**Deanship of Graduate Studies**
**Al-Quds University**

# Mapping Techniques and Visualization of Statistical Indicators

**Haitham H. Zeidan**

**M.Sc. Thesis**

**Jerusalem-Palestine**

**1435 AH / 2014 AD**

# Mapping Techniques
# and
# Visualization of Statistical Indicators

**Prepared By:**
**Haitham H. Zeidan**

**B.Sc.: Computer Science - Al-Quds University - Palestine**

**Supervisors: Dr. Jad Najjar and Dr. Rashid Jayousi**

**A thesis submitted in partial fulfillment of requirements for the Master's Degree in Computer Science / Department of Computer Science / Faculty of Graduate Studies – Al-Quds University.**

**1435 AH / 2014 AD**

**Al-Quds University**
**Deanship of Graduate Studies**
**Computer Science Department**

# Thesis Approval

# Mapping Techniques and Visualization of Statistical Indicators

**Prepared By: Haitham H. Zeidan**
**Registration No: 21010324**

**Supervisors: Dr. Jad Najjar and Dr. Rashid Jayousi**

Master thesis submitted and accepted on: / /2014
Examining Committee Members (Name & Signature):

| | |
|---|---|
| 1- Head of Committee: Dr. Jad Najjar | Signature: ……..….. |
| 2- Co-Supervisor: Dr. Rashid Jayousi | Signature: ……..….. |
| 2- Internal Examiner: Dr. Badie Sartawi | Signature: ……..…….. |
| 3- External Examiner: Dr. Mohammad Matar | Signature: ……..…….. |

Jerusalem – Palestine

**1435 AH/ 2014 AD**

# Dedication

This work is dedicated…

To my parents for their love, endless support and encouragement…

To my beloved wife, without her caring support it would not have been possible…

To my children: Baraa', Lamar and Noor Al-Deen

To my brothers, sisters, friends and colleagues…

To all of you I say a big

“Thank you” for being example of love and care.

Haitham H. Zeidan

## Declaration

I certify that this thesis submitted for the Master's Degree, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of it) has not been submitted for a higher degree to any other university or institution.

Signed……..

Haitham H. Zeidan

Date: / /2014

## Acknowledgments

First and foremost Praise be to the Almighty Allah, Lord of all creatures, the Most Gracious and Most Merciful, for all the blessings I have experienced throughout my life.

My sincere thanks for my Supervisors: Dr. Jad Najjar and Dr. Rashid Jayousi, for their sincere efforts, interest and time they have spent to guide my research.

Thanks also to the Examining Committee Members: Dr. Jad Najjar, Dr. Rashid Jayousi, Dr. Badie Sartawi and Dr. Mohammad Matar.

I am very grateful to all professors at Al-Quds University, Computer Science Department, for their support and dedication.

I am extremely grateful and indebted to my colleagues in the Palestinian Central Bureau of Statistics (PCBS), for helping me during the study in preparing the statistical data.

My thanks are extended to all those, who have participated in the evaluation process. Thank you for the dedicated time.

I thank my mother for her love and support, my father for the encouragement and motivation, and other members of my family for being there.

Finally, and most importantly, I would like to thank my wife. Her support, encouragement, quiet patience and unwavering love were undeniable.

## Abstract

The aim of this research is to introduce a new mapping algorithm and visualization system towards enhancing mapping and integration of statistical indicators. This study aims to investigate how data integration and data visualization could be used to increase readability and interoperability of statistical data. Statistical data has gained a growing interest from policy makers, city planners, researchers and ordinary citizens as well. From an official statistical point of view, data integration is of major importance as a means of using available information more efficiently that improve the quality of the products of any National Statistical Office. By using integration methods, the value added that can be extracted from the existing stock of information is greatly augmented. In this case, data integration is a way of generating a comprehensive statistical database as a sound foundation for deliberate decision making. We implemented and proposed statistical indicators schema and mapping algorithm which is conceptually simple and is based on hamming distance [47], edit (Levenshtein) distance [48] and ontology. Also we build graphical user interface (GUI) to import indicators with data values from different sources. The performance and accuracy of this algorithm was measured by experiment, we performed experiments on many indicators chosen from different countries. We started to import the data and indicators from different sources to our target schema which contains indicators, units and subgroups. During data importing using the proposed algorithm, the exact matched indicators, units and subgroups will be mapped automatically to the indicators, units and subgroups in the schema. In case of no exact matching 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. The data and indicators were stored in the data table of our schema as Resource Description Framework (RDF), this helped us to visualize statistical indicators using visualization techniques. Evaluation for the visualization system was also done by interviews to analyze user requirements. Each user has his/her own concerns and expectations from the system. In order to identify user needs, interviews with expert users have been done. The first interview aimed to find users' requirements and discover probable problems that users face without visualizing the results, while the second interview aimed to evaluate our final system. The evaluation showed high levels of effectiveness and efficiency of the system. Moreover, majority of participants in the evaluation were satisfied with the system, because it helped them to accomplish all tasks successfully.

Our recommendations for official statistics are to focus on data integration to produce new statistics and to maximize the benefit of the existing statistics and also enable greater level of research. This could enhance official statistics by increasing knowledge on characteristics of the society, economy and environment, as data integration could offer a less time consuming and less costly alternative, although it does still require a significant level of time and resources, also data integration has an additional advantage when it comes to reducing respondents' burden by making use in a more effective manner of the existing data sources. To increase the quality of data and accuracy of data integration a very careful review of the statistical data indicators from official statistics should be carried out for each of the common variables to be used in matching process. Often, even if similar, we could not establish exact equivalence of concepts and slight differences could lead to large discrepancies in the data.

## تقنيات التكامل والتصور البصري المرئي للمؤشرات الاحصائية

**إعداد: هيثم حسني زيدان جوابرة**

**إشراف: د. جاد النجار ، د. رشيد الجيوسي**

**ملخص:**

أصبحت البيانات والمؤشرات الاحصائية تسترعي اهتمام العديد من صانعي السياسات والإحصائيين والباحثين والمواطنين العاديين، ومن هذا المنطلق، جاءت هذه الدراسة لبناء خوارزمية ونظام بصري لتوظيفه في تكامل البيانات الاحصائية لتحسين جودة البيانات وزيادة ثقة المستخدمين لها، وكذلك لزيادة التفاعل والفهم لهذه البيانات.

تعتبر المكاتب الإحصائية الرسمية قضايا تكامل البيانات وتحسين جودة المؤشرات التي توفرها تلك البيانات غاية مهمة لاستخدام تلك البيانات والمؤشرات بشكل أكثر كفاءة، ولأن تكامل البيانات تعتبر وسيلة فاعلة لتوليد قاعدة بيانات إحصائية شاملة كأساس لاتخاذ القرارات السليمة، فقد حاولت هذه الدراسة بناء خوارزمية تستند على Hamming Distance and Edit Distance، وتم أيضا إضافة المعاني والدلالات لبعض المؤشرات الى قاعدة البيانات النهائية؛ لضمان تكامل البيانات والمؤشرات الاحصائية وزيادة دقة الخوارزمية في عملية تكامل البيانات والمؤشرات.

ومن خلال التجربة العملية تم فحص دقة الخوارزمية التي تم بناؤها وذلك بإدخال عدة مؤشرات الى قاعدة البيانات بالاعتماد على الخوارزمية، ومن خلال إدخال المؤشرات أظهرت النتائج أن دقة وكفاءة الخوارزمية تحسنت بإضافة الدلالات والمعاني الى قاعدة البيانات، وفي مرحلة لاحقة لإدخال وتكامل المؤشرات تم تخزينها داخل قاعدة البيانات، ومن ثم تم استخدام تقنيات بصرية تفاعلية لعرض المؤشرات.

ولقياس درجة رضا المستخدمين عن النظام المقترح، تم اجراء مقابلات مع المستخدمين لتحديد احتياجاتهم وتم اجراء مقابلات مع المستخدمين مرة أخرى بعد بناء النظام البصري لتقييم النظام من ناحية الكفاءة والفعالية ومدى تلبية النظام لاحتياجات المستخدمين، وقد أظهرت نتائج المقابلات كفاءة وفعالية النظام من وجهة نظر الميتخدمين.

ولعل من اهم التوصيات التي خرجت بها الدراسة للمراكز الاحصائية هي التركيز على تكامل البيانات؛ لإنتاج مؤشرات جديدة متكاملة، وذات جودة عالية، وتعزيز قيمة البيانات والمؤشرات الموجودة من خلال التركيز على تكامل البيانات والمؤشرات، وبالتالي تمكين المستخدمين من إجراء المزيد من البحوث، كما أن تكامل البيانات والمؤشرات الاحصائية يقلل الوقت ويقلل التكلفة لدى المراكز الاحصائية، ولزيادة جودة ودقة تكامل البيانات والمؤشرات الاحصائية فقد اوصت الدراسة بإجراء مراجعة دقيقة للمؤشرات الاحصائية من قبل المراكز الاحصائية قبل اجراء عملية التكامل.

## Table of Contents

## List of Tables

## List of Figures

## List of Appendices

# Chapter One

## 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) [1].

The major trends of Web 2.0 are collaboration and sharing, The term ‘Web 2.0’ has become undisputed linked with developments such as blogs, wikis, social networking and collaborative software development. Web 2.0 can make dramatic impact on developing interactive and collaborative visual analytics tools for the Internet. Tools are needed to advance humans ability to exchange gained knowledge and develop a shared understanding with other people [2].

In this study, we introduced a new mapping algorithm and visualization system to enhance integration and presentation of official statistics based on mapping algorithm and visualization techniques. Our system aims to provide techniques that help in mapping and presenting results in a meaningful and intuitive way while allowing to interact with the data.

### 1.1 Purpose and Motivation

Due to increase in complexity and heterogeneity of statistical data, an increased need for mapping techniques and sophisticated visualization technology arises, also an increasing need for a common schema to integrate massive data and visualize findings, so that viewers can easily derive an insight of data. A new schema and mapping algorithm were introduced, as well as a mapping algorithm based on hamming distance, editing distance and ontology was established. This algorithm is aiming at enhancing integration and mapping of official statistics.

### 1.2 Objectives

The objective of this research is to introduce mapping algorithm and a new visualization system 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 interaction and visualization techniques.

### 1.3 Background

In this section, the statistical indicators used in the research are discussed with emphasis on their sources and characteristics, in addition, a detailed description of the visualization techniques and String Comparator Metrics are discussed and explored.

#### 1.3.1 Statistical Indicators

A statistical indicator [69] is a data element that represents statistical data for a specified time, place, age, sex, location, and other characteristics, and is corrected for at

least one dimension (usually size) to allow for meaningful comparisons [69]. A simple aggregation such as the number of accidents, total income or women Members of Parliament, is not in itself an indicator, as it is not comparable between populations. However, if these values are standardized, e.g. number of accidents per thousand of population, average income, or women Members of Parliament as a percentage of the total, these values meet the criteria for an indicator.

### 1.3.2 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 as typographical error may lead to many incorrect disagreements. Several methods for dealing with this problem have been developed including: string comparators of mappings from a pair of strings to the interval [0, 1] and measuring the degree of compliance of the compared strings. [43]. String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage, discriminate analysis and logistic regression. By concluding, the simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator.

An early string comparator is the Damerau-Levenstein (D-L) Metric [48], which is in fact only one particular comparator metric from the class of edit distance metrics. The basic idea is the fact that any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another divided by the maximum length of the two compared strings is a measure of the difference between them which is easily converted to a string comparator rating the degree of agreement of the two strings. For a discussion of several enhancements of the D-L metric see Hall and Dowling [49].

Jaro [50,51] introduced a string comparator that is more straightforward to implement and more closely related to the type of human decisions in comparing strings than the D-L metric. Basically, it accounts for the proportion of common characters in both strings and the number of transpositions that have to be made to create the sequence of common characters of one string from the sequence of common characters of the other string. Several enhancements to the Jaro comparator have been developed, in particular by Winkler [46].

### 1.3.3 Hamming Distance

One of the earliest and most natural metrics is the hamming distance [47], 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 binary strings "a" and "b", the Hamming distance is equal to the number of ones (population count) in a XOR b. The metric space of length-n binary strings, with the Hamming distance, is known as the Hamming cube; it is equivalent as a metric space to the set of distances between vertices in a hypercube graph. One can also view a binary string of length “n” as a vector in “$R^{n}$” by treating each symbol in the string as a real coordinate; with this embedding, the strings form the vertices of an n-dimensional hypercube.

### 1.3.4 Edit (Levenshtein) Distance

Edit distance [48] 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.

The edit distance ed(x, y) between strings $x=x_1 \ldots x_m$ and $y=y_1 \ldots y_n$, where $x, y \in \Sigma^*$ is the minimum cost of a sequence of editing steps required to convert x into y. The alphabet $\Sigma$ of possible characters *ch* gives $\Sigma^*$, the set of all possible sequences of $ch \in \Sigma$. Edit distance can be calculated using dynamic programming [70]. Dynamic programming is a method of solving a large problem by regarding the problem as the sum of the solution to its recursively solved sub problems. Dynamic programming is different to recursion however. In order to avoid recalculating the solutions to sub problems, dynamic programming makes use of a technique called Memoization, whereby the solutions to sub problems are stored once calculated, to save recalculation.

To compute the edit distance ed(x,y) between strings "x" and "y", a matrix $M_{1\ldots m+1,1\ldots n+1}$ is constructed where $M_{i,j}$ is the minimum number of edit operations needed to match $x_{1\ldots i}$ to $y_{1\ldots j}$. 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.

$$M_{1,1} \leftarrow 0$$

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

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 $M_{m+1,n+1}$.

For example, the Levenshtein distance between "kitten" and "sitting" is 3, since the following three edits change one into the other, and there is no way to do it with fewer than three edits: kitten → sitten (substitution of "s" for "k"), sitten → sittin (substitution of "i" for "e"), sittin → sitting (insertion of "g" at the end).

Table (1.1) 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 1.1: 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** |

### 1.3.5 Ontology

An ontology typically provides a vocabulary that describes a domain of interest and a specification of the meaning of terms used in the vocabulary [71]. Depending on the precision of this specification, the notion of ontology encompasses several data and

conceptual models, including, sets of terms, classifications, database schemas, or fully axiomatized theories [54]. 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 Extensible Markup Language (XML), does not reduce heterogeneity: it raises heterogeneity problems at a higher level.

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 [54]. Diverse solutions for matching have been proposed in the last decades [55, 56]. Several recent surveys [57–58] and books [54, 59] have been written on the topic as well.

### 1.3.6 Data Visualization

Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. It is also the process of transforming objects, concepts, and numbers into a form that is visible to the human eyes. Data visualization is all about understanding ratios and relationships among numbers. Not about understanding individual numbers, but about understanding the patterns, trends, and relationships that exist in groups of numbers. Interaction techniques like Searching, Comparing, Re-visualizing, Sorting, and Filtering will help also in understanding that data in visualization process [72].

According to [74], Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces. Today, data is produced at an incredible rate and the ability to collect and store the data is increasing at a faster rate than the ability to analyze it. Over the last decades, a large number of automatic data analysis methods have been developed. However, the complex nature of many problems makes it indispensable to include human intelligence at an early stage in the data analysis process. Visual Analytics methods allow decision makers to combine their human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems. Using advanced visual interfaces, humans may directly interact with the data analysis capabilities of today's computer, allowing them to make well-informed decisions in complex situations, visual Analytics can be seen as an integral approach combining visualization, human factors, and data analysis. Fig. (1.1) illustrates the research areas related to Visual Analytics.

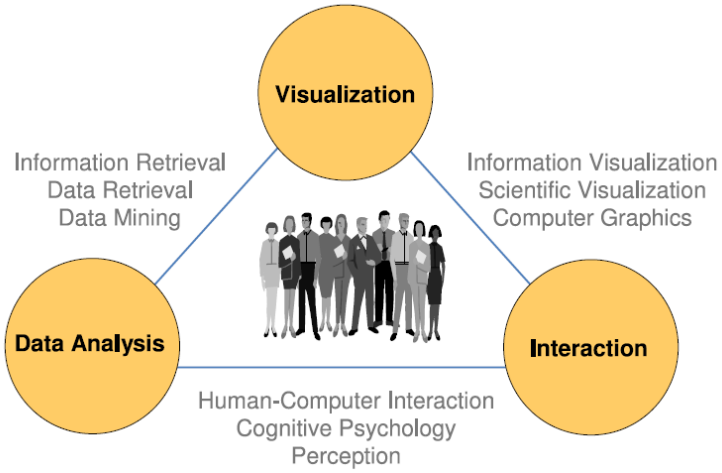



Figure 1.1: Visual analytics: integral approach combining visualization, human factors, and data analysis [74].

Besides visualization and data analysis, especially human factors, including the areas of cognition and perception, play an important role in the communication between the human and the computer, as well as in the decision-making process. With respect to visualization, Visual Analytics relates to the areas of Information Visualization and Computer Graphics,

and with respect to data analysis, it profits from methodologies developed in the fields of information retrieval, data management & knowledge representation as well as data mining.

### 1.3.7 Data Visualization Process

*The basic idea of data visualization is to visually represent the information, Allowing the human to directly interact with the data to gain insight, draw conclusions, and ultimately make better decisions. The visual representation of the information reduces complex cognitive work needed to perform certain tasks. "People use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data to provide timely, defensible, and understandable assessments"* [74].

Data visualization process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Fig (1.3) shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the Visual Analytics Process.

In many application scenarios, heterogeneous data sources need to be integrated before visual or automatic analysis methods can be applied. Therefore, the first step is often to pre-process and transform the data to derive different representations for further exploration (as indicated by the Transformation arrow in Fig. (1.2)). Other typical pre-processing tasks include data cleaning, normalization, grouping, or integration of heterogeneous data sources.

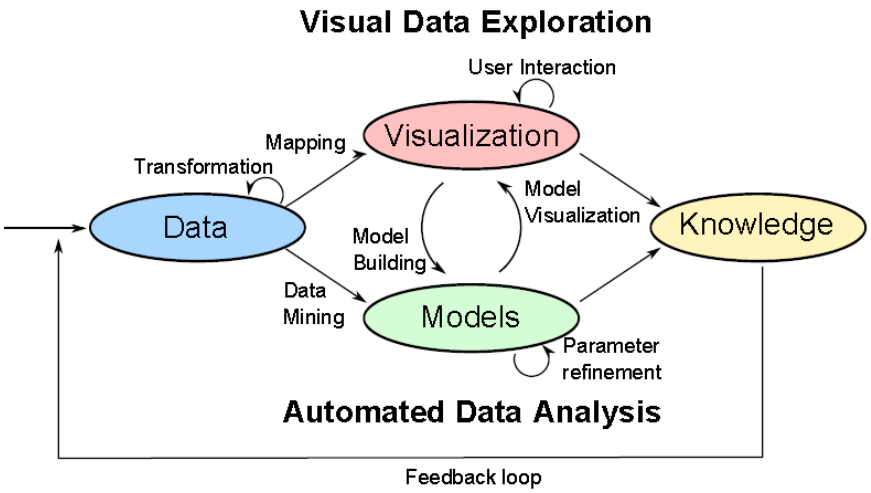



Figure 1.2: Data visualization process is characterized through interaction between data, visualizations, models about the data, and the users in order to discover knowledge [73].

After the transformation, the analyst may choose between applying visual or automatic analysis methods. If an automated analysis is used first, data mining methods are applied to generate models of the original data, data mining is the process to discover interesting knowledge from large amounts of data, it is an interdisciplinary field with contributions from many areas, such as statistics, machine learning, information retrieval, pattern recognition and bioinformatics, it is widely used in many domains, such as retail, finance, telecommunication and social media. The main techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining, and also some new techniques such as social network analysis and sentiment analysis.

Once a model is created the analyst has to evaluate and refine the models, which can best be done by interacting with the data. Visualizations allow the analysts to interact with the automatic methods by modifying parameters or selecting other analysis algorithms. Model visualization can then be used to evaluate the findings of the generated models. Alternating between visual and automatic methods is characteristic for the Visual Analytics process and leads to a continuous refinement and verification of preliminary

results. Misleading results in an intermediate step can thus be discovered at an early stage, leading to better results and a higher confidence. If a visual data exploration is performed first, the user has to confirm the generated hypotheses by an automated analysis. User interaction with the visualization is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. Findings in the visualizations can be used to steer model building in the automatic analysis. In summary, in the visual analytics process knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts.

The visual analytics process aims at tightly coupling automated analysis methods and interactive visual representations. The classic way of visually exploring data as defined by the Information Seeking Mantra ("Overview first, Zoom/Filter, Details on demand") [61] therefore needs to be extended to the Visual Analytics Mantra [62] "Analyze First, Show the Important, Zoom, Filter, and Analyze Further, and Details on Demand".

### 1.3.8 Multi-faceted Data: Characteristics and Challenges

The work in this study was motivated by a number of visualization challenges that arise from the heterogeneous nature of data. Such data are usually given with a strong inherent reference to space and time and results from data acquisition method.

When talking about multi-faceted data, we consider the following:

1) Time-dependent data that represents dynamically changing phenomena;

2) Multivariate data consisting of different attributes (data variants) such as age or sex; and

3) Multi-modal data from different data sources.

The visualization and analysis of time-varying data is challenging too [63, 64] and [65]. Analysts want to investigate how their data change over time. They want to uncover spatial and temporal patterns, understand major data trends, and detect anomalies such as outliers. One common goal is to integrate data from multiple time steps in a single image, for instance, by using one spatial axis in the visualization to represent time. Automated analysis methods are often applied in order to abstract time-related data characteristics, for instance, by computing statistical aggregates such as temporal mean values or standard deviations [66]. When designing an analysis framework for time-varying data one also has to consider different data characteristics [63, 64].

# Chapter Two

# Literature Review

Many studies have been done world wide 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.

## 2.1 Mapping Statistical Indicators and Data Integration

A project of micro-founded indicators was developed [4], 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. [4] showed the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making.

The data integration project surveyed available methods of data integration, to provide a critical assessment of different data integration methods with a focus primarily on

statistical issues and to provide an overview of statistical indicators for quality measures of multi-source databases. All these activities have been seen in view of the concrete application within DIECOFIS. Denk and Oropallo [5] surveyed the available methods of data integration. Denk, Inglese, and Calza [6] discussed the relative merits of the methods in the context of their application to national statistics databases. Quality indicators for assessing multi-source databases were provided by Denk, Inglese and Oropallo [7].

In [4] study the main emphasis was on statistical data integration methodology and quality indicators for the assessment of different approaches and applications. However, some technical considerations need to be addressed for multi-source database integration. In addition to technical considerations, semantic discrepancies and similarities of data sources need to be analyzed before the application of statistical methods generally and integration methods in particular: Data source integration as a prerequisite of dataset integration. A metadata oriented approach for the detection and formalised representation of semantic heterogeneities was proposed, following the ideas and concepts of IDARESA (e.g., IDARESA, [8], [9], Denk and Froeschl [10], and Denk, Froeschl and Grossmann, [11]). Froeschl [12] discusses the possible contributions of meta-computing, i.e., the processing of metadata alongside the accompanied statistical data as well as procedures for controlling the integration process based on metadata, to the integration of statistical data and metadata.

Filippo Oropallo and Francesca Inglese [53] 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.

These studies focused on The Jaro–Winkler distance metric that is designed and best suited for short strings such as person names without using ontology in mapping process, these studies also focused more on administrative records from different sources but in our study we focused on statistical indicators from different sources, we focused on hamming distance for the strings with same length and also we used edit distance for strings with different and long length, we built new algorithm based on hamming and edit distance and we added ontology to our algorithm to improve the accuracy of mapping and integration, to test the accuracy of our algorithm we performed experiments on many indicators, the indicators chosen from different countries.

### 2.2 Data Visualization

Over the last years, a number of information visualisation projects and products have been created from scratch. Several applications have been developed for specific data structures and visualisations. A thorough overview of existing toolkits and libraries can be found in [13]. All toolkits and libraries are developed for a specific purpose.

For instance, the Prefuse toolkit Heer et al. [14], implemented in java, is more oriented towards data whose structure corresponds to that of a graph. Among them we can include networks, hierarchies, trees, etc.

The Geovista studio is mainly oriented towards geoscientific analysis. However the toolkit can also be reused in other projects where the focus lies in some other domain Takatsuka and Gahegan [15].

Other toolkits focus more on one specific visualisation technique. The Tree- Map Java Library [16] and the HCIL Treemap 4.0 toolkit [17], both focus on visualisations of treemap algorithms. However the first one can visualise squarified cushion treemaps where the latter can visualise ordered and quantum treemaps [18].

The InfoVis Cyber infrastructure is a central resource unit that provides access to a comprehensive set of software packages easing the exploration, modification, comparison, and extension of data mining and information visualisation algorithms. InfoVis website is complemented with a series of learning modules about the different aspects of data mining and information visualisation, software, databases and the available computing resources [19].

Patrik Lundblad et al. studied [20]. A framework and class library (GAV Flash) implemented in Adobe ActionScript, import data through Excel – data model, create a story and visualization using analytic tools (dynamic query, filter, regional categorization, profiles, highlight), and dynamic color scale, then share the story. The result was a statistical geovisual analytics application for exploring and publishing statistical data on the web, developed with the GAV Flash toolkit, based on a recommendation from the visual analytics (VA) research program.

Mikael Jern [21], build web-enabled application platform that is emerging as a de facto standard in the statistics community for exploring and communicating statistics data, using storytelling mechanism.

Mikael Jern, et al. [22], build tools for interactively analyzing and communicating gained insights and discoveries about spatial-temporal and multivariate OECD regional data. GeoAnalytics Visualization (GAV) component toolkit is based on the principles behind the visual analytics re-search program, using adobe flash basic graphics and flex 3 for user interface design (a collection of high-performance interactive visualization web-enabled components based on common methods from the information and geovisualization research domain).

But the above studies have limitations, For instance, there is no integration and mapping of heterogeneous data into a common schema. Applications for visualizing statistical data are still rare. Moreover, existing applications also suffer the drawbacks that mentioned above. In our study interaction techniques were implemented for visualizations in our system: Searching feature allows the user to find what he are looking for almost instantly, Comparing looking for both similarities and differences. Compared can be single values or series of values (in case of pattern analysis), which provide a selection of graphs that support full spectrum of commonly needed comparison (comparison of single values as well as comparison of patterns), Re-visualizing involves changing visual representation: switching from one type of visualization to another. Being able to do it quickly and easy is essential. Re-visualizing helps to use the strengths of every type of visualizations, Sorting feature helps to put values in order, and Filtering is the act of reducing the data that are viewed to a subset of what is currently there.

### 2.2.1 Toolkits for spatial-temporal multivariate data

InfoVis Toolkit [23], CommonGIS [24], GeoVista [25], VIS-STAMP [26], GAV [27] and CGV [28] are examples of exploratory data analysis (EDA) tools that all have evolved from research and can leverage visualization and computational methods to search for space-time and multivariate patterns. While the benefits of visual analytics tools are many, it has been a challenge to adapt these tools to the Internet and reach a broader user community. Web-enabled tools are therefore needed for applications explicitly designed with the purpose of visualizing, exploring and communicating spatial-temporal and multivariate data through web environment. Such tools should also employ data transformers and data providers, layout mechanisms, interaction, time animation and storytelling suited for a geovisual analytics‟ task. Protovis [29, 30] and perfuse flare [31] are examples of web-enabled tools. They include a collection of interactive visualization components. Nevertheless, they do not support linking mechanisms, for example through selection or filtering, to create multiple linked views.

Tableau [32] and its predecessor Polaris [33] are other examples of a popular web-enabled tools applied to business analytics. In addition to a collection of interactive visualizations, they support multiple linked views, storytelling mechanism and geographical maps.

### 2.2.2 Tools for storytelling

The importance of a capacity to snapshot exploratory data analysis (EDA) sessions [34] and then reuse them for presentation and evaluation was early demonstrated by MacEachren [35] and [36] and incorporated features to capture and reuse interactions and integrate them into electronic documents. CCMaps [37] presents a mapping tool that allows users to save snapshot events and reuse them for presentation purposes. Another effort was made by Visual Inquiry Toolkit [38] that allows users to place pertinent clusters

into a “patternbasket” to be reused in the visualization process. Robinson [39] describes a method they call “Re-Visualization” and a related tool ReVise that captures and reuses analysis sessions. Keel [40] describes a visual analytics system of computational agents that supports the exchange of task relevant information and incremental discoveries of relationships and knowledge among team members commonly referred to as sense-making. Many Eyes [41] is an interesting storytelling approach implemented for a public web site where novice users can upload their own data, create dynamic visualizations and participate in discussions. Nevertheless, Many Eyes does not support creating multiple-linked views and time animation. (See Appendix 1) that shows the summary of literature review contributions.

# Chapter Three

## System Design and Case Study

To achieve the objectives of this research and build the mapping algorithm and visualization system towards enhancing mapping and integration of official statistics, we focused on Millennium Development Goals (MDGs) indicators [42] from different surveys and sources. The original indicators are included in heterogeneous sources and files. As a case study we covered and focused on indicators from Palestinian Central Bureau of Statistics (PCBS) [1], Department of Statistics (Jordan) [44] and from Central Agency for Public Mobilization and Statistics (Egypt) [45]. Selecting indicators from different countries will help us to test the accuracy of our mapping algorithm and integration of data during the import process of these indicators based on our schema.

These statistical indicators were included in different sectors and sub sectors like economy, education, environment, health, information and communication, nutrition, population and women. many interviews with the managers of Palestinian Central Bureau of Statistics (PCBS) departments and divisions were conducted to know the structure of indicators, subgroups and units and to help us in building the final schema.

### 3.1 Data Analysis

Statistical data are sets of often numeric observations which typically have time associated with them, see Fig. (3.1). 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.

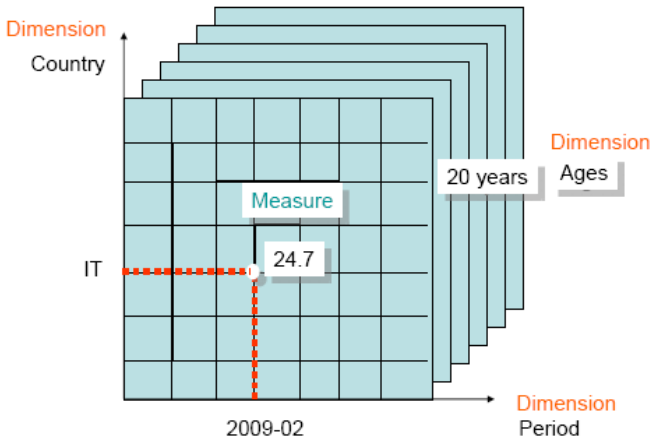



Figure 3.1: Multidimensional 'Cube' of data.

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 (3.1) below gives several examples of these subgroups.

Table 3.1: The subgroup dimension values for the subgroup diminution “Sex” are “Male” and “Female”.

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

We repeat the process many times as needed to enter and associate units and subgroups with all indicators. Each indicator should be correctly associated with its unit(s) and subgroup(s).

After defining indicators and their associated units and subgroups, we define the sectors and sub sectors for our indicators, we then link the indicator-unit-subgroup (IUS) combinations to different sectors and sub-sectors, see Fig. (3.2). Linking I-U-S combinations to sectors and sub-sectors is important, as it allows users to quickly find indicators in our final system. Failure to link indicators to at least one classification will mean that users will not be able to search for those indicators by sector in our final system.

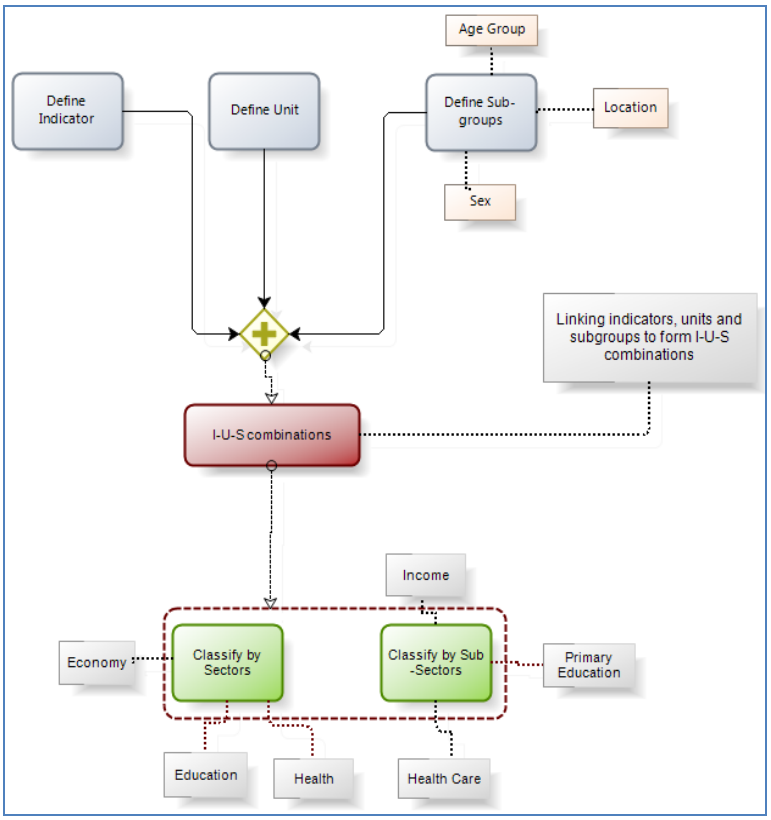



Figure 3.2: indicator-unit-subgroup (IUS) combinations.

### 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 (3.3) 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.

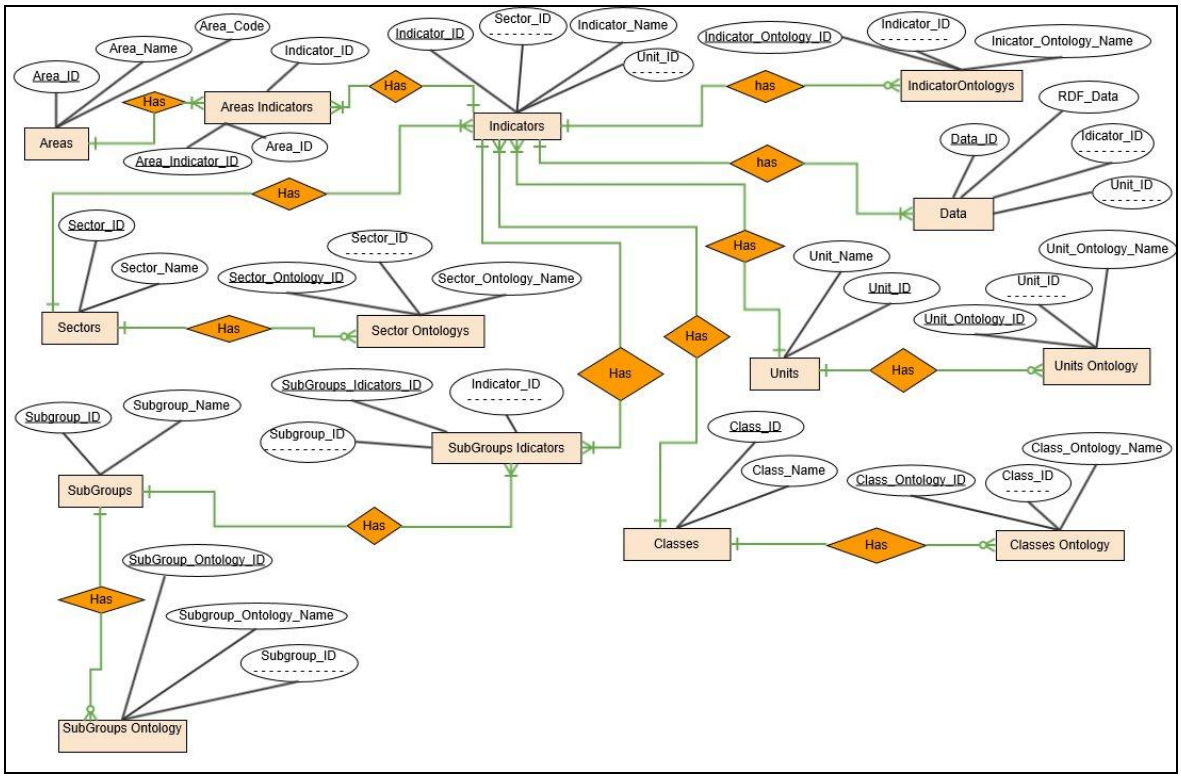


Figure 3.3: Entity Relationship Diagram.

Depending on entity relationship diagram we created SQL Server Database Schema (logical design) as shown in Fig. (3.4), 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 indicators, or units maybe not the same matching but with the same meaning.

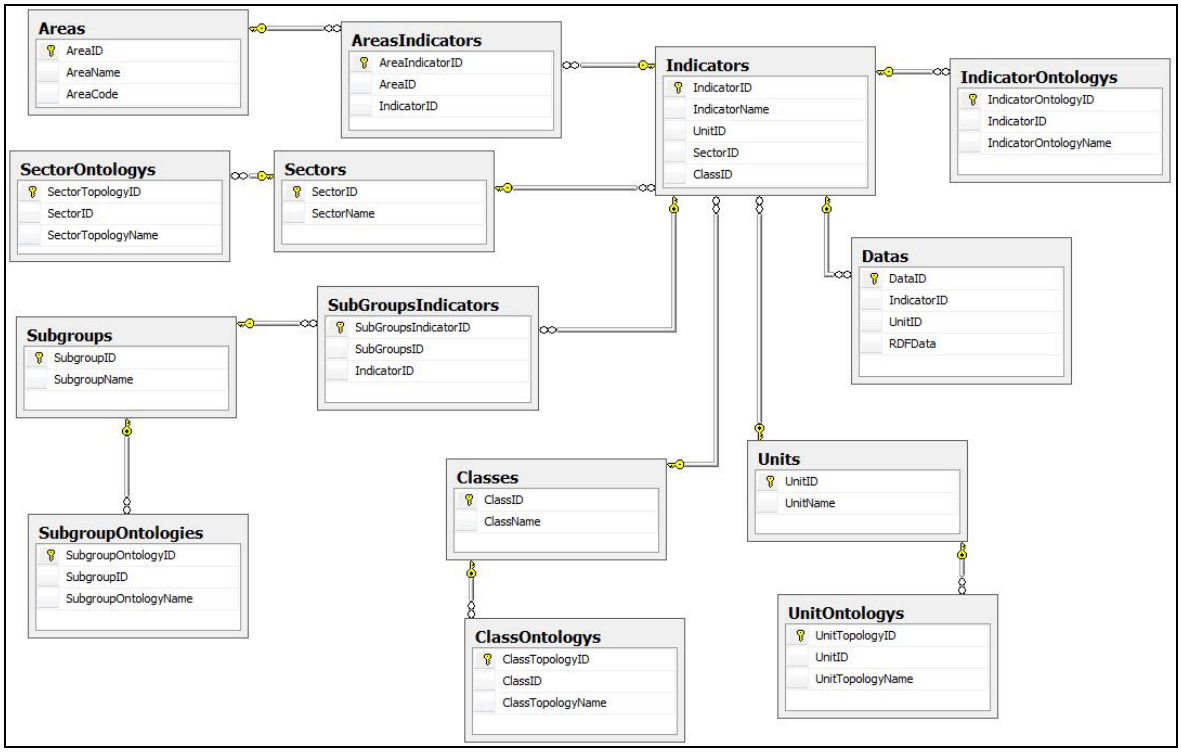



Figure 3.4: Database Schema Snapshot.

As an example "urbanization level" indicator and "level of urbanization" indicator, without using ontology the "Urbanization Level" indicator was mapped to "Population Size", but when using ontology it will check first the ontology lookup table to give the correct indicator according to the meaning not just according to the mapping algorithm using hamming distance and edit distance. Also "%" unit and "Percent" unit are the same unit in meaning, the mapping algorithm that we will discus it in details in the next section will not solve the meaning during mapping process when we import indicators, so adding the ontology tables to our schema will solve and increase the accuracy of mapping.

The schema contains also data table to store the data and values of indicators during import process for different indicators from different sources after mapping process, the data will be stored in this table as Resource Description Framework (RDF) data, the aim for saving data as RDF is to simplify the presentation and dissemination of the data to the end user using visualization techniques in the system. RDF is the standard for encoding metadata and other knowledge on the semantic web, Fig (3.5) shows part of our RDF/XML template, we validate our RDF template and data depending on the world wide web consortium (W3C) markup validation service, which is a free service that helps check the validity of Web documents [52].

```
<rdf:RDF  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:xlsRow="http://www.w3.org/1999/02/RowDescription">
  <xlsRow:xlsRow>
    <xlsRow:Time>...</xlsRow:Time>
    <xlsRow:AreaID>...</xlsRow:AreaID>
    <xlsRow:AreaName>...</xlsRow:AreaName>
    <xlsRow:DataValue>...</xlsRow:DataValue>
    <xlsRow:Subgroup>...</xlsRow:Subgroup>
    <xlsRow:Source>...</xlsRow:Source>
    <xlsRow:Denominator>...</xlsRow:Denominator>
  </xlsRow:xlsRow>
 </rdf:RDF>
```

Figure 3.5: RDF Template Snapshot.

During import statistical data and after the mapping, the data will be stored and saved as RDF/XML in the RDFData field in data table of our schema, this will help to visualize the data. To protect the data stored in the database from unauthorized access, misuse, destruction and loss, the access to database limited to the administrator, administrator must be aware of the different users who access the system and their requirements. Consideration must be given to the different requirements of the users and their access privileges to personal data should fully reflect these requirements. Also in order to maximize security and protection of data, we built a second database backup. The

objective: two redundantly designed data centers that can back up each other in the event of a fault, the second data center is not just restricted to backup alone, but also handles part of productive operations. In the event of a fault, one of the locations takes over the functions of the other. In this way, we squeeze all we can out of our investment and also ensure maximum protection of the whole IT infrastructure against failures and loss of data.

### 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, 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. (3.6) shows the pseudocode of mapping algorithm.

```
If (S1) in ontology return matching from ontology table
If (S1) not in ontology table use hamming distance and edit distance
def hamming_distance(s1, s2):
    #Return the Hamming distance between equal-length sequences
    if len(s1) != len(s2):
        raise ValueError("Undefined for sequences of unequal length")
    return sum(ch1 != ch2 for ch1, ch2 in zip(s1, s2))

int LevenshteinDistance(char s[1..m], char t[1..n])
{
  // for all i and j, d[i,j] will hold the Levenshtein distance between
  // the first i characters of s and the first j characters of t;
  // note that d has (m+1)*(n+1) values
  declare int d[0..m, 0..n]

  clear all elements in d // set each element to zero

  // source prefixes can be transformed into empty string by
  // dropping all characters
  for i from 1 to m
    {
      d[i, 0] := i
    }

  // target prefixes can be reached from empty source prefix
  // by inserting every characters
  for j from 1 to n
    {
      d[0, j] := j
```

```
   }

  for j from 1 to n
   {
    for i from 1 to m
      {
       if s[i] = t[j] then
         d[i, j] := d[i-1, j-1]       // no operation required
       else
         d[i, j] := minimum
                (
                  d[i-1, j] + 1,  // a deletion
                  d[i, j-1] + 1,  // an insertion
                  d[i-1, j-1] + 1 // a substitution
                )
      }
   }

  return d[m, n]
}
```

Figure 3.6: Pseudocode of Mapping Algorithm.

Depending on the definition of hamming, edit (Levenshtein) distance and by adding the ontology we implemented our mapping algorithm, (See Appendix 2) that shows the algorithm code in details, Also we build GUI to import indicators with data values from different sources. Fig(3.7) shows the summery steps of mapping algorithm.

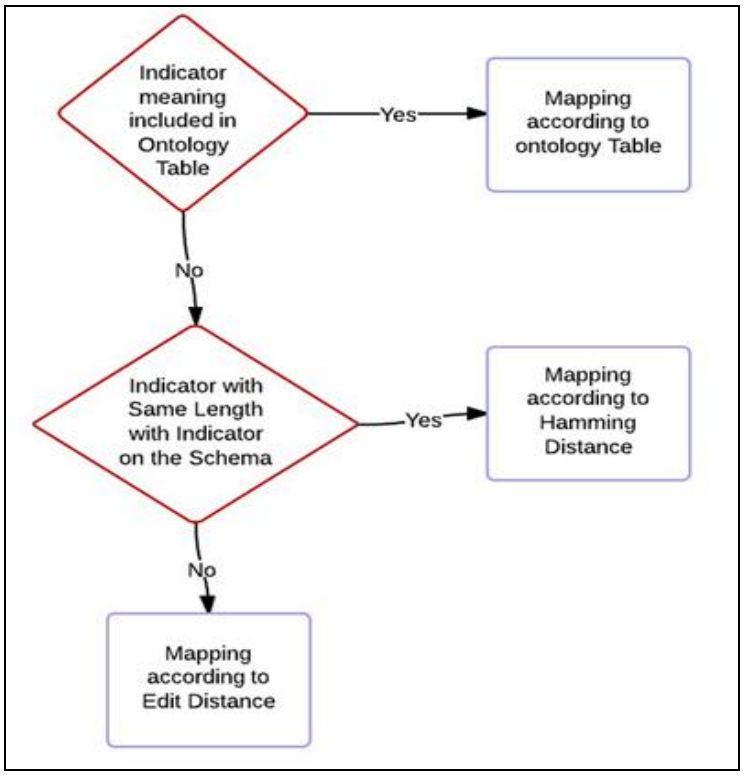
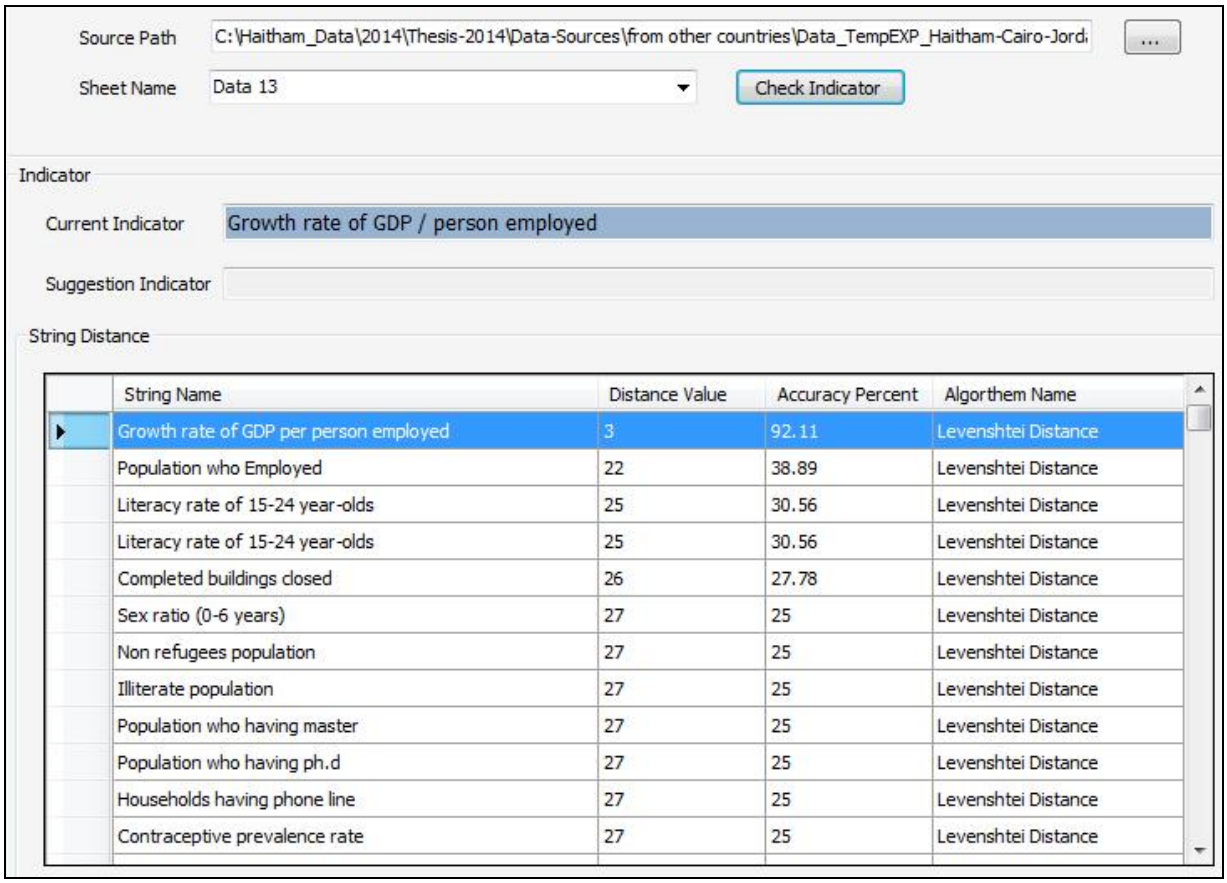


Figure 3.7: Summery Steps of Mapping Algorithm.

### 3.3.1 Indicators Mapping

Fig. (3.8) illustrates the importing and mapping process of indicators and data values. Using our algorithm we started to import the data and indicators from different sources files to our target schema which contains indicators, Units and Subgroups. during the data import using our algorithm, the exact matched indicators, units and subgroups will be mapped automatically to 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. (3.8) shows that when we import "Growth rate of GDP/person employed" indicator from one of our sources to the schema, 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. (3.8) 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 calculate the accuracy in percent using the formula: percent = (largerString.Length - editDistance) / largerString.Length)*100.



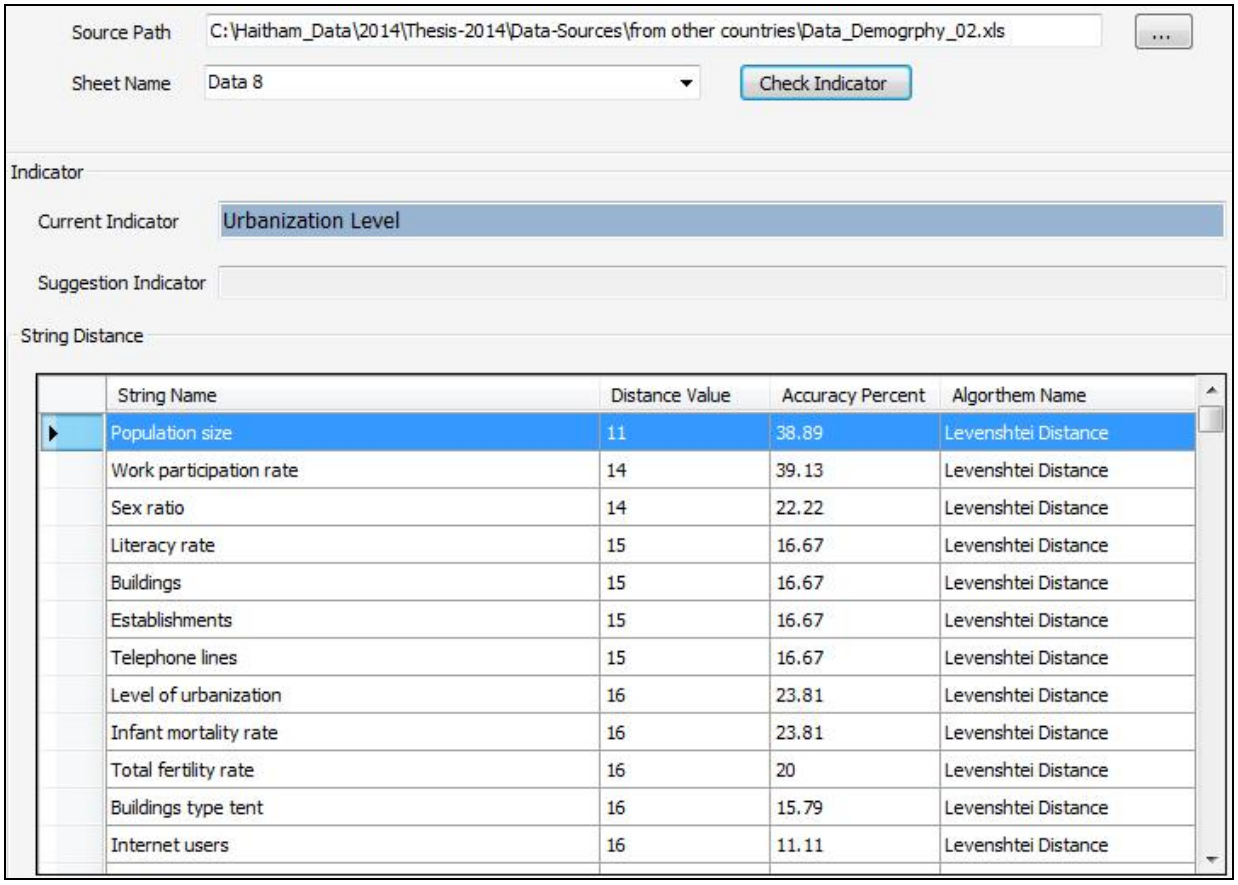

| String Name | Distance Value | Accuracy Percent | Algorthem Name |
|---|---|---|---|
| Growth rate of GDP per person employed | 3 | 92.11 | Levenshtei Distance |
| Population who Employed | 22 | 38.89 | Levenshtei Distance |
| Literacy rate of 15-24 year-olds | 25 | 30.56 | Levenshtei Distance |
| Literacy rate of 15-24 year-olds | 25 | 30.56 | Levenshtei Distance |
| Completed buildings closed | 26 | 27.78 | Levenshtei Distance |
| Sex ratio (0-6 years) | 27 | 25 | Levenshtei Distance |
| Non refugees population | 27 | 25 | Levenshtei Distance |
| Illiterate population | 27 | 25 | Levenshtei Distance |
| Population who having master | 27 | 25 | Levenshtei Distance |
| Population who having ph.d | 27 | 25 | Levenshtei Distance |
| Households having phone line | 27 | 25 | Levenshtei Distance |
| Contraceptive prevalence rate | 27 | 25 | Levenshtei Distance |

Figure 3.8: Example of Importing and Mapping "Growth rate of GDP /person employed" Indicator.

### 3.3.2 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.(3.9) illustrate that when we import "urbanization level" indicator the nearest indicator to this indicator as

shown in Fig.(3.9) is "Population Size" indicator, this is false matching since the two indicators not the same.




| String Name | Distance Value | Accuracy Percent | Algorthem Name |
|---|---|---|---|
| Population size | 11 | 38.89 | Levenshtei Distance |
| Work participation rate | 14 | 39.13 | Levenshtei Distance |
| Sex ratio | 14 | 22.22 | Levenshtei Distance |
| Literacy rate | 15 | 16.67 | Levenshtei Distance |
| Buildings | 15 | 16.67 | Levenshtei Distance |
| Establishments | 15 | 16.67 | Levenshtei Distance |
| Telephone lines | 15 | 16.67 | Levenshtei Distance |
| Level of urbanization | 16 | 23.81 | Levenshtei Distance |
| Infant mortality rate | 16 | 23.81 | Levenshtei Distance |
| Total fertility rate | 16 | 20 | Levenshtei Distance |
| Buildings type tent | 16 | 15.79 | Levenshtei Distance |
| Internet users | 16 | 11.11 | Levenshtei Distance |

Figure 3.9: 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. Fig.(3.10).

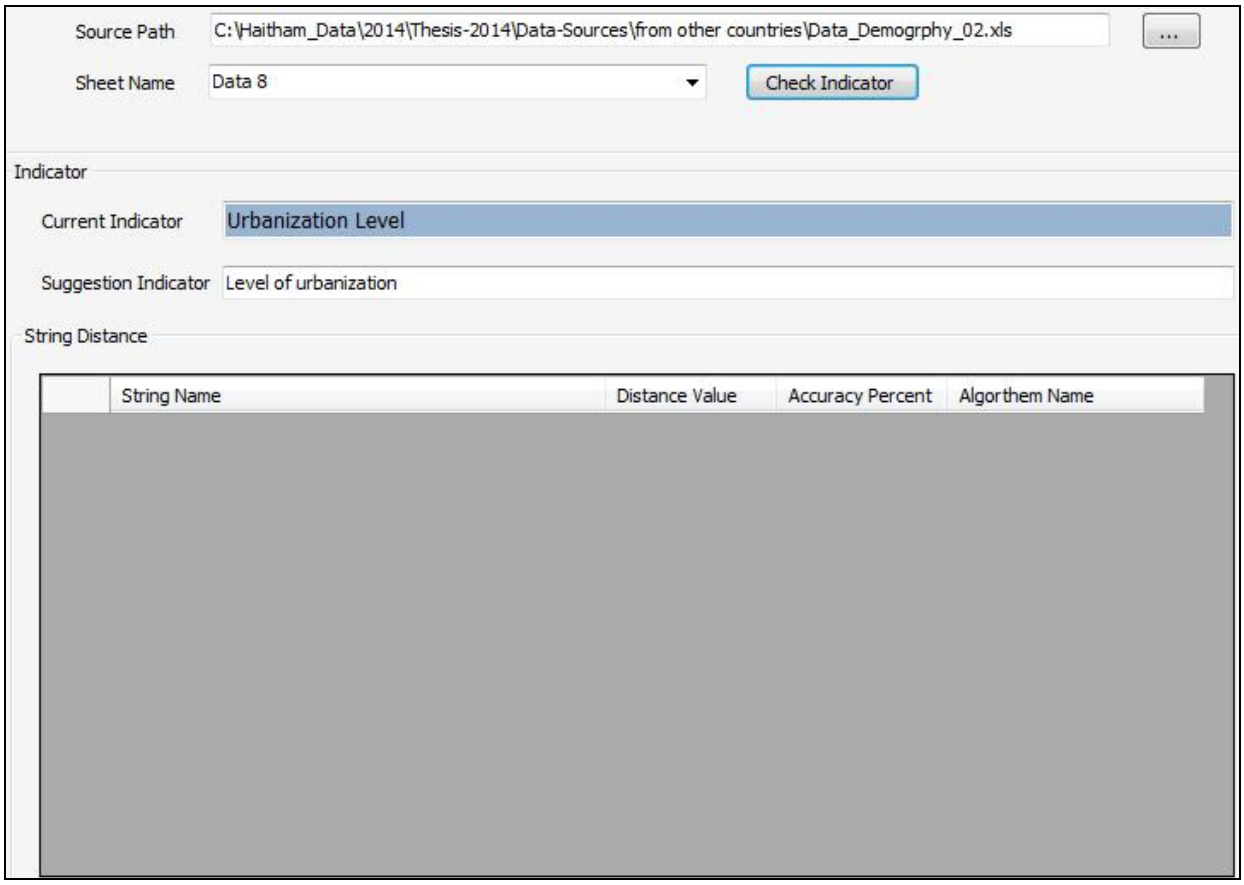


Figure 3.10: Example of Importing and Mapping "Urbanization Level" Indicator with Ontology.

### 3.3.3 Sectors, Classes, and Units Mapping

From the GUI after importing the indicator we imported the sector, class, unit and Subgroups for each indicator, using our algorithm we found the exact matching, nearest matching with minimum edit distance, and ontology matching for the sector, class, unit and subgroups as we did with indicator.

As an example if we import source file contains "health" sector, "safe motherhood" class, and "Births/Woman" unit, the algorithm returns exact matching to "health" sector from the schema, exact matching to "safe motherhood" class from the schema, and nearest matching to "Births/Woman" unit to "Births per woman" unit in the schema with minimum edit distance equal 5 and accuracy between the two units 68.75%. Fig.(3.11).

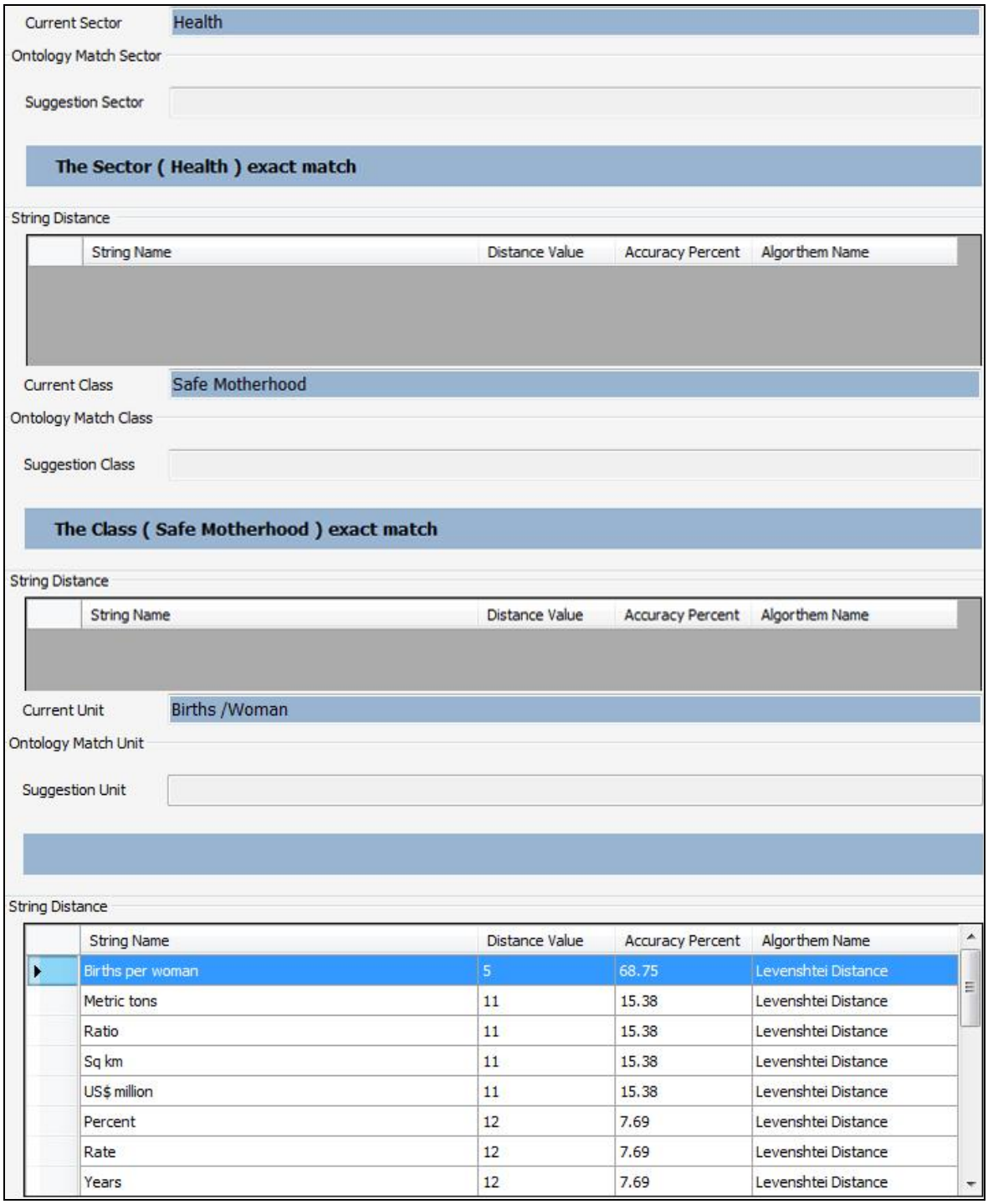


Figure 3.11: Example of Importing and Mapping "Health" Sector, "Safe Motherhood" class, and "Births/Woman" Unit.

If we import another source file contains "%" unit instead of "percent", the algorithm will return nearest matching with minimum edit distance for "%" unit from the source file to "US$" unit in the schema with minimum edit distance equal 3 and the accuracy between the two units 0%. As shown in Fig.(3.12). this is false mapping since "%" unit not the same

"US$" unit, this happen without using ontology, if we have "%" in unit ontology lookup table, the algorithm first check the ontology table and return the correct mapping to percent. Fig.(3.13).

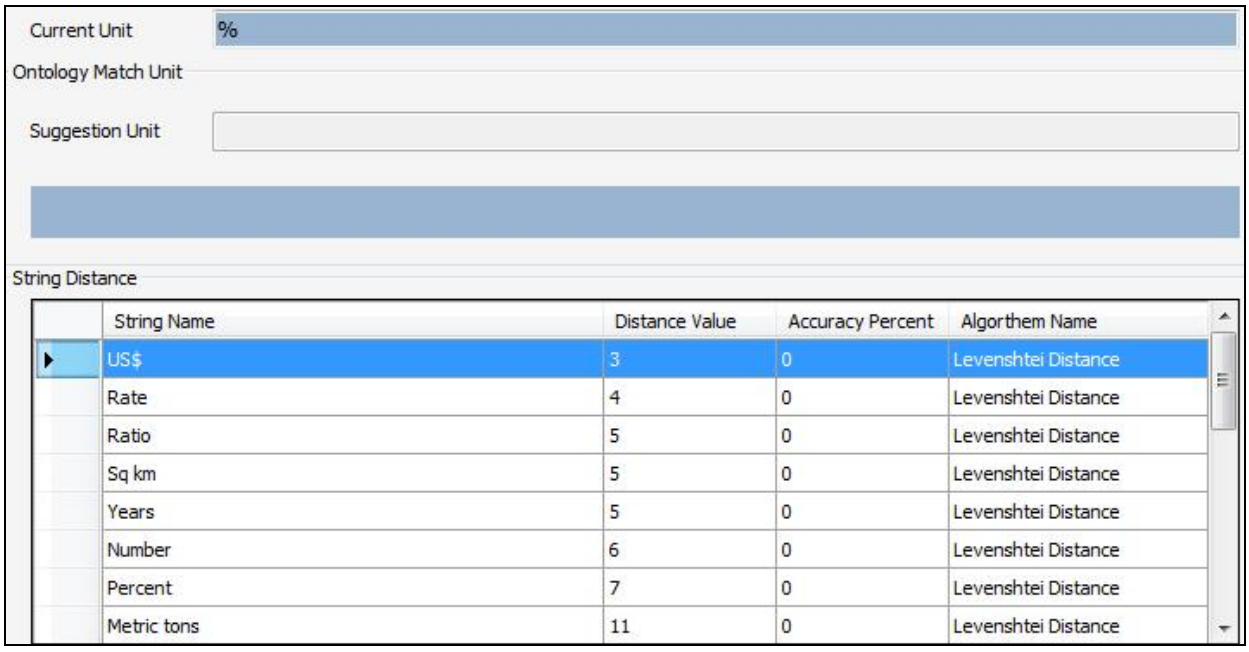



Figure 3.12: Example of Importing % Unit without using Ontology.

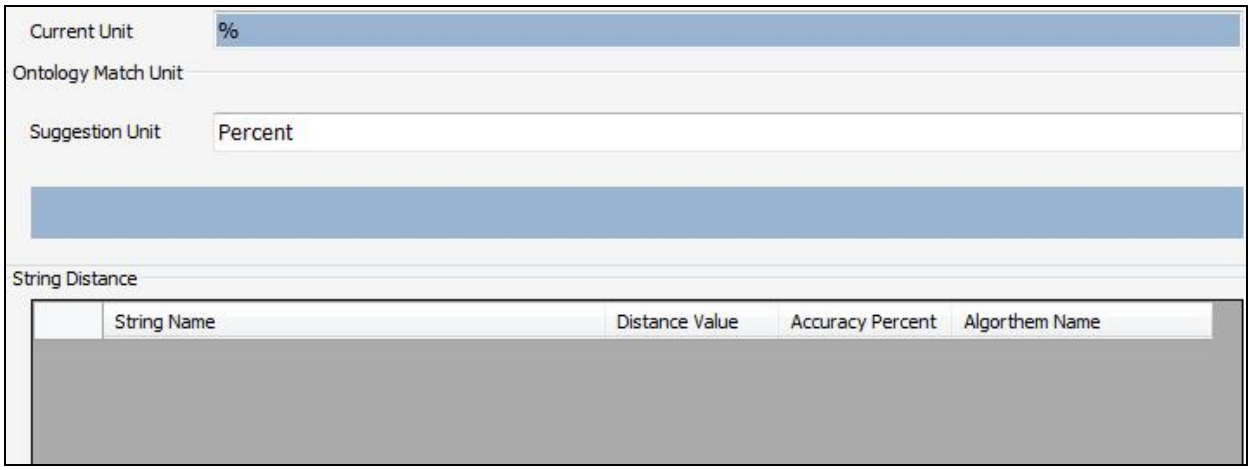



Figure 3.13: Example of Importing % Unit with Ontology.

### 3.4 Data Visualization

The data and indicators after importing and mapping stored in the data table of our schema as RDF, this helps us to build the visualization system. in our system we used different visualization techniques pie-, bar-, column-, line, dot chart, scatter plot and tree maps, …etc. Using these techniques makes ability for the user to interact with data.

Our proposed system in Fig. (3.14) consists of three main parts: Data Analysis, Data Mapping, and Data Visualization. Data analysis is defining indicators, units and subgroups, linking indicators, units and subgroups to form Indicator-Unit-Subgroup combinations, Categorizing Indicator-Unit-Subgroup combinations under various classifications, building the schema, importing indicators from different files. Our mapping algorithm used to map indicators, units, subgroups with the reference file (schema), data visualization is defining and using visualization techniques (scatter plot, bubble chart, map chart, line graphs, stack graphs, bye chart, table lens, histogram, parallel axes plot, time graph, data table, tree map).

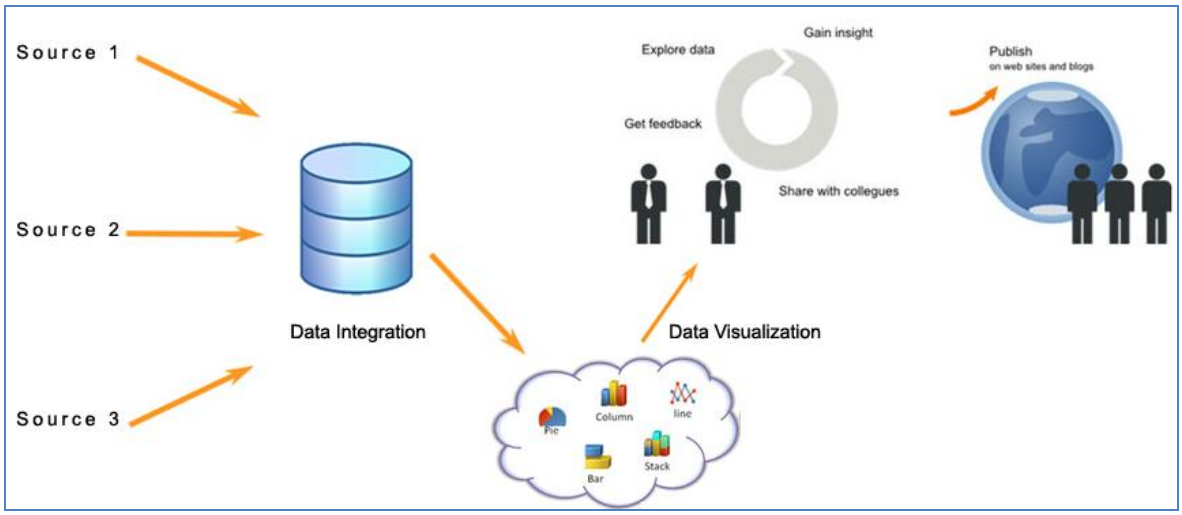



Figure 3.14: Mapping and Visualization System Architecture.

Easy integration of existing information visualization techniques and third-party libraries should be possible. As there are more than hundred different visualization algorithms, implemented in different libraries and different programming languages, the system must have means to quickly select a visualisation technique for a given data collection.

Research on different visualisation techniques and algorithms is a continuous process. For example, the use of a tree-map visualisation was suggested in [3] as a compact

visualisation of directory tree structures. Numerous extensions and implementations of this idea have been created during the following years. We want to be able to use the different implementations of this technique and its extensions for every possible data collection, without having to write code for well-documented algorithms from scratch.

### 3.5 Visualization Implementation

The visualization of our system were implemented based on the Microsoft platform and .NET Framework using the software development tools of Microsoft Visual Studio, including .NET and ASP.NET., Highcharts Java Script libraries [67], Highcharts is a charting library written in pure JavaScript, offering an easy way of adding interactive charts to web site or web application. Highcharts currently supports line, spline, area, areaspline, column, bar, pie, scatter, angular gauges, arearange, areasplinerange, column range and polar chart types. Flex visualization libraries [68].data visualization library consists of Charting, Advanced Data Grid, OLAP Data Grid, and Automation components. Flex Charting has numerous chart types which we can use to build powerful, interactive charts quickly. It also has an extensible API for customization. Advanced Data Grid is another powerful component which enables grouping, aggregation, and display of hierarchical data.

# Chapter Four

## System Design Evaluation

To test and evaluate the accuracy of the mapping algorithm in practice, we performed experiments on many indicators, indicators chosen from different countries.

### 4.1 Mapping Results without Ontology

We test the algorithm without using ontology, by importing different indicators and their units, sectors, and subgroups.

Table (4.1), Table (4.2) and Table (4.3) (see appendix 3) present the summary of results for importing different indicators, units, and subgroups respectively as testing indicators, units and subgroups for the mapping algorithm.

Table (4.1 in appendix 3) shows the summary of the results of mapping when we import indicators, the algorithm return the best and nearest mapping for each indicator according to the hamming and edit distance for each indicator with indicators in the schema, the results shown that the accuracy of the algorithm mapping is 67% as shown in Fig.(4.1), we can increase the accuracy of the algorithm by decreasing the false mapping when using ontology as we will see in the next section.

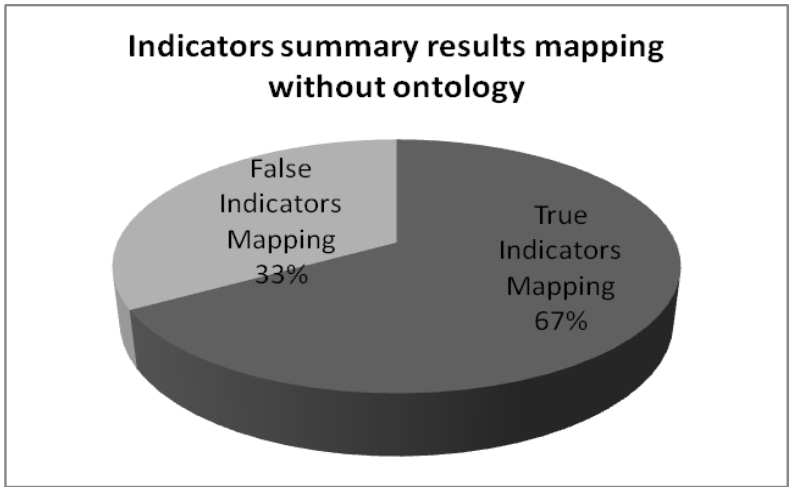



Figure 4.1: Algorithm Indicators Mapping Accuracy without Ontology.

Table (4.2 in appendix 3) shows the summary of the results of mapping when we import units of 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, the results shown that we have about 4 unit with false mapping, as the results shown in Table (4.2 in appendix 3), the different in writing the unit with same meaning causes false mapping for units, as an example in Table (4.2 in appendix 3) "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.(4.2) shows that the algorithm accuracy for units mapping without ontology is 82%, the algorithm accuracy can be improved by using ontology this will be discussed in the next section.

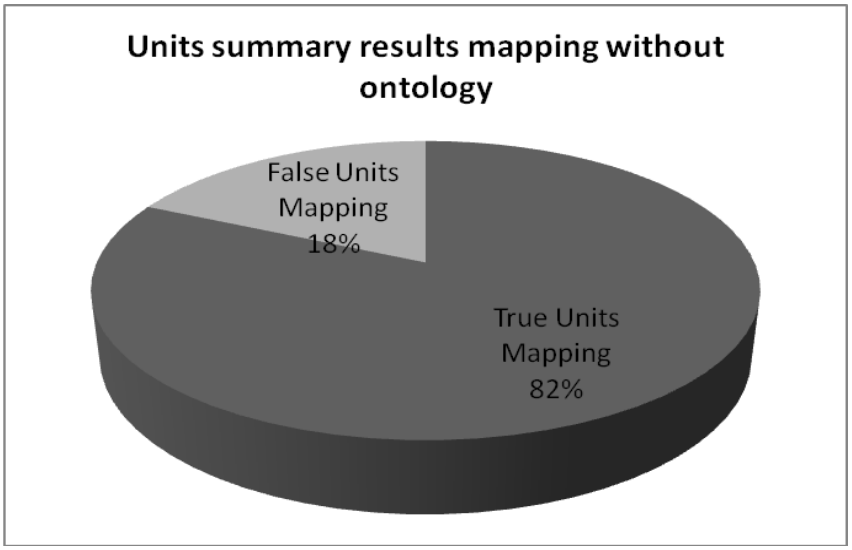



Figure 4.2: Algorithm Units Mapping Accuracy without Ontology.

Table (4.3 in appendix 3) shows the summary of the results of mapping when we import subgroups of indicators, the algorithm calculate the best and nearest mapping for each subgroup according to our algorithm for each subgroup with subgroups in our schema, the results shows that we have about 10 subgroups with false mapping. Fig.(4.3) 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.

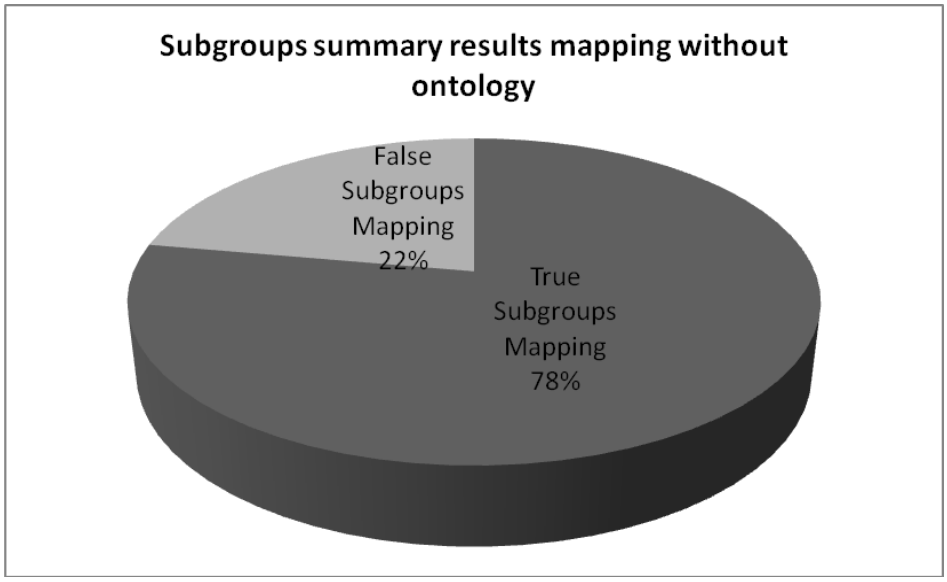



Figure 4.3: Algorithm Subgroups Mapping Accuracy without Ontology.

Fig.(4.4) 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 indicators is 82% and the accuracy of the algorithm for mapping subgroups of indicators is 78%.

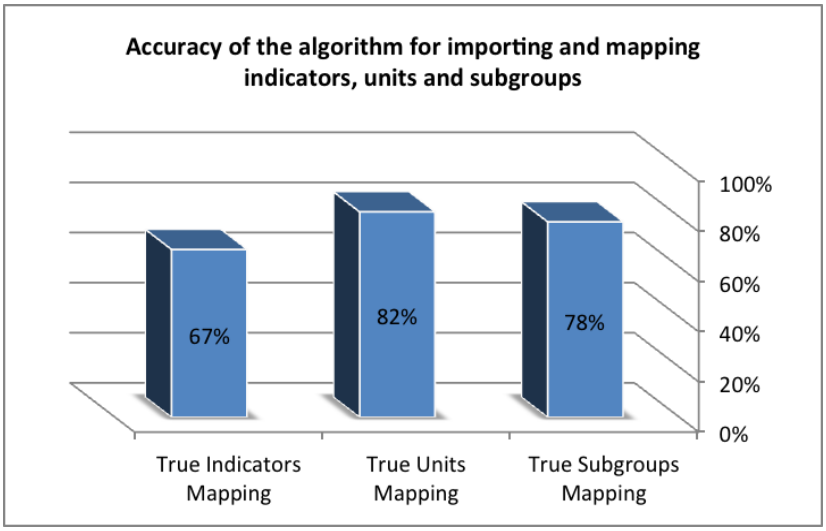



Figure 4.4: Algorithm Indicators, Units and Subgroups Mapping Accuracy without Ontology.

## 4.2 Mapping Results with Ontology

We improve the accuracy of our mapping 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.

Table (4.4), Table (4.5) and Table (4.6) (see appendix 3) present the summary of results for importing different indicators, units, and subgroups respectively as testing indicators, units and subgroups for the mapping algorithm, but this time the algorithm

check first the ontology tables in the schema before calculating the edit distance and hamming distance to return better mapping results.

Table (4.4 in appendix 3) shows the summary of the results of mapping when we import indicators, the algorithm locking for the meaning terms in 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 hamming and edit distance for each indicator with indicators in the schema, the results shown that there are about 6 indicators with false mapping, Fig.(4.5) 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 Table (4.1 in appendix 3) and Fig(4.1).

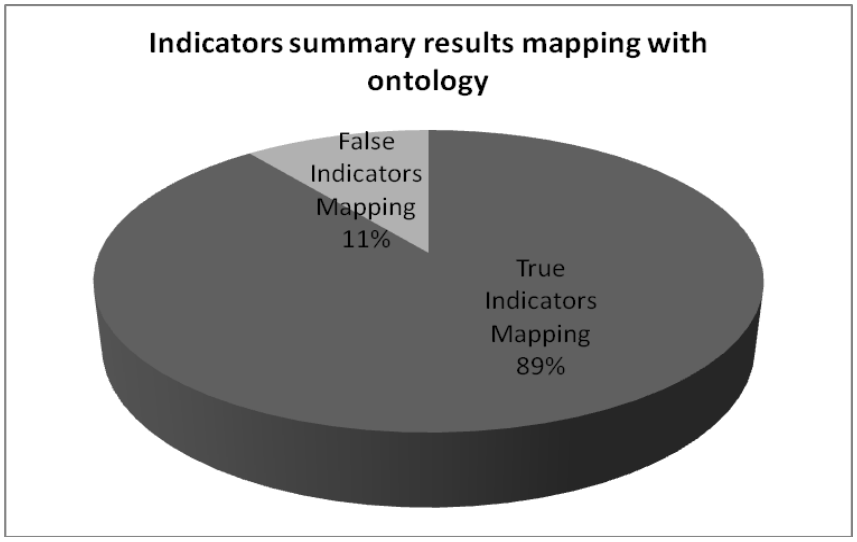



Figure 4.5: Algorithm Indicators Mapping Accuracy with Ontology.

Table (4.5 in appendix 3) shows the summary results of mapping when we import units of 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, the results shown that there is one unit with false mapping from 22 units imported, 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.(4.6) 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 Table (4.2 in appendix 3) and Fig.(4.2).

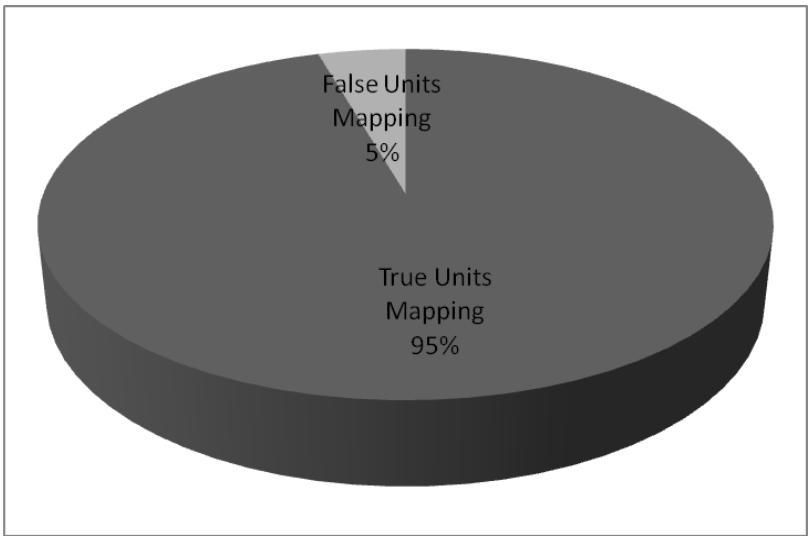



Figure 4.6: Algorithm Units Mapping Accuracy with Ontology.

Table (4.6 in appendix 3) presents the summary of the results of mapping when we import subgroups of indicators, by using ontology the algorithm check first the ontology table of subgroups, as an example as shown in Table (4.6 in appendix 3) 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 as shown in the previous section in Table (4.3 in appendix 3), 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 as shown in Table (4.3 in appendix 3). the accuracy of the algorithm using ontology for subgroups is higher than the accuracy of the algorithm without ontology as shown in Table (4.6 in appendix 3) results, Fig.(4.7) 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 Table (4.3 in appendix 3) and Fig.(4.3) the accuracy was 78%.

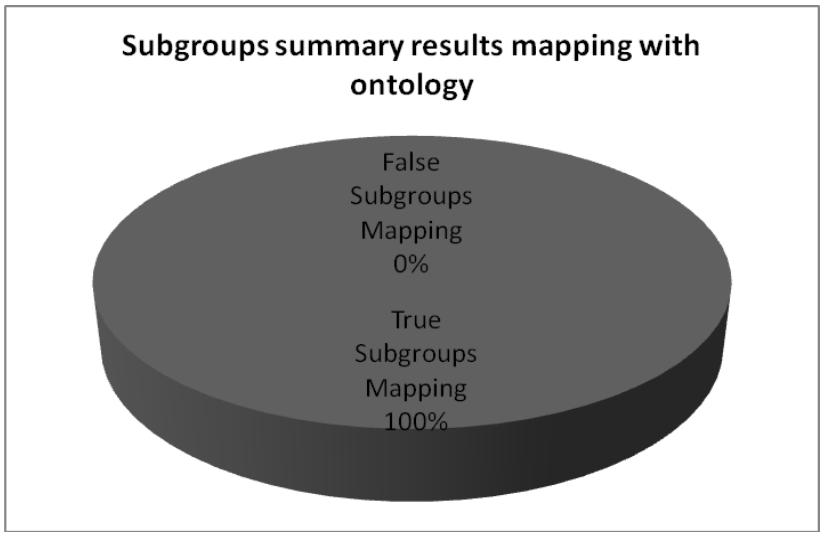



Figure 4.7: Algorithm Subgroups Mapping Accuracy with Ontology.

Fig.(4.8) 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 indicators is 95% and the accuracy of the algorithm for mapping subgroups of indicators is 100%.

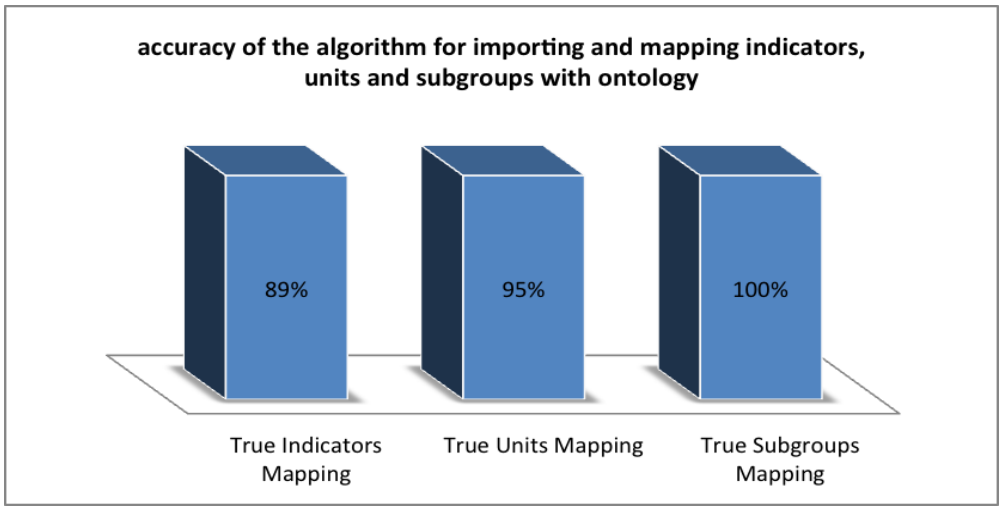



Figure 4.8: Algorithm Indicators, Units and Subgroups Mapping Accuracy with Ontology.

Fig.(4.9) summarize the accuracy of the algorithm according to our results without ontology and with ontology for importing indicators, units and subgroups.

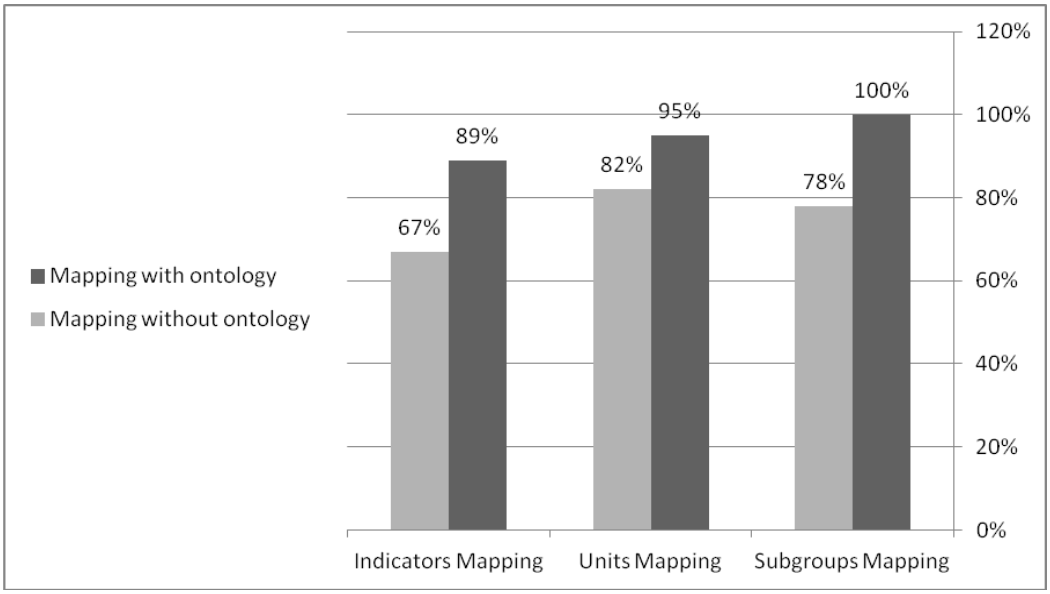



Figure 4.9: Algorithm Indicators, Units and Subgroups Mapping Accuracy with Ontology and without Ontology.

Fig.(4.9) 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.

### 4.3 Interview with expert users to specify their requirements

This section describes the approach used to specify users requirements and to evaluate the system. The first step is done by making interview to get requirements before implementation phase of visualization techniques for our system from the end users (Statisticians, researchers, decision makers). The second step is realized by making interview to get and find out the requirements and feedback of end-users for our final system design and visualization techniques that we used. Further sections describe the process of evaluation in greater detail.

We make interviews to analyze the user requirements. Each type of users has their own concerns and expectations from the system. In order to identify user needs, interview with expert users have been done. The interviews aim to find the user requirements and discover probable problems which users face without visualizing the results. a usability test was performed, This step is done by making interview to get requirements before implementation phase of our system visualization from the end users, We collected feedbacks from interview with 13 participants. Participants were selected to be representative of the intended user community, including Statisticians, researchers, decision makers so that they could specify user requirements related to visualization, design appropriate functionalities, and develop a visualization system for statistical data. (see appendix 4), we tried to specify user requirements related to visualization, design appropriate functionalities, to build visualization system according to the users requirements. Our findings from this interview are concluded as below:

**Req1. How do you use statistical data in your work?** All of those people are Statisticians, researchers, decision makers so that they could specify user requirements related to visualization, design appropriate functionalities, and develop a visualization system for statistical data.

**Req2. Which information from visualization system, are you interested in?** The users interested to see in a visualized model: economy, information and communication, health, nutrition, education, and women, eleven of the users mentioned that interested to see economy information in the system Fig.(4.10). All users agreed that visualization can help them to have a better and clearer understanding of the results.

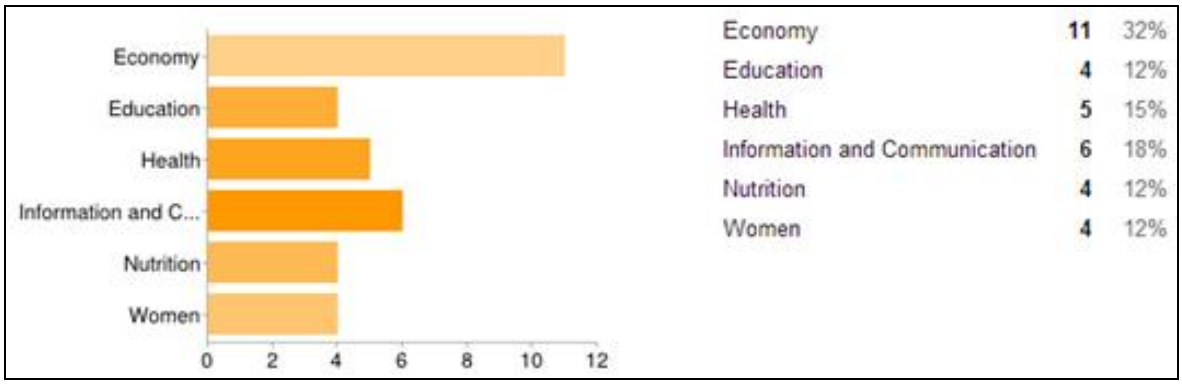



Figure 4.10: Information in the system that the users interested in.

**Req3. Which visualization techniques more convenient way for you to explore the results?** All the users reported that they liked each of the visualizations. most of the users liked line charts, pie chart, bar chart, column chart, and map chart. nine of the users mentioned that the line chart a more convenient way to explore the results than other techniques Fig.(4.11). the treemap was met with some reservation, but in the end, users found that they could think of different uses for it. one user specified that the treemap was his favourite visualization technique and one user specified that the scatter plot was his favourite.

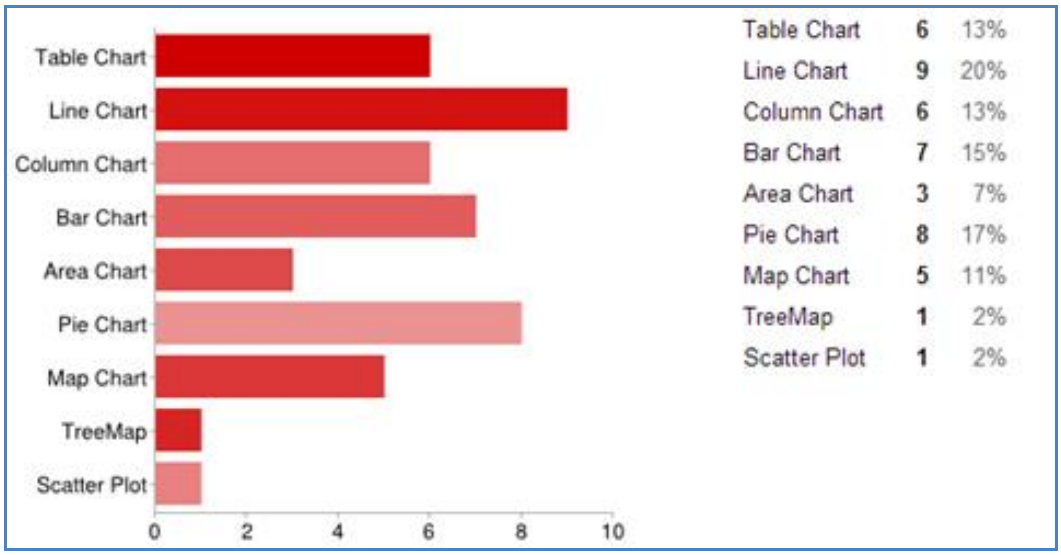


Figure 4.11: Convenient visualization techniques for users to explore the results.

**Req4. Do you think visualizing statistical data results can help you in your work?** We can conclude from the interview that both expert and end users are interested in a visualization statistical data and they think it could facilitate their work Fig.(4.12), we conclude that we should visualize the input related to each output as well. Also, the users believe that comparing different scenarios could be an important feature of the system. the users also feel that visualization of results could contribute to their work.

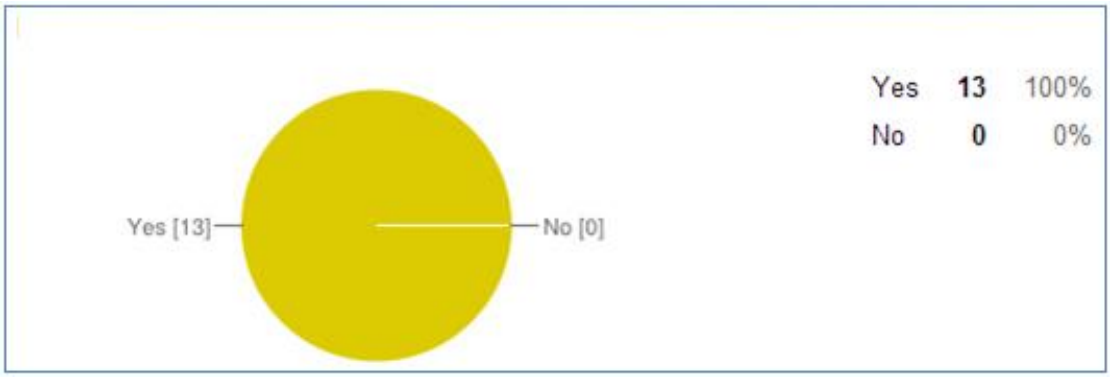


Figure 4.12: Users who think that visualizing results can help them in their work.

**Req5. Can visualization help you in understanding the results? If yes, what is your suggestion for designing of a visualize model?** Yes, the visualization would be facilitated. If the results could be appropriately visualized it would also facilitate for other people to better understand the results Fig.(4.13). And they suggested: to use different

types of graphs and charts describing the statistical data and comparing different time series, visualize model should support different languages, using maps and colours, adding features to download and share the visualization output and they suggested also to use animation for time series and using colour gradient, to be more flexible when sorting according to time series data and going through items and sub items.

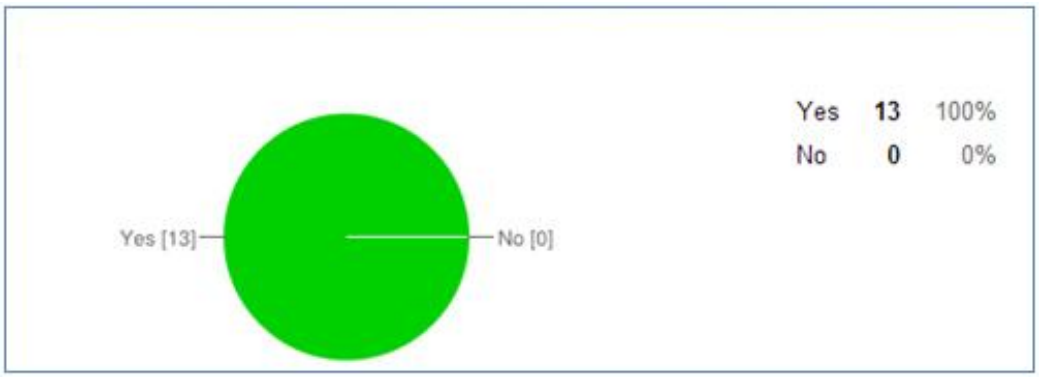



Figure 4.13: Users who think that visualization can help them in understanding the results.

**Req6. Are you interested in comparing, filtering, and sorting different scenarios of statistical data in the system?** 11 users reported that they liked comparing different scenarios of statistical data this could be an important feature of the system. they think these features can help them in evaluating the output results and facilitate the decision making process Fig.(4.14).

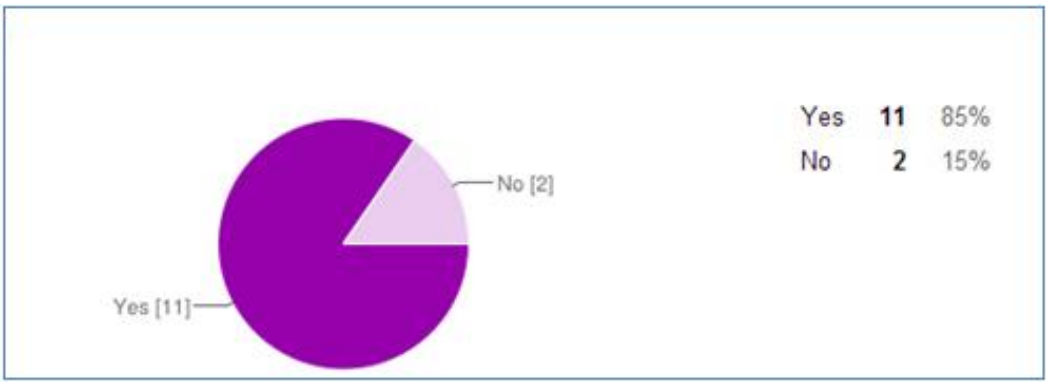



Figure 4.14: Users who are interested in comparing, filtering, and sorting different scenarios of statistical data in the system.

## 4.4 Interview with end users to evaluate the visualization and interaction techniques in the system

The first interview and feedback questionnaire was aimed to get requirements before visualization implementation. The second interview and feedback questionnaire was aimed to get a feedback about the whole system and visualization techniques used and implemented in general. a structured interview with pre-defined questions performed. 12 participants were asked to use the visualization techniques of the system to perform a number of preselected tasks that are related to participants work context. Participants were asked to think aloud while carrying out the tasks. At the end participants were asked to fill in a feedback questionnaire on the overall use of the system and estimate their answers from 1-"absolutely disagree" to 5-"absolutely agree". All questions were carefully selected in order to make conclusions about effectiveness, efficiency of the system, and the level of the user satisfaction (see appendix 5). All of those people are Statisticians, researchers, decision makers so that they could really estimate whether the new visualization system capability are effective and efficient enough for solving certain tasks from that research field. In the interview there were also **open questions**, where the participants of the evaluation could write their opinions and comments. The answers on those questions are very important.

Summarizing the results of the interview done with the help of feedback questionnaire, the following conclusions about effectiveness, efficiency, and user satisfaction of the new visualization system:

- Almost 83% of the participants were absolutely agree, that in order to solve the tasks the system was easy to use as shown in Fig.(4.15). Also those people stated that they felt confident about the results they have got after some tasks were accomplished Fig.(4.16).

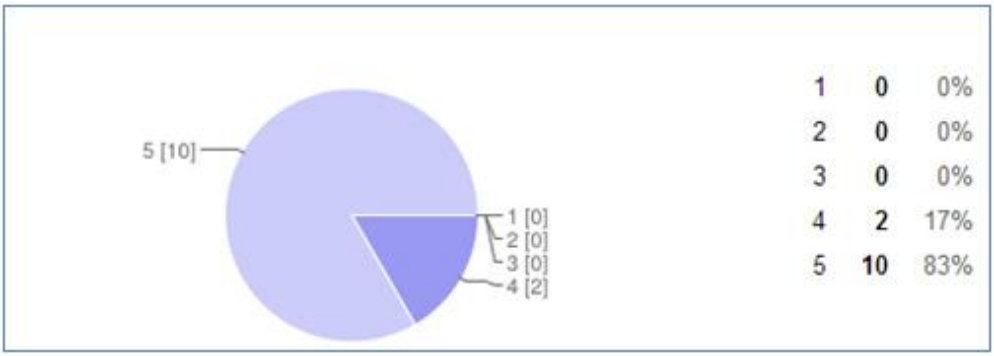



Figure 4.15: Users who are absolutely agree that the system easy to use.

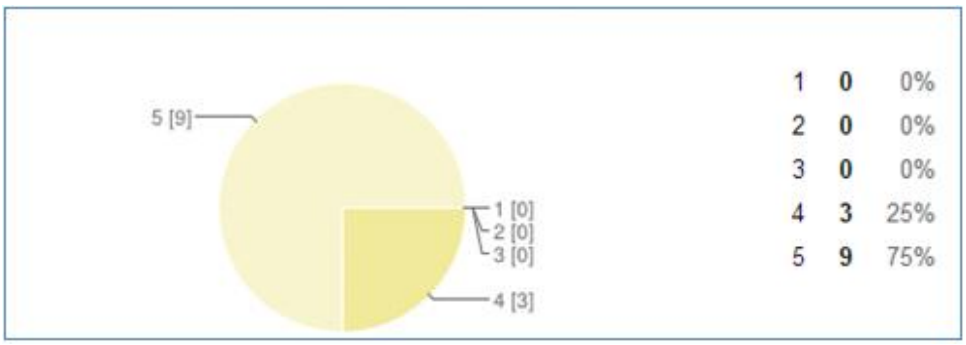



Figure 4.16: Users who feel very confident using the system.

.

- Almost 67% of the participants absolutely agree and liked the integration of the different views of the structure of the data. and 33% of the participants agree Fig.(4.17)

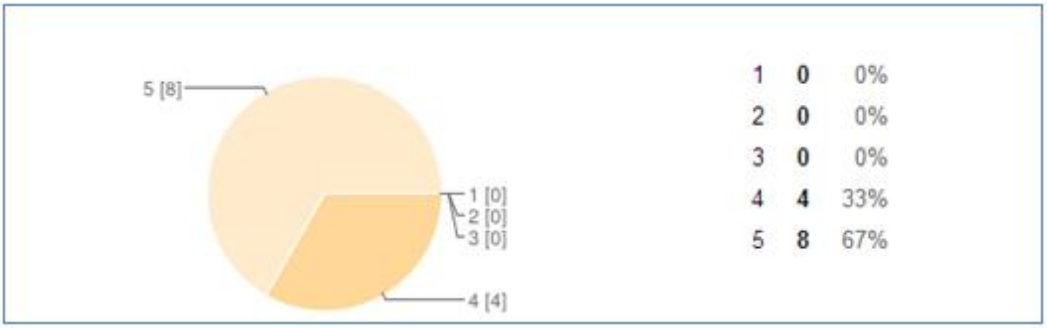



Figure 4.17: Users who are absolutely agree that functions and techniques in the system are well integrated.

- Almost 42% of the participants of the evaluation were absolutely disagree that they should spend much time to complete the tasks or that they were often confused during the tasks accomplishment, and 42% also disagree Fig.(4.18) and Fig.(4.19).

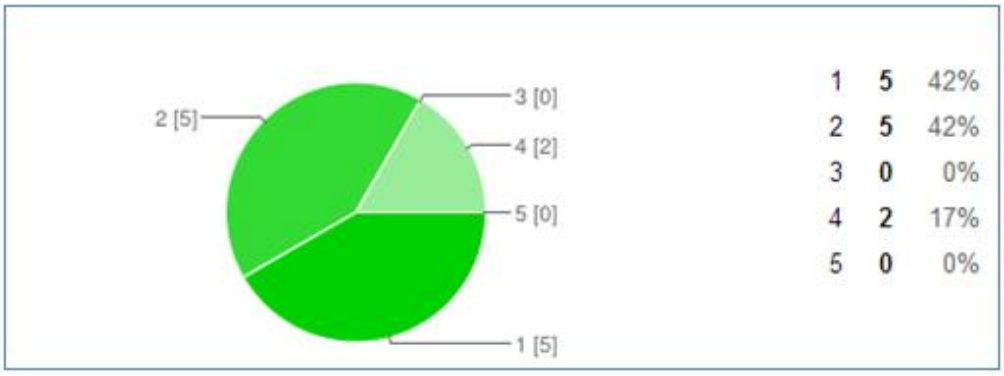



Figure 4.18: Users who are absolutely disagree and disagree that spend much time in order to accomplish the tasks.

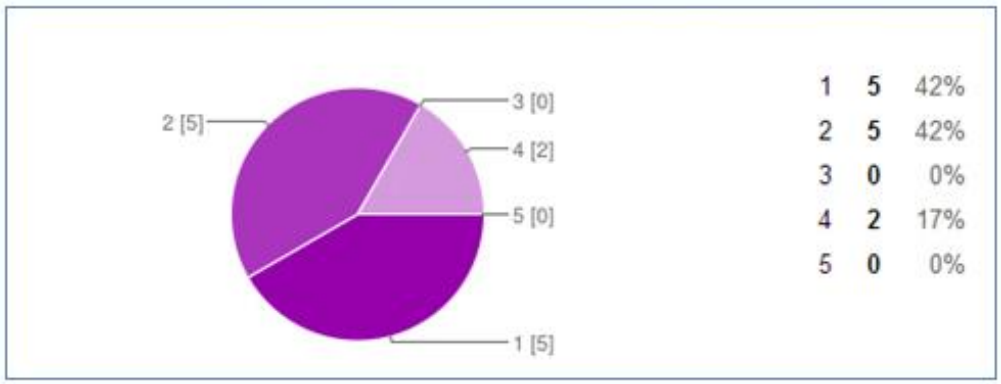



Figure 4.19: Users who are absolutely disagree and disagree that they were often confused during the tasks accomplishment.

Such answers show a high level for both effectiveness and efficiency of the system. Moreover the majority of the evaluation participants were satisfied with the system, because it has helped them to accomplish all tasks successfully. As an approval to the conclusions, stated above, it is important to mention that almost the half of the participants have given their positive feedback also in comments. They have appreciated:

- Well designed GUI,
- Good usability, the system is very exciting because it's easy to use, Where it's can help everyone according to their education levels because it contains more than one type of view, as well as filtering, comparison and filtering, fast response of the system,
- Nice visualization representation.

**Also there were some requests for additional functionality:**

- It would be great if the system can do some basic statistical analysis,
- It will be very useful if there is an option to add more than one indicator in the same chart (where its applicable and could be executed), like when we need to show more than one variable in the same line chart (like if we need to show the values through time series for Number of household and the average household size at the same year, or the population and the number of unemployed individuals),
- Supported of different languages,
- Animation and color gradient,
- More tooltips in order to explain the meaning of indicators

The analysis of the answers concerning the system in general has also shown a high level of effectiveness, efficiency, as well as of the user satisfaction. All of the participants were agree that the system is easy to use and that the various functions in the system are well integrated. More over most of the participants of the evaluation have stated that they felt confident using the system.

Thought the overall feedback to the visualization system was positive, there are several issues, which can be improved. The next section describes very shortly those improvements in need.

### 4.5 Possible Improvements

After the evaluation it was clear, that the appearance of the visualization should be optimized. the improvements related to the visualizations could be:

- Increasing the size of a visualization, so that it could be also possible to make the font size bigger and, therefore, more readable,

- Implementing double-click for the graphical elements of a visualization. For example, in order to filter a single result,
- Implementing the help option, where the end user could get more detailed information about visualization and available traits.
- Supporting of different languages, also  It would be great if it can do some basic statistical analysis.
- For the GUI it would be a good idea to implement more tooltips for better user experience, when the end user is acquainted more quickly with the new system.

# Chapter Five

## Conclusion and Future Work

This research aimed to introduce a new mapping algorithm and visualization system for mapping statistical indicators based on common schema, heterogeneous data from different data sources integrated using the created algorithm before visual methods applied, we suggested new mapping algorithm based on hamming distance, edit distance and ontology, using our algorithm we enhanced integration and mapping of statistical indicators from different sources, the data after importing saved as RDF 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.

We enhanced presentation of official statistics based on dynamic visual user interfaces. this system has been introduced to provide techniques that make humans capable of presenting results in a meaningful and intuitive way while allowing to interact with the data. Visualizing statistical data help promote the use of statistical data for improved

planning and policy making. All visualizations were implemented with the help of Highcharts Java Script libraries and Flex visualization libraries, which has helped to make highly interactive and hence responsive for the end user visualizations.

The system was successfully evaluated by different users and experts, statisticians, researchers, decision makers. The evaluation results have shown high levels for effectiveness, efficiency, and for the user satisfaction of the system.

Future work includes focus more on data mapping using ontology. Main line of future research involves extending mapping algorithm to handle more sophisticated mappings between ontologies (i.e., non 1-1 mappings), also to focus more on mapping indicators from different sources since we focused as a case study on importing data from different excel sources (files), future work includes extending the system design to do some basic statistical analysis, this feature will support and help the decision makers, also to extend the system to integrate it with statistics tools like SPSS® and SAS® to get the benefits from these tools to enhance statistical analysis features, future work includes also improving collaboration with visualization system. 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.

**Self Reflection**

The work done in this thesis will help me to improve and enhance dissemination of statistical data in my work in Palestinian central bureau of statistics (PCBS) as a head of electronic dissemination division, the trends in PCBS according to the best practices and international standards to use visualization techniques and integration of heterogeneous data to present and disseminate statistical data, I am the head of visualization techniques

committee team in PCBS because of the experience that I get from this study work and results.

Also the trends in PCBS is to harmonize all statistical data from different ministries and agencies by building common statistical business register, this require mapping and integration the data from different ministries, I am also member of statistical business register committee to follow up on the mapping and integration issues.

The thesis help me to participate and publish papers in many statistical conferences that focus on quality of data, data visualization, data integration and dissemination.

# References


1. Palestinian Central Bureau of Statistics (PCBS): http://www.pcbs.gov.ps.
2. Thomas J. and Cook K. (2005): Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press.
3. Card S.K., Mackinlay J.D., and Shneiderman (eds) B. (1999): Readings in Information Visualization, Morgan Kaufmann Publishers.
4. Michaela D. and Peter H. (2004): Data Integration and Record Matching: An Austrian: Contribution to Research in Official Statistics, Austrian Journal of Statistics, Vol. 32 (2003), Number 4, 305-321
5. Denk and Oropallo F. (2002): Overview of the Issues in Multi-Source Databases. DIECOFIS Deliverable 1.1, ISTAT, Rome.
6. Denk, Inglese F. and Calza M.G. (2003): Assessment of Different Approaches for the Integration of Sample Surveys. DIECOFIS Deliverable 1.2, ISTAT, Rome.
7. Denk, Inglese F. and Oropallo F. (2003): Report on Statistical Indicators for the Assessment of Multi-source Databases. DIECOFIS Deliverable 1.3, ISTAT, Rome.
8. IDARESA. (1997): The Data Model – Final Version, Deliverable 3.4.2, Dept. of Statistics, University of Vienna.
9. IDARESA. (1998): IDARESA Tandem Structures, TPR–viu–3.4.2/3, Dept. of Statistics, University of Vienna.
10. Denk and Froeschl K.A. (2000): The IDARESA Data Mediation Architecture for Statistical Aggregates. Research in Official Statistics. 3(1):7-38.
11. Denk, Froeschl K.A. and Grossmann W. (2002): Statistical Composites: A Transformation bound Representation of Statistical Datasets. In J. Kennedy, editor, Proc. 14th Int. Conf. Scientific and Statistical Database Management (Edinburgh, UK), pages 217- 226. IEEE Computer Society Press, Los Alamitos.
12. Froeschl A. (2004): A Sketch of Statistical Meta-Computing as a Data Integration Framework. Austrian Journal of Statistics, Volume 33 (2004), Number 1&2, 173-194
13. Fekete, J.D. (2004): The infovis toolkit. In INFOVIS '04: Pro-ceedings of the IEEE Symposium on Information Visualization, pages 167-174, Washington, DC, USA. IEEE Computer Society.
14. Heer, J., Card, S. K., and Landay, J. A. (2005): prefuse a toolkit for interactive information visualization. In CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 421- 430, New York, NY, USA. ACM.
15. Takatsuka, M. and Gahegan, M. (2002): Geovista studio: A codeless visual programming environment for geoscientific data analysis and visualization. The Journal of Computers & Geosciences, 28(10):pp.1131-1144.
16. Treemap java library (April 2008): http://treemap.sourceforge.net/.
17. Treemap 4.1.1 toolkit. (April 2008): http://www.cs.umd.edu/hcil/treemap/.
18. Bederson, B. B., Shneiderman, B., and Wattenberg, M. (2002): Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. ACM Trans. Graph., 21(4):833-854.
19. Shashikant, P., Mane, K., and Borner, K. (2004): A Toolkit for Large Scale Network Analysis. SLIS SLISWP-04-02, Indiana University.
20. Patrik L. et al. (2012): Web-Enabled Visualization Toolkit for Geovisual Analytics, Proceedings of SPIE, the International Society for Optical Engineering: SPIE: Electronic Imaging Science and Technology, Visualization and Data Analysis, Information Visualization, (11), 1, 22-42.

21. Mikael Jern. (2010): Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress, the 7th international conference on Cooperative design, visualization, and engineering, pp. 189-198.
22. Mikael Jern, et al. (2008): Geovisual Analytics Web-enabled Tools for Dissemination of OECD Regional Statistics.
23. Fekete J and Plaisant C. (2004): InfoVis Toolkit. In: Proceedings of the 10th IEEE symposium in Information Visualization, 167-174.
24. Voss H, Andrienko N and Andrienko G. (2000): Commongis - common access to geographically referenced data. ERCIM News, 41: 44–46.
25. GeoVista Studio: http://www.geovistastudio.psu.edu.
26. Guo D, Chen J, MacEachren AM and Liao K. (2006): A visualization system for space-time and multivariate patterns (VISSTAMP). IEEE Transactions on Visualization and Computer Graphics, 12(6): 1461-1474.
27. Jern M, Åström T. and Johansson S. (2012): GeoAnalytics tools applied to large geospatial datasets. In: Proceedings of the 12th International Conference Information Visualisation (IV08), 362-372.
28. Tominski C, Abello J and Schumann H. (2009): CGV-An interactive graph visualization system. Computers and Graphics, 33(6): 660-678.
29. Bostock M. and Heer J. (2009): Protovis: A graphical toolkit for visualization. IEEE Transactions on Visualization and Computer Graphics; 15(6): 1121-1128.
30. Protovis: http://vis.stanford.edu/protovis/
31. Flare: http://flare.prefuse.org/
32. Tableau: http://www.tableausoftware.com/
33. Stolte C, Tang D. and Hanrahan P. (2002): Polaris: A system for query, analysis, and visualization of multidimensional relational databases. IEEE Transactions on Visualization and Computer Graphics; 8(1): 52–65.
34. Shrinivasan YB and Wijk JJ. (2008): Supporting the analytical reasoning process in information visualization. In: Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, 1237-1246.
35. MacEachren AM, Brewer I. and Steiner E. (2001): Geovisualization to mediate collaborative work: Tools to support different-place knowledge construction and decision-making. In: Proceedings of the 20th International Cartographic Conference, 6-10.
36. Jern M. (2001): Smart documents for web-enabled collaboration. In: Vince JA and Earnshaw RA (eds) Digital Content Creation. New York, Springer-Verlag, pp.140-162.
37. Carr DB, White D. and MacEachren AM. (2005): Conditioned choropleth maps and hypothesis generation. Annals of the Association of American Geographers; 95(1): 32-53.
38. Guo D., Chen J., MacEachren AM and K. Liao. (2006): A visualization system for space-time and multivariate patterns (VISSTAMP). IEEE Transactions on Visualization and Computer Graphics; 12(6): 1461-1474.
39. Robinson A. (2006): Re-visualization: Interactive visualization of the progress of visual analysis. In: Proceedings of GIScience Workshop Visual Analytics & Spatial Decision Support, 1-21.
40. Keel P. (2006): Collaborative visual analytics: Inferring from the spatial organisation and collaborative use of information. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST), 137-144.

41. Viégas FB., Wattenberg M., Ham FV, Kriss J. and McKeon M. (2007): Many Eyes: a Site for Visualization at Internet Scale. IEEE Transactions on Visualization and Computer Graphics, 13(6): 1121-1128.
42. Millennium Development Goals (MDGs): http://pcbs.gov.ps/site/994/default.aspx
43. Winkler. (1990): String Comparator Metrics and Enhanced Decision Rules in the Fellegi- Sunter Model of Record Linkage. In Proc. Section on Survey Research Methods, pages 354-359. American Statistical Association.
44. Department of Statistics (Jordan): http://www.dos.gov.jo/dos_home_a/main/index.htm
45. Central Agency for Public Mobilization and Statistics (Egypt): http://capmas.gov.eg/
46. Porter and Winkler W. (1997): Approximate String Comparison and its Effect on an Advanced Record Linkage System, RR97-02, U.S. Bureau of the Census, Available at http://www.census.gov/srd/www/byyear.html.
47. Wikipedia, the free encyclopedia (2014): Hamming Distance: (http://en.wikipedia.org/wiki/Hamming_distance)
48. Wikipedia, the free encyclopedia (2014): Edit Distance: (http://en.wikipedia.org/wiki/Edit_distance)
49. Hall and Dowling G.R. (1980): Approximate String Matching. ACM Computing Surveys. 12(4):381-402.
50. Winkler. (1985): Preprocessing of Lists and String Comparison. In B. Kilss, W. Alvey, editors, Record Linkage Techniques, pages 181-187. FCSM, Washington, DC.
51. Winkler. (1990): String Comparator Metrics and Enhanced Decision Rules in the Fellegi- Sunter Model of Record Linkage. In Proc. Section on Survey Research Methods, pages 354-359. American Statistical Association.
52. W3C Markup Validation Service: http://www.w3.org/RDF/Validator/
53. 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.
54. Euzenat J. and Shvaiko P. (2007): Ontology matching. Springer, Approx. 340 p., 67 illus., Hardcover, ISBN 978-3-540-49611-3
55. 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.
56. Spaccapietra S. and Parent C. (1991): "Conflicts and correspondence assertions in interoperable databases," SIGMOD Record, vol. 20, no. 4, pp. 49–54.
57. Rahm E. and Bernstein P. (2001): "A survey of approaches to automatic schema matching," The VLDB Journal, vol. 10, no. 4, pp. 334–350.
58. 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.
59. Bellahsene Z., Bonifati A. and Rahm E. (2011): Schema Matching and Mapping. Springer.
60. Thomas J., Cook K. (2005): Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE Press, Los Alamitos.
61. Shneiderman, B. (1996): The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE Symposium on Visual Languages, pp. 336–343.
62. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H. (2006): Challenges in visual data analysis. In: Information Visualization (IV 2006), London, United Kingdom, July 5-7. IEEE, Los Alamitos.
63. Aigner W., Miksch S., Müller W., Schumann H., and Tominski C. (2007): Visualizing time-oriented data: A systematic view. Computers & Graphics, 31(3):401–409.

64. Aigner W., Miksch S., Müller W., Schumann H., and Tominski C. (2008): Visual methods for analyzing time-oriented data. IEEE Trans. Visualization and Computer Graphics, 14(1):47–60.
65. Müller W. and Schumann H. (2003): Visualization methods for time-dependent data – an overview. In Proc. Winter Simulation Conf., pages 737–745.
66. Andrienko N. and Andrienko G. (2006): Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach. Springer.
67. Highcharts library written in pure JavaScript: http://www.highcharts.com.
68. Flex visualization libraries: https://code.google.com/p/flexlib
69. Economic Commission for Europe of the United Nations (UNECE). (2000), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva.
70. Navarro, G. and Raffinot, M. (2002): Flexible pattern matching in strings: practical on-line search algorithms for texts and biological sequences. Cambridge University Press New York, NY, USA, Pages 221.
71. Gil, Y. et al. (2005): A String Metric for Ontology Alignment, ISWC 2005, LNCS 3729, pp. 624–637, 2005.
72. Data Visualization: http://www.cs.uic.edu/~kzhao/Papers/00_course_Data_visualization.pdf
73. Keim, A., Andrienko G. (2008): Visual analytics: Definition, process, and challenges. Information Visualization, 154–175.
74. Daniel A. Keim et al, (2008): Visual Analytics: Combining Automated Discovery with Interactive Visualizations, Discovery Science, pp. 2-14.

## Appendix 1: Summary of literature review contributions

| Study Title | Authors, Year | Description of study design methodology | Type of research | Limitations | Intervention | Results |
|---|---|---|---|---|---|---|
| Data Integration: Techniques and Evaluation | Michaela Denk and Peter Hackl, 2004 | Building statistical Data integration framework, record matching, and the generation of multi-source databases that are used as a basis of micro simulations are a core issue of the project, An empirical study has been designed that compares the applicability of various integration procedures. | Analysis, Evaluation | semantic discrepancies and similarities of data sources need to be analyzed before the application of statistical methods generally and integration methods in particular: Data source integration as a prerequisite of dataset integration. | The project showed the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making. | A system of micro-founded indicators was built, 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. |
| The Development of an Integrated and Systematized Information System for Economic and Policy Impact Analysis | Filippo Oropallo and Francesca Inglese, 2004 | database integration was concerned with the creation of the Survey datasets, Administrative datasets, Business Register. The integration of administrative and survey data was performed by exact matching when the same unit was identified | Analysis | Evaluation needed and using ontology in addition to integration using matching. | addressing the integration problems that have been faced in reconciling administrative and survey sources and combining them into one multi-source database. | Creation of a multi-source integrated and systematised data base of enterprises data. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | otherwise it was performed by statistical matching techniques | | | | |
| Web-Enabled Visualization Toolkit for Geovisual Analytics | Patrik Lundblad et al. - 2012 | A framework and class library (GAV Flash) implemented in Adobe ActionScript, import statistical data through Excel – data model, create a story and visualization using analytic tools (dynamic query, filter, regional categorization, profiles, highlight), and dynamic colour scale, then share the story. | Analysis | No integration and mapping of heterogeneous data into a common schema | Interactive visualization (filtering, selection, brushing, zooming), Storytelling and dealing with large datasets. | A statistics geovisual analytics application for exploring and publishing statistical data on the web, developed with the GAV Flash toolkit was built. based on a recommendation from the visual analytics (VA) research program. |
| Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress | Mikael Jern - 2010 | Web-enabled application based on Adobe's object-oriented language ActionScript and includes a collection of common geo- and information Visualization representations and Flash class library | Analysis | No integration and mapping of heterogeneous data into a common schema | Web-enabled application for exploring and communicating statistics data using storytelling mechanism | Web-enabled application platform that is emerging as a de facto standard in the statistics community for exploring and communicating statistics data |
| Geovisual Analytics Web-enabled Tools | Mikael Jern, et al. 2008 | GeoAnalytics Visualization (GAV) component toolkit is | Analysis, Evaluation | include a more comprehensive user task | Tools for interactively analyzing and | exchange of statistical measures and knowledge through innovative Geovisual |

| | | | | | | |
|---|---|---|---|---|---|---|
| for Dissemination of OECD Regional Statistics | | based on the principles behind the Visual Analytics re-search program, using Adobes Flash basic graphics and Flex 3 for user interface design (a collection of high-performance interactive visualization web-enabled components based on common methods from the information and geovisualization research domain). | | analysis. explore trends over time (yearly time series) for the indicators in the regional database. | communicating gained insights and discoveries about spatial-temporal and multivariate OECD regional data. | Analytics techniques. |
| Visualization, Road Weather | Patrik, et al., 2011 | The data visualized by RoadVis consists of weather observations from 770 automated observation stations around Sweden collected every half hour as well as a long 24 hour and a short six hour forecasts for each station that is made every hour. Using threading and dynamic queries the application is constantly updated with the latest data in the background | Analysis | | Analyzing and communicating information about road weather conditions, particularly during the Swedish winter months. | Framework that is used to analyze and make decisions, often in time-critical situations, on the large and ever-increasing amounts of time-varying and geospatial digital weather information related to emergency scenarios. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | and the analyst can still use the application during their vital work to keep the roads safe. | | | | |
| Information Visualization in Climate Research | Christian Tominski, et al. 2010 | A survey that conducted to evaluate the application of interactive visualization methods and to identify the problems related to establishing such methods in scientific practice, 76 participants. | Evaluation | Extend existing solutions and to Integrate additional tools. | Illustrate how interactive visualization tools can be successfully applied to accomplish climate research tasks. | Climate researchers have begun to recognize information visualization as a valuable tool. Based on a list of requirements. |
| The InfoVis Toolkit | Jean-Daniel Fekete 2004 | Create specific data structures first, then apply or create visualizations, The InfoVis Toolkit consists of approximately 30,000 lines of Java and a 300K Jar file. It is currently licensed under the QPL and available at: http://www.lri.fr/~fekete/InfovisToolkit. | Analysis | Implement mechanisms to support animation and continuous monitoring for time-oriented visualizations needed. | A toolkit that supports the development and extension of 2D Information Visualization components and applications using Java and Swing. | InfoVis Toolkit, designed to support the creation, extension and integration of advanced 2D Information Visualization components into interactive Java Swing applications. The InfoVis Toolkit provides specific data structures to achieve a fast action/feedback loop required by dynamic queries. |
| Prefuse: a toolkit for interactive information visualization | Jeffrey Heer, et. al. 2005 | Create framework for creating dynamic visualizations, representing abstract data, mapping data into an intermediate, visualizable | Analysis | Introduce more powerful operations for manipulating source data, provide | A framework of higher-level abstractions for presentation, navigation, and batch processing of | Prefuse, a software framework for creating dynamic visualizations of both structured and unstructured data. prefuse includes a library of |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | form, and then using these visual analogues to provide interactive displays, usability studies and uage surveys. | | additional components. | interactive objects that simplifies visualization creation while affording the freedom to explore new designs. | layout algorithms, navigation and interaction techniques, integrated search, and more. |
| GeoVISTA Studio: A Codeless Visual Programming Environment For Geoscientific Data Analysis And Visualization | Masahiro Takatsuka and Mark Gahegan 2002 | Two example applications are presented to illustrate the potential of the Studio environment for exploring and better understanding large, complex geographical datasets and for supporting complex visual and computational analysis. | Analysis | Use this environment to seek a deeper understanding of the kinds of tools required for effective knowledge construction, including how these tools should best interact with each other, and with the user, to provide a coordinated system of analysis. | GeoVISTA Studio project to improve geoscientific analysis by providing an environment that operationally integrates a wide range of analysis activities, including those both computationally and visually based, also for exploratory data analysis, knowledge discovery, and other data modeling and visualization issues. | GeoVISTA: Supporting complex visual and computational analysis, GeoVISTA Studio a codeless visual programming environment that supports rapid construction of sophisticated geoscientific data analysis and visualization programs. |
| A Toolkit for Large Scale Network Analysis | Shashikant. et al., 2004 | The code library is programmed in Java with Perl-CGI for the front-end providing fast, efficient | Analysis | Providing a large number of popular and useful algorithms to be | The toolkit and the associated web interface provide an extremely user-friendly manner to | a toolkit for large scale network analysis. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | and scalable system. The underlying classes and methods are written in a manner so as to facilitate the easy extension of the library. | | applied to large-scale networks and a major visualization component. | obtain network analysis results. | |
| Treemaps for space-onstrained visualization of hierarchies. | Ben Shneiderman, 2009. | Tree structured node-link diagrams grew too large to be useful, so he explored ways to show a tree in a space-constrained layout. splitting the screen into rectangles in alternating horizontal and vertical directions as you traverse down the levels. Using recursive algorithm. | Analysis | Evaluation the implemented algorithms. | a compact visualization of directory tree structures. Since the 80 Megabyte hard disk in the HCIL was shared by 14 users it was difficult to determine how and where space was used. Finding large files that could be deleted, or even determining which users consumed the largest shares of disk space were difficult tasks. | Producing a compact visualization of directory tree structures. 5 treemap algorithms implemented. |

## Appendix 2: Mapping Algorithm C# code using Hamming and Edit Distance.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;

namespace DistanceBetweenStrings
{
   class Algorithms
   {
        /// <summary>
   /// Compute the distance between two strings.
   /// </summary>
   public static int LevenshteinDistanceCompute(string s, string t)
   {
         int n = s.Length;
         int m = t.Length;
         int[,] d = new int[n + 1, m + 1];

         // Step 1
         if (n == 0)
         {
            return m;
         }

         if (m == 0)
         {
            return n;
         }

         // Step 2
         for (int i = 0; i <= n; d[i, 0] = i++)
         {

         }

         for (int j = 0; j <= m; d[0, j] = j++)
         {
         }

         // Step 3
         for (int i = 1; i <= n; i++)
         {
            //Step 4
            for (int j = 1; j <= m; j++)
            {
                  // Step 5
                  int cost = (t[j - 1] == s[i - 1]) ? 0 : 1;
```

```
                // Step 6
                d[i, j] = Math.Min(
                    Math.Min(d[i - 1, j] + 1, d[i, j - 1] + 1),
                    d[i - 1, j - 1] + cost);
            }
        }
        // Step 7
        return d[n, m];
    }
    public static int HammingDistanceCompute(string s, string t)
    {
     return  s.Zip(t, (c1, c2) => c1 == c2 ? 0 : 1).Sum();
    }
    }
}
```

## Appendix 2: Mapping Algorithm C# code to check indicator length and return the distance value (Cont.)

```
using System;
using System.Collections.Generic;
using System.Data;
using BusinessObjects;
using DataObjects;
using System.Transactions;
using DistanceBetweenStrings;

namespace ServiceLibrary
{
    class IndicatorServiceLibrary
    {
      static readonly IndicatorDao indicatorDao = new IndicatorDao();

      public bool CheckIndicator(string indicatorName)
       {
          return indicatorDao.CheckIndicatorByName(indicatorName);

       }
       public Indicator GetIndicatorByName(string indicatorName)
       {
          return indicatorDao.GetIndicatorByName(indicatorName);
       }
       public Indicator GetIndicatorByID(long ID)
       {
          return indicatorDao.GetIndicatorByID(ID);
       }
       public List<DistanceString> GetDistanceStrings(string currentIndicator)
       {
```

```
        List<DistanceString> distanceStrings =new List<DistanceString>();
        var indicators= indicatorDao.GetIndicators();

        foreach (Indicator indicator in indicators)
        {
           if (currentIndicator.Length == indicator.IndicatorName.Length)
           {
              var
distanceValue=Algorithms.HammingDistanceCompute(currentIndicator,indicator.Indicato
rName);
              distanceStrings.Add(
               new DistanceString
              {
                 stringName=indicator.IndicatorName,
                 DistanceValue = distanceValue,
                 AccuracyPercent =
float.Parse(((float.Parse((indicator.IndicatorName.Length - distanceValue).ToString())
                              / (float)(indicator.IndicatorName.Length)) *
100).ToString("F2")),
                 AlgorthemName="Hamming Distance"
              });

           }
           else
           {
              var distanceValue =
Algorithms.LevenshteinDistanceCompute(currentIndicator,indicator.IndicatorName);
              var LargDistance= (indicator.IndicatorName.Length >=
currentIndicator.Length ) ?
                           indicator.IndicatorName.Length :
                           currentIndicator.Length;
               distanceStrings.Add(
               new DistanceString
              {
                 stringName=indicator.IndicatorName,
                 DistanceValue = distanceValue,
                 AccuracyPercent = float.Parse(((float.Parse((LargDistance -
distanceValue).ToString())
                              / (float)(LargDistance)) * 100).ToString("F2")),
                 AlgorthemName="Levenshtei Distance"
              });

           }

        }
        return distanceStrings;

     }
  }
}
```

## Appendix 2: Mapping Algorithm C# ontology code (Cont.)

```
using System;
using System.Collections.Generic;
using System.Data;
using BusinessObjects;

namespace DataObjects
{
   class IndicatorOntologyDao
   {
      public List<IndicatorOntology> GetIndicatorOntologys(string sortExpression =
"IndicatorOntologyId ASC")
      {
         String SqlQuery =
          @"SELECT IndicatorOntologyID,IndicatorID, IndicatorOntologyName
               FROM [IndicatorOntologys]".OrderBy(sortExpression);

         return Db.ReadList(SqlQuery, Make);
      }

      public IndicatorOntology GetIndicatorOntologyByID(long indicatorOntologyID)
      {
         String SqlQuery =
          @"SELECT IndicatorOntologyID,IndicatorID, IndicatorOntologyName
               FROM [IndicatorOntologys]
              WHERE IndicatorOntologyID=@IndicatorOntologyID";

         object[] parms = { "@IndicatorOntologyID", indicatorOntologyID };
         return Db.Read(SqlQuery, Make, parms);
      }
      public IndicatorOntology GetIndicatorOntologyByName(string
indicatorOntologyName)
      {
         String SqlQuery =
          @"SELECT IndicatorOntologyID,IndicatorID, IndicatorOntologyName
               FROM [IndicatorOntologys]
              WHERE IndicatorOntologyName=@IndicatorOntologyName";

         object[] parms = { "@IndicatorOntologyName", indicatorOntologyName };
         return Db.Read(SqlQuery, Make, parms);
      }

      public bool CheckIndicatorOntologyByName(string indicatorOntologyName)
      {
         String SqlQuery =
          @"SELECT IndicatorOntologyID,IndicatorID, IndicatorOntologyName
               FROM [IndicatorOntologys]
              WHERE IndicatorOntologyName=@IndicatorOntologyName";
```

```
    object[] parms = { "@IndicatorOntologyName", indicatorOntologyName };
    return Db.Check(SqlQuery, parms);
}

public void InsertIndicatorOntology(IndicatorOntology indicatorOntology)
{
    string sql =
    @"INSERT INTO [UnitOntology] (IndicatorID,IndicatorOntologyName )

      VALUES (@IndicatorID,@IndicatorOntologyName)";

    Db.Insert(sql, Take(indicatorOntology));

}

public void UpdateIndicatorOntology(IndicatorOntology indicatorOntology)
{
    string sql =
    @"UPDATE [UnitOntology]
          SET IndicatorID = @IndicatorID
         ,IndicatorOntologyName = @IndicatorOntologyName";

    Db.Update(sql, Take(indicatorOntology));
}

public void DeleteIndicatorOntology(IndicatorOntology indicatorOntology)
{
    string sql =
    @"DELETE FROM [UnitOntology]
      WHERE IndicatorOntologyId = @IndicatorOntologyId";

    Db.Update(sql, Take(indicatorOntology));
}

/// <summary>
/// Creates a new Incident object based on DatIndicatorOntologyder.
/// </summary>
private static Func<IDataReader, IndicatorOntology> Make = reader =>
  new IndicatorOntology
  {
     IndicatorOntologyID = reader["IndicatorOntologyID"].AsLong(),
     IndicatorID = reader["IndicatorID"].AsLong(),
     IndicatorOntologyName = reader["IndicatorOntologyName"].AsString()
  };

/// <summary>
/// Creates query parameters list from IndicatorOntology object
/// </summary>
/// <param name="indicatorOntology">IndicatorOntology.</param>
/// <returns>Name value parameter list.</returns>
```

```
        private object[] Take(IndicatorOntology indicatorOntology)
        {
            return new object[]
            {
                "@IndicatorOntologyID", indicatorOntology.IndicatorOntologyID,
                "@IndicatorID", indicatorOntology.IndicatorID,
                "@IndicatorOntologyName", indicatorOntology.IndicatorOntologyName,

            };
        }

    }
}

namespace ServiceLibrary
{
    class IndicatorOntologyServiceLibrary
    {
        static readonly IndicatorOntologyDao indicatorOntologyDao = new
IndicatorOntologyDao();

        public bool CheckIndicatorOntology(string indicatorOntologyName)
        {
            return
indicatorOntologyDao.CheckIndicatorOntologyByName(indicatorOntologyName);

        }
        public List<IndicatorOntology> GetIndicatorOntologys(string sortExpression =
"IndicatorOntologyID ASC")
        {
          return   indicatorOntologyDao.GetIndicatorOntologys(sortExpression);
        }

        public IndicatorOntology GetIndicatorOntologyByID(long indicatorOntologyID)
        {
            return indicatorOntologyDao.GetIndicatorOntologyByID(indicatorOntologyID);
        }

        public IndicatorOntology GetIndicatorOntologyByName(string
indicatorOntologyName)
        {
            return
indicatorOntologyDao.GetIndicatorOntologyByName(indicatorOntologyName);
        }

        public void InsertIndicatorOntology(IndicatorOntology indicatorOntology)
        {
            indicatorOntologyDao.InsertIndicatorOntology(indicatorOntology);

        }
```

```
        public void UpdateIndicatorOntology(IndicatorOntology indicatorOntology)
        {
            indicatorOntologyDao.UpdateIndicatorOntology(indicatorOntology);
        }

        public void DeleteIndicatorOntology(IndicatorOntology indicatorOntology)
        {
            indicatorOntologyDao.DeleteIndicatorOntology(indicatorOntology);
        }
    }
}
```

# Appendix 3: Summary of Mapping Results Tables

Table 4.1: Summary of mapping results for importing different indicators (random indicators from different countries).

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | True Mapping should be | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Non- refugees | Internet Users | False | Non- refugees population | 9 | 35.71 |
| Urbanization Level | Population Size | False | Level of urbanization | 11 | 38.89 |
| Proportion of land area covered by forest | Land area covered by forest | True | | 15 | 63.41 |
| Growth rate of GDP /person employed | Growth rate of GDP per person employed | True | | 3 | 92.11 |
| Antenatal care coverage | Adult Literacy Rate | False | Antenatal care coverage for at least one visit | 15 | 34.78 |
| Antenatal care coverage for at least 1 visit | Antenatal care coverage for at least one visit | True | | 3 | 93.48 |
| Net enrolment ratio in primary education | Net enrolment ratio in basic education | True | | 7 | 82.5 |
| Fixed Tel. lines | Telephone lines | True | | 9 | 43.75 |
| Growth rate of GDP | Sex Ratio | False | Growth rate of GDP per person employed | 13 | 27.78 |
| Growth rate of GDP per person employed | Growth rate of GDP per person employed | True | | Exact Match | 100 |
| Proportion of 1 year-old children immunised against measles | Children 1 year-old immunized against measles | True | | 21 | 64.41 |
| Proportion of the population using improved drinking water sources, urban | Proportion of population using an improved drinking water source | True | | 15 | 79.45 |

Table 4.1: Summary of mapping results for importing different indicators (random indicators from different countries). (Cont.)

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | True Mapping should be | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Emp-to-pop. Ratio | Separated Population | False | Employment-to-population ratio | 12 | 40 |
| Marine areas protected, sq. km. | Land area covered by forest | False | Proportion of terrestrial and marine areas protected | 20 | 35.48 |
| Marine areas protected to territorial waters, percentage | Land area covered by forest | False | Proportion of terrestrial and marine areas protected | 37 | 33.93 |
| Non- refugees Palestinian population | Non- refugees population | True | | 12 | 66.67 |
| Terrestrial and Marine areas protected, sq. km. | Proportion of terrestrial and marine areas protected | True | | 24 | 53.85 |
| Population aged 25 to 29 yr | Population aged 25-29 years | True | | 8 | 70.37 |
| Widowed | Buildings | False | Widowed population | 7 | 22.22 |
| Never married | Sex ratio | False | Never married population | 9 | 30.77 |
| Married | Sex ratio | False | Married population | 8 | 11.11 |
| Ratio of literate female to male | Ratio of literate women to men age 15-24 years | True | | 21 | 54.35 |
| Gap in literacy rate by residence | Gap in literacy rate by residence (Urban-Rural) | True | | 14 | 70.21 |
| Children attending school | Children aged 6-10 years attending school | True | | 16 | 60.98 |
| Child workers | Child workers ratio | False | Child workers population | 6 | 68.42 |

Table 4.1: Summary of mapping results for importing different indicators (random indicators from different countries). (Cont.)

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | True Mapping should be | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Age dependency | Age dependency ratio | True | | 6 | 70 |
| Population aged 0-4 yrs | Population aged 0-4 years | True | | 2 | 92 |
| Widowed Palestinian population | Widowed population | True | | 12 | 60 |
| Type of household private | Type of household private /nuclear | False | Type of household private /One person | 9 | 73.53 |
| Fertility rate | Total fertility rate | True | | 6 | 70 |
| Type of household private Per One person household | Type of household private /One person | True | | 15 | 70 |
| Population aged above 80 years | Population aged 80+ years | True | | 7 | 76.67 |
| Population aged 0-4 yr | Population aged 0-4 years | True | | 3 | 88 |
| Mean age at marriage | Mean age at marriage | True | | Exact Match | 100 |
| Never married Palestinian population | Never married population | True | | 12 | 66.67 |
| Married Palestinian population | Married population | True | | 12 | 60 |
| Households headed by men | Households headed by males | True | | 3 | 88.46 |
| Households headed by women | Households headed by males | False | Households headed by Females | 4 | 84.62 |
| Seats in national parliament | Seats held by women in national parliament | True | | 14 | 66.67 |
| Current contraceptive use among married women 15-49 years old, modern methods, percentage | Proportion of pupils starting grade 1 who reach last grade of primary | False | Contraceptive prevalence rate | 63 | 29.21 |
| Current contraceptive use among married women 15-49 years old, condom, percentage | Contraceptive prevalence rate | True | | 58 | 28.4 |
| Unmet need for family planning, spacing, percentage | Unmet need for family planning | True | | 21 | 58.82 |
| Unmet need for family planning, | Unmet need for family planning | True | | 22 | 57.69 |

| | | | | | |
|---|---|---|---|---|---|
| limiting, percentage | | | | | |

Table 4.1: Summary of mapping results for importing different indicators (random indicators from different countries). (Cont.)

| **Imported Indicator** | **Mapping to Indicator in the Schema** | **Mapping (True or False)** | **True Mapping should be** | **Minimum Distance Value** | **Accuracy Percent** |
|---|---|---|---|---|---|
| Proportion of births attended by skilled health personnel | Births attended by skilled health personnel | True | | 15 | 73.68 |
| Mortality rate | Infant mortality rate | False | Under-five mortality rate | 7 | 66.67 |
| Infant mortality | Infant mortality rate | True | | 5 | 76.19 |
| Under-5 mortality rate | Under-five mortality rate | True | | 4 | 84 |
| Literacy rate of 15-24 year-olds, women and men | Literacy rate of 15-24 year-olds | True | | 15 | 68.09 |
| Employment to population ratio | Employment-to-population ratio | True | | 2 | 93.33 |
| Prevalence of underweight children under-five years of age | Prevalence of underweight (moderate) | False | Needs ontology | 28 | 51.72 |
| Prevalence of underweight children under-five years of age | Prevalence of underweight (severe) | False | Needs ontology | 28 | 51.72 |
| Prevalence of underweight-moderate | Prevalence of underweight (moderate) | True | | 3 | 91.67 |
| Prevalence of underweight | Prevalence of underweight (severe) | False | Needs ontology | 9 | 73.53 |
| household size | Average household size | True | | 8 | 63.64 |
| Population | Population size | True | | 5 | 66.67 |
| Persons completed education | Persons age 15+ completed primary education | False | Needs ontology | 16 | 62.79 |
| Population aged 40 to 44 yr. | Population aged 40-44 years | True | | 7 | 75 |

Table 4.2: Summary of mapping results for importing different indicators units (random units for different indicators).

| Imported Unit | Mapping to Unit in the Schema | Mapping (True or False) | Minimum Edit Distance Value | Accuracy Percent |
|---|---|---|---|---|
| Birth per woman | Birth per woman | True | Exact Match | 100 |
| Birth/woman | Birth per woman | True | 5 | 68.57 |
| Percent | Percent | True | Exact Match | 100 |
| % | US$ | False | 3 | 0 |
| Percentage | Percent | True | 3 | 70 |
| Metric tons | Metric tons | True | Exact Match | 100 |
| Tons | Rate | False | 4 | 0 |
| Per 100 population | Per 100 population | True | Exact Match | |
| Per 100 pop. | Per 100 population | True | 7 | 61.11 |
| Per 100,000 population | Per 100,000 population | True | Exact Match | 100 |
| Per 100000 population | Per 100,000 population | True | 1 | 95.45 |
| Per 100,000 pop. | Per 100,000 population | True | 7 | 68.18 |
| Deaths/1000 live births | Deaths per 1000 live births | True | 5 | 81.48 |
| Number | Number | True | Exact Match | 100 |
| Num. | Number | True | 3 | 50 |
| Percentage points | Percent | True | 10 | 41.18 |
| Females per 100 males | Females per 100 males | True | Exact Match | 100 |
| Females/100 males | Females per 100 males | True | 5 | 76.19 |
| F/100 M | Number | False | 7 | 0 |
| Persons/sq km | Persons per sq km | True | 5 | 70.59 |
| Years | Years | True | Exact Match | 100 |
| yr | US$ | False | 3 | 0 |

Table 4.3: Summary of mapping results for importing different indicators subgroups (random subgroups for different indicators)

| Imported Subgroup | Mapping to Subgroup in the Schema | Mapping (True or False) | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|
| Female | Female | True | Exact Match | 100 |
| Male | Male | True | Exact Match | 100 |
| F | Male | False | 4 | 0 |
| M | Male | True | 3 | 25 |
| Total | Total | True | Exact Match | 100 |
| <5 yr | <5 yr | True | Exact Match | 100 |
| <5 year | <5 yr | True | 2 | 71.43 |
| 0-14 yr | 0-14 yr | True | Exact Match | 100 |
| 0-14 year | 0-14 yr | True | 2 | 77.78 |
| 1 yr | 1 yr | True | Exact Match | 100 |

Table 4.3: Summary of mapping results for importing different indicators subgroups (random subgroups for different indicators) (Cont.)

| Imported Subgroup | Mapping to Subgroup in the Schema | Mapping (True or False) | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|
| 1 year | 1 yr | True | 2 | 66.67 |
| One yr | <5 yr | False | 3 | 50 |
| One year | <5 yr | False | 5 | 37.5 |
| Female 1 yr | Female 1 yr | True | Exact Match | 100 |
| Female 1 year | Female 1 yr | True | 2 | 84.62 |
| F 1 yr | 1 yr | False | 2 | 66.67 |
| Female One yr | Female 1 yr | True | 3 | 76.92 |
| Male 1 yr | Male 1 yr | True | Exact Match | 100 |
| M 1 yr | 1 yr | False | 2 | 66.67 |
| Male One yr | Male 1 yr | True | 3 | 72.73 |
| Male 1 year | Male 1 yr | True | 2 | 81.82 |
| M 1 year | 1 yr | False | 4 | 50 |
| 1 yr Rural | 1 yr Rural | True | Exact Match | 100 |
| 1 year Rural | 1 yr Rural | True | 2 | 83.33 |
| One yr Rural | 1 yr Rural | True | 3 | 75 |
| One year Rural | 1 yr Rural | True | 5 | 64.29 |
| Female <5 yr | Female <5 yr | True | Exact Match | 100 |
| Female <5 year | Female <5 yr | True | 2 | 85.71 |
| F <5 yr | <5 yr | False | 3 | 62.5 |
| F <5 year | <5 yr | False | 4 | 55.56 |
| Male <5 yr | Male <5 yr | True | Exact Match | 100 |
| M <5 yr | <5 yr | False | 2 | 71.43 |
| Male <5 year | Male <5 yr | True | 2 | 83.33 |
| Male less than 5 yr | Male <5 yr | True | 10 | 47.37 |
| Male less than 5 year | Male <5 yr | True | 12 | 42.86 |
| 15-49 yr | 15-49 yr | True | Exact Match | 100 |
| 15-49 year | 15-49 yr | True | 2 | 80 |
| 15 to 49 yr | 15-49 yr | True | 4 | 63.64 |
| 15-19 yr Total | 15-19 yr Total | True | Exact Match | 100 |
| 15-19 year Total | 15-19 yr Total | True | 2 | 87.5 |
| Female 15-49 yr Rural | Female 15-49 yr Rural | True | Exact Match | 100 |
| Female 15-49 year Rural | Female 15-49 yr Rural | True | 2 | 91.3 |
| F 15-49 yr Rural | 15-49 yr Rural | False | 3 | 81.25 |
| Urban Male 12+ yr | Urban Male 12+ yr | True | Exact Match | 100 |
| Urban Male above 12 yr | Urban Male 12+ yr | True | 7 | 68.18 |

Table 4.4: Summary of mapping results for importing different indicators (random indicators from different countries) after adding ontology.

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | Ontology Mapping | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Non- refugees | Ontology Used | True | Non- refugees population | Ontology Used | 100 |
| Urbanization Level | Ontology Used | True | Level of urbanization | Ontology Used | 100 |
| Proportion of land area covered by forest | Land area covered by forest | True | | 15 | 63.41 |
| Growth rate of GDP /person employed | Growth rate of GDP per person employed | True | | 3 | 92.11 |
| Antenatal care coverage | Ontology Used | True | Antenatal care coverage for at least one visit | Ontology Used | 100 |
| Antenatal care coverage for at least 1 visit | Antenatal care coverage for at least one visit | True | | 3 | 93.48 |
| Net enrolment ratio in primary education | Net enrolment ratio in basic education | True | | 7 | 82.5 |
| Fixed Tel. lines | Telephone lines | True | | 9 | 43.75 |
| Growth rate of GDP | Ontology Used | True | Growth rate of GDP per person employed | Ontology Used | 100 |
| Growth rate of GDP per person employed | Growth rate of GDP per person employed | True | | Exact Match | 100 |
| Proportion of 1 year-old children immunised against measles | Children 1 year-old immunized against measles | True | | 21 | 64.41 |
| Proportion of the population using improved drinking water sources, urban | Proportion of population using an improved drinking water source | True | | 15 | 79.45 |
| Emp-to-pop. Ratio | Ontology Used | True | Employment-to-population ratio | Ontology Used | 100 |

Table 4.4: Summary of mapping results for importing different indicators (random indicators from different countries) after adding ontology. (Cont.)

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | Ontology Mapping | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Marine areas protected, sq. km. | Ontology Used | True | Proportion of terrestrial and marine areas protected | Ontology Used | 100 |
| Marine areas protected to territorial waters, percentage | Ontology Used | True | Proportion of terrestrial and marine areas protected | Ontology Used | 100 |
| Non- refugees Palestinian population | Non- refugees population | True | | 12 | 66.67 |
| Terrestrial and Marine areas protected, sq. km. | Proportion of terrestrial and marine areas protected | True | | 24 | 53.85 |
| Population aged 25 to 29 yr | Population aged 25-29 years | True | | 8 | 70.37 |
| Widowed | Buildings | False | Widowed population | 7 | 22.22 |
| Never married | Sex ratio | False | Never married population | 9 | 30.77 |
| Married | Sex ratio | False | Married population | 8 | 11.11 |
| Ratio of literate female to male | Ratio of literate women to men age 15-24 years | True | | 21 | 54.35 |
| Gap in literacy rate by residence | Gap in literacy rate by residence (Urban-Rural) | True | | 14 | 70.21 |
| Children attending school | Children aged 6-10 years attending school | True | | 16 | 60.98 |
| Child workers | Child workers ratio | False | Child workers population | 6 | 68.42 |
| Age dependency | Age dependency ratio | True | | 6 | 70 |
| Population aged 0-4 yrs | Population aged 0-4 years | True | | 2 | 92 |

Table 4.4: Summary of mapping results for importing different indicators (random indicators from different countries) after adding ontology. (Cont.)

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | Ontology Mapping | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Widowed Palestinian population | Widowed population | True | | 12 | 60 |
| Type of household private | Ontology Used | True | Type of household private /One person | Ontology Used | 100 |
| Fertility rate | Total fertility rate | True | | 6 | 70 |
| Type of household private Per One person household | Type of household private /One person | True | | 15 | 70 |
| Population aged above 80 years | Population aged 80+ years | True | | 7 | 76.67 |
| Population aged 0-4 yr | Population aged 0-4 years | True | | 3 | 88 |
| Mean age at marriage | Mean age at marriage | True | | Exact Match | 100 |
| Never married Palestinian population | Never married population | True | | 12 | 66.67 |
| Married Palestinian population | Married population | True | | 12 | 60 |
| Households headed by men | Households headed by males | True | | 3 | 88.46 |
| Households headed by women | Households headed by males | False | Households headed by Females | 4 | 84.62 |
| Seats in national parliament | Seats held by women in national parliament | True | | 14 | 66.67 |
| Current contraceptive use among married women 15-49 years old, modern methods, percentage | Proportion of pupils starting grade 1 who reach last grade of primary | False | Contraceptive prevalence rate | 63 | 29.21 |
| Current contraceptive use among married women 15-49 years old, condom, percentage | Contraceptive prevalence rate | True | | 58 | 28.4 |
| Unmet need for family planning, spacing, percentage | Unmet need for family planning | True | | 21 | 58.82 |
| Unmet need for family planning, limiting, percentage | Unmet need for family planning | True | | 22 | 57.69 |

Table 4.4: Summary of mapping results for importing different indicators (random indicators from different countries) after adding ontology. (Cont.)

| Imported Indicator | Mapping to Indicator in the Schema | Mapping (True or False) | Ontology Mapping | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|---|
| Proportion of births attended by skilled health personnel | Births attended by skilled health personnel | True | | 15 | 73.68 |
| Mortality rate | Ontology Used | True | Under-five mortality rate | Ontology Used | 100 |
| Infant mortality | Infant mortality rate | True | | 5 | 76.19 |
| Under-5 mortality rate | Under-five mortality rate | True | | 4 | 84 |
| Literacy rate of 15-24 year-olds, women and men | Literacy rate of 15-24 year-olds | True | | 15 | 68.09 |
| Employment to population ratio | Employment-to-population ratio | True | | 2 | 93.33 |
| Prevalence of underweight children under-five years of age | Ontology Used | True | Prevalence of underweight (moderate) | Ontology Used | 100 |
| Prevalence of underweight-moderate | Prevalence of underweight (moderate) | True | | 3 | 91.67 |
| Prevalence of underweight | Ontology Used | True | Prevalence of underweight (moderate) | Ontology Used | 100 |
| household size | Average household size | True | | 8 | 63.64 |
| Population | Population size | True | | 5 | 66.67 |
| Persons completed education | Ontology Used | True | Persons age 15+ completed primary education | Ontology Used | 100 |
| Population aged 40 to 44 yr. | Population aged 40-44 years | True | | 7 | 75 |

Table 4.5: Summary of mapping results for importing different units (random units for different indicators) after adding ontology.

| Imported Unit | Mapping to Unit in the Schema | Mapping (True or False) | Minimum Edit Distance Value | Accuracy Percent |
|---|---|---|---|---|
| Birth per woman | Birth per woman | True | Exact Match | 100 |
| Birth/woman | Birth per woman | True | 5 | 68.57 |
| Percent | Percent | True | Exact Match | 100 |
| % | Percent | True | Ontology Used | 100 |
| Percentage | Percent | True | 3 | 70 |
| Metric tons | Metric tons | True | Exact Match | 100 |
| Tons | Rate | False | 4 | 0 |
| Per 100 population | Per 100 population | True | Exact Match | |
| Per 100 pop. | Per 100 population | True | 7 | 61.11 |
| Per 100,000 population | Per 100,000 population | True | Exact Match | 100 |
| Per 100000 population | Per 100,000 population | True | 1 | 95.45 |
| Per 100,000 pop. | Per 100,000 population | True | 7 | 68.18 |
| Deaths/1000 live births | Deaths per 1000 live births | True | 5 | 81.48 |
| Number | Number | True | Exact Match | 100 |
| Num. | Number | True | 3 | 50 |
| Percentage points | Percent | True | 10 | 41.18 |
| Females per 100 males | Females per 100 males | True | Exact Match | 100 |
| Females/100 males | Females per 100 males | True | 5 | 76.19 |
| F/100 M | Females/100 males | True | Ontology Used | 100 |
| Persons/sq km | Persons per sq km | True | 5 | 70.59 |
| Years | Years | True | Exact Match | 100 |
| yr | Years | True | Ontology Used | 100 |

Table 4.6: Summary of mapping results for importing different indicators subgroups (random subgroups for different indicators) after adding ontology.

| Imported Subgroup | Mapping to Subgroup in the Schema | Mapping (True or False) | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|
| Female | Female | True | Exact Match | 100 |
| Male | Male | True | Exact Match | 100 |
| F | Female | True | Ontology Used | 100 |
| M | Male | True | 3 | 25 |
| Total | Total | True | Exact Match | 100 |
| <5 yr | <5 yr | True | Exact Match | 100 |
| <5 year | <5 yr | True | 2 | 71.43 |
| 0-14 yr | 0-14 yr | True | Exact Match | 100 |
| 0-14 year | 0-14 yr | True | 2 | 77.78 |
| 1 yr | 1 yr | True | Exact Match | 100 |

Table 4.6: Summary of mapping results for importing different indicators subgroups (random subgroups for different indicators) after adding ontology. (Cont.)

| Imported Subgroup | Mapping to Subgroup in the Schema | Mapping (True or False) | Minimum Distance Value | Accuracy Percent |
|---|---|---|---|---|
| 1 year | 1 yr | True | 2 | 66.67 |
| One yr | 1 yr | True | Ontology Used | 100 |
| One year | 1 yr | True | Ontology Used | 100 |
| Female 1 yr | Female 1 yr | True | Exact Match | 100 |
| Female 1 year | Female 1 yr | True | 2 | 84.62 |
| F 1 yr | Female 1 yr | True | Ontology Used | 100 |
| Female One yr | Female 1 yr | True | 3 | 76.92 |
| Male 1 yr | Male 1 yr | True | Exact Match | 100 |
| M 1 yr | Male 1 yr | True | Ontology Used | 100 |
| Male One yr | Male 1 yr | True | 3 | 72.73 |
| Male 1 year | Male 1 yr | True | 2 | 81.82 |
| M 1 year | Male 1 yr | True | Ontology Used | 100 |
| 1 yr Rural | 1 yr Rural | True | Exact Match | 100 |
| 1 year Rural | 1 yr Rural | True | 2 | 83.33 |
| One yr Rural | 1 yr Rural | True | 3 | 75 |
| One year Rural | 1 yr Rural | True | 5 | 64.29 |
| Female <5 yr | Female <5 yr | True | Exact Match | 100 |
| Female <5 year | Female <5 yr | True | 2 | 85.71 |
| F <5 yr | Female <5 yr | True | Ontology Used | 100 |
| F <5 year | Female <5 yr | True | Ontology Used | 100 |
| Male <5 yr | Male <5 yr | True | Exact Match | 100 |
| M <5 yr | Male <5 yr | True | Ontology Used | 100 |
| Male <5 year | Male <5 yr | True | 2 | 83.33 |
| Male less than 5 yr | Male <5 yr | True | 10 | 47.37 |
| Male less than 5 year | Male <5 yr | True | 12 | 42.86 |
| 15-49 yr | 15-49 yr | True | Exact Match | 100 |
| 15-49 year | 15-49 yr | True | 2 | 80 |
| 15 to 49 yr | 15-49 yr | True | 4 | 63.64 |
| 15-19 yr Total | 15-19 yr Total | True | Exact Match | 100 |
| 15-19 year Total | 15-19 yr Total | True | 2 | 87.5 |
| Female 15-49 yr Rural | Female 15-49 yr Rural | True | Exact Match | 100 |
| Female 15-49 year Rural | Female 15-49 yr Rural | True | 2 | 91.3 |
| F 15-49 yr Rural | Female 15-49 yr Rural | True | Ontology Used | 100 |
| Urban Male 12+ yr | Urban Male 12+ yr | True | Exact Match | 100 |
| Urban Male above 12 yr | Urban Male 12+ yr | True | 7 | 68.18 |

# Appendix 4: Interview with expert users to specify users requirements

This interview is being conducted by Haitham Zeidan. I am student of Al-Quds University, doing my master thesis in computer Science under the supervision of Dr. Jihad Najjar and Dr. Rashid Jayousi. The aim of this research is to introduce a new mapping algorithm and visualization system towards enhancing mapping and integration of statistical indicators. This questionnaire is designed to find out the requirements before implementation phase of our system. Your contribution by answering this questionnaire is much appreciated.

* Required

**How do you use statistical data in your work? ***

- ○ I am mainly interested in statistical data results
- ○ I am work with the statistical data
- ○ I am decision maker
- ○ I am Researcher

**Which information from visualization system, are you interested in? ***

- ☐ Economy
- ☐ Education
- ☐ Health
- ☐ Information and Communication
- ☐ Nutrition
- ☐ Women

**Which visualization techniques more convenient way for you to explore the results? ***

- ☐ Table Chart
- ☐ Line Chart
- ☐ Column Chart
- ☐ Bar Chart
- ☐ Area Chart
- ☐ Pie Chart
- ☐ Map Chart
- ☐ TreeMap
- ☐ Scatter Plot

**Do you thing visualizing statistical data results can help you in your work? ***

- ○ Yes

- ○ No

**Can visualization help you in understanding the results? If yes, what is your suggestion for designing of a visualize model?** *

- ○ Yes
- ○ No

**What is your suggestion for designing of a visualize model?**

**Are you interested in comparing, filtering, and sorting different scenarios of statistical data in the system?** *

- ○ Yes
- ○ No

Submit

Never submit passwords through Google Forms.

# Appendix 5: Interview with end users to evaluate the system design

This interview is being conducted by Haitham Zeidan. I am student of Al-Quds University, doing my master thesis in computer Science under the supervision of Dr. Jihad Najjar and Dr. Rashid Jayousi. The aim of this research is to introduce a new mapping algorithm and visualization system towards enhancing mapping and integration of statistical indicators. This questionnaire is designed to find out the requirements of end-users of our system. Your contribution by answering this questionnaire is much appreciated.

* Required

**I think the system is easy to use** *

from 1-"strongly disagree" to 5-"strongly agree"

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

**I find the various functions and techniques in the system are well integrated** *

from 1-"strongly disagree" to 5-"strongly agree"

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

**I think there are too much inconsistency in the system** *

from 1-"strongly disagree" to 5-"strongly agree"

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

**I feel very confident using the visualization system.** *

from 1-"strongly disagree" to 5-"strongly agree"

- ○ 1
- ○ 2
- ○ 3
- ○ 4

- 5

**I had to spend much time in order to accomplish the task** *
from 1-"strongly disagree" to 5-"strongly agree"

- 1
- 2
- 3
- 4
- 5

**I was often confused and frustrated during the task accomplishment** *
from 1-"strongly disagree" to 5-"strongly agree"

- 1
- 2
- 3
- 4
- 5

**In case you were confused or frustrated during the task accomplishment, what was the reason for it? Please write your answer:**

**I am satisfied with the system, because it has helped me to accomplish the task successfully.** *
from 1-"strongly disagree" to 5-"strongly agree"

- 1
- 2
- 3
- 4
- 5

**Interaction techniques like sorting, filtering, comparing, and re-visualizing data help me to get better results?** *
from 1-"strongly disagree" to 5-"strongly agree"

- 1
- 2
- 3
- 4

5

**Please write your comment here: what did you like/dislike most of all, and what should be changed?**

Submit

Never submit passwords through Google Forms.

100%: You made it.

Powered by Google Forms

This content is neither created nor endorsed by Google.

Report Abuse - Terms of Service - Additional Terms