WDI Project Report

Mariam Arustashvili, Jusztina Judák, Petra Révész, Maria Schlüter, and Stefan Alexander Stingl

University of Mannheim
Web Data Integration Project
https://github.com/StayFN/WebDataIntegrationProj

1 Project Goal and Overview

Our data integration project aimes at harmonizing science-based targets (SBTs) information with revenue-related data from a diverse array of companies, to offer insights into the climate goals of companies, alongside validating and expanding core information about these companies. This project is bound to catch the interest of investors looking to understand what's driving the market. It's also a valuable resource for company executives who want to spot their main competitors in the industry. For this goal, we integrated three different datasets. In the following, we will give an overview of each of them, then describe the integrated schema and how the datasets were mapped to it. Then we go into the different identity resolution strategies and their results, followed by the data fusion and quality evaluation of results.

2 Data

For our project, we used three datasets - "Companies taking action", "Forbes Top 2000 Global Companies" and "DBpedia Companies". In the following, we will give some basic information about each dataset, describe the schema and profile of each of them, as well as the pre-processing that was applied, if any. An overview of the size and attributes of each dataset can be found in Table 1.

2.1 Dataset 1: "Companies taking action"

Our first dataset, in the following referred to as SBTI, was downloaded from the Science-Based Target Initiative website[1]. It contains companies and their target-related data. Science-based targets offer companies a clear roadmap for reducing greenhouse gas emissions. These targets are considered 'science-based' when they align with the latest climate science recommendations, specifically aiming to limit global warming to no more than 1.5°C above pre-industrial levels in accordance with the Paris Agreement goals. As it can be seen in Table 1, the dataset contains 6157 entities and 21 different attributes. Among these attributes, eight exhibit a significant occurrence of missing values, surpassing 50%. These attributes include *ISIN*, *LEI*, *Long Term - Target Status*, *Long Term - Target Classification*, *Long Term - Target Year*, *Net-Zero Year*, *BA1.5 Date*, and *Extension*. However, there are no missing values in the term of key attributes as *Company Name*, *Industry*, *Location*, which we later used for identity resolution. The main purpose of using this dataset was to integrate data about greenhouse gas emission reduction target set by a company. The SBT attributes are divided into Near Term (achieved by 2030) and Long Term (Achieved by 2050). The Status contains three classes, which are *Committed*, *Target Set* and *Removed*. The Classification values mean the temperature alignment, which can be 1.5°C, well-below 2°C, and 2°C.

2.2 Dataset 2: "Forbes Top 2000 Global Companies"

Our second dataset, in the following being referred to as Forbes, is available on Kaggle[2]. It encompasses information about the top 1999 Forbes companies, as per the 2022 ranking. In addition to the standard company identification data like *company name*, *country*, *founding year*, and *industry*, it also includes additional attributes such as *revenue*, *profits*, *assets*, and *market value*. Companies are classified into 29 distinct industries. In the case of this dataset, we do not have any missing values.

[1] https://sciencebasedtargets.org/companies-taking-action

[2] https://www.kaggle.com/datasets/rakkesharv/forbes-2000-global-companies

2.3 Dataset 3: DBpedia Companies

Lastly, we worked with our data collection scraped from DBpedia, which resulted in over 10,000 items. The dataset adds further information about key persons in the companies, net income and company type. It's essential to underscore that this dataset isn't exhaustive, as it contains numerous empty columns, adding a compelling aspect to our web data integration project. We only queried for companies with available revenue data. Due to limitations in the SPARQL endpoint, we couldn't retrieve all companies with a single query. Instead, multiple queries were run, utilizing the "OFFSET" and "LIMIT" keywords. The resulting .html files were combined and converted to .csv format. We preprocessed the data, including converting the RDF Links and Labels to string and numeric values. The *Label* column represents the company names. The companies are categorized into 2418 industry categories and 447 types of organizational structures. From the 10 attributes, the level of missing values is especially high in the case of *Founders* and *Country.* Each of them exceeds 9000 (over 80%) missing values. Additionally, the dataset also contains a *Key People* attribute. In the case of the other two attributes, we also have a lot of missing values, *Assets* (7323; 68%) and *Net Income* (6023; 56%).

Working with DBpedia data introduces some difficulties regarding the values that are provided, that are worth mentioning. Some values in DBpedia are just wrong (for example, in revenue/assets/netIncome, the units are different - millions, billions, or thousands. This is nearly impossible to automatically fix since we don't know the unit. For example, for PetroBank the revenue is 1.575331 (here the EX from scientific notation is missing so we don't know what power of 10 it is. This just gets converted to 1 since it is just a decimal number. It probably is 1.575331E8, but we can't just guess.

Table 1: Dataset Information. Attributes with > 30% missing values are marked with *.

Dataset	Source	Format	#Entities	#Attr.	Attribute names
Companies taking action (SBTI)	Science Based Targets	xlsx	6,157	21	Company Name, ISIN*, LEI*, Near term - Target Status, Near term - Target Classification, Near term - Target Year, Long term - Target Status*, Long term - Target Classification*, Long term - Target Year*, Net-Zero Committed, Net-Zero Year*, Organization Type, BA1.5, BA1.5 Date*, Location, Region, Sector, Date, Target, Target Classification, Extension*
Forbes Top 2000 Global Companies	Kaggle (originally Forbes)	csv	1,999	11	Ranking, Organization Name, Industry, Country, Year founded, CEO, Revenue, Profits, Assets, Market Value, Total Employees
DBpedia Companies	DPBedia Query dataset query	csv (converted 2 html files)	10,720	10	Label, Industries, KeyPeople, Founders*, Country*, Revenue, Assets*, NetIncome*, FoundingYear, Type

3 Data Translation

3.1 Integrated Schema

The integrated schema was developed based on the attributes present in all or at least two datasets, while also keeping the attributes that were just contained in one dataset. From the original attributes, we excluded *BA1.5*, *BA1.5 Date*, and *Extension* because they weren't well-documented. All attributes used can be seen in Table 2. Firstly we identified which attributes are overlapping in order to create the schema. The attributes that only appear in one dataset were directly mapped. However, for some records a different name was chosen for the integrated schema due to ambiguities in the naming. For example, in SBTI, *Organization*

Type refers to the company size (in categories), while for DBpedia its *Type* attribute refers to the company's legal type. Also, if some attribute loses its meaning without the dataset context, its name was converted in the integrated schema, e.g., the Forbes *Ranking* attribute was named *Forbes2022Rating*. Both SBTI and Forbes have a company size attribute, however, one is numeric (number of employees), and one is categorical. Since we don't know the criteria used for the categories, we decided to map them to two distinct size attributes. Then, the overlapping attributes were analysed. The overlapping attributes *company name*, *industries*, *founding year*, *country*, *revenue*, *assets* and *profit* all represent comparable values across datasets. For the different person attributes - *keyPeople*, *founder* and *ceo* - we wanted the integrated schema to contain all person names and their roles (if they are a founder or ceo) which could be derived from the *founder* and *ceo* attribute, so that there is no information loss, but at the same time a person doesn't occur several times (e.g. once in a *keyPeople* element and once in a *founder* element). Table 2 shows the 28 final attributes in the integrated schema. Furthermore, the integrated XML schema with some example values can be found on Github.

3.2 Transformations

The mapping of the source schemata to the integrated schemata were done using Altova Mapforce[3] for each dataset. For the SBTI dataset, there weren't any transformations necessary. All attributes were mapped directly to their corresponding schema attributes. For Forbes, the main transformation needed was turning values of Revenue, Assets, and Profit into unitary values instead of billions. Apart from that, we needed to set the CEO to True if it was known and put its value in PersonName.

For DBpedia, we did extensive preprocessing beforehand, specifically converting the different currencies of assets, revenue and profit to USD, similar to Forbes and extracted string and numeric values. After these preprocessing steps, we could directly map all attributes except the Industries, People, and Founders attributes to their schema correspondences. We tokenized industries (list type) using a comma as a delimiter and mapped each industry to a single schema element. For DBpedia, we didn't know the specific role of a person in the KeyPeople list attribute (e.g., if it is CEO, founder, or has another role); also the persons list seemed incomplete in many cases, so that it wasn't possible to infer the distinct roles combining e.g. the founders and Forbes ceo attribute. To deal with the attributes People and Founders, we decided to merge the two lists, removing the duplicate names. Then, the resulting list was tokenized and all distinct names were mapped to the PersonName attribute. Afterward, the Founders list was used to create a Boolean value. For this, we used the contains function in Mapforce.

4 Identity Resolution

4.1 Goldstandard

For our goldstandard we used 500 candidate pairs per dataset combination. Approximately 50% of them were labeled as non-matches, 20% as matches, and 30% as corner cases. As a starting point to collect the samples, we used Fuzzy string matching[4], which calculates the similarity of two strings by using Levenshtein distance, resulting in scores between 0 (least similar) and 100 (identical strings). This strategy was applied to compare the company names of the datasets pairwise. Then, for each dataset combination, we proceeded as follows:

1. To get example *matches*, we sampled the target number of examples from the pairs with a fuzzy score of 100, meaning they have identical company names and thus are assumed to refer to the same entity. The sampling was done randomly to avoid selection bias, e.g., through alphabetical order.
2. To get examples for *non-matches* and *"match although dissimilar" corner cases*, we first filtered for the pairs with a fuzzy score below 85. These samples were also randomly shuffled to avoid selection bias, making sure that, for example, not only samples with very low similarity scores were in the non-matching goldstandard. Then, from the resulting list, each pair was manually checked if it was matching. If it was not matching, it was added to the list of non-matches; if it was matching, it was added to the list of

[3] https://www.altova.com/mapforce

[4] https://pypi.org/project/fuzzywuzzy/

Table 2: Overlap of Integrated Schema: Attributes, Data Types, and original attributes names. Preprocessing before the mapping is indicated in brackets with P, transformations during mapping with T. Otherwise mapping was done directly. Attributes used for identity resolution are highlighted in gray.

Integrated schema attribute	**Type**	**SBTI attribute**	**Forbes attribute**	**DBpedia attribute**
CompanyName	Str	Company Name	Organizat. Name	Label (P: clean)
ISIN	Str	ISIN	-	-
LEI	Str	LEI	-	-
Forbes2022Rating	Int	-	Ranking	-
Industries	Str/list	Sector	Industry	Industries (P: extract strings from urls; T: tokenize)
FoundedYear	Date	-	Year founded	FoundingYear
Country	Str	Location	Country	Country
Region	Str	Region	-	-
PersonName	Str/list	-	CEO	KeyPeople + Founders (P: combine both attributes; T: tokenize)
Founder	Bool	-	-	Founders (T: Bool if contained in KeyPeople-Founders list)
CEO	Bool	-	CEO (T: True if person is CEO)	-
Revenue	Float	-	Revenue	Revenue (P: currencies & notation)
Assets	Float	-	Assets	Assets (P: currencies & notation)
Profit	Float	-	Profits	NetIncome (P: currencies & notation)
MarketValue	Float	-	Market Value	-
SizeEmployees	Int	-	Total Employees	-
SizeCategory	Str	Organization Type	-	-
LegalType	Str	-	-	Type
SustGoalDescription	Str	Target	-	-
SustGoalClassification	Str	Target Classification	-	-
SustGoalStatus_NearTerm	Str	Near term - Target Status	-	-
SustGoalStatus_LongTerm	Str	Long term - Target Status	-	-
SustGoalClassification_NearTerm	Str	Near term - Target Class.	-	-
SustGoalClassification_LongTerm	Str	Long term - Target Class.	-	-
SustGoalYear_NearTerm	Date	Near term- Target Year	-	-
SustGoalYear_LongTerm	Date	Long term - Target Year	-	-
NetZeroCommited	Bool	Net-Zero Committed	-	-
NetZeroCommitedYear	Date	Net-Zero Committed Year	-	-

corner cases (since it is matching, although the name is rather dissimilar). This was done until the target number of non-matches and an appropriate number of matching corner cases were found.

3. To get examples of *"non-matching although quite similar" corner* cases we filtered for the samples with a fuzzy score between 85 and 99, again shuffling the resulting samples. Then, they were manually checked and added to the list of corner cases, if they were non-matches (despite having a quite high similarity score). This was again done until the target number of corner cases was reached.

While, in general, the decision on whether two samples were matched was based on the name, there are multiple cases where it is ambiguous. We decided to consider a company's country subgroup as a separate company, e.g., Vodafone does not match Vodafone Turkey (see also examples given in Table 3. However, if a company changed its corporate form over time, we consider it a match, e.g., International Airlines and International Airlines Group. If searching for these companies on the web led to the same webpage, we assumed them to be identical, as it would have been too hard to do further research about the company's history of corporate forms. Also, due to time constraints, we didn't compare the other attributes in addition to the name for 1500 samples. The difficulties arising from this approach will be discussed in the final part of this report.

Finally, the matching, non-matching, and corner case samples were appended into one csv file per dataset combination, each row having the format "df1_id, df2_id, match". To make sure that no examples get a higher weight in the goldstandard than others, the resulting list was checked for duplicates. Some examples can be found in table 3. To use this goldstandard for the identity resolution, we applied a train-test split of 70/30 for the machine learning approach. For comparability, we also used the same test split for the evaluation of the simple linear combination approach.

Table 3: Identity resolution goldstandard examples

	Company1	**Company2**	**Match**
Matching	China Minsheng Bank	China Minsheng Bank	True
	Unicharm	Unicharm	True
	American Water Works	American Water Works	True
Non-matches	ERM (consultancy)	Tata Consultancy Services	False
	Gategroup	PulteGroup	False
	GMO Internet	ASM International	False
Corner cases	Vodafone Turkey	Vodafone	False
	TMX Group	TMF Group	False
	Evergrande Group	China Evergrande Group	True

4.2 Used Attributes

As can be seen in Table 2, three of the attributes (*company name*, *industries* the company is associated with, and the company's *country*) are present in all three datasets. Another five attributes (*founding year*, *person names*, *revenue*, *assets*, and *profit*) are contained in two datasets. However, since it was not possible to use the *persons* attribute for the identity resolution, only the other five attributes were used. In the following, the matching rules for each of them will be described.

4.3 Comparators

In a data integration project, the implementation of identity resolution matching rules play an important role in ensuring accurate and reliable results. For the reconciliation of company names, we experimented with various string matching algorithms, including Levenshtein, Jaro, and Jaro-Winkler. These algorithms were applied to assess the similarity between *company names* in all three datasets and facilitate the identification of corresponding entities in our dataset. In the case of company headquarters (*country*), we adopted a diverse approach by experimenting with Jaro, exact matching, Jaro-Winkler, and Levenshtein after a preprocessing step. This evaluation allowed us to determine the most effective comparator for country data, which was Jaro Wrinkler. For the *FoundedYear* attribute we opted for a straightforward exact match approach. This decision was driven by the nature of this attribute, where precision in matching is paramount for accurate data integration. The values of the *assets*, *profit* and *revenue* attributes in DBpedia and Forbes vary a lot over time and also show a larger range and thus a large variance in the differences of absolute values. Thus, only a relative comparison makes sense for these attributes, so that we decided to base the similarity on the percental difference between values, trying out different thresholds for the maximum percentage difference. For *Industries*, we utilized Levenshtein Edit Distance to measure the similarity between industry names. The comparator processes industry names by converting them to lowercase and trimming spaces, ensuring standardization. It then calculates the similarity between the two companies' industries using an Overlap Coefficient. This coefficient is calculated by finding the proportion of the intersection size (considering industries with a similarity score above a defined threshold) to the smaller of the two industry sets. This approach allows for a nuanced comparison that accounts for slight variations in industry naming while focusing on the core similarities.

4.4 Blocking

We employed standard blocking, specifically using the *CompanyName* attribute as a blocking key. This approach was crucial for managing the computational resources effectively. Without blocking, our attempts at running the identity resolution process resulted in an out-of-memory exception, even with 64GB of RAM. This limitation was particularly evident in handling the DBpedia combinations, which were too large and thus required even more memory resources. While we were able to process the SBTI-Forbes combination, the sheer volume of data in the DBpedia combinations prevented us from running these processes to completion. Consequently, due to these computational constraints, we were unable to draw meaningful conclusions regarding the impact of blocking on the final correspondences. This underscores the importance of efficient data management techniques, like blocking, in handling large-scale datasets for identity resolution tasks.

4.5 Linear Combination Results

Our first approach for creating the matching rules was a simple linear combination of the comparators, with an emphasis on optimizing F1 scores. For this, we utilized weighted comparator features, including Company Name, Profit, Revenue, Assets, Industry, and Founded Year. Each feature was assigned a varying weight, allowing for precise adjustments to the matching algorithm. In each scenario, different thresholds were strategically selected to achieve an optimal balance between precision and recall, thereby enhancing the F1 score. This deliberate and calculated approach was vital for effectively managing the precision-recall trade-off.

The final weights and thresholds can be seen in Table 4. During our experiments Company Name emerged as the predominant factor, having a high weight of 0.5 to 0.7, as compared to the other attributes which only got weights of about 0.1 to 0.2. A reason could be that names have the highest density, are most predictive, and also persistent over time, while especially the financial attributes assets, profit, and revenue are quite sensitive to changes over time. Furthermore, it is important to note that our process of determining the weights and combinations of these features was manual, with adjustments made in the direction that most significantly improved the results. However, this method inherently suggests the possibility that the results obtained might represent local optima. There remains the potential for discovering more optimal combinations, leading to even better results, as the current configurations may not necessarily be the global optima. Thus, in addition we tried machine learning based approaches, which will be described in the following section.

4.6 Machine Learning Results

In our analysis, we selected three distinct models from the Weka package: Logistic Regression, Bayes, and AdaBoostM1, to ensure a robust and comprehensive approach to identity resolution. These models were chosen for their diverse methodologies in pattern recognition and their ability to model complex relationships within our datasets. Logistic Regression was employed for its proficiency in handling binary classification

Table 4: Matching Analysis for the linear combination approach

Matching	**Comparator[Weight]**	**Threshold**	**Pre.**	**Rec.**	**F1**
DBpedia - Forbes	CompanyName[0.5], Profit[0.1], Revenue[0.1], Assets[0.1], Industry[0.1], FoundedYear[0.1]	0.4	0.96	0.80	0.87
DBpedia - SBTI	CompanyName[0.6], Industry[0.2], Country[0.2]	0.50	0.85	0.83	0.84
SBTI - Forbes	CompanyName[0.7], Industry[0.1], Country[0.2]	0.80	1.00	0.67	0.80

tasks and its interpretability, which is crucial for understanding the influence of different attributes on the matching decision. The Bayes model offered a probabilistic perspective, advantageous for scenarios with uncertainty and where the conditional independence assumption holds. Lastly, AdaBoostM1 was included for its capacity to boost the performance of weak classifiers, thus creating a strong composite model. Throughout the modeling process, our primary objective was to optimize for the F1 score, a balanced measure that considers both precision and recall. This metric was pivotal in scenarios where the cost of false positives and false negatives is equally important. While default settings were predominantly used, we experimented with AdaBoostM1's parameters, such as the number of iterations. Although increasing iterations occasionally improved precision, it adversely affected recall, leading to a lower overall F1 score. Therefore, we adhered to default options where they yielded the most favorable F1 outcomes.

Table 5: Matching analysis for the machine learning approach. The best models are highlighted gray.

Matching	**Model**	**Matching Rule**	**Thresh.**	**Pre.**	**Rec.**	**F1**
DBpedia - Forbes	Logistic Regression	$-12.37 + \text{Revenue} \times 1.12$ $+ \text{CompanyName} \times 12.47$ $+ \text{Industry} \times -0.24$ $+ \text{Country} \times -0.23$ $+ \text{Assets} \times 0.83$ $+ \text{Profit} \times 0.52$	0.10	0.96	0.80	0.87
DBpedia - Forbes	Bayes	See Github (Link)	0.001	1.00	0.8	0.89
DBpedia - Forbes	AdaBoostM1	See Github (Link)	0.10	0.98	0.76	0.86
DBpedia - SBTI	Logistic Regression	$-18.66 + \text{CompanyName} \times 2$	0.7	1.00	0.81	0.90
DBpedia - SBTI	Bayes	See Github (Link)	0.6	1.00	0.81	0.90
DBpedia - SBTI	AdaBoostM1	See Github (Link)	0.80	0.95	0.81	0.88
SBTI - Forbes	Logistic Regression	$-19.62 + \text{CompanyName} \times 1$ $+ \text{Country} \times 4.78$	0.5	1.00	0.67	0.80
SBTI - Forbes	Bayes	See Github (Link)	0.2	1.00	0.67	0.80
SBTI - Forbes	AdaBoostM1	See Github (Link)	0.7	0.97	0.67	0.79

4.7 Comparison

The machine learning models generally achieved comparable, if not slightly better, F1 scores than the linear combination approach. For instance, in the DBpedia - Forbes matching, the machine learning models, especially Bayes, reached an F1 score as high as 0.89, slightly surpassing the linear combination's peak score of 0.87. This trend is consistent across other dataset matchings, where machine learning models either matched or marginally improved upon the F1 scores of the linear combination method.

These results suggest that while the linear combination approach offers the advantage of fine-tuned control over feature weighting, the machine learning models excel in extracting complex, non-linear relationships within the data, often leading to a more optimized balance between precision and recall. It's important to note, however, that the success of machine learning models could partly be attributed to their inherent ability to adapt and learn from the data, potentially revealing patterns that may not be immediately apparent through manual feature weighting.

4.8 Examples

In the context of our project, we encountered some particularly nuanced case that underscores the complexities inherent in entity resolution. Our Identity Resolution models identified the London Stock Exchange and the London Stock Exchange Group as distinct entities, which seemed to be a false negative, when compared to our gold standard. However, a deeper examination of the organization's history and structural evolution revealed a more intricate picture.

The London Stock Exchange Group was, in fact, the entity formerly known as the London Stock Exchange until 2007. After acquiring an Italian stock exchange, the organization underwent significant restructuring, resulting in the formation of the London Stock Exchange Group. This new entity includes multiple exchanges, one of which is the original London Stock Exchange. Both the Group and the current Exchange trace their roots back to the historical London Stock Exchange established in 1801, the Group has been recognized as a separate legal entity since 2007.[5] [6] Despite this formal distinction, our entity resolution criteria dictate that renaming or changing corporate form does not necessarily create a separate entity. Therefore, for our analysis, we have concluded that the London Stock Exchange and the London Stock Exchange Group are the same entity, adhering to our predefined rules that favor historical continuity and corporate lineage over changes of name or corporate form. This instance highlights the challenges in defining a gold standard for entity resolution, as applying the same rule across different cases can lead to varying outcomes. Creating a specific exception rule for each unique scenario is not feasible.

DBpedia: London Stock Exchange Group

```
<Company>
    <ID>DBpedia_3577</ID>
    <CompanyName>London Stock Exchange
       Group</CompanyName>
    <Industries>
        <Industry>Financial services</
           Industry>
    </Industries>
    <FoundedYear>2007</FoundedYear>
    ...
    <Revenue>6502</Revenue>
    <Profit>3263</Profit>
    ...
</Company>
```

Forbes: London Stock Exchange

```
<Company>
    <ID>Forbes_500</ID>
    <CompanyName>London Stock Exchange</
       CompanyName>
    ...
    <Industries>
        <Industry>Diversified Financials<
           /Industry>
    </Industries>
    <FoundedYear>1802</FoundedYear>
    <Country>United Kingdom</Country>
    ...
    <Revenue>9270000000</Revenue>
    ...
</Company>
```

In a less contentious example of a false negative, we can examine J Sainsbury, the renowned retail chain. Officially registered as J Sainsbury, they operate under the trading name Sainsbury's. Unfortunately, a discrepancy in DBpedia's data contributes to a lower similarity score for the company name, as it fails to recognize this common trading name variation. Moreover, while both entities operate within the same industry, our current similarity assessment method relies solely on edit distance and overlap, making it interpret them as dissimilar despite their semantic similarity. Another contributing factor is the outdated revenue information in DBpedia, which falls outside our predefined similarity range. Consequently, our models

[5] http://news.bbc.co.uk/1/hi/business/6233196.stm

[6] https://www.londonstockexchange.com/discover/lseg/our-history

incorrectly classifies this record as a non-match, highlighting the need for more robust and context-aware similarity measures in our matching process.

DBpedia: Sainsbury’s

```
<Company>
    <ID>DBpedia_9365</ID>
    <CompanyName>Sainsbury’s</CompanyName
        >
    <Industries>
        <Industry>Retail</Industry>
    </Industries>
    <KeyPersons>
        <Person>
            <Name>John James Sainsbury</
                Name>
            <Founder>true</Founder>
        </Person>
        <Person>
            <Name>Martin Scicluna (
                businessman)</Name>
            <Founder>false</Founder>
        </Person>
        <Person>
            <Name>Simon Roberts (
                businessman)</Name>
            <Founder>false</Founder>
        </Person>
    </KeyPersons>
    <Revenue>29048000000</Revenue>
    <LegalType>Public limited company</
        LegalType>
</Company>
```

Forbes: J Sainsbury

```
<Company>
    <ID>Forbes_1136</ID>
    <CompanyName>J Sainsbury</CompanyName
        >
    <Forbes2022Rating>1136</
        Forbes2022Rating>
    <Industries>
        <Industry>Food Markets</Industry>
    </Industries>
    <FoundedYear>1922</FoundedYear>
    <Country>United Kingdom</Country>
    <KeyPersons>
        <Person>
            <Name>Simon John Roberts</
                Name>
            <Ceo>true</Ceo>
        </Person>
    </KeyPersons>
    <Revenue>40820000000</Revenue>
    <Assets>34620000000</Assets>
    <Profit>390000000</Profit>
    <MarketValue>7170000000</MarketValue>
    <SizeEmployees>160500</SizeEmployees>
</Company>
```

5 Data Fusion

For our data fusion task we used all our three datasets, namely SBTI, DBpedia and Forbes. Three attributes were contained in all datasets (Company name, Country and Industry) and 5 additional attributes came from two datasets, DBpedia and Forbes (Profit, Revenue, Asset, Key People, Year Founded). As provenance data we experimented with different quality scores.

The table 6 shows the density of our datasets before and after data fusion. As one can see, SBTI and Forbes contain no empty records, only DBpedia datasets has a lower density with an average 73%. After the fusion task we increased the density of Year Founded to 96% and to 79% for Assets. Since two of our dataset (SBTI and Forbes) had no missing values is Company name, Country and Industry, we ended up having a 100% dense result for these attributes. It is important to note the overall fused density results contain all attributes (including Net Zero Targets, Net Zero Committed Year, which contains many missing values, since it only originates from SBTI dataset).

5.1 Gold Standard

In the pursuit of establishing a data fusion gold standard, 15 companies were selected randomly from the matched pairs. This gold standard aims to contain the true data of the conflicting records, including company names, location, industry, key personnel (CEOs), and financial metrics such as revenue, assets, and profit etc. The data collection process focused on information available from external sources on the internet and official company websites and financial reports from the year 2021. It is important to mention that especially for the financial values, there were non-consistent information from different sources so through the assessment process, we used a range to evaluate these attributes. For our fusion task, we used all three

Table 6: Attribute densities before fusion

Attributes	Forbes	SBTI	DBpedia	Fused
Company name	1.00	1.00	1.00	1.00
Country	1.00	1.00	0.11	1.00
Industry	1.00	1.00	1.00	1.00
Assets	1.00	0	0.32	0.79
Revenue	1.00	0	1.00	1.00
Profit	1.00	0	1.00	1.00
Year Founded	1.00	0	0.89	0.96
...	...	...	...	
Overall	0.43	0.39	0.28	0.54

These are only the attributes that we used for data fusion. Overall densities include other columns as well, that are present in only one of the datasets and in our target schema.

datasets. The resulting data fusion gold standard, comprising 15 randomly selected companies, stands as a trustful benchmark for evaluating and advancing data fusion methodologies

5.2 Fusion Rules

Table 7 presents the outcomes of our experimentation with various fusion rules. For text-based attributes we tested three fusion approaches, favour source, shortest and longest string. The order of favourite sources was as follows: SBTI (most favourite), Forbes, DBpedia (least favourite). This order was chosen because SBTI is the most reliable and DBpedia the least reliable, because the data can be edited by anyone and could also be old. Company names can vary widely in terms of abbreviations, legal entities, and naming conventions. The shortest string fusion rule is effective in capturing the core identity of a company by focusing on the shortest representation, which often contains essential information. In case of founded year we experimented with favour source fusion strategy. For evaluating the Company name and Country attributes we applied exact match when comparing the fused record with the gold standard. Regarding financial metrics, where variations may be expected due to seasonal or cyclical factors, using the mean is a suitable approach for capturing the typical or average performance over a given period. For these metrics, evaluation could be difficult, because we found different values for our gold standard, compared to our Forbes and DBpedia data. For this reason, we used a 5% range for the evaluation rule, to compare our fused values with the gold standard. The Industry is also common attribute across the data sources, we used union to keep all values. The attribute is represented as a list and contains numerous values which are not consistent between the datasets because of the different naming conventions. Since naming of the industry differ in synonyms we further analyzed our fused results manually.

For our trial experiment, we employed correspondences characterized by the highest F1 score obtained through identity resolution. However, these correspondences led to the identification of 49 distinct groups, with the largest group comprising 87 entities. The challenge arose from utilizing Jaro-Winkler with a low distance value, causing numerous company names to be matched based on shared initial word tags. Consequently, we revisited the identity resolution process, seeking a balanced approach. We established a new stricter set of matching rules that, while yielding a lower F1 score, significantly improved the quality of the identified groups. In the next section, Optimal Identity Resolution Correspondences, we will provide a detailed account of the revised identity resolution methodology, the specific changes made to the matching rules, and the subsequent impact on the accuracy and reliability of the identified groups. This examination will provide insight into the iterative nature of the research process and the strategic adjustments made to achieve more meaningful results.

Table 7: Fusion Rules with the best Identity Resolution Results

Matching	**Fusion Strategy**	**Accuracy**	**Consistency**	**Density**
Company Name	Shortest String	57%	48%	100%
	Longest String	36%	48%	100%
	Favour Source	43%[7]	48%	100%
Country	Shortest String	43%	71%	100%
	Favour Source	64%[8]	71%	100%
Founded Year	Favour source, Evaluation equals	57%	73%	97%
Asset	Mean, Evaluation 5%	93%	94%	84%
Revenue	Mean, Evaluation 5%	50%	92%	100%
Profit	Mean, Evaluation 5%	57%	90%	89%

In the final fusion stage, we employed the adjusted correspondences while selecting fusion rules based on the highest accuracy values obtained in the preceding experiment. Table 8 presents the ultimate fusion rules along with accuracy, consistency, and density values exclusively associated with the application of each respective fusion rule.

For attributes originating from a single data source, where conflicts do not arise, we implemented additional fusion rules to incorporate these values into the fused dataset. In such instances, we converted the records into strings and applied the fusion rule based on the longest string.

For the final fusion, we applied all the fusion rules, and it resulted in 85% accuracy.

Table 8: Fusion Rules with Optimized Identity Resolution Results

Matching	**Fusion Strategy**	**Accuracy**	**Consistency**	**Density**
Company Name	Shortest String	100%[9]	86%	100%
Country	Shortest String	100%	86%	100%
Industry	Union	86%	86%	86%
Founded Year	Favour source	57%	73%	97%
Asset	Mean, Evaluation 5%	56%	94%	84%
Revenue	Mean, Evaluation 5%	50%	92%	100%
Profit	Mean, Evaluation 5%	57%	90%	89%

5.3 Optimal Identity Resolution Correspondences

After our identity resolution result evaluation, we discovered that our data correspondences were too permissive (matching some companies to many others) and therefore not effective for the data fusion task. To understand the reason of having larger groups, we looked into the correspondences and saw several examples. Here are a couple examples of larger groups: [Matrox, Marc's, Mastercard, Martinsa-Fadesa, MasTec,...], [TELUS, Telstra, Telia, Ternium, Telenor, Textron, Terna, Tenaris, Teleperformance, Terumo, Teradyne, Tenneco,...]

Analyzing those examples showed us that the problem was caused by the Jaro-Wrinkler comparator used for Company name, which had a high weight (70-80%). Therefore, for companies that were lacking data in other comparator attributes such as profit, revenue, asset, country, the matching only relied on the company name itself. The settings of Jaro-Wrinkler had the common fix parameter setup for 5 with a 10% of scaling factor and was considered to be match over 70% similarity score. To tackle this challenge, two students randomly experimented with different parameters and fusion strategies for an hour. During this time, we found the following correspondence results as the most appropriate for our use case. These results were optimal for our use-case as we reduced our maximum group size of 87 to 4 with a relatively high Precision, Recall and F1 scope. Table 9 gives an overview about the parameters and distribution groups and size for the fusion task.

Table 9: Identity Resolution results

(a) IR results using optimal comparators and weights

Dataset	Comparator and Weights	Precision	Recall	F1
Forbes & DBpedia	Company - Levenshtein[0.8], FoundedYear[0.1], Country - Jaro[0.1]	97.56%	72.73%	83.33%
DBpedia & SBTI	Industry[0.2], Company - Levenshtein[0.6], Country - Jaro[0.2]	87.50%	72.92%	79.55%
Forbes & SBTI	Industry[0.1], Company - JaroWrinkler[0.7], Country[0.2]	100%	67%	80%

(b) Resulting group Sizes

Distr. groups	Distr. size
2	1124
3	125
4	7

6 Overall Results

Our largest dataset was DBpedia with 10,720 records, SBTI contained 6157, while Forbes had 1999 entities. From these datasets we got 719 correspondences from DBpedia and Forbes, 534 from DBpedia and SBTI and 323 from Forbes-SBTI. Using these correspondences, our final fused dataset consisted of 1256 different records.

The final dataset demonstrated an overall accuracy of 85%, as assessed against our gold standard comprising 15 companies. It is important to note that this accuracy is influenced by all fusion evaluations. In the fusion process, we incorporated all attributes, including those which are only in one dataset. During the evaluation phase, we consistently set parameters to true since there were no conflicts, potentially contributing to a more favorable impact on our accuracy. Evaluating industries and financial data seemed to challenging task since there are many different taxonomies for industries (Automobiles and Components vs. Consumer Durables) and there is no gold truth for these records. Therefore we analyzed the fused industries manually. Furthermore, attributes such as assets, revenue and profit differs a lot not only in our datasets but also on the web, especially because we don't know which year the DBpedia values originated from. These challenges highlight the limitations in achieving precise and consistent data fusion, underscoring the complexity of aligning diverse datasets with varying temporal contexts and data quality standards. Our final fused dataset encompasses information on 1256 distinct companies, incorporating financial information and sustainability measures. Notably, we achieved an increase in overall density, elevating it from 43% to 54%. As this comprehensive density encompasses all attributes, we find considerable satisfaction in the quality of our fused dataset. By successfully integrating three datasets with a combined total of 28 columns, our results provide a solid foundation for investors seeking to understand and evaluate sustainable companies. The enriched dataset serves as a valuable resource, empowering investors with comprehensive insights into the financial and sustainability dimensions of potential investment opportunities considering the existing limitations of our results.

Table 10: Contributions table

Data Integration Step	**Contributors**
DBpedia querying	Stefan
Schema Mapping Forbes	Mariam
Schema Mapping DBpedia	Stefan, Maria
Schema Mapping SBTI	Petra, Jusztina
Identity Resolution	Mariam, Stefan, Maria, Petra, Jusztina (Each of us wrote different comparators)
Data Fusion	Mariam, Stefan, Maria, Petra, Jusztina (Each of us wrote different fusing rules)
Experimenting with different weights, comparators, fusion strategies	Mariam, Stefan, Maria, Petra, Jusztina
Writing a report	Mariam, Stefan, Maria, Petra, Jusztina