Interoperable Visualization Framework towards enhancing mapping and integration of official statistics

Haitham Zeidan
Palestinian Central Bureau of Statistics (PCBS)
Jerusalem, Palestine
Haitham@pcbs.gov.ps

Abstract

The aim of this research is to introduce a new interoperable visual analytics framework Towards Enhancing Presentation of Official Statistics. This paper aims to investigate how data integration and information visualization could be used to increase readability and interoperability of statistical data. Statistical data has gained many interests from policy makers, city planners, researchers and ordinary citizens as well. from an official statistics' point of view, data integration is of major interest as a means of using available information more efficiently and improving the quality of a statistical agency's products.

We implemented and proposed statistical indicators schema and mapping algorithm which is conceptually simple and is based on hamming distance [1] and edit (Levenshtein) distance [2] mapping methods in addition to the ontology. Also we build GUI to import the indicators with data values from different sources. The performance and accuracy of this algorithm was measured by experiment, we started to import the data and indicators from different sources to our target schema which contains the indicators, Units and Subgroups. during the data import using our algorithm, the exact matched indicators, units and subgroups will be mapped automatically to the indicators, units, and subgroups in the schema, in case that we import not exact matched indicator, units or subgroups the algorithm will calculate the edit distance (minimum operations needed) for mapping the imported indicator with the nearest indicator in the schema, the same thing will happen for units or subgroups, the results showed that the accuracy of the algorithm increased by adding ontology, ontology matching is a solution to the semantic heterogeneity problem.

1. Introduction

Official statistics are statistics published by government agencies or other public bodies such as international organisations. These statistics provide quantitative and qualitative information on all major areas of citizens' lives, such as economic and social development, living conditions, health, education and the environment. Official statistics can be found on web sites of national statistical agencies such as the Palestinian central bureau of statistics (PCBS) [3]. "Official Statistics" are the data which are collected and disseminated by a set of governmental and international organizations to provide the factual basis for making policy

and supporting research. Therefore, there is a need for a common schema to integrate that massive data and visualize the findings, so that viewers can easily derive an insight into data. The objective of the integration is to perform the data harmonization acquired from different sources and files, mapping algorithm is used to map indicators.

Due to the increasing complexity and heterogeneity of statistical data, an increasing need for sophisticated visualization technology and integration arises, We introduce a new interoperable visual analytics framework to enhance integration, mapping, dissemination and presentation of official statistics based on a mapping algorithm that uses hamming distance, edit distance and ontology. Information Visualization has an important role in different contexts. In fact, it has been used in different fields and it is an expanding area of knowledge [4].

1.1 Objectives

The objective of this study is to introduce a new interoperable visual analytics framework for:

- Mapping, grouping, and integrating heterogeneous data and statistical indicators into a common schema.
- Mapping indicators by building mapping algorithm using hamming distance, edit distance and ontology.
- Enhancing presentation of official statistics based on visual analytical approach that combines both data analysis and interactive visualization.

1.2 String Comparator Metrics

When comparing values of string variables like names or addresses, it usually does not make sense to just discern total agreement and disagreement. Typographical error may lead to many incorrect disagreements. Several methods for dealing with this problem have been developed: string comparators are mappings from a pair of strings to the interval [0, 1] measuring the degree of compliance of the compared strings [5]. String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage, discriminate analysis or logistic regression. The simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator.

1.3 Hamming Distance

One of the earliest and most natural metrics is the hamming distance [1], where the distance between two strings is the number of mismatching characters. in information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. In another way, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other.

For instance, the Hamming distance between "toned" and "roses" is 3, between “1011101” and “1001001” is 2, and between “2173896” and “2233796” is 3. For a fixed length n, the Hamming distance is a metric on the vector space of the words of length n, as it fulfills the conditions of non-negativity, identity of indiscernible and symmetry, and it can be shown by complete induction that it satisfies the triangle inequality as well. For instance, the Hamming distance between two words "a" and "b" can also be seen as the Hamming weight of "a−b" for an appropriate choice of the − operator.

1.4 Edit (Levenshtein) Distance

Edit distance [2] is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. Edit distances find applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that have a low distance to the word in question. In bioinformatics, it can be used to quantify the similarity of macromolecules such as DNA, which can be viewed as strings of the letters A, C, G and T.

To compute the edit distance ed(x,y) between strings x and y, a matrix $M_{1...m+1,1...n+1}$ is constructed where $M_{i,j}$ is the minimum number of edit operations needed to match $x_{1...i}$ to $y_{1...j}$. Each matrix element $M_{i,j}$ is calculated as per Equation 1, where $\delta(a,b) = 0$ if a = b and 1 otherwise. The matrix element $M_{1,1}$ is the edit distance between two empty strings.

$$M_{i,j} \leftarrow \min \begin{cases} M_{1,1<0} \\ M_{j-1,j} + 1 \\ M_{i,j-1} + 1 \\ M_{i-1,j-1} + \delta(x_i, y_i) \end{cases}$$

Equation 1: Edit distance ed(x,y) between strings x and y.

The algorithm considers the last characters, x_i and y_j. If they are equal, then $x_{1..i}$ can be converted into $y_{1..j}$ at a cost of $M_{i-1,j-1}$. If they are not equal, xi can be converted to y_j by substitution at a cost of $M_{i-1,j-1} + 1$, or xi can be deleted at a cost of $M_{i-1,j} + 1$ or y_j can be appended to x at a cost of $M_{i,j-1} + 1$. The minimum edit distance between x and y is given by the matrix entry at position $M_{m+1,n+1}$.

Table (1) is an example of the matrix produced to calculate the edit distance between the strings "DFGDGBDEGGAB" and "DGGGDGBDEFGAB". The edit distance between these strings given as Mm+1,n+1 is 3.

Table 1: Edit distance matrix for the strings "DFGDGBDEGGAB" and "DGGGDGBDEFGAB" with the minimum edit distance position highlighted.

		D	G	G	G	D	G	B	D	E	F	G	A	B
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
D	1	0	1	2	3	4	5	6	7	8	9	10	11	12
F	2	1	1	2	3	4	5	6	7	8	8	9	10	11
G	3	2	1	1	2	3	4	5	6	7	8	8	9	10
D	4	3	2	2	2	2	3	4	5	6	7	8	9	10
G	5	4	3	2	2	3	2	3	4	5	6	7	8	9
B	6	5	4	3	3	3	3	2	3	4	5	6	7	8
D	7	6	5	4	4	3	4	3	2	3	4	5	6	7
E	8	7	6	5	5	4	4	4	3	2	3	4	5	6
G	9	8	7	6	5	5	4	5	4	3	3	3	4	5
G	10	9	8	7	6	6	5	5	5	4	4	3	4	5
A	11	10	9	8	7	7	6	6	6	5	5	4	3	4
B	12	11	10	9	8	8	7	6	7	6	6	5	4	**3**

1.5 Ontology

Ontology matching is a solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of ontologies. These correspondences can be used for various tasks, such as ontology merging, query answering, or data translation. Thus, matching ontologies enables the knowledge and data expressed with respect to the matched ontologies to interoperate [6]. Diverse solutions for matching have been proposed in the last decades [7, 8]. Several recent surveys [9–10] and books [6, 11] have been written on the topic as well.

An ontology typically provides a vocabulary that describes a domain of interest and a specification of the meaning of terms used in the vocabulary. Depending on the precision of this specification, the notion of ontology encompasses several data and conceptual models, including, sets of terms, classifications, thesauri, database schemas, or fully axiomatized

theories [6]. Ontologies tend to be put everywhere. They are viewed as the silver bullet for many applications, such as information integration, peer-to-peer systems, electronic commerce, semantic web services, social networks, and so on. They, indeed, are a practical means to conceptualize what is expressed in a computer format. However, in open or evolving systems, such as the semantic web, different parties would, in general, adopt different ontologies. Thus, just using ontologies, like just using XML, does not reduce heterogeneity: it raises heterogeneity problems at a higher level.

2. Related Work

Many studies have been done worldwide on data integration and data visualization. Applications of data integration and visualization were used in several sectors, especially in Transportation, Statistics, Scientific research, Digital libraries, financial data analysis, and Market studies.

Michaela Denk and Peter Hackl, 2004 [12] was develop a project of micro-founded indicators. It aimed at (i) assembling a wide ranging system of statistical information including data from economic, tax and social insurance sources into an integrated multi-source enterprise database, and (ii) creating micro-simulation models for enterprise taxation in two European countries, Italy and the UK, with a view to eventually producing an "EU demonstrator" as a foundation for the development of similar models in the whole EU. For the creation of such a multi-source database of enterprise data as a basis of micro simulations, data integration, mainly record matching, was a core issue of the project. Michaela Denk and Peter Hackl, 2004 [12] showed the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making.

Filippo Oropallo and Francesca Inglese [13] addressed the integration problems that have been faced in reconciling administrative and survey sources and combining them into one multi-source database. they showed the architecture of the integration process that has been adopted and the exploitation of the integrated database for economic and policy impact analysis at a micro level. The integration of administrative and survey data was performed by exact matching when the same unit was identified otherwise it was performed by statistical matching techniques. To apply these techniques, matching variables were required: one quite apparent option was to use firm characteristics as provided by the business register. The development of the Enterprise Integrated and Systematized Information System (EISIS) opens new possibility in micro simulation analysis to study the tax burden and the economic

performance of enterprises through the construction of micro-founded indicators. IT (Information Technology) features of the whole process were also described that were the formalization of the integration process and the structure of the user friendly interface of the integration software. Confidentiality was satisfied by remote processing on a protected server that was only accessible to granted users of the National Statistical Institute.

3. Methodology

To achieve the objectives of this research and build the Visual Analytics Framework, we collected statistical data indicators (MDGS) from different surveys and spreadsheets. The original data and indicators are included in heterogeneous sources and files. We cover the indicators coming from Palestinian Central Bureau of Statistics (PCBS) [3], Department of Statistics (Jordan) [14] and from Central Agency for Public Mobilization and Statistics (Egypt) [15]. The goal of selecting indicators from different countries is to test the accuracy of our mapping algorithm and integration of data during the import process of these indicators based on our schema.

3.1 Data Analysis

Statistical data are sets of often numeric observations which typically have time associated with them, see Fig. (1). They are associated with a set of metadata values, representing specific Concepts, which act as identifiers and descriptors of the data. These metadata values and Concepts can be understood as the named Dimensions of a multi-dimensional co-ordinate system, describing what is often called a 'cube' of data.

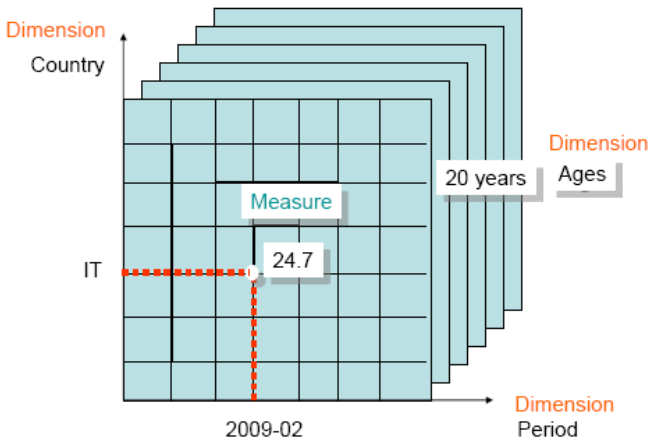

Figure 1: Multidimensional 'Cube' of data

After defining and preparing indicators, we define the unit for each indicator and associate each indicator with the correct unit, then we define the subgroups for each indicator. A subgroup is a subset within a sample or population identified by some common dimension such as sex, age or location.

Subgroup dimensions refer to broad subgroup categories such as sex, location, age. Under each subgroup dimension come various subgroup dimension values. For example, for the subgroup dimension "Sex", the subgroup dimension values are "Male" and "Female". Finally, subgroups consist of a combination of one or more subgroup dimension values, such as "Male 5-9 yr Urban". Table (2) below gives several examples of these subgroups.

Table 2: The subgroup dimension values for the subgroup diminution "Sex" are "Male" and "Female".

Subgroup dimensions	Subgroup dimension values	Subgroups
Sex	Male, Female	Male Female Urban Rural Male Urban Female Urban Male Rural Female Rural Male Urban 0-4 yr Female Urban 0-4 yr Male Rural 0-4 yr Female Rural 0-4 yr
Age	0-4 yr, 5-9 yr, 10-14 yr	
Location	Urban, Rural, Total	

3.2 Entity-Relationship (ER) diagram and Database Schema

We built the entity relationship diagram (Conceptual data model) which is a graphical representation of entities and their relationships to each other, Fig (2) shows the ER Diagram that we built to organize the data within the database, the ER Diagram shows the relationships between all statistical data entities and display the attributes for each entity, the attributes with underline are the primary keys, and the attributes with Dashed line are foreign keys, the entities of our ER Diagram are: Area, Indicators, AreasIndicators, IndicatorOntology, Sectors, Sectorontologys, Units, UnitsOntology, SubGroups, SubGroupsOntology, subGroupsIndicators, Classes, ClassesOntology. We described these entities farther using attributes, as an example indicators entity contains Indicator_ID (primary key), Sector_ID (foreign key), Indicator_Name, Unit_ID (foreign key) as attributes, the relationships between entities represented in the diagram, Indicators entity has many-to-one relationship with

Sectors entity, many-to-one with Units entity, one-to-many relationship with IndicatorOntologys, one-to-many with data entity, and many-to-many relationship with SubGroups, we added two one-to-many relationships, one-to-many between Indicators entity and SubGroupsIndicators, and one-to-many relationship between SubGroups and SubGroupsIndicators, Indicators entity has also many-to-many relationship with Areas entity, we added two one-to-many relationships, one between Indicators entity and Areasindicators, and the other relationship between Areas entity and AreasIndicators. Units entity has one-to-many relationship with UnitsOntology, Sectors entity has one-to-many relationship with SectorOntologys entity, SubGroups entity has one-to-many relationship with SubGroupsOntology, and Classes entity has one-to-many relationship with ClassesOntology.

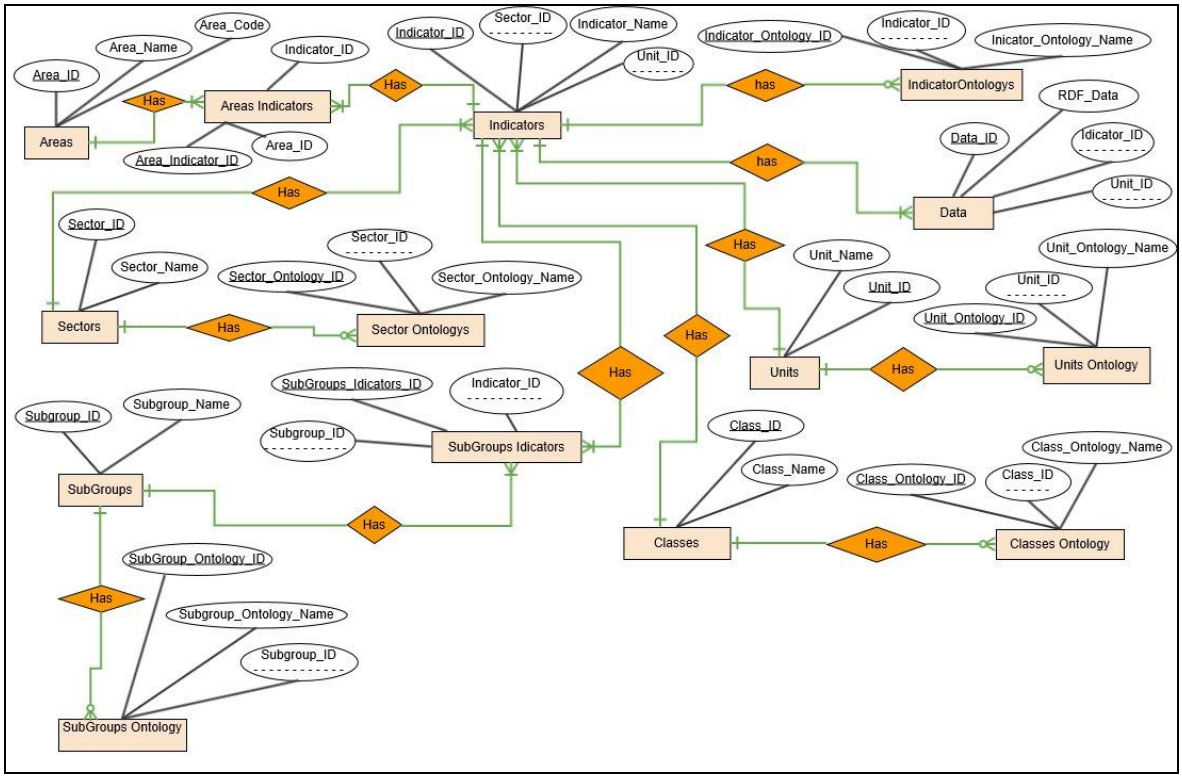

Figure 2: Entity Relationship Diagram.

Depending on the entity relationship diagram we created SQL Server database Schema (logical design), our schema consist of indicators table, Units table, Subgroups table, Sectors table, Areas table, and classes table. The definition of each Indicator ,Unit, Subgroup, Sector, Area and Class entered to our schema tables. also our schema contains ontology lookup tables for all the schema tables (IndicatorsOntology, UnitsOntology, SubGroupsOntology, SectorsOntology, and ClassOntology). These ontology's tables will help us to increase the algorithm accuracy mapping during importing process of indicators and data to our schema from different sources, since the indicators, or units maybe not the same matching but with the same meaning.

3.3 Data Mapping Algorithm

After building our schema, we build mapping algorithm in C# using hamming distance and edit (Levenshtein) distance and by adding ontology to our algorithm also, edit distance can be considered a generalization of the Hamming distance, which is used for strings of the same length and only considers substitution edits. Fig(3) shows the summery steps of mapping algorithm. Also we build GUI to import the indicators with data values from different sources.

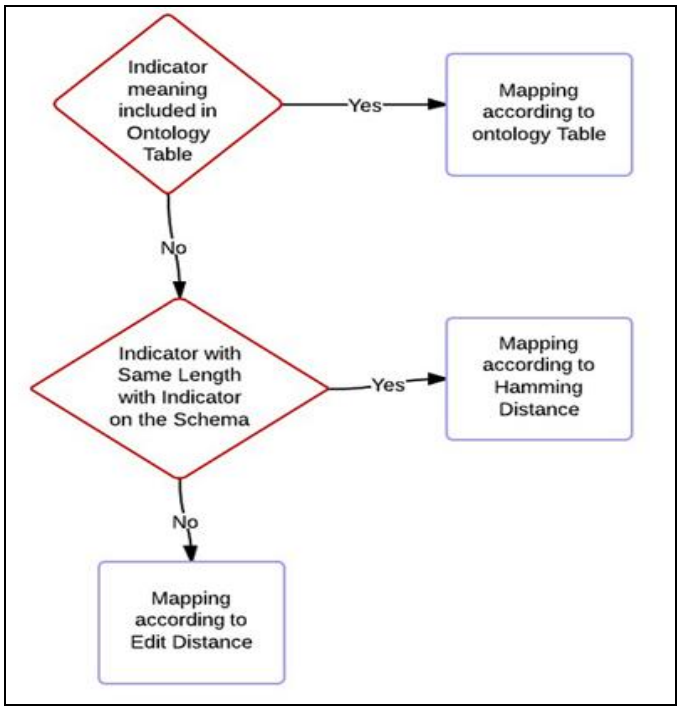

Figure 3: Summery Steps of Mapping Algorithm.

3.4 Indicators Mapping without Ontology

Using our algorithm we started to import the data and indicators from different sources files to our target schema which contains the indicators, Units and Subgroups. during the data import using our algorithm, the exact matched indicators, units and subgroups will be mapped automatically to the indicators, units, and subgroups in the schema, in case that we import not exact matched indicator, units or subgroups the algorithm will calculate the edit distance (minimum operations needed) for mapping the imported indicator with the nearest indicator in the schema, the same thing will happen for units or subgroups, the algorithm will calculate the nearest indicator, unit and subgroups from the schema for unmatched indicators, units and subgroups, Fig. (4) shows that when we import "Growth rate of GDP/person employed" indicator from one of our sources to the schema as an example, the algorithm try to find the exact mapping first, if there are no exact matching the algorithm will calculate the nearest matching indicator according to the minimum edit distance, in this case as shown in the Fig.

(4) below the imported indicator matched to "Growth rate of GDP per person employed" indicator, that's true mapping and the distance as shown between the two indicators is 3 with accuracy 92% between the two indicators. We calculated the accuracy in percent using the formula: percent = (largerString.Length - editDistance) / largerString.Length)*100.

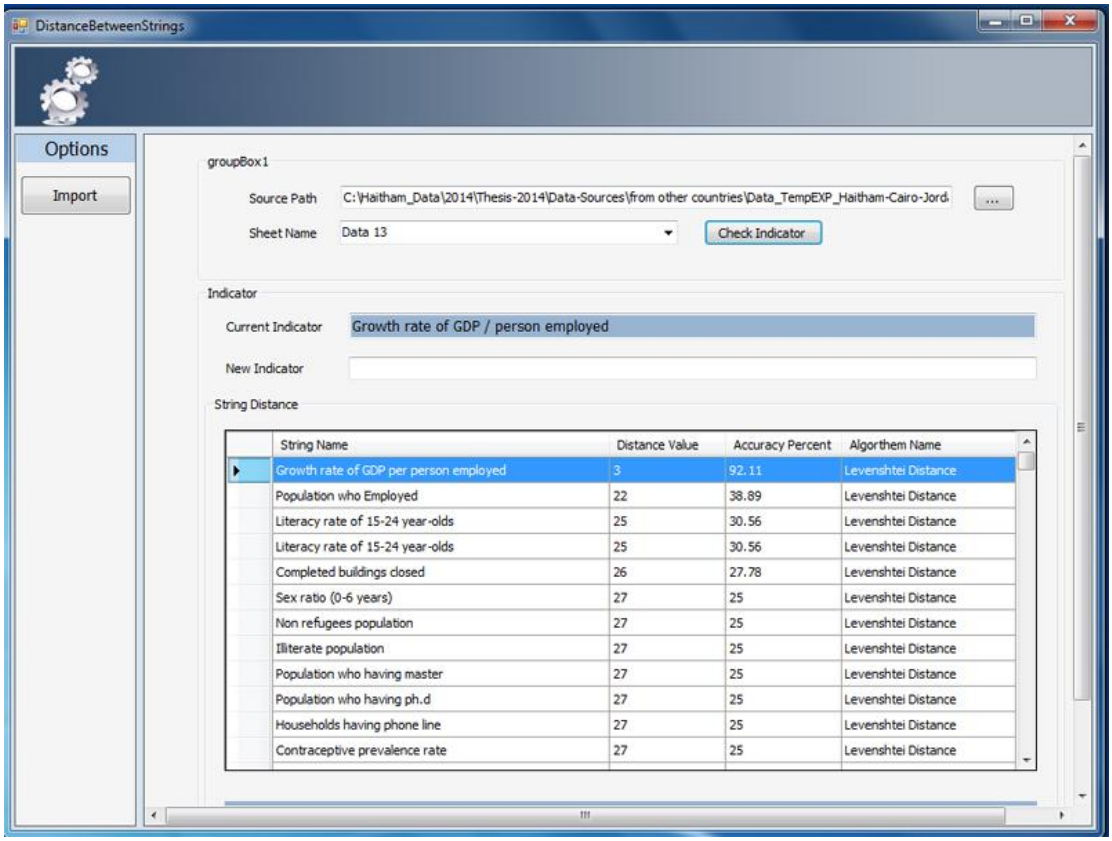

Figure 4: Example of Importing and Mapping "Growth rate of GDP /person employed" Indicator.

3.4.1 Indicators Mapping with Ontology

As an example if we import source file with "urbanization level" indicator to our schema, the algorithm find the exact matching for this indicator from the schema, if not exist it will calculate the minimum edit distance and nearest indicator using edit distance to match the Imported indicator according to the minimum distance, Fig.(5) illustrate that when we import "urbanization level" indicator the nearest indicator to this indicator is "Population Size" indicator, this is false matching since the two indicators not the same.

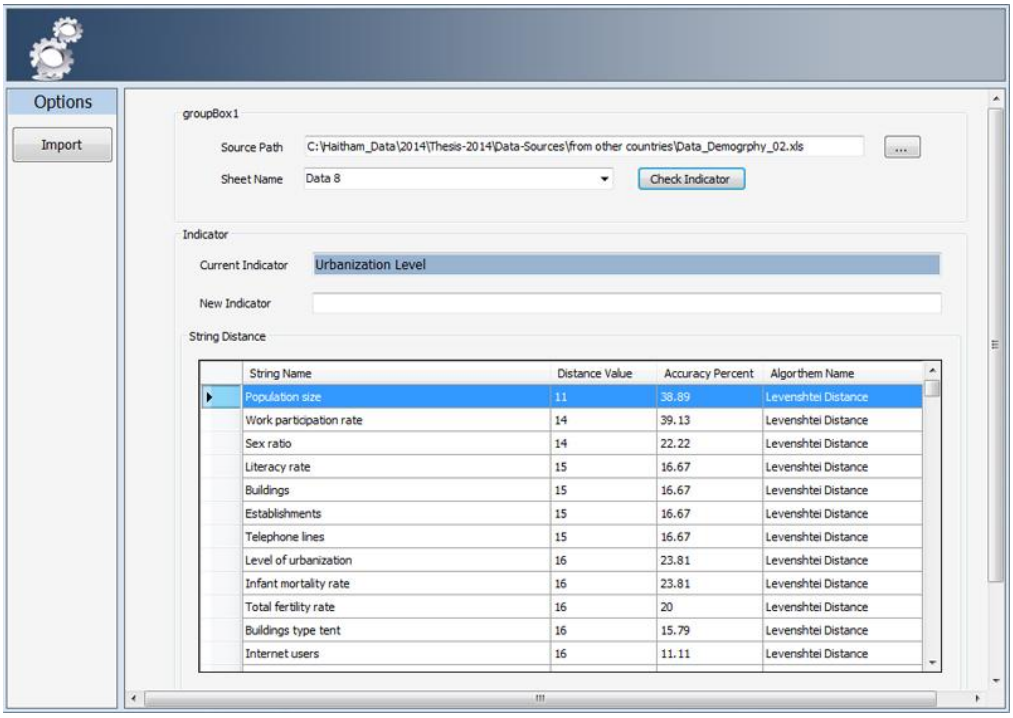

Figure 5: Example of Importing and Mapping "Urbanization Level" Indicator without Ontology.

To solve this issue we added ontology lookup tables to our schema to increase the accuracy of mapping, in this case when using ontology, our algorithm will check first the ontology lookup table for indicators and it will return the ontology matched indicator from the ontology table and will return the true ontology mapping, in this case the "urbanization level" indicator will mapped to "level of urbanization" indicator from ontology table as shown in Fig.(6).

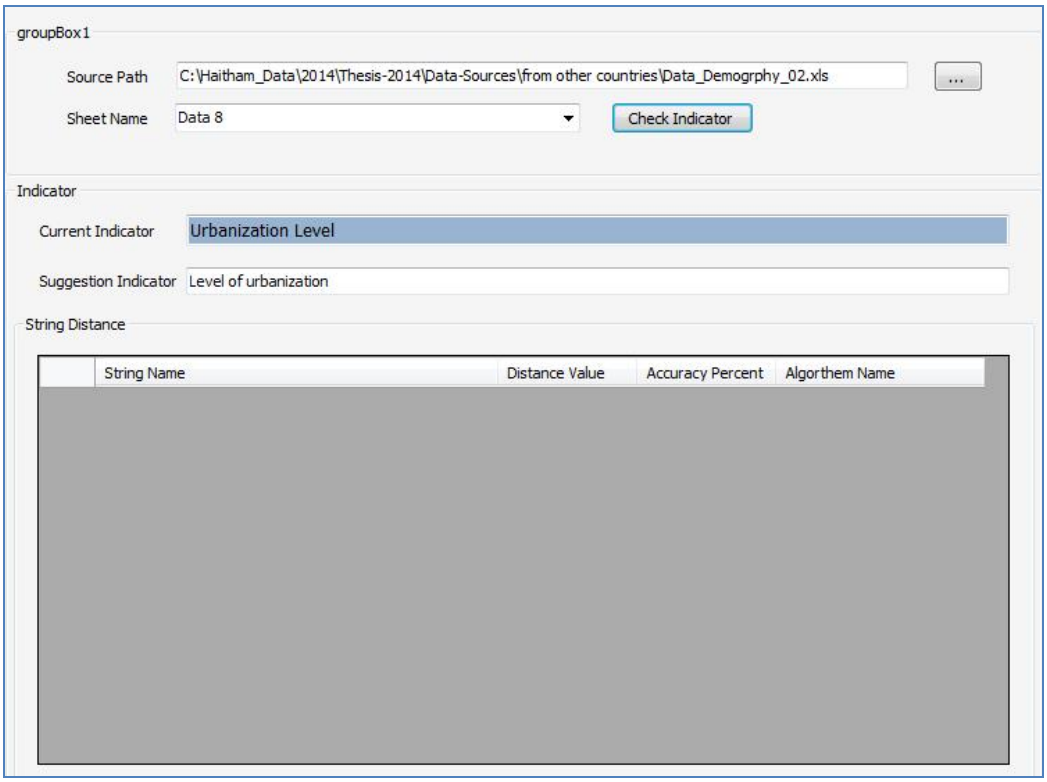

Figure 6: Example of Importing and Mapping "Urbanization Level" Indicator with Ontology.

3.5 Experimental Results

To test and evaluate the accuracy of the mapping algorithm in practice, we performed some experiments on many indicators, the indicators chosen from different countries since each country indicators different from others in the name of the indicators, units and subgroups, this will help us to test the algorithm accuracy.

3.5.1 Mapping Results without Ontology

We test the algorithm without using ontology, by importing different indicators and their units, sectors, and subgroups. the results shown that the accuracy of the algorithm is 67% as shown in Fig.(7) since there was indicators with false mapping, but we can increase the accuracy of the algorithm by decreasing the false mapping when using ontology as we will see in the next section. when we import units of the indicators, the algorithm return the best and nearest mapping for each unit according to the edit distance and hamming distance for each unit with units in our schema, as the results shown the false mapping for some units since the different in writing the unit with same meaning, as an example "Percentage" unit mapped to "percent" unit with minimum edit distance equal 3 and this mapping is true, but "%" unit mapped to "US$" unit with minimum edit distance equal 3 and this mapping is false, since "%" unit means "percent", also "years" unit exact mapping with "years" unit from the schema, but importing "yr." unit mapped to "US$" unit, this mapping false since "yr." unit mean "years" unit, because of that we added ontology to our algorithm as we will see in the next section.

Fig.(7) shows that the algorithm accuracy for units mapping without ontology is 82%, when we import subgroups of the indicators, the algorithm calculate the best and nearest mapping for each subgroup according to our algorithm for each subgroup with subgroups in our schema, Fig.(7) shows the accuracy of the algorithm for importing subgroups is 78%, the accuracy can be increased by using ontology, this will be discussed in details in the next section.

Fig.(7) summarize the accuracy of the algorithm for importing and mapping indicators, units and subgroups depending on hamming distance and edit distance without using ontology, algorithm accuracy for mapping indicators 67%, accuracy for mapping units of the indicators is 82% and the accuracy of the algorithm for mapping subgroups of the indicators is 78%.

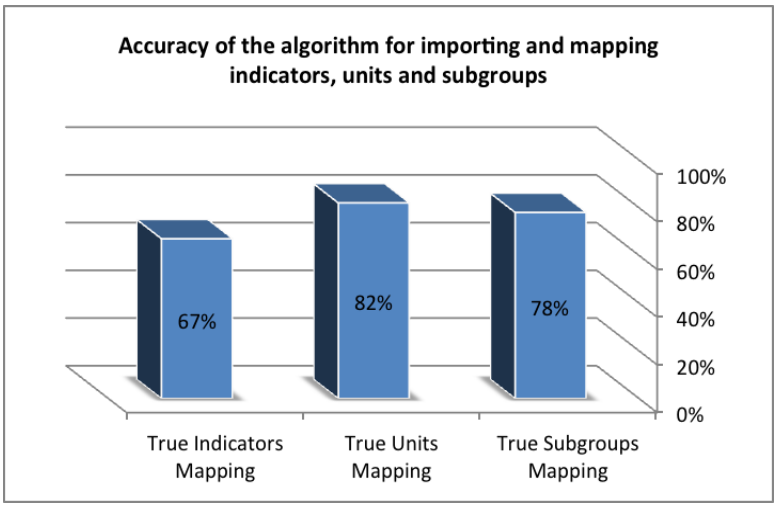

Figure 7: Algorithm Indicators, Units and Subgroups Mapping Accuracy without Ontology.

3.5.2 Mapping Results with Ontology

We improved the accuracy of our algorithm by adding ontology implementation to the algorithm code, and we added some indicators, units and subgroups meaning in ontology tables inside the schema, then we test the algorithm using ontology in addition to the hamming and edit distance implementation by importing different indicators, units, and subgroups.

when we import indicators, the algorithm locking for the meaning terms in the indicators ontology table inside the schema to return the meaning of the imported indicator, if it is included in the meaning terms and ontology table, the indicator will be mapped, if not the algorithm will return the best and nearest mapping for each indicator according to the edit distance for each indicator with indicators in the schema, Fig.(8) shows the accuracy of the algorithm using ontology is 89%. we can see that the accuracy increased by using ontology comparing with the accuracy of the algorithm without using ontology, it was 67% as shown in the previous results in Fig(7).

when we import units of the indicators, the algorithm locking for the meaning of units in the terms inside unit ontology table in the schema to return the meaning of the imported unit, if it is included in the meaning terms, the unit will be mapped, if not the algorithm will return the best and nearest mapping for each unit according to the edit distance for each unit with units in the schema, As an example importing "years" unit exact mapping with "years" unit from the schema, importing "yr." unit mapped to "years" unit by using ontology and return the mapping from unit ontology table, also "%" unit mapped to "percent" by using ontology. The

results showed that the accuracy of the algorithm with ontology higher than the accuracy without ontology.

Fig.(8) shows that the accuracy of the algorithm for units mapping with ontology is 95%, we can see that the accuracy increased by using ontology comparing with the accuracy of the algorithm for units mapping without using ontology it was 82% as shown in the previous results in Fig.(7).

when we import subgroups of the indicators, by using ontology the algorithm check first the ontology table of subgroups, as an example when we import "F" subgroup which means "Female" in the subgroups ontology table, the "F" subgroup mapped to "Female" subgroup, "F" subgroup was mapped to "male" subgroup without ontology, and importing "One yr" subgroup mapped to "1 yr" using ontology since we have "one yr" which means "1 yr" in subgroups ontology table. But "one yr" mapped to "<5 yr" without ontology since the nearest unit according to edit distance to "one yr" was "<5 yr" without ontology. the accuracy of the algorithm using ontology for subgroups is higher than the accuracy of the algorithm without ontology, Fig.(8) shows the accuracy of the algorithm this time is 100% comparing with the accuracy of the algorithm for mapping subgroups without ontology as shown in Fig.(7) the accuracy was 78%.

Fig.(8) summarize the accuracy of the algorithm for importing and mapping indicators, units and subgroups depending on ontology in addition to hamming distance and edit distance, algorithm accuracy for mapping indicators 89%, accuracy for mapping units of the indicators is 95% and the accuracy of the algorithm for mapping subgroups of the indicators is 100%.

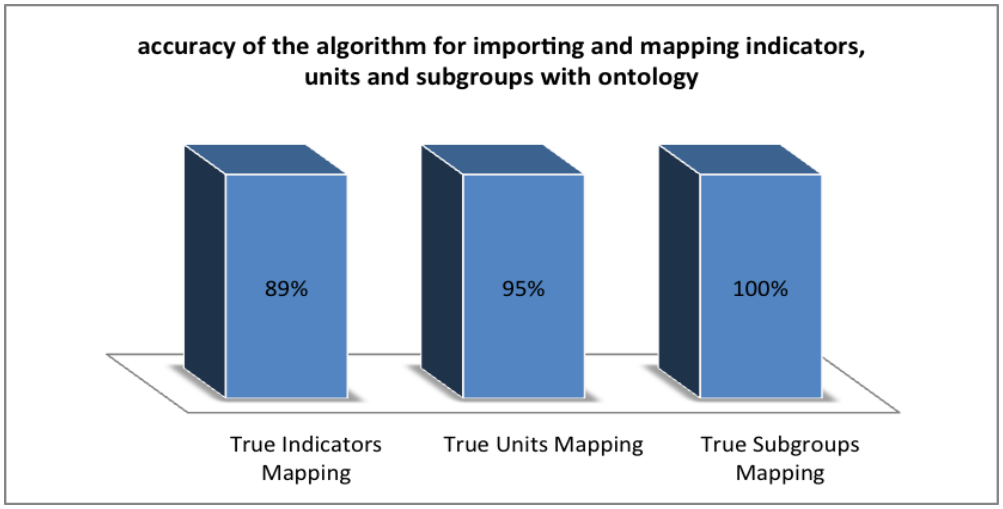

Figure 8: Algorithm Indicators, Units and Subgroups Mapping Accuracy with Ontology.

Fig.(9) summarize the accuracy of the algorithm according to our results without ontology and with ontology for importing indicators, units and subgroups. and shows that the accuracy of the algorithm when importing and mapping indicators without ontology is 67% and with ontology is 89%, for units mapping it is 82% without ontology and 95% with ontology, and for subgroups it is 78% without ontology and 100% with ontology. In general we can conclude that adding the ontology to our algorithm in addition to using of hamming distance and edit distance improved the algorithm accuracy for mapping indicators, units and subgroups.

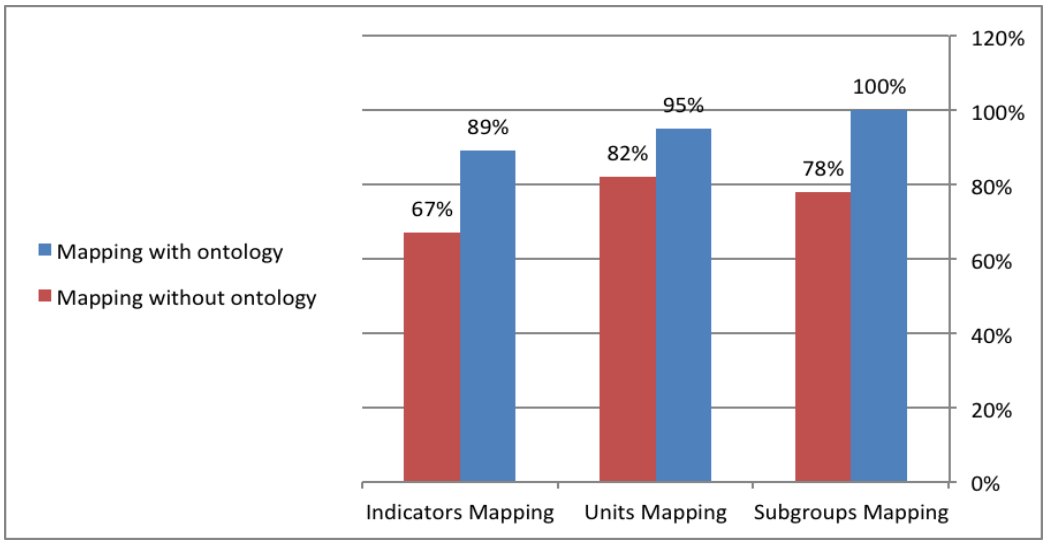

Figure 9: Algorithm Indicators, Units and Subgroups Mapping Accuracy with Ontology and without Ontology.

4. Conclusion and Future Work

This research aimed to introduce a new interoperable visual analytics framework for Collecting, mapping, processing, and disseminating statistical data based on common schema, heterogeneous data from different data sources integrated using the created algorithm, we suggested new mapping algorithm based on hamming distance, edit distance and ontology, using our algorithm we enhanced integration and mapping of statistical data indicators from different sources, the data after importing saved in the schema that we created, the schema included ontology tables to improve and increase the accuracy of the mapping algorithm. We tested the accuracy of the algorithm, experimental results shown high accuracy of mapping for the algorithm by adding the ontology to the algorithm. the accuracy of the algorithm when importing and mapping indicators without ontology was 67% and with ontology the accuracy was 89%, for units mapping the accuracy was 82% without ontology and 95% with ontology, and for subgroups the accuracy was 78% without ontology and 100% with ontology. In

general we can conclude that adding the ontology to our algorithm in addition to using of hamming distance and edit distance improved the algorithm accuracy for mapping indicators, units and subgroups.

Future work includes focus more on data mapping using ontology. our main line of future research involves extending our mapping algorithm to handle more sophisticated mappings between ontologies (i.e., non 1-1 mappings), also to focus more on collecting data from different sources since we focused as a case study on importing data from different excel sources (files), future work includes also improving collaboration with visual analytics framework. Additional methods are required to support the users in finding good views on the data and in determining appropriate visualization techniques. we have to consider the 3D visualization of uncertain graph structures with uncertain attributes, which we think is a formidable challenge.

5. References

[1] Hamming Distance: http://en.wikipedia.org/wiki/Hamming_distance

[2] Levenstein, V. (1966), Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Sov. Phys. Dokl. 10:707-710.

[3] Palestinian Central Bureau of Statistics: http://www.pcbs.gov.ps

[4] Card, S.K., Mackinlay, J.D., and Shneiderman (eds), B. (1999), Readings in Information Visualization, Morgan Kaufmann Publishers.

[5] Winkler, W.E. (1990), String Comparator Metrics and Enhanced Decision Rules in the Fellegi- Sunter Model of Record Linkage. In Proc. Section on Survey Research Methods, American Statistical Association, pages 354-359.

[6] Euzenat, J. and Shvaiko, P. (2007), Ontology matching, Springer.

[7] Batini, C., Lenzerini, M., and Navathe, S. (1986), "A comparative analysis of methodologies for database schema integration," ACM Computing Surveys, vol. 18, no. 4, pp. 323–364.

[8] Spaccapietra, S. and Parent, C. (1991), "Conflicts and correspondence assertions in interoperable databases," SIGMOD Record, vol. 20, no. 4, pp. 49–54.

[9] Rahm, E. and Bernstein, P. (2001), "A survey of approaches to automatic schema matching," The VLDB Journal, vol. 10, no. 4, pp. 334–350.

[10] Gal, A. and Shvaiko, P. (2009), "Advances in ontology matching," in Advances in Web Semantics I, T. S. Dillon, E. Chang, R. Meersman, and K. Sycara, Eds. Springer, pp. 176–198.

[11] Bellahsene, Z., Bonifati, A., and Rahm, E. (2011), Schema Matching and Mapping. Springer.

[12] Michaela, D. and Peter, H. (2004), Data Integration: Techniques and Evaluation, The DIECOFIS Project: Progress and Lessons. Austrian Journal of Statistics.

[13] Filippo, O. and Francesca, I. (2004), The Development of an Integrated and Systematized Information System for Economic and Policy Impact Analysis. Austrian Journal of Statistics, Volume 33 (2004), Number 1&2, 211-235.

[14] Department of Statistics (Jordan): http://www.dos.gov.jo.

[15] Central Agency for Public Mobilization and Statistics (Egypt): http://capmas.gov.eg.