESSMENT

Principle #	Criteria	Conformance	Note
F1	Globally unique identifier	Yes	Kaggle
F1	Persistent identifier	Yes	Kaggle
F2	Has metadata	Partially	Present but baren
F2	Has rich metadata	No	Barely any data
F3	Has data identifier in metadata	No	
F3	Has data identifier with clear predicate in metadata	No	
F4	Indexed/registered in search resource	Yes	
A1	Has resolution protocol	Yes	Kaggle

A1.1	Protocol is open, free and universally implementable	Yes	
A1.2	Protocol has authentication and authorisation mechanisms	Yes	Kaggle
A2	Metadata has longer persistency than data	No	No reference to metadata being available elsewhere
I1	(meta)data uses broadly applicable language	Yes	There's barely any language
I2	Vocabulary that follows FAIR principles	No	
I3	Qualified references to other data	No	None at all
R1	Described with plurality of relevant attributes	No	
R1.1	Clear usage license	Partial	One dataset yes, other no
R1.2	Clear provenance	Partial	Little information available
R1.3	Domain relevant community standards	-	

3.2 SEMANTIC DATA MODEL

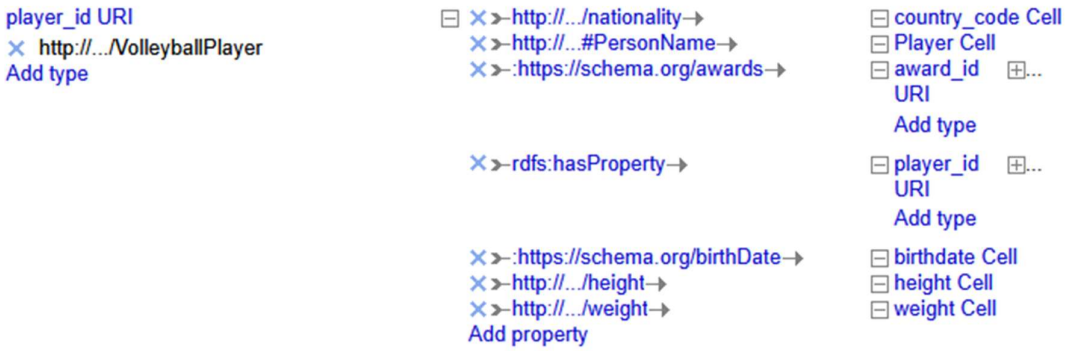
URI	Use
DBpedia Ontology (http://dbpedia.org/ontology/)	Core concepts like VolleyballPlayer and Country.
Schema.org (https://schema.org/)	General properties like Person.
Sport Schema (https://sportschema.org/ontologies/)	Sport specific concepts like VolleyballMatch and various statistics.
DCAT (http://www.w3.org/ns/dcat#)	Describe the dataset catalog.
Dublin Core (http://purl.org/dc/terms/)	Metadata properties like title, description, and creator.
FOAF (http://xmlns.com/foaf/0.1/)	Describe people (the dataset creators).
SHACL (http://www.w3.org/ns/shacl#)	For defining data constraints.



3.3 DATA TRIPLIFICATION

The goal of data triplification is to transform the concatenated tabular data into Resource Description Framework (RDF) triples, making the data machine-readable and explicitly linkable to the external vocabularies defined in the semantic data model. This process is crucial for achieving the Interoperability (I) principles of FAIR, particularly I1 and I3.

Due to the project's requirement to enrich the VNL player dataset by merging it with a vastly larger one, the resulting data structure contains multiple rows for the same player (e.g., one for each award or transfer). Consequently, we could not set the Subject URI to the row index of the merged dataset, as this would incorrectly identify a single player entity as multiple dist inct resources. The core challenge, therefore, was ensuring that the unique identity of the volleyball player served as the constant Subject for all related triples.



Screenshot of our RDF skeleton from OpenRefine

<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#player_p705> a <http://dbpedia.org/ontology/VolleyballPlayer>;
<http://purl.org/healthcarevocab/v1#PersonName> "Ichikawa Yuki" .
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#statistics_p705>
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#attack_percentage> "15.8"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#block_percentage> "1.13"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#serve_percentage> "1.4"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#set_percentage> "0.07"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#dig_percentage> "4.8"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#receive_percentage> "5.6"^^<http://www.w3.org/2001/XMLSchema#double>;
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#Player_position> "OH" .
<https://w3id.org/FAIR-course-UT/2025-2026/group2/data#player_p705> rdfs:hasProperty <https://w3id.org/FAIR-course-UT/2025-2026/group2/data#statistics_p705>.

Example of RDF triples

3.4 SEMANTIC METADATA MODEL

Component	Vocabularies	Use
Metadata	DCAT, Dublin Core, FOAF dcat:Dataset, dct:title, dct:creator	Identity, scope and author. DCAT as main container. Dublin Core for descriptive properties. FOAF for attributions.
Data Validation	SHACL sh:NodeShape, sh:property, sh:targetClass	Ensures data quality and consistency. Defines rules.

3.5 METADATA AND DATA PUBLICATION

Data (RDF Triples):
https://raw.githubusercontent.com/s2515091/FAIR_data/refs/heads/main/turtle_data_of_whole_dataset.ttl

Metadata: https://raw.githubusercontent.com/s2515091/FAIR_data/refs/heads/main/metadata.ttl

To the repository where everything is stored: https://github.com/s2515091/FAIR_data

3.6 FINAL FAIRNESS ASSESSMENT

Principle #	Criteria	Conformance	Note
F1	Globally unique identifier	Yes	
F1	Persistent identifier	Yes	
F2	Has metadata	Yes	
F2	Has rich metadata	Yes	Enriched and relevant
F3	Has data identifier in metadata	Yes	
F3	Has data identifier with clear predicate in metadata	Yes	

F4	Indexed/registered in search resource	Yes	
A1	Has resolution protocol	Yes	
A1.1	Protocol is open, free and universally implementable	Yes	
A1.2	Protocol has authentication and authorisation mechanisms	Yes	
A2	Metadata has longer persistency than data	Yes	
I1	(meta)data uses broadly applicable language	Yes	
I2	Vocabulary that follows FAIR principles	Yes	
I3	Qualified references to other data	Yes	
R1	Described with plurality of relevant attributes	Yes	
R1.1	Clear usage license	Yes	
R1.2	Clear provenance	Yes	
R1.3	Domain relevant community standards	-	

4. DISCUSSIONS AND CONCLUSIONS

The FAIRification project successfully improved the VNL player data, but the most difficult part was the initial data cleanup and merging.

The toughest step was the Data Preparation (Section 2.2) because the two datasets had no easy way to connect, which is a major issue for making data Findable (F1). We had to use complex matching methods, like checking player names with NLP, just to guess which records belonged together. This proved difficult because:

1. Missing Players: Some top players were surprisingly absent from the main "world" dataset.
2. Confusing Names: Different spellings, naming conventions or languages for player names made it very hard for the computer to match them automatically.

After we finally linked the data (mostly using awards and transfers), we created a merged table where one player might appear on many rows. This structure caused problems when we got to Data Triplification (Section 2.4). We couldn't use the row number as the player's unique ID because that would incorrectly treat one player as many different people. Therefore, we decided to use the Player IDs as the basis of our URIs, which lead to a weird (and slightly complex) RDF skeleton in OpenRefine.

In the end, the project taught us how to systematically turn messy, siloed information into clean, useful data that can be understood by computers and connected to the rest of the web. This reinforced how important interoperability is for future data projects in the industry.

UNIVERSITY OF TWENTE
Drienerlolaan 5
7522 NB Enschede

P.O.Box 217
7500 AE Enschede

P +31 (0)53 489 9111

info@utwente.nl
www.utwente.nl