Visual Analytics Framework Towards Enhancing Presentation of Official Statistics

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

Abstract

Statistical data has gained many interests from policy makers, city planners, researchers and ordinary citizens as well. Statistical data is normally available in different sources with different formats (HTML, PDF, Excel, Tables, Graphs, ..etc) this makes it difficult for analyst and the decision maker to interact with the data and to take decision.

Applications for visualizing statistical data are still rare. Moreover, these seldom applications also suffer the following drawbacks: (1) some are standalone and only for expert usages; (2) some do not support interactive functionalities such as selection, hovering, zooming, filtering and linking which are essential for visual analytics; (3)some can only provide an overview of statistical data, and (4) some do not support multiple representation for the statistical data. In this paper, our aim is to address these challenges.

This paper will investigate (address) interoperable visual analytics framework Towards Enhancing Presentation of Official Statistics. Nowadays by increasing importance of information in all sectors, illustrating data in a communicative format helps decision makers to understand and analyze effectively large amount of information in a short time. Information visualization, as a way of presenting different data types in a more understandable form, is growing increasingly in various areas. Also this paper will investigate to collect, map, group, or integrate of heterogeneous data into a common schema. investigate how information visualization could be used to increase readability and usability of statistical data.

With our framework, we could collect, disseminate and visualize statistical data, statistical data can be easily and timely presented as tables, charts, and maps. It helps promote the use of statistical data for improved planning and policy making.

1. Introduction

Official statistics are published by government agencies or other public bodies such as international organizations. They provide quantitative or qualitative information in 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 Palestinian central bureau of statistics (PCBS) [1].

The major tenets 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 that advances humans ability to exchange gained knowledge and develop a shared understanding with other people [2].

The “participative web” [3] is increasingly utilized by intelligent web services which empower developers to customize web-enabled visualization applications that contribute to collaboration and communicate visual content Figure 1.

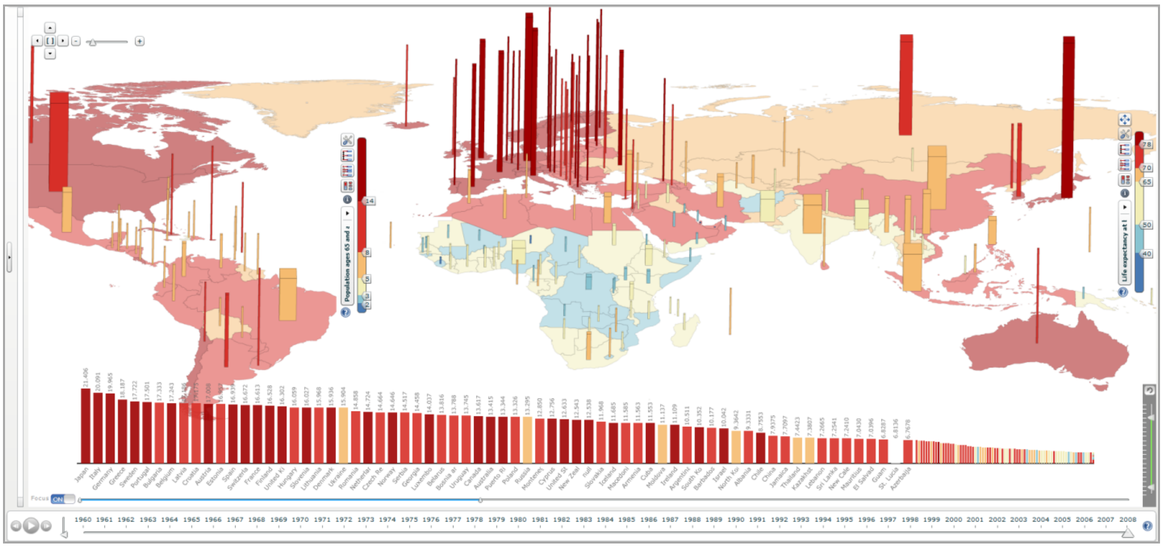


Figure 1: Web-enabled 3D view of the world animating statistical indicators from the World Databank simultaneously in two linked views for 1960-2010.The map and histogram color representing Life expectancy is displayed in the choropleth map; the height of the 3D bars shows percentage of age group 65+ and bar width represents the total population.

In this study we introduce a new interoperable visual analytics framework to enhance dissemination and presentation of official statistics based on dynamic data analytics.

Our visual analytics framework aims to provide techniques that make humans capable of analyzing data by presenting results in a meaningful and intuitive way while allowing to interact with the data.

2. Objectives

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

- Enhancing presentation of official statistics based on visual analytical approach that combines both data analysis and interactive visualization.
- Collecting, mapping, grouping, or integrating of heterogeneous data into a common schema.
- Exchanging statistical data between ministries and agencies based on the framework and the common schema.

3. Related Work

Many studies have been done across countries on data Visualization. Applications of data Visualization were used in a large number of fields, especially for transportation, statistics, Scientific research, Digital libraries and Financial data analysis, market studies.

Patrik Lundblad et al. 2012 study [4]. 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 statistics geovisual analytics application for exploring and publishing statistical data on the web and developed with the GAV Flash toolkit. based on a recommendation from the visual analytics (VA) research program.

Mikael Jern 2010 [5], build web-enabled application platform that is emerging as a de facto standard in the statistics community for exploring and communicating statistics data. Web-enabled application for exploring and communicating statistics data using storytelling mechanism.

Mikael Jern, et al. 2008 [6], 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 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).

Other toolkits focus more on one specific visualisation technique. The Tree- Map Java Library [7] and the HCIL Treemap 4.0 toolkit [8] 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 [9].

The InfoVis Cyberinfrastructure 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. Its 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 [10].

But the above studies have limitations no integration and mapping of heterogeneous data into a common schema. Applications for visualizing statistical data are still rare. Moreover, these seldom applications also suffer the drawbacks that mentioned above.

4. Data Sources

Palestinian Central Bureau of Statistics (PCBS), annually conducted surveys as (Demographic Survey, Demographic and Health Survey, Palestinian Family Health Survey, Transportation & Communication Statistics in the Palestinian Territory Survey, Gender Statistics Survey, Household Expenditure and Consumption Survey, Transportation & Communication Statistics in the Palestinian Territory Survey, Education Census Survey, National Accounts Statistics Survey, Computer, Internet and Mobile Phone Survey, Household Environment Survey, Households Culture Survey, Households Survey on Information and Communications Technology Survey, Housing and Housing Condition Survey, Labour Force Survey, Land Use Statistics in the Palestinian Territory Survey, Mass Media Survey, National Accounts Statistics Survey, Population, Housing and Establishment Census 1997, Population, Housing and Establishment Census 2007. The basic goal of these surveys is to provide a necessary indicators for formulating national policies at various levels, there are many goals of these indicators called Millennium Development Goals (MDGs) [11]:

They consist of eight major goals:

- Eradicating poverty and hunger
- Achieving universal primary education

- Promoting gender equality and empower women
- Reducing child mortality
- Improving maternal health
- Combating HIV/AIDS, malaria and other diseases
- Ensuring environmental sustainability
- Developing a global partnership for development

5. Methodology

To achieve the objectives of this research, we started to prepare and clean the statistical data indicators from different surveys and sources in order to build the Visual Analytics Framework.

5.1 Data Preparation

We focus in our study on Millennium Development Goals (MDGs) indicators, these statistical data and indicators were included in different sectors and subsectors like economy, education, environment, health, information and communication, nutrition, women.

After specifying the list of indicators and sources identified to be included in the database and by collecting indicators from different sources by using Devinfo tool, It is an open source tool for organizing, storing and presenting data in a uniform way. The tool help us to import data and indicators from different sources (Microsoft Access, Microsoft SQL Server, Microsoft Excel, XML, Spreadsheets, Geographic Databases) Figure 2.

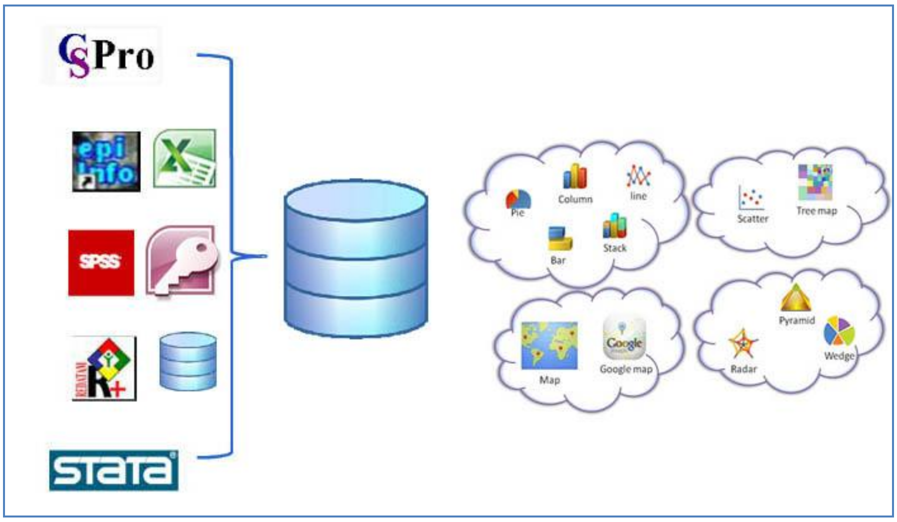


Figure 2: Collecting Statistical Data from Different Sources

5.2 Data Analysis

Statistical data are sets of often numeric observations which typically have time associated with them Figure 3. 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.

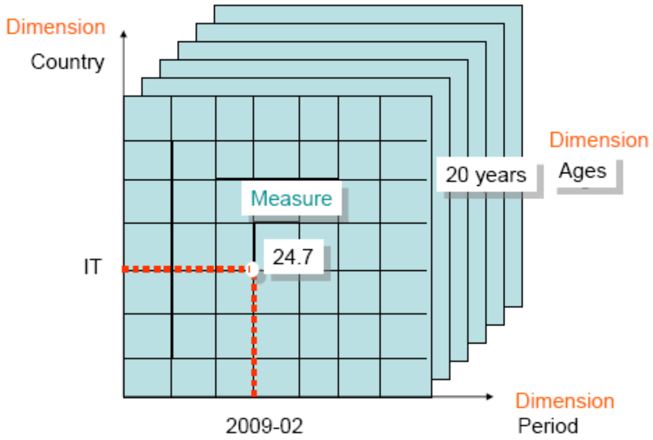


Figure 3: Multidimensional 'Cube' of data

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 1 below gives several examples.

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

Table 1: For example, for the subgroup dimension "Sex", the subgroup dimension values are "Male" and "Female"

We repeat the process many times as needed to enter and associate units and subgroups with all the indicators. each indicator should correctly associated with its unit(s) and subgroup(s) Figure 4.

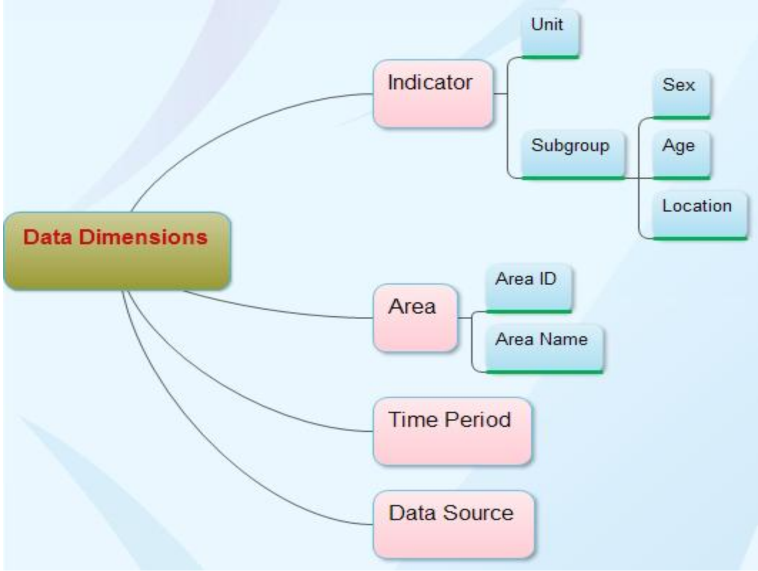


Figure 4: Statistical Data Dimensions

After defining the indicators and their associated units and subgroups, we define the sectors and subsectors for our indicators, then we link the indicator-unit-subgroup (IUS) combinations to different sectors and sub-sectors Figure 5. linking I-U-S combinations to sectors and sub-sectors is important, as it allows users to quickly find indicators in our final framework. 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 framework.

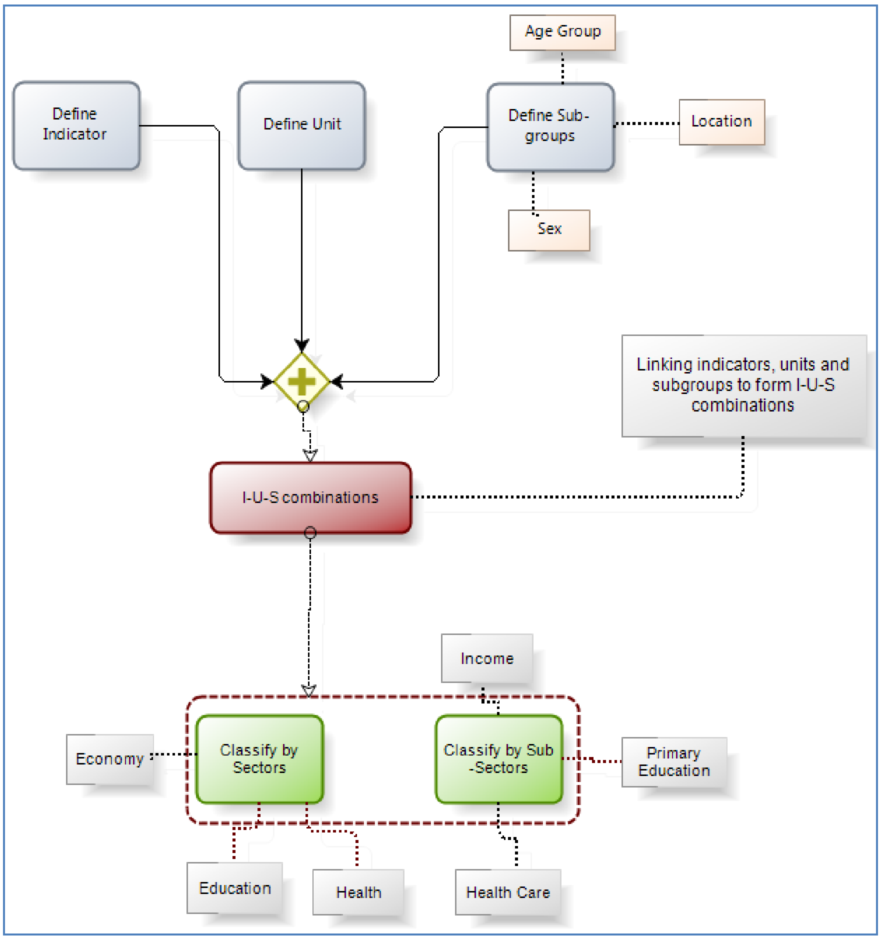
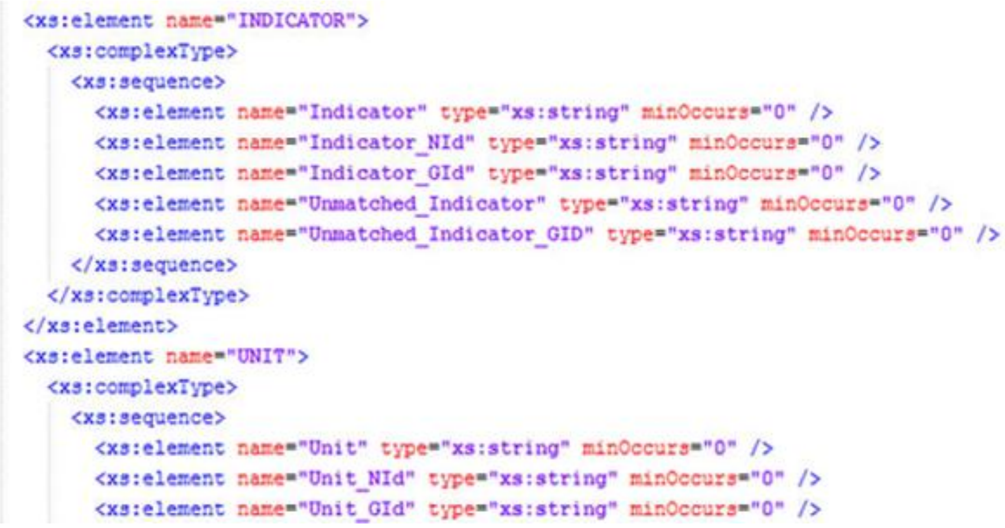


Figure 5: indicator-unit-subgroup (IUS) combinations

We created XML Schema Figure 6 with the list of indicators to be entered into the framework. The definition of each indicator is entered. Each indicator is linked to a unit of measurement and a subgroup. The XML Schema provides the structural model for building a framework database, by specifying the elements against which data can be entered into the database.



```
<xs:element name="INDICATOR">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Indicator" type="xs:string" minOccurs="0" />
      <xs:element name="Indicator_NId" type="xs:string" minOccurs="0" />
      <xs:element name="Indicator_GId" type="xs:string" minOccurs="0" />
      <xs:element name="Unmatched_Indicator" type="xs:string" minOccurs="0" />
      <xs:element name="Unmatched_Indicator_GID" type="xs:string" minOccurs="0" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="UNIT">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Unit" type="xs:string" minOccurs="0" />
      <xs:element name="Unit_NId" type="xs:string" minOccurs="0" />
      <xs:element name="Unit_GId" type="xs:string" minOccurs="0" />
```

Figure 6: Small set of XML Schema snapshot code

5.3 Data Mapping

After building our schema we started to import the data and indicators from different sources (Microsoft Access, Microsoft SQL Server, Microsoft Excel, XML, Spreadsheets, Geographic Databases) to our target schema which contains the indicators names, each indicator is linked to a unit of measurement and a subgroup.

During the data import using devinfo tool, the matched indicators, units and subgroups will be mapped automatically, the different indicators, units or subgroups should be mapped manually, the tools will give us suggested for unmatched indicators, units and subgroups Table 2.

In the first step, we select source file(s) to be imported, at second step, selecting a reference file (target schema) against which we wish to import the data. The data will be imported based on the indicators, units, subgroups and areas defined in the reference file (schema). Third step is mapping data elements, matching the different elements of the source file(s) with those of the reference file (schema). In each step, we can view the unmatched elements. When we have finished mapping all data elements, the data will be imported from the source file against the selected reference template or database, The tools will generate a log file displaying details of the import, map and validation process.

Unmapped Indicator	**Suggested Indicator to map**
Employment to population	Employment-to-population ratio
Growth rate of GDP	Growth rate of GDP per person employed
Unmapped Unit	**Suggested Unit to map**
%	Present
Per women	Births per women
Unmapped Subgroups	**Suggested Subgroups to map**
F	Female
M	Male
Female 15-49 year	Female 15-49 yr

Table 2: Examples of Indicators, Units and Subgroups Mapping

6. Framework Architecture

Our proposed framework in Figure 7 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 I-U-S combinations, Categorizing I-U-S combinations under various classifications, building XML Schema, importing Data from Different Sources. Data Mapping used to map indicators, units, subgroups, I-U-S, the different elements of the source file(s) with those of the reference file (schema), Data Visualization Techniques and Methods (Scatter Plot, Bubble Chart, Map Chart, Line Graphs, Stack Graphs, Bye Chart, table lens, histogram, parallel axes plot, time graph, data table, Tree Map) used.

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 framework 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 [12] 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.

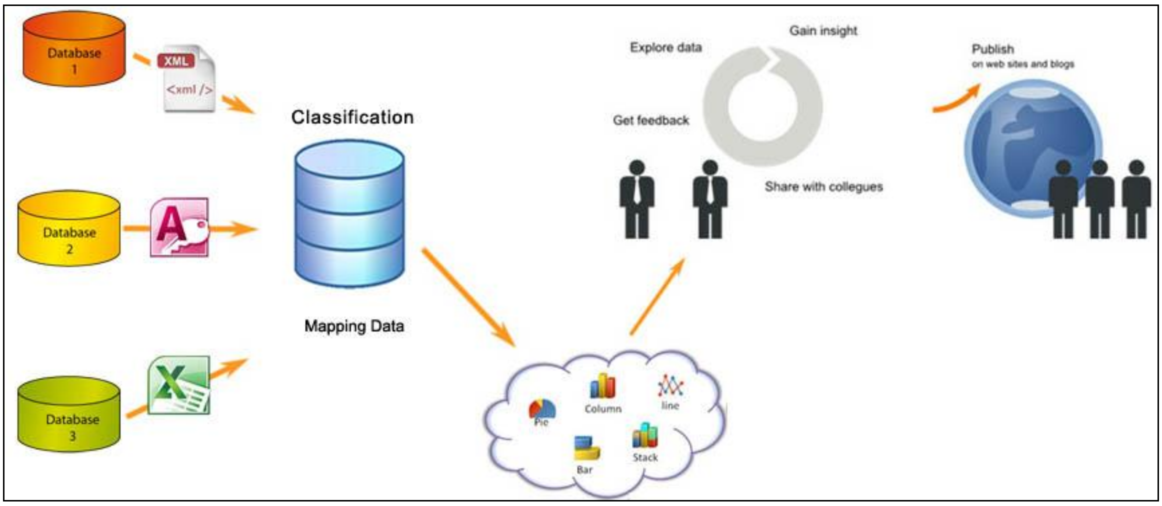


Figure 7: Visual Analytics Framework Architecture

6.1 Implementation

The framework implementation is 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 [13], 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 [14].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.

6.2 Framework Supported Tasks

The framework supports several task to do visualization. The following sections will provide an overview of these tasks.

- **Quick data search:** The Quick Data Search feature allows us to find what we are looking for almost instantly.
- **Advanced search feature:** The Advanced Search feature to select both indicators and areas at the same time.
- **Visualizing search results:** visualize search results in table, graph, chart and map formats. depending on search results, we can visualize data either for the main area selected or for the individual sub-areas.
- **Sharing search results:** share search results with others (via email, Twitter, Facebook, embedded code, or direct page URL).

7. Conclusion and Future Work

Collecting, disseminating and visualizing statistical data help promote the use of statistical data for improved planning and policy making.

In this paper, we proposed a new interoperable visual analytics framework to Collect, process, and disseminate statistical data and metadata based on common schema, heterogeneous data from different data sources integrated before visual methods applied, we enhanced presentation of official statistics based on dynamic visual user interfaces and the principles for Visual Analytics. this framework has been introduced to provide techniques that make humans capable of analyzing data by presenting results in a meaningful and intuitive way while allowing to interact with the data.

Future work includes evaluation of our framework by different users and experts from national statistics organizations, regional researchers, municipal planners, the feedback from users will help us to improve our framework, in addition to that we will focus more on data mapping to be automatic mapping.

8. References

[1] Palestinian central bureau of statistics (PCBS): http://www.pcbs.gov.ps.

[2] J. Thomas and K. Cook. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, 2005.

[3] Andrienko G., Andrienko N., Demsar U., Dransch D., Dykes J., Fabrikant S. I., Jern M., et al. “Space, Time and Visual Analytics”. International Journal of Geographical Information Science, 24(10), 1577-1600. Taylor & Francis. 2010

[4] Patrik Lundblad et al. Web-Enabled Visualization Toolkit for Geovisual Analytics. 2012.

[5] Mikael Jern. Explore, Collaborate and Publish Official Statistics for Measuring Regional Progress. 2010.

[6] Mikael Jern, et al. Geovisual Analytics Web-enabled Tools for Dissemination of OECD Regional Statistics. 2008.

[7] Treemap java library: http://treemap.sourceforge.net/. last retrieved: April 2008.

[8] Treemap 4.1.1 toolkit. http://www.cs.umd.edu/hcil/treemap/. last retrieved: April 2008.

[9] Bederson, B. B., Shneiderman, B., and Wattenberg, M.: Ordered and quantum treemaps: Making e_ective use of 2d space to display hierarchies. ACM Trans. Graph., 21(4):833-854, 2002.

[10] Shashikant, P., Mane, K., and Borner, K.: A Toolkit for Large Scale Network Analysis. SLIS SLISWP-04-02, Indiana University, 2004.

[11] Millennium Development Goals (MDGs) http://pcbs.gov.ps/Portals/_Rainbow/Documents/MDGsPal_2011_English.pdf. 22/11/2012

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

[13] Highcharts library written in pure JavaScript: http://www.highcharts.com.

[14] Flex visualization libraries: https://code.google.com/p/flexlib.