Improving Schema Matching with Linked Data

Ahmad Assaf, Eldad Louw, Aline Senart, Corentin Follenfat and David Trastour
SAP Research, SAP Labs France SAS
805 avenue du Dr. Maurice Donat, BP 1216
06254 Mougins Cedex, France
firstname.lastname@sap.com

Raphaël Troncy
EURECOM
06904 Sophia Antipolis Cedex, France
raphael.troncy@eurecom.fr

ABSTRACT

With today's public data sets containing billions of data items, more and more companies are looking to integrate external data with their traditional enterprise data to improve business intelligence analysis. These distributed data sources however exhibit heterogeneous data formats and terminologies and may contain noisy data. In this paper, we present a novel framework that enables business users to semi-automatically perform data integration on potentially noisy tabular data. This framework offers an extension to Google Refine with novel schema matching algorithms leveraging Freebase rich types. First experiments show that using Linked Data to map cell values with instances and column headers with types improves significantly the quality of the matching results and therefore should lead to more informed decisions.

1. INTRODUCTION

Companies have traditionally performed business analysis based on transactional data stored in legacy relational databases. The enterprise data available for decision makers was typically relationship management or enterprise resource planning data. However social media feeds, weblogs, sensor data, or data published by governments or international organizations are nowadays becoming increasingly available [22].

The quality and amount of structured knowledge available make it now feasible for companies to mine this huge amount of public data and integrate it in their next-generation enterprise information management systems. Analyzing this new type of data within the context of existing enterprise data should bring them new or more accurate business insights and allow better recognition of sales and market opportunities [89].

These new distributed sources, however, raise tremendous challenges. They have inherently different file formats, access protocols or query languages. They possess their own data model with different ways of representing and storing the data. Data across these sources may be noisy (e.g. duplicate or inconsistent), uncertain or be semantically similar yet different [14]. Integration and provision of a unified view for these heterogeneous and complex data structures therefore require powerful tools to map and organize the data.

In this paper, we present a framework that enables business users to semi-automatically combine potentially noisy data residing in heterogeneous silos. Semantically related data is identified and appropriate mappings are suggested to users. On user acceptance, data is aggregated and can be visualized directly or exported to Business Intelligence reporting tools. The framework is composed of a set of extensions to Google Refine server and a plug-in to its user interface[1]. Google Refine was selected for its extensibility as well as good cleansing and transformation capabilities [27].

We first map cell values with instances and column headers with types from popular data sets from the Linked Open Data Cloud. To perform the matching, we use the Auto Mapping Core (also called AMC [113]) that combines the results of various similarity algorithms. The novelty of our approach resides in our exploitation of Linked Data to improve the schema matching process. We developed specific algorithms on rich types from vector algebra and statistics. The AMC generates a list of high-quality mappings from these algorithms allowing better data integration.

First experiments show that Linked Data increases significantly the number of mappings suggested to the user. Schemas can also be discovered if column headers are not defined and can be improved when they are not named or typed correctly. Finally, data reconciliation can be performed regardless of data source languages or ambiguity. All these enhancements allow business users to get more valuable and higher-quality data and consequently to take more informed decisions.

The rest of the paper is organized as follows. Section 2 presents some related work. Section 3 describes the framework that we have designed for business users to combine data from heterogeneous sources. Section 4 validates our approach and shows the value of the framework through experiments. Finally, Section 5 concludes the paper and discusses future work.

[1]http://code.google.com/p/google-refine/

2. RELATED WORK

While schema matching has always been an active research area in data integration, new challenges are faced today by the increasing size, number and complexity of data sources and their distribution over the network. Data sets are not always correctly typed or labeled and that hinders the matching process.

In the past, some work has tried to improve existing data schemas [105] but literature mainly covers automatic or semi-automatic labeling of anonymous data sets through Web extraction. Examples include [119] that automatically labels news articles with a tree structure analysis or [138] that defines heuristics based on distance and alignment of a data value and its label. These approaches are however restricting label candidates to Web content from which the data was extracted. [33] goes a step further by launching speculative queries to standard Web search engines to enlarge the set of potential candidate labels. More recently, [94] applies machine learning techniques to respectively annotate table rows as entities, columns as their types and pairs of columns as relationships, referring to the YAGO ontology. The work presented aims however at leveraging such annotations to assist semantic search queries construction and not at improving schema matching.

With the emergence of the Semantic Web, new work in the area has tried to exploit Linked Data repositories. The authors of [127] present techniques to automatically infer a semantic model on tabular data by getting top candidates from Wikitology [51] and classifying them with the Google page ranking algorithm. Since the authors' goal is to export the resulting table data as Linked Data and not to improve schema matching, some columns can be labeled incorrectly, and acronyms and languages are not well handled [127]. In the Helix project [65], a tagging mechanism is used to add semantic information on tabular data. A sample of instances values for each column is taken and a set of tags with scores are gathered from online sources such as Freebase[2]. Tags are then correlated to infer annotations for the column. The mechanism is quite similar to ours but the resulting tags for the column are independent of the existing column name and sampling might not always provide a representative population of the instance values.

3. PROPOSITION

Google Refine (formerly Freebase Gridworks) is a tool designed to quickly and efficiently process, clean and eventually enrich large amounts of data with existing knowledge bases such as Freebase. The tool has however some limitations: it was initially designed for data cleansing on only one data set at a time, with no possibility to compose columns from different data sets. Moreover, Google Refine has some strict assumptions over the input of spreadsheets which makes it difficult to identify primitive and complex data types. The AMC is a novel framework that supports the construction and execution of new matching components or algorithms. AMC contains several matching components that can be plugged and used, like string matchers (Levenshtein, JaroWinkler... etc.), data types matchers and path matchers. It also provides a set of combination and selection algorithms to produce optimized results (weighted average, average, sigmoid... etc.). In this section, we describe in detail our framework allowing data mashup from several sources. We first present our framework architecture, then the activity flow and finally our approach to schema matching.

3.1 Framework Architecture

Google Refine makes use of a modular web application framework similar to OSGi called Butterfly[3]. The server-side written in Java maintains states of the data (undo/redo history, long-running processes, etc.) while the client-side implemented in Javascript maintains states of the user interface (facets and their selections, view pagination, etc.). Communication between the client and server is done through REST web services.

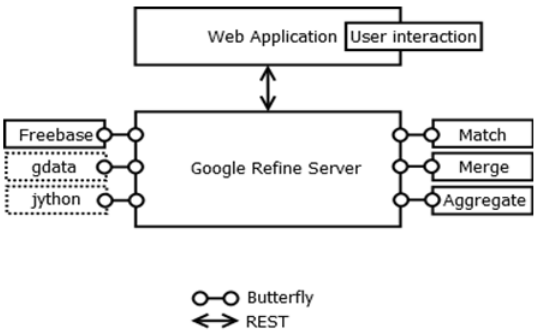

Figure 1: Framework Architecture

As depicted in 1, our framework leverages Google Refine and defines three new Butterfly modules to extend the server's functionality (namely Match, Merge and Aggregate modules) and one JavaScript extension to capture user interaction with these new data matching capabilities.

3.2 Activity Flow

This section presents the sequence of activities and interdependencies between these activities when using our framework. 2 gives an outline of these activities.

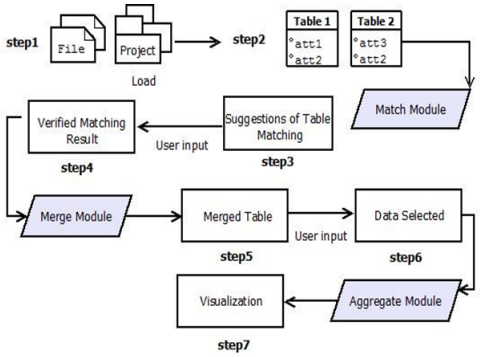

Figure 2: Activity Workflow

The data sets to match can be contained in files (e.g. csv, Excel spreadsheets, etc.) or defined in Google Refine projects (step 1). The inputs for the match module are

[2]http://www.freebase.com/

[3]http://code.google.com/p/simile-butterfly/

the source and target files and/or projects that contain the data sets. These projects are imported into the internal data structure (called schema) of the AMC [112] (step 2). The AMC then uses a set of built-in algorithms to calculate similarities between the source and target schemas on an element basis, i.e. column names in the case of spreadsheets or relational databases. The output is a set of similarities, each containing a triple consisting of source schema element, target element, and similarity between the two.

These results are presented to the user in tabular form (step 3) such that s/he can check, correct, and potentially complete the mappings (step 4).

Once the user has completed the matching of columns, the merge information is sent back to Google Refine, which calls the merge module. This module creates a new project, which contains a union of the two projects where the matched columns of the target project are appended to the corresponding source columns (step 5). The user can then select the columns that s/he wants to merge and visualize by dragging and dropping the required columns onto the fields that represent the x and y axes (step 6).

Once the selection has been performed, the aggregation module merges the filtered columns and the result can then be visualized (step 7). As aggregation operations can quickly become complex, our default aggregation module can be replaced by more advanced analytics on tabular data. The integration of such a tool is part of future work.

3.3 Schema Matching

Schema matching is typically used in business to business integration, metamodel matching, as well as Extract, Transform, Load (ETL) processes. For non-IT specialists the typical way of comparing financial data from two different years or quarters, for example, would be to copy and paste the data from one Excel spreadsheet into another one, thus creating reduncancies and potentially introducing copy-and-paste errors. By using schema matching techniques it is possible to support this process semi-automatically, i.e. to determine which columns are similar and propose them to the user for integration. This integration can then be done with appropriate business intelligence tools to provide visualisations.

One of the problems in performing the integration is the quality of data. The columns may contain data that is noisy or incorrect. There may also be no column headers to provide suitable information for matching. A number of approaches exploit the similarities of headers or similarities of types of column data. We propose a new approach that exploits semantic rich typing provided by popular datasets from the Linked Data cloud.

3.4 Data Reconciliation

Reconciliation enables entity resolution, i.e. matching cells with corresponding typed entities in case of tabular data. Google Refine already supports reconciliation with Freebase but requires confirmation from the user. For medium to large data sets, this can be very time-consuming. To reconcile data, we therefore first identify the columns that are candidates for reconciliation by skipping the columns containing numerical values or dates. We then use the Freebase search API to query for each cell of the source and target columns the list of typed entities candidates. Results are cached in order to be retrieved by our similarity algorithms.

3.5 Matching Unnamed and Untyped Columns

The AMC has the ability to combine the results of different matching algorithms. Its default built-in matching algorithms work on column headers and produce an overall similarity score between the compared schema elements. It has been proven that combining different algorithms greatly increases the quality of matching results [113][133]. However, when headers are missing or ambiguous, the AMC can only exploit domain intersection and inclusion algorithms based on column data. We have therefore implemented three new similarity algorithms that leverage the rich types retrieved from Linked Data in order to enhance the matching results of unnamed or untyped columns. They are presented below.

3.5.1 Cosine Similarity

The first algorithm that we implemented is based on vector algebra. Let v be the vector of ranked candidate types returned by Freebase for each cell value of a column. Then:

$$v := \sum_{i=1}^{K} a_i * \overrightarrow{t_i}$$

where a_i is the score of the entry and $\overrightarrow{t_i}$ is the type returned by Freebase. The vector notation is chosen to indicate that each distinct answer determines one dimension in the space of results.

Each cell value has now a weighted result set that can be used for aggregation to produce a result vector for the whole column. The column result V is then given by:

$$V? \sum_{i=1}^{n} v_i$$

We compare the result vector of candidate types from the source column with the result vector of candidate types from the target column. Let W be the result vector for the target column, then the similarity s between the columns pair can be calculated using the absolute value of the cosine similarity function:

$$s? \frac{|(V * W)|}{\|V\| * \|W\|}$$

3.5.2 Pearson Product-Moment Correlation Coefficient (PPMCC)

The second algorithm that we implemented is PPMCC, a statistical measure of the linear independence between two variables (x, y) [86]. In our method, x is an array that represents the total scores for the source column rich types, y is an array that represents the mapped values between the source and the target columns. The values present in x but not in y are represented by zeros. We have:

$$SourceColumn\ [\{R_1, C_{sr1}\}, \{R_2, C_{sr2}\}, \{R_3, C_{sr3}\} \ldots\ \{R_n, C_{srn}\}]$$

$$TargetColumn\ [\{R_1, C_{tr1}\}, \{R_2, C_{tr2}\}, \{R_3, C_{tr3}\} \ldots\ \{R_n, C_{trn}\}]$$

Where $R_1,\ R_2, \ldots, R_n$ are different rich type values retrieved from Freebase, $C_{sr1},\ C_{sr2}, \ldots,\ C_{srn}$ are the sum of scores for each corresponding r occurrence in the source column, and $C_{tr1},\ C_{tr2},\ \ldots,\ C_{trn}$ are the sum of scores for each corresponding r occurrence in the target column.

The input for PPMC consists of two arrays that represent the values from the source and target columns, where the

source column is the column with the largest set of rich types found. For example:

$$X = [C_{sr1}, C_{sr2}, C_{sr4}, \ldots, C_{srn}]$$

$$Y = [0, C_{tr2}, C_{tr4}, \ldots, C_{trn}]$$

Then the sample correlation coefficient (r) is calculated using:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Based on a sample paired data(x_i, y_i), the sample PPMCC is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

Where $\left(\frac{x_i - \overline{x}}{s_x}\right)$, $\overline{x}$ and s_x are the standard score, sample mean and sample standard deviation, respectively.

3.5.3 Spearman's Rank Correlation Coefficient

The last algorithm that we implemented to match unnamed and untyped columns is Spearman's rank correlation coefficient. It applies a rank transformation on the input data and computes PPMCC afterwards on the ranked data. In our experiments we used Natural Ranking with default strategies for handling ties and NaN values. The ranking algorithm is however configurable and can be enhanced by using more sophisticated measures.

3.6 Column Labeling

We showed in the previous section how to match unnamed and untyped columns. Column labeling is however beneficial as the results of our previous algorithms can be combined with traditional header matching techniques to improve the quality of matching.

Rich types retrieved from Freebase are independent from each other. We need to find a method that will determine normalized score for each type in the set by balancing the proportion of high scores with the lower ones. We used Wilson score interval for a Bernoulli parameter that is presented in the following equation:

$$w = \left(\widehat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\left[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n\right]/n}\right)/\left(1 + z^2_{\alpha/2}/n\right)$$

Here $\hat{p}$ is the average score for each rich type, n is the total number of scores and $z_{\alpha/2}$ is the score level; in our case it is 1.96 to reflect a score level of 0.95.

3.7 Handling Non-String Values

So far, we have covered several methods to identify the similarity between "String" values, but how about other numeral values such as dates, money, distance,etc.? For this purpose, we have implemented some basic type identifier that can recognize dates, money, numerical values, numerals used as identifiers. This will help us in better match corresponding entries. Adjusting AMC's combination algorithms can be of great importance at this stage. For example, assigning weights to different matchers and tweaking the configuration can yield more accurate results.

4. EXPERIMENTS

We present in this section results from experiments we conducted using the different methods described above. To appreciate the value of our approach, we have used a real life scenario that exposes common problems faced by the management in SAP. The data we have used come from two different SAP systems: the Event tracker and the Travel Expense Manager.

The Event Tracker provides an overview of events (Conferences, Internal events, etc.) that SAP Research employees contribute to or host. The entries in this system contain as much information as necessary to give an overview of the activity like the activity type and title, travel destination, travel costs divided into several sub categories (conference fees, accommodation, transportation and others), and duration related information (departure, return dates). Entries in the Event Tracker are generally entered in batches as employees fill in their planned events that they wish to attend or contribute to at the beginning of each year. Afterwards, managers can either accept or reject these planned events according to their allocated budget.

On the other hand, the Travel Expense Manager contains the actual expenses data for the successfully accepted events. This system is used by employees to enter their actual trip details in order to claim their expenses. It contains more detailed information and aggregated views of the events, such as the total cost, duration calculated in days, currency exchange rates and lots of internal system tags and identifiers.

Matching reports from these two systems is of great benefit to managers to organize and monitor their allocated budget. They mainly want to:

1. Find the number of the actual (accepted) travels compared with the total number of entered events.
2. Calculate the deviation between the estimated and actual cost of each event.

However, matching from these two sources can face several difficulties that can be classified in two categories: column headers and cells. Global labels (or column headers as we are dealing with spreadsheet files) can have the following problems:

1. Missing labels: importing files into Google Refine with empty headers will result in assigning that column a dummy name by concatenating the word "column" with a number starting from 0.
2. Dummy labels or semantically unrelated names: this is a common problem especially from the data coming from the Travel Expense Manager. This can be applied to columns that are labeled according to the corresponding database table (i.e. lbl_dst to denote destination label). Moreover, column labels do not often convey the semantic type of the underlying data.

The second category of difficulties is at cell (single entry) level:

1. Detecting different date formats: we have found out that dates field coming from the two systems have different formats. Moreover, the built-in type detection in Google Refine converts detected date into another third format.

2. Entries from different people can be made in different languages.

3. Entries in the two systems can be incomplete, an entry can be shortened automatically by the system. For example, selecting a country in the Travel Expense Manager will result in filling out that country code in the exported report (i.e. France = FR).

4. Inaccurate entries: this is one of the most common problems. Users enter sometimes several values in some fields that correspond to the same entity. For example, in the destination column, users can enter the country, the airport at the destination, the city or even the exact location of the event (i.e. office location).

The data used in our evaluation consists of around 60 columns and more than 1000 rows. Our source data set will be the data coming from Event Tracker, and our target data set will be the data from the Travel Expense Manager.

By manually examining the two data sets, we have found out that most of the column headers in the source table exist and adequately present the data. However, we have noticed few missing labels in the target table and few ambiguous column headers. We have detected several entries in several languages: the main language is English but we have also identified French, German. Destination field had entries in several formats: we have noticed airport names, airports by their IATA code, country codes, and cities.

Running AMC with its default matchers returns the matching results shown in Table 1.

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
Begins On	Trip Begins On	0.8333334
Ends On	Trip Ends On	0.8
Total	Total Cost	0.7333335
Trip	Trip Destination	0.72727275
Amount	Receipt Amount	0.7142875
Pd by Comp	Paid by Company	0.6904762
Period	Period Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55
Curr.	Currency	0.5
Crcy	Currency	0.5

Table 1: Similarity Scores Using the AMC Default Matching Algorithms

The AMC has perfectly matched the two columns labeled "Reason for Trip" using name and data type similarity calculations (the type here was identified as a String). Moreover, it has computed several similarities for columns based on the pre-implemented String matchers that were applied on the column headers and the primitive data types of the cells (Integer, Double, Float, etc.). However, there is no alignment found between the other columns since their headers are not related to each other, although the actual cell values can be similar. AMC's default configuration has a threshold of 50%, so any similarity score below that will not be shown.

The Cosine Similarity algorithm combined with the AMC default matchers produces the results shown in Table 2.

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
tr_dst		0.9496432
Begins On	Trip Begins On	0.9166667
Ends On	Trip Ends On	0.9
Amount	Receipt Amount	0.8571428
Curr.	Currency	0.75
Crcy	Currency	0.75
Total	Total Cost	0.7333335
Trip	Trip Destination	0.7321428
Pd by Comp	Paid by Company	0.6904762
Period	Period Number	0.6666667
Trip	Trip Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55

Table 2: Similarity Scores Using the AMC Default Matching Algorithms + Cosine Similarity Method

We notice that we have an increased number of matches (+2), and that the similarity score for several matches has improved. For example, the "tr_dst" column is now aligned to the blank header. This shows that our approach allows performing schema matching on columns with no headers.

For simplicity reason we have used the default combination algorithm for AMC which is an average of the applied algorithms (AMC's native and Cosine). We should also note that we have configured AMC's matchers to identify a "SIMILARTY_UNKOWN" value for columns that could not be matched successfully, which will allow other matchers to perform better. For example, our semantic matchers will skip columns that do not convey semantic meaning thus not affecting the score of other matchers. Moreover, the relatively high similarity score of "tr_dst" column is explained by the fact that the native AMC matching algorithm has skipped that column as it does not have a valid header, and the results are solely those of the Cosine matcher. Likewise, the Cosine matcher skips checking the "Cost" columns as they contain numeric values, and the implemented numerical matchers with the AMC's native matcher results are taken into account. Our numerical matchers' implementation gives a perfect similarity score for columns that are identified as date or money or IDs. However, this can be improved in the future as we can have different date hierarchy and numbers as IDs can present different entities. Combining this approach with the semantic and string matchers was found to yield good matching results.

The (PPMCC) Similarity algorithm combined with the AMC default matchers produces the results shown in Table 3.

We notice that by plugging the Spearman method, the number of matches and similarity results have decreased (-4). After Several experiments we have found that this method does not work well with noisy data sets. For instance, the similarity results returned by Cosine, Pearson's and Spearman's matchers for the {tr_dst, empty header} pair is much higher: 95%, 97% and 43% respectively.

To properly measure the impact of each algorithm, we have tested the three algorithms (Cosine, PPMCC and Spearman) alone by de-activating the AMC's default matchers on the above data set. We have noticed that generally, the Cosine and PPMCC matchers perform well, resulting in more

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
tr_dst		0.97351624
Begins On	Trip Begins On	0.833334
Ends On	Trip Ends On	0.8
Total	Total Cost	0.7333335
Trip	Trip Destination	0.7321428
Amount	Receipt Amount	0.7142857
Curr.	Currency	0.7041873
Crcy	Currency	0.6931407
Pd by Comp	Paid by Company	0.6904762
Period	Period Number	0.6666667
Trip	Trip Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55

Table 3: Similarity Scores Using the AMC Default Matching Algorithms+ the PPMCC Similarity Method

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
Begins On	Trip Begins On	0.8333334
Ends On	Trip Ends On	0.8
Total	Total Cost	0.7333335
Amount	Receipt Amount	0.7142857
Pd by Comp	Paid by Company	0.6904762
Currency2	Curr.	0.6689202
Trip	Trip Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55

Table 4: Similarity Scores Using the AMC Default Matching Algorithms + Spearman Similarity Method

matching and better similarity score. However, the Spearman method was successful in finding more matches but with a lower similarity score than the others.

To better evaluate the three algorithms, we have tested them on four different data sets extracted from the Travel Expense Manager and Event Tracker systems. We ensured that the different experiments will cover all the cases needed to properly evaluate the matcher dealing with all the problems mentioned earlier.

We have found that generally the Cosine method is the best performing algorithm compared to the other two especially when dealing with noisy data sets. This was noticed particularly in our fourth experiment as the Cosine algorithm performed around 20% better than the other two methods. After investigating the dataset, we have found that several columns contained noisy and unrelated data. For example, in a "City" column, we had values such as "reference book" or "NOT_KNOWN".

To gain better similarity results we decided to combine several matching algorithms together. By doing so, we would benefit from the power of the AMC's string matchers that will work on column headers and our numeral and semantic matchers.

The Cosine and PPMCC Similarity algorithms combined with the AMC default matchers produces the results shown in Table 4.

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
tr_dst		0.96351624
Curr.	Currency	0.79221311
Crcy	Currency	0.78173274
Begins On	Trip Begins On	0.77777785
Ends On	Trip Ends On	0.76666665
Amount	Receipt Amount	0.7380952
Total	Total Cost	0.7333335
Trip Country/Group	Ctr2	0.7194848
Pd by Comp	Paid by Company	0.6904762
Period	Period Number	0.6666667
Trip	Trip Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55

Table 5: Similarity Scores Using the Combination of Cosine, PPMCC and AMC's defaults

The combination of the above mentioned algorithms have enhanced generally the similarity scores for the group. Moreover, we notice that the column "Trip Country/Group" was matched with "Ctr2". This match was not computed singularly by any of the previous algorithms. However, we notice that the match {Trip, Trip Destination} is now missing, probably as the similarity score is below the defined threshold.

Now, we will try and group all the mentioned algorithms. The combination of all Similarity algorithms with the AMC default matchers produces the results shown in Table 5.

Source Column	Target Column	Similarity Score
Reason for Trip	Reason for Trip	1
tr_dst		0.8779132
Curr.	Currency	0.80033726
Crcy	Currency	0.79380125
Begins On	Trip Begins On	0.7708334
Trip Country/Group	Ctr2	0.767311
Ends On	Trip Ends On	0.7625
Amount	Receipt Amount	0.7410714
Total	Total Cost	0.7333335
Trip	Trip Destination	0.7321428
Pd by Comp	Paid by Company	0.6904762
Period	Period Number	0.6666667
Trip	Trip Number	0.6666667
Pers.No.	Sequential no.	0.5555556
M/Km	Total Miles/Km	0.55

We notice that now we have an increased number of matches (15 compared to 14 in the previous trials). The column {Trip, Trip Destination} is matched again and the newly previously matched column {Trip Country/Group, Ctr2} has a higher similarity score. We have found that combining matching algorithms resulted in higher number of matches. Several tuning methods can be applied in order to enhance the similarity score as well. Trying other combination algorithms instead of the naiive average will be an essential part of our future work.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a framework enabling mashup of potentially noisy enterprise and external data. The implementation is based on Google Refine and uses Freebase to annotate data with rich types. As a result, the matching process of heterogeneous data sources is improved. Our preliminary evaluation shows that for datasets where mappings were relevant yet not proposed, our framework provides higher quality matching results. Additionally, the number of matches discovered is increased when Linked Data is used in most datasets. We plan in future work to evaluate the framework on larger datasets using rigorous statistical analysis of [49]. We also consider integrating additional linked open data sources of semantic types such as DBpedia [26] or YAGO [126] and evaluate our matching results against instance-based ontology alignment benchmarks such as OAEI[4] or ISLab[5]. Another future work will be to generalize our approach on data schemas to data classification. The same way the AMC helps identifying the best matches for two datasets, we plan to use it for identifying the best statistical classifiers for a sole dataset, based on normalized scores.

6. REFERENCES

[1] Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.

[2] M. Acosta, A. Zaveri, E. Simperl, and D. Kontokostas. Crowdsourcing Linked Data quality assessment. In *12th International Semantic Web Conference (ISWC)*, 2013.

[3] A. Ahmad, S. Aline, and T. Raphaël. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In *24th World Wide Web Conference (WWW), Demos Track*, Florence, Italy, 2015.

[4] H. Aidan, H. Andreas, and D. Stefan. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.

[5] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.

[6] M. Alistair and B. Sean. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 2009. `http://www.w3.org/TR/skos-reference/`.

[7] J. Anja, C. Richard, and B. Chrstian. State of the lod cloud. `http://lod-cloud.net/state/`.

[8] F. Annika. Quality Characteristics of Linked Data Publishing Datasources. Master's thesis, Humboldt-UniversitÃd't zu Berlin, 2010.

[9] I. Antoine and S. Ed. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, 2009.

[10] A. Assaf, E. Louw, A. Senart, C. Follenfant, R. Troncy, and D. Trastour. RUBIX: a framework for improving data integration with linked data. In *International Workshop on Open Data (WOD'12)*, pages 13–21, 2012.

[11] A. Assaf and A. Senart. Data Quality Principles in the Semantic Web. In *6th International Conference on Semantic Computing ICSC '12*, 2012.

[12] A. Assaf, A. Senart, and R. Troncy. SNARC - An Approach for Aggregating and Recommending Contextualized Social Content. In *The Semantic Web: ESWC 2013 Satellite Events, Revised Selected Papers*, pages 319–326, 2013.

[13] S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.

[14] C. Avitha, G. S. Sadasivam, and S. N. Shenoy. Ontology Based Semantic Integration of Heterogeneous Databases. *European Journal of Scientific Research*, page 115, 2011.

[15] S. Ben. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[16] H. Bernhard and P. Niko. DSNotify: Detecting and Fixing Broken Links in Linked Data Sets. In *8th International Workshop on Web Semantics*, 2009.

[17] S. Besiki, G. Les, T. M. B., and S. L. C. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 2007.

[18] C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Jorunal of Web Semantics*, 7(1), 2009.

[19] C. Böhm, G. Kasneci, and F. Naumann. Latent Topics in Graph-structured Data. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2663–2666, Maui, Hawaii, USA, 2012.

[20] C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with ProLOD. In *26th International Conference on Data Engineering Workshops (ICDEW)*, 2010.

[21] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ACM International Conference on Management of Data (SIGMOD)*, 2008.

[22] D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[23] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th International Conference on World Wide Web (WWW'98)*, 1998.

[24] C. Buil-Aranda and A. Hogan. SPARQL Web-Querying Infrastructure: Ready for Action? In *12th International Semantic Web Conference (ISWC)*, 2013.

[25] B. Christian. Evolving the Web into a Global Data Space. In *28th British National Conference on*

[4]http://oaei.ontologymatching.org/2011/instance/index.html
[5]http://islab.dico.unimi.it/iimb/

Advances in Databases, 2011.

[26] B. Christian, L. Jens, K. Georgi, A. Sören, B. Christian, C. Richard, and H. Sebastian. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.

[27] B. Christian, H. T, and B.-L. T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[28] M. Christian, H. Bernhard, and I. Antoine. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.

[29] B. Christoph, L. Johannes, and N. Felix. Creating voiD Descriptions for Web-scale Data. *Journal of Web Semantics*, 9(3):339–345, 2011.

[30] M. Cornolti, P. Ferragina, and M. Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In *22nd World Wide Web Conference (WWW)*, 2013.

[31] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *5th European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Spain, 2008.

[32] R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing Linked Datasets with the VoID Vocabulary. W3C Note, 2011. `http://www.w3.org/TR/void/`.

[33] A. S. da Silva, D. Barbosa, J. M. B. Cavalcanti, and M. A. S. Sevalho. Labeling Data Extracted from the Web. In *On The Move Confederated International Conferences*, pages 1099–1116, 2007.

[34] M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semantic Web Journal*, 2011.

[35] R. Dave. The Organization Ontology. W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-org`.

[36] J. Debattista, C. Lange, and S. Auer. daQ, an Ontology for Dataset Quality Information. In *Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)*, 2014.

[37] R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. In *7th European Semantic Web Conference (ESWC)*, 2010.

[38] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked Open Data to Support Content-based Recommender Systems. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.

[39] C. Didier, U. Ricardo, B. Andreas, and L. Jens. CROCUS: Cluster-based ontology data cleansing. In *2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice*, 2014.

[40] B. Diego, F. Sergio, and F. Iv"an. Cooking HTTP content negotiation with Vapour. In *4th Workshop on Scripting for the Semantic Web (SFSW'08)*, 2008.

[41] K. Dimitris, Z. Amrapali, A. Sören, and L. J. TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. *4th Conference on Knowledge Engineering and Semantic Web*, 2013.

[42] L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. In *13st ACM International Conference on Information and Knowledge Management (CIKM)*, 2004.

[43] N. Douglas. Developing Spatial Data Infrastructures: The SDI Cookbook, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf`.

[44] P. Emmanuel, B. Christian, K. David, and L. Ryan. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *5th International Semantic Web Conference (ISWC'06)*, pages 158–171, 2006.

[45] D.-A. Ernesto, D. Lucas, S.-T. Lars, and N. Wolfgang. Real-time top-n recommendation in social streams. In *6th ACM conference on Recommender systems - RecSys*, 2012.

[46] S. Evren, S. Michael, and W. Evan. Opening, Closing Worlds - On Integrity Constraints. In *5th OWLED Workshop on OWL: Experiences and Directions*, 2008.

[47] B. Eytan, R. Itamar, M. Cameron, and A. Lada. The role of social networks in information diffusion. In *21th International Conference on World Wide Web (WWW'12)*, 2012.

[48] M. Fadi and E. John. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-dcat/`.

[49] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recogn. Lett.*, 2006.

[50] B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In *11th European Semantic Web Conference (ESWC)*, 2014.

[51] T. Finin, Z. Syed, J. Mayfield, P. Mcnamee, and C. Piatko. Using Wikitology for Cross-Document Entity Coreference Resolution. In *AAAI Spring Symposium on Learning*, 2009.

[52] G. Flouris, Y. Roussakis, and M. Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. In *2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12)*, 2012.

[53] B. Forchhammer, A. Jentzsch, and F. Naumann. LODOP - Multi-Query Optimization for Linked Data Profiling Queries. In *International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES)*, Heraklion, Greece, 2014.

[54] P. Frischmuth, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C. Marquardt. Towards Linked Data based Enterprise Information Integration. In *Workshop on Semantic Web Enterprise Adoption and Best Practice Co-located with 12th International Semantic Web Conference (ISWC'13)*, 2013.

[55] P. Frischmuth, J. Klímek, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C.-M. Marquardt. Linked Data in Enterprise Information Integration. 2012.

[56] M. Frosterus, E. Hyvönen, and J. Laitio. Creating and Publishing Semantic Metadata about Linked and Open Datasets. In *Linking Government Data*. 2011.

[57] M. Frosterus, E. Hyvönen, and J. Laitio.

DataFinland - A Semantic Portal for Open and Linked Datasets. In *8th Extended Semantic Web Conference (ESWC)*, pages 243–254, 2011.

[58] C. Fürber and M. Hepp. SWIQA - A Semantic Web information quality assessment framework. 2011.

[59] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of The ACM*, 30(11):964–971, 1987.

[60] T. Giovanni, C. Richard, C. Michele, D. Szymon, D. Renaud, and D. Stefan. Sig.ma: Live views on the Web of data. *Journal of Web Semantics*, 8(4), 2010.

[61] G. Gouriten and P. Senellart. API Blender: A Uniform Interface to Social Platform APIs. In *21th International Conference on World Wide Web (WWW'12)*, 2012.

[62] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing Linked Data Mappings Using Network Measures. In *9th European Semantic Web Conference (ESWC)*, 2012.

[63] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data Summaries for On-demand Queries over Linked Data. In *19th World Wide Web Conference (WWW)*, 2010.

[64] A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. In *8th International Semantic Web Conference (ISWC)*, 2009.

[65] O. Hassanzadeh, S. Duan, A. Fokoue, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Helix: Online Enterprise Data Analytics. In *20th International Conference Companion on World Wide Web (WWW'11)*, pages 225–228, 2011.

[66] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. 2010.

[67] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.

[68] K. Houda, A. Ghislain, R. Giuseppe, and R. Troncy. Aggregating Social Media for Enhancing Conference Experiences. In *1st Internationl Workshop on Real-Time Analysis and Mining of Social Streams*, 2012.

[69] R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.

[70] C. Iván and B. Alejandro. Semantic contextualisation of social tag-based profiles and item recommendations. In *12th Internationl Conference on E-Commerce and Web Technolgoies*, 2011.

[71] P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data, 2013. `http://knoesis.wright.edu/faculty/pascal/pub/nomoneylod.pdf`.

[72] M. James and D. E. Almasi. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey Business Technology Office, 2001.

[73] L. Jens and S. Soeren. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research*, 2009.

[74] A. Jentzsch. Profiling the Web of Data. In *13th International Semantic Web Conference (ISWC), Doctoral Consortium*, Trentino, Italy, 2014.

[75] D. Jeremy, L. Santiago, L. Christoph, and A. Sören. LUZZU - A framework for linked data quality assessment. *CoRR*, abs/1412.3750, 2014.

[76] J. M. Juran and A. B. Godfrey. *Juran's quality handbook*. McGraw Hill, 1999.

[77] K. B. K., S. D. M., and W. R. Y. Information quality benchmarks: product and service performance. *Communications of the ACM*, 2002.

[78] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *10th European Semantic Web Conference (ESWC)*, 2013.

[79] H. Kang and B. Shneiderman. MediaFinder: an interface for dynamic personal media management with semantic regions. In *Conference on Human Factors in Computing Systems (CHI)*, pages 764–765. ACM, 2003.

[80] C. Keet, M. del Carmen Suárez-Figueroa, and M. Poveda-Villalón. The Current Landscape of Pitfalls in Ontologies. In *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2013.

[81] S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *7th Extended Semantic Web Conference (ESWC)*, pages 272–287, Heraklion, Greece, 2010.

[82] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Journal*, 1999.

[83] M. Konrath, T. Gottron, S. Staab, and A. Scherp. SchemEX - Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. *Journal of Web Semantics*, 16, 2012.

[84] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven Evaluation of Linked Data Quality. In *23rd International Conference on World Wide Web (WWW'14)*, 2014.

[85] Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.

[86] C. J. Kowalski. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society*, 1972.

[87] S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic Domain Identification for Linked Open Data. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 205–212, 2013.

[88] A. Langegger and W. Woss. RDFStats - An Extensible RDF Statistics Generator and Library. In *20th International Workshop on Database and Expert Systems Application (DEXA)*, pages 79–83, 2009.

[89] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 2011.

[90] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, 1998.

[91] M. Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.

[92] J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *12th th ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*, 2006.

[93] H. Li. Data Profiling for Semantic Web Data. In *International Conference on Web Information Systems and Mining (WISM)*, pages 472–479, 2012.

[94] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationships. *VLDB Endowment*, pages 1338–1347, 2010.

[95] E. Mäkelä. Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets. In *11th European Semantic Web Conference (ESWC), Demo Track*, Heraklion, Greece, 2014.

[96] P. Marco and G. Siva. Investigating topic models for social media user recommendation. In *20th International Conference on World Wide Web (WWW'11)*, 2011.

[97] N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery Hub: On-the-fly Linked Data Exploratory Search. In *The 9th International Conference on Semantic Systems*, 2013.

[98] B. Martin, B. Ciro, E. Ivan, F. Markus, K. Dimitris, and H. Sebastian. DataID: Towards Semantically Rich Metadata for Complex Datasets. In *10th International Conference on Semantic Systems*, 2014.

[99] V. Mateja. LODGrefine - LOD-enabled Google Refine in Action. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.

[100] S. Max, B. Christian, and P. Heiko. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, 2014.

[101] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.

[102] H. Michael, H. Wolfgang, R. Yves, F. Lee, and A. Danny. SCOVO: Using Statistics on the Web of Data. In *ESWC*, 2009.

[103] N. Mihindukulasooriya, R. Garcia-Castro, and M. E. Gutiérrez. Linked Data Platform as a novel approach for Enterprise Application Integration. In *4th International Workshop on Consuming Linked Data (COLD'13)*, 2013.

[104] B. Mike. Deconstructing the Google Knowledge Graph. `http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph`.

[105] R. J. Miller and P. Andritsos. Schema Discovery. *IEEE Data Engineering Bulletin*, 26:40–45, 2003.

[106] T. Nickolai, U. J, and D. Renaud. DING! Dataset ranking using formal descriptions. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.

[107] A. Nikolov, M. d'Aquin, and E. Motta. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. In *Joint International Semantic Technology Conference (JIST)*, 2011.

[108] H. Olaf and Z. Jun. Using web data provenance for quality assessment. In *8th International Semantic Web Conference (ISWC)*, 2009.

[109] S. Osma and M. Christian. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.

[110] S. Osma and H. Eero. Improving the quality of SKOS vocabularies with skosify. In *The 18th International Conference on Knowledge Engineering and Knowledge Management*, 2012.

[111] H. Patricia. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Getty Research Institute, 2010.

[112] E. Peukert, J. Eberius, and E. Rahm. AMC - A framework for modelling and comparing matching systems as matching processes. In *IEEE 27th International Conference on Data Engineering (ICDE'11)*, 2011.

[113] E. Peukert, J. Eberius, and E. Rahm. A Self-Configuring Schema Matching System. In *IEEE 28th International Conference on Data Engineering (ICDE'12)*, 2012.

[114] A. Phil and S. Gofran. Asset Description Metadata Schema (ADMS). W3C Working Group Note, 2013. `http://www.w3.org/TR/vocab-adms`.

[115] M. PN, M. Hannes, and B. Christian. Sieve: linked data quality assessment and fusion. 2012.

[116] M. Poveda-Villalón, M. Suárez-Figueroa, and A. GÃşmez-Pérez. Validating Ontologies with OOPS! In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2012.

[117] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[118] N. Press. Understanding Metadata. Technical report, National Information Standards Organization, 2004.

[119] D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender. Automatic Web News Extraction Using Tree Edit Distance. In *13th International World Wide Web Conference (WWW'04)*, pages 502–601, 2004.

[120] I. Renato and M. James. vCard Ontology - for describing People and Organizations. W3C Interest Group Note, 2014. `http://www.w3.org/TR/vcard-rdf`.

[121] E. Ruckhaus, O. Baldizan, and M.-E. Vidal. Analyzing Linked Data Quality with LiQuate. In *11th European Semantic Web Conference (ESWC)*, 2014.

[122] A. Rula and A. Zaveri. Methodology for Assessment of Linked Data Quality. In *1st Workshop on Linked Data Quality (LDQ)*, 2014.

[123] D. Soergel. Thesauri and ontologies in digital libraries. In *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.

[124] C. Soumen, D. B. E., S. K. Ravi, R. Prabhakar, R. Sridhar, T. Andrew, G. David, and K. Jon. Mining the web's link structure. *Computer*, 1999.

[125] T. Steiner and S. Mirea. SEKI@home or Crowdsourcing an Open Knowledge Graph. In *1st International Workshop on Knowledge Extraction & Consolidation from Social Media (KECSM'12)*, Boston, USA, 2012.

[126] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW)*, 2007.

[127] Z. Syed, T. Finin, V. Mulwad, and A. Joshi. Exploiting a Web of Semantic Data for Interpreting Tables. In *2nd Web Science Conference*, 2010.

[128] J. Tao, L. Ding, and D. L. McGuinness. Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems. In *42nd Hawaii International Conference on System Sciences, HICSS'09*, pages 1–10, 2009.

[129] B.-L. Tim. Linked Data - Design Issues. W3C Personal Notes, 2006. `http://www.w3.org/DesignIssues/LinkedData`.

[130] L. Timothy, S. Satya, and M. Deborah. PROV-O: The PROV Ontology. W3C Recommendation, 2013. `http://www.w3.org/TR/prov-o`.

[131] A. Tonon, M. Catasta, G. Demartini, P. Cudré-Mauroux, and K. Aberer. TRank: Ranking Entity Types Using the Web of Data. In *12th International Semantic Web Conference (ISWC)*, 2013.

[132] G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the open linked data. In *6th International Semantic Web Conference (ISWC)*, 2007.

[133] S. Umberto and T. Raphaël. oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies. In *6th International Conference on Web Information Systems Engineering*, 2005.

[134] R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL - General Entity Annotation Benchmark Framework. In *24th World Wide Web Conference (WWW)*, 2015.

[135] Z. Valentina and C. L. Social ranking: uncovering relevant content using tag-based recommender systems. In *2nd ACM conference on Recommender systems - RecSys*, 2008.

[136] G. Vickery. Review of Recent Studies on PSI-use and Related Market Developments. Technical report, EC DG Information Society, 2011.

[137] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI Workshop: Ontologies and Information*, pages 108–117, 2001.

[138] J. Wang and F. H. Lochovsky. Data Extraction and Label Assignment for Web Databases. In *12th International World Wide Web Conference (WWW'03)*, pages 187–196, 2003.

[139] W. R. Y. and S. D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996.

[140] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.