Web Data Integration

Project Report

presented by
Fabian Rösch (1981845)
Jan Herbers (1664742)
Kai Reffert (1980476)
Nico Sharei (1986818)
Nursultan Mamatov (1983726)

submitted to the
Data and Web Science Group
Prof. Dr. Christian Bizer
University of Mannheim

December 2023

Contents

List of Tables

1 Phase I: Data Collection and Data Translation

1.1 Use Case

We decided on fusing data sets about football players from a wide variety of professional leagues to obtain complete information about biographical and football-related features. Our merged data set enables the user to compare player attributes among 18 different professional leagues. Analysts can use this data to gain insights, make predictions, and draw meaningful conclusions about the world of football both in the virtual realm and in real-world football.

Note: For the sake of faster run-time and due to the methodological nature of this project, we excluded most attributes that were present in only one data set (>100 variables). Including these would significantly expand the information provided with attributes ranging from a wide variety of performance capabilities (e.g. passing, dribbling and defending) to overall occurrences and data on players' national team careers and even recent injuries.

1.2 Our Data Sets

We chose three different data sets about male football players, which will be described in the following section.

The first data set "EA Sports FC 24 complete player dataset", sourced from the platform Kaggle (Leone, 2023), offers a rich and extensive collection of data pertaining to FIFA's Career Mode, spanning from FIFA 15 to EA Sports FC 24. The specific data set we utilize focuses on male players and contains data for the edition EA Sports FC 24, encompassing a total of 109 distinct attributes (referred to as *EA* in the following). In terms of player specific data, the data set provides a thorough examination of player positions, as well as an extensive array of attributes, such as Attacking, Skills, Defense, Mentality, and more, offering a holistic view of their performance capabilities. Additionally, it includes personal player details like nationality, club affiliation, date of birth, wage, and other pertinent information.

The second data set (referred to as *TM*) is obtained from transfermarkt.com (Transfermarkt, 2023) through web scraping and offers an extensive collection of player-related data. Besides biographical information, it includes a player's current market value and wage as of September 2023. Transfermarkt is generally regarded as one of the most trustworthy sources for this kind of data in the football world.

The last data set, the "Football Manager complete data" from Kaggle (Bai, 2023) provides an all-encompassing repository of information pertaining to every player available in FM 23 (referred to as *FM*). It includes comparable attributes to the other two data sets.

A summary of meta data on the three data sets can be found in Table 1.

Data set	Format	Class	Number of Entities	Number of Attributes	List of Attributes
EA	csv	players	10714	13	short_name, player_positions, value_eur, wage_eur, age, dob, height_cm, weight_kg, club_name,league_name, club_jersey_number, nationality_name, preferred_foot
FM	csv	players	17439	12	Name, DoB, Nat, Division, Club, Preferred Foot, Position, Height, Weight, Age, Transfer Value, Wage
TM	csv	players	9867	7	player_name, birthday, number, current_mv, team, nationality, positions

Table 1: Input Data Sets

1.3 Data Translation

The level of preprocessing that was required for schema matching differed among the three data sets.

Regarding the EA data, we first selected relevant variables and the 18 most relevant leagues for our project due to computational limitations. Lastly, we converted the *player_positions* attribute to a list attribute. From the FM data we also selected variables and the relevant leagues and first stripped the data from leading and trailing white space. In order to facilitate identity resolution, we used the 15 different positions from the EA data as baseline and manually converted those in the football manager data to this format. Further preprocessing regarding the removal of strings such as units (e.g., "kg") in numerical columns or age in parentheses in date columns was performed on the *weight*, *height*, *date of birth*, *wage*, *preferred foot* and *transfer value* attributes to enable better identity resolution and

data fusion. For the *transfer value* attribute, the data set initially contained ranges of (estimated) market values from which we took the average for better comparability with the rest of our data. We lose some precision here, since some ranges were quite large (i.e., the market value could in reality be much larger/smaller than the mean) while others were already rather precise (i.e., our mean is a good indication of the actual value). We also recoded "not for sale" statements as missing because the variable is supposed to contain market value which is independent of being available for a transfer. The preprocessing of our third data set (TM) contained the following operations. First, we had to additionally scrape the team names by their *team_id* from Transfermarkt to replace the *team_id* column with a *team_name* column. Second, if a player had multiple nationalities listed, we only kept the first one, as we were not interested in a list of values. Third, we combined the columns of *main_position* and *other_positions* to obtain a single list attribute. In addition, we also renamed the positions, so that they are consistent with the positions listed in the EA data set. Fourth, similar to previous procedures, we also removed strings from *market_value* (e.g. *500k*). Lastly, the *jersey_number* columns contained integer values and special characters such as *"-"*, therefore we replaced the special characters by -1, since -1 is a jersey number that was not present in our data set before and is an integer value.

Preprocessing was mostly done in python. We used MapForce to harmonize date formats, add data set-specific indices, and to convert the three preprocessed data sets to our target schema containing 13 attributes. The coverage of the data sets would have enabled us to include a larger number of attributes, but for computational purposes we reduced it to those variables that occur in multiple data sets.

Attribute Name	Attribute Type	Data set in Which Attribute Is Found
name	string	EA, FM, TM
birthdate	date	EA, FM, TM
age	int	EA, FM
nationality	string	EA, FM, TM
positions	list	EA, FM, TM
jersey_number	string	EA, TM
club	string	EA, FM, TM
league	string	EA, FM
weight	int	EA, FM
height	int	EA, FM
preferred_foot	string	EA, FM
current_market_value	int	EA, FM, TM
wage	int	EA, FM

Table 2: Integrated Schema

2 Phase II: Identity Resolution

2.1 Creation of the Gold Standard

Our data sets contain a rather large number of entities. Therefore – in the pairwise combination of data sets – we took samples of size 900/1500/2000 (depending on how many rows were necessary to obtain desired number of exact matches) from one data set and tried to find the best possible match among all entries in the respective other data set. Each data set was sampled once and functioned as comparison to another data sample once as well.

We experimented with different similarity measures to obtain matches for the Gold standard (GS). A combination of Levenshtein similarity (Levenshtein et al., 1966) on the players' full name and tokenized Jaccard similarity (Jaccard, 1901) on the birth date worked best as indicated by manual assessment as well as the share of matches among corner cases with high similarity scores. We opted for

tokenizing the birth date (day, month and year as individual tokens) to capture mistakes in the order of month and day that occur in the data. The similarity function's weights vary slightly by data set combination. We considered our matching rule good enough for the gold standard to not include other measures on any other columns to keep it as simple as possible.

These results of our similarity function are split up into perfect matches, corner cases and non-matches according to data set specific thresholds. Perfect matches are matches with a score equal to one. To figure out non-obvious corner-cases, we treated matches with a score higher than the threshold (FM-EA: .55, FM-TM: .5, EA-TM: .4,) and lower than 1 as corner cases. Out of these corner cases, we took the 200-500 rows with the lowest scores. This procedure leads us to have more difficult corner cases. Those correspondences were manually checked and set as true or false matches.

To obtain non-matches, entries which were neither matches nor corner cases (i.e., whose highest similarity score with the best match in the other data set was still low) were randomly matched with an entry from the other data set. Due to the small probability that the correct entity in the other data set was randomly chosen, it got manually verified.

In the end, perfect matches, corner cases and non-matches were combined in a ratio of 1:2:2 (sizes: FM-EA: 500 entities; FM-TM: 500; EA-TM: 475). For Machine Learning, the gold standards for linear combination are used as test sets. Extra training sets have been created by employing the same procedure (sizes: FM-EA: 140; FM-TM: 750; EA-TM: 290).

2.2 Matching Rules

Various blockers and matching rules were experimented with across the three combinations of data sets (TM-FM, EA-FM and EA-TM). A selection of results can be seen in the Tables 3, 4 and 5 below. The best linear combination or machine learning model for the respective table is shaded in gray.

Matching Rule	Blocker	Th	P	R	F1	#Corr	Time
Rule1:.5*Name +.5*DoB(5)	no Blocker	.7	-	-	-	-	out of memory
Rule1:.45*Name +.55*DoB(5)	SHBName(110)	.7	.83	.94	.88	8258	9s
Rule1:.45*Name +.55*DoB(5)	Standard-Name(5)	.7	.84	.91	.88	8042	5s
Rule1:.5*Name +.5*DoB(5)	SHBName(110)	.7	.84	.94	.89	8208	9s
ML:RandomTree	SNBName(110)	.7	.96	.94	.95	8373	94s
ML:RandomForest	SNBName(110)	.3	.88	.94	.91	8729	115s
ML:SimpleLogistic	SNBName(110)	.5	.84	.94	.89	8284	93s

Table 3: Outcomes for the combination of EA-TM

Matching Rule	Blocker	Th	P	R	F1	#Corr	Time
Rule1:.5*Name +.5*DoB(20)	no Blocker	.7	-	-	-	-	out of memory
Rule3:.5*Name+ .5*DoB(20)	Standard-Name(1)	.7	1	.42	.59	1396	2s
Rule1:.5*Name +.5*DoB(20)	Standard-Name(1)	.7	1	.92	.96	6542	4s
Rule1:.5*Name +.5*DoB(20)	SNBName(110)	.7	.99	.90	.94	6365	9s
Rule2	SNBName(110)	.7	.99	.90	.94	6356	11s
ML:SimpleLogistic	Standard-Name(1)	.9	.97	.91	.94	6565	31s
ML:RandomTree	Standard-Name(1)	.9	.99	.93	.96	7480	30s
ML:RandomForest	Standard-Name(1)	.7	.99	.92	.95	6490	36s

Table 4: Outcomes for the combination of EA-FM

Matching Rule	**Blocker**	**Th**	**P**	**R**	**F1**	**#Corr**	**Time**
Rule1:.5*Name +.5*DoB(20)	no Blocker	.7	-	-	-	-	out of memory
Rule3:.5*Name +.5*DoB(100)	Standard-Name(1)	.7	.99	.58	.73	5901	2s
Rule1:.6*Name +.4*DoB(20)	SNBName(110)	.7	.98	.73	.84	6223	10s
Rule1:.6*Name +.4*DoB(20)	Standard-Name(1)	.7	.99	.75	.85	6399	5s
ML:SimpleLogistic	SNBName(110)	.7	1	.81	.89	6445	178s
ML:RandomTree	Standard-Name(1)	.7	.99	.79	.88	6463	45s

Table 5: Outcomes for the combination of FM-TM

Rules:

- Rule1: Monge-Elkan similarity on name + absolute difference in days in birth date
- Rule2: Monge-Elkan similarity on name + absolute difference in days in birth date + relative difference in height + relative difference in weight (.5 * Name + .3 * DoB(20) + .1 * Height(.3) + .1 * Weight(.3))
- Rule3: Jaccard similarity on name + absolute difference of () days in birth date

We came to the conclusion that using a combination of name and birthdate similarity returned the best results both in terms of F1-score as well as group size distributions for linear combinations of matching rules. Using a relative distance measure of 30 % on height and weight was somewhat useful as an addition, while the data in all other columns was not accurate enough to employ similarity measures on it. Clubs and leagues for example differed quite a bit since the data was not collected in the same months and wage and transfer value are only estimations that are subject to substantial variations. Likewise, more than 100 different nationalities were spelled differently across all three data sets.

We eventually compared different similarity measures on the name and decided on using Monge-Elkan similarity with Levenshtein inside because it yielded the best results among linear combinations (rule 1/2). While Jaccard similarity was

equally precise, recall was significantly lower. We suppose that this is due to the order of some first and last names being reversed, especially for players of Chinese heritage in the *Transfermarkt* data set (e.g., Chen Wei - Wei Chen). Furthermore, a tokenized measure is needed here to capture the abbreviated first names in the EA data. Since a fraction of birth dates in the data was incorrect by a few days, we opted for a similarity in terms of an absolute difference in days. Preprocessing to lower case did generally not improve our results as that is not an issue in the data. After obtaining suitable matching rules, we provided them to several machine learning methods e.g., *RandomTree*, *RandomForest* or *NaiveBayesMultinomial*. Out of those, we again only portray the best performing models in Tables 3, 4 and 5.

Besides considering F1-score and the number of correspondences to select the best rules we also looked at the concrete group size distributions for the above combinations of data sets and opted for those rules that produce the least larges groups while optimizing the other criteria. This is why the highest number of correspondences or the best F1-score might not necessarily indicate the best rule. We suppose this might be due to the gold standard leaving room for optimization. But we did not achieve an improvement here, despite already having adapted the GS multiple times (e.g., we added more positive corner cases for FM-TM). Our final best rules – in combination with the best blocking method – are shaded gray in the tables.

Our best rules still allow for improvement. Notably, in the combinations of FM-TM and EA-FM, most selected rules and models achieved almost perfect precision but had a lower recall, which indicates a higher rate of false negative matches. As mentioned, there are certain players with wrong birthdates or switched day and month digits. We tried lowering the weight of birthday comparators, which resulted in worse overall results. We also experimented with increasing the window of the absolute difference in days. The impact of the parameter was not significantly noticeable until 100. An increase beyond this increased recall, but unfortunately lowered precision and resulted in significantly larger group sizes.

2.3 Blocking Methods

First, using no blocker at all was not feasible on our hardware and resulted in memory errors for all three combinations of data sets. We experimented with the Standard Record Blocker and the Sorted Neighbourhood Blocker (SNB) using the first letter of the name attribute. Using birth year, club, or league as blocking keys results in significantly lower F1-scores and is therefore not documented here. It became clear that name is the most consistent attribute among all, especially if reduced to the first letter(s). When it comes to reduction ratios, there was no

significant difference between SNB and standard blockers (.98 - .99). In the combination of EA-FM, the latter produced slightly better results and in the other two combinations, SNB performed best. SNB exhibited consistently lower run times.

The lower recall discussed above could also be due to the blocking strategy. When using name as blocking key, Chinese players whose order of first and last names in the data is falsely reversed, do not appear in a window together. Nevertheless, other potential blocking keys are even more unsuited as mentioned above and therefore, we decided on accepting this potential flaw.

3 Phase III: Data Fusion

3.1 Preparation

In this phase, the attributes of the players will be fused using the correspondences found in the previous chapter. Therefore, all three data sets are used to create the new fused one. The group size distribution can be seen in Table 6. An overview of matching results including scores of density, consistency, and accuracy as well as conflict resolution and evaluation methods can be found in Table 7.

Group Size	**Frequency**
2	3348
3	5976
4	15
5	10
6	7

Table 6: Overall group size distribution

We added provenance data which includes scores for the trustworthiness and publication dates. The TM data turns out to have very accurate values and thus got the highest score of 2.0. The lowest trustworthiness was assigned to the football manager data set, as certain attributes were not in right format (e.g., nationality) and the data is not quite up to date. The EA data got a score of 1.5. It is generally very accurate but first names and clubs are sometimes abbreviated. Publication dates are '2023-09-22' for EA data, '2022-11-07' for FM and '2023-09-01' for TM.

The gold standard comprises 15 randomly chosen entities that appeared in at

Attribute	Density			Consistency	Accuracy	Conflict Resolution	Evaluation Method
	EA	FM	TM				
name	1.0	1.0	1.0	0.53	1.0	Longest String	Levenshtein
					1.0	Favour Source	
birthdate	1.0	1.0	1.0	0.99	1.0	Voting	Equal
					1.0	Favour Source	
nationality	1.0	1.0	1.0	0.64	0.87	Favour Source	Levenshtein
					0.80	Longest String	
positions	1.0	1.0	1.0	0.45	1.0	Intersection 2-Sources	Equal
					0.67	Intersection	
jersey_number	1.0	0.0	1.0	0.97	1.0	Favour Source	Equal
					1.0	Most Recent	
club	1.0	1.0	0.97	0.39	0.40	Longest String	Levenshtein
					1.0	Most Recent	
league	1.0	1.0	0.0	0.48	0.93	Most Recent	Levenshtein
					0.53	Longest String	
weight	1.0	1.0	0.0	1.0	0.93	Average	Percentage
					0.93	Median	
height	1.0	1.0	0.0	1.0	0.93	Average	Percentage
					0.93	Median	
preferred_foot	1.0	1.0	0.0	0.95	1.0	Most Recent	Levenshtein

Table 7: Attribute Densities, Consistencies, and Accuracies

least two of the data sets, most players are included in all three. The correct information were sourced from Wikipedia and websites of the respective leagues of the players (Serie A, 2023; Bundesliga, 2023) with an overview of biographical information about each player. We decided not to include the attributes market value and wage in our gold standard as they are mere estimations that vary substantially by source and especially for lesser known teams and players and there is simply not one *true* value. Nevertheless, they are fused by average and included in the final data set for informational purposes. At this stage, we also decided not to fuse the attribute age as it contains redundant information (DoB).

3.2 Conflict Resolution and Evaluation Methods

In the process of resolving conflicts, each attribute underwent testing using two conflict resolution methods that were deemed to be the most intuitive. These methods were applied individually to the gold standard, and their effectiveness was evaluated based on output accuracy scores. The results of this evaluation are presented in Table 7.

For textual data, the primary conflict resolution rule experimented with is the Longest String method, chosen for its ability to provide the most information. Therefore, this method is used on name as in the EA data some first names are abbreviated and middle names left out. An alternative here would be favouring the most trustworthy source according to our experiments. On the whole data set though, this might not yield equally promising results. An exception to using the longest string is made for the league and club attributes, where encoding the player's current club is crucial. The FM data set, dated 2022-11-07, represents the oldest set of data. Given that player club affiliations may have changed since then, it is crucial to prioritize information from the most recent data set. This is exemplified by significantly higher accuracy scores.

In the EA data set, the nationalities are encoded via country codes *(e.g., GER)*, so we use the most trustworthy source TM because it is in the right format. This leaves some room for improvement. The nationality value is wrong for one of the players included in the GS (TM: Guadeloupe instead of France). Using the shortest string would always select the country code (e.g., FRA) while longest string results in something like 'China PR' for some cases and thus slightly lower accuracy. Had proper harmonization of formats been done prior to identity resolution, voting would have been more appropriate. Evaluation for the string attributes is conducted using the Levenshtein similarity measure, chosen for its focus on exact semantic similarity.

Numerical values are differentiated by continuous or relatively static scale. For example, attributes like height and weight, where a deviation of one is inconse-

quential, are averaged (median yields same results as there are only two data sets with this attribute). These attributes are deemed accurate with a deviation of under two percent. In contrast, for attributes like jersey number, even a minor change can have significant implications. In such cases, the entry from the most recent data set is utilized, considering potential club changes that might affect this value (e.g., if the jersey number is already assigned to another player). Only exact matches are accepted here.

The positions attribute, being the sole list attribute, utilizes the Intersection k-Sources method with a value of two, as opposed to a standard intersection. This choice is made to address the challenge of numerous non-matches when considering all three data sets for this attribute as exemplified by a perfect accuracy of one. Here, again only exact matches are accepted. The birth date attribute, the only date attribute, undergoes conflict resolution via Voting resulting in a perfect accuracy on the GS. The most trustworthy source seems to work too, as the inaccuracies in the data do not stem from the TM data set. The absolute equality measure is employed for this purpose as well.

As the preferred foot attribute exists only in two data sets, we opted for choosing the most recent one (same as most trustworthy in this case).

4 Summary

Our fused data set contains 9356 elements, which corresponds to an eventual decrease of elements compared to our largest data set (FM) with 17439 entries.

The density of our largest data set (FM) is 91%, whereas our final fused data set exhibits an improved density of 100%. All differences come from the *jersey_number*, *current_market_value* and *wage* columns in the FM data set, which exhibited an attribute density of 0%, 95% and 99% respectively.

Lastly, our resulting fused data set achieved an accuracy of 97% with respect to the gold standard, see Table 7 for attribute specific accuracies.

Bibliography

Jin Bai. Football Manager 2023 Dataset. https://www.kaggle.com/datasets/platinum22/foot-ball-manager-2023-dataset, 2023.

Bundesliga. Alle Spieler der Bundesliga. https://www.bundesliga.com/de/bundesliga/spieler, 2023. Accessed 26-11-2023.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Stefano Leone. Ea Sports FC 24 complete player dataset. https://www.kaggle.com/datasets/stefanoleone992/ea-sports-fc-24-complete-player-dataset?select=male_players.csv, 2023. Accessed 30-09-2023.

Vladimir I. Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Serie A. All the players. https://www.legaseriea.it/en/serie-a/calciatori, 2023. Accessed 26-11-2023.

Transfermarkt. https://www.transfermarkt.com/, 2023. Accessed 30-09-2023.