TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

Minor Project Report on

DATA EXTRACTION, VISUALIZATION AND PROCESSING

By:

Santosh Ghimire  (066 BCT 533)

Subit Raj Pokharel (066 BCT 538)

Sudip Kafle (066 BCT 539)

A PROJECT WAS SUBMITTED TO THE DEPARTMENT OF ELECTRONICS
AND COMPUTER ENGINEERING IN PARTIAL FULLFILLMENT OF THE
REQUIREMENT FOR THE BACHELOR’S DEGREE IN
COMPUTER ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

LALITPUR, NEPAL

SEPTEMBER, 2012
TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

Minor Project Report on
DATA EXTRACTION, VISUALIZATION AND PROCESSING

By:
Santosh Ghimire (066 BCT 533)
Subit Raj Pokharel (066 BCT 538)
Sudip Kafle (066 BCT 539)

A PROJECT WAS SUBMITTED TO THE DEPARTMENT OF ELECTRONICS
AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE BACHELOR’S DEGREE IN
COMPUTER ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
LALITPUR, NEPAL

SEPTEMBER, 2012
PAGE OF APPROVAL

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS
DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a project report entitled "Data Extraction, Visualization and Processing" submitted by Santosh Ghimire, Subit Raj Pokahrel and Sudip Kafle in partial fulfilment of the requirements for the Bachelor’s degree in Computer Engineering.

_________________________________________________
Supervisor, name of Supervisor
Title
Name of the Department

__________________________________________________
Internal Examiner, name of Internal Examiner
Title
Name of the Organization, he belongs to

________________________________________________
External Examiner, name of External Examiner
Title
Name of the Organization, he belongs to

--------------------------------------------------------------------------------
Coordinator, Name of Coordinator
Title
Name of the coordinating committee

DATE OF APPROVAL:
COPYRIGHT

The author has agreed that the Library, Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering may make this report freely available for inspection. Moreover, the author has agreed that permission for extensive copying of this project report for scholarly purpose may be granted by the supervisors who supervised the project work recorded herein or, in their absence, by the Head of the Department wherein the project report was done. It is understood that the recognition will be given to the author of this report and to the Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering in any use of the material of this project report. Copying or publication or the other use of this report for financial gain without approval of to the Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering and author’s written permission is prohibited.

Request for permission to copy or to make any other use of the material in this report in whole or in part should be addressed to:

Head
Department of Electronics and Computer Engineering
Pulchowk Campus, Institute of Engineering
Lalitpur, Kathmandu
Nepal
ACKNOWLEDGEMENT

We extend our sincere gratitude to Dr. Arun Timalsina and Er. Shayak Raj Giri for providing the necessary guidelines and support for understating the feasibility and technical aspects for the project. Our special thanks go to Er. Diwaker Ghimire of Young Innovations Pvt. Ltd. (YIPL) for guiding us through the project and helping us understand its complexity. Similarly, we also like to thank the Central Bureau of Statistics (CBS) and Election Commission (EC) of Nepal Government for providing us with the necessary data. We would also like to thank Pulchowk Campus and Department of Electronics and Computer Engineering for providing us the opportunity to build up the project. Last but not the least; we would like to thank our friends and seniors for helping us in the project. Thanks for making the difference.

Santosh Ghimire (066 BCT 533)

Subit Raj Pokharel (066 BCT 538)

Sudip Kafle (066 BCT 539)
ABSTRACT

The project is aimed to create a tool to get a quick analysis and visualization of data with the use of statistical methodology. A web-based application is created which will then be applied to analyze census and election data. Further, it aims to process the population and election result of the country.

The project as a whole is divided into three phases – Data extraction, visualization and Processing. Scripts written in C# extract data from different file formats including XML, PDF and HTML and save it to our database. The data is then visualized using Maps, Charts and Tag Clouds. Finally, processing is done with a tool to find facts from raw data. Along with it, the data has been utilized to find the best model of Federalism in Nepal.

The project is a descriptive website which presents the data in an informative and interesting way. It is a user friendly system which helps to visualize the census data in an interactive form and also allows the admin for managing the data.
# CONTENTS

| PAGE OF APPROVAL | ii |
| COPYRIGHT | iv |
| ACKNOWLEDGEMENT | v |
| ABSTRACT | vi |
| LIST OF FIGURES | ix |
| LIST OF SYMBOLS AND ABBREVIATIONS | x |

1. INTRODUCTION ................................................................. 1
   1.1. Background ................................................................. 1
   1.2. Objectives ................................................................. 1
   1.3. Scope ................................................................. 1

2. LITERATURE REVIEW .......................................................... 2

3. METHODOLOGY .......................................................................... 3
   3.1. Collection of Data ........................................................... 3
   3.2. Programming Languages .................................................... 3
   3.3. Developing Environment ................................................... 3
   3.4. Other Technologies/APIs .................................................. 4

4. SYSTEM DESCRIPTION .............................................................. 5
   4.1. Extraction and Management of Data ..................................... 9
      4.1.1. Extraction of Data ................................................... 9
      4.1.2. Admin Section ....................................................... 9
   4.2. Visualization of Data ....................................................... 12
      4.2.1. Data Visualization Using Google Map ......................... 12
      4.2.2. Tag Cloud ............................................................ 14
   4.3. Data Processing ............................................................ 16
      4.3.1. Analyzing the Models of Federalism ......................... 16
4.3.2. Facts Finder........................................................................................................... 17
5. OUTPUT......................................................................................................................... 19
6. CHALLENGES.............................................................................................................. 23
7. CONCLUSION............................................................................................................... 24
8. FUTURE ENHANCEMENTS......................................................................................... 25
9. BIBLIOGRAPHY............................................................................................................ 26
LIST OF FIGURES

Figure 4.1 System Block Diagram ................................................................. 5
Figure 4.2 Database Schema Diagram of the System ....................................... 6
Figure 4.3 Class Diagram of the system .......................................................... 7
Figure 4.4 Use Case Diagram of the system .................................................... 8
Figure 4.5 Activity diagram for Data Visualization on Map .............................. 12
Figure 4.6 State Chart diagram for finding data of Federal states ..................... 16
Figure 4.7 Sequence diagram of Facts Finder ................................................ 17
Figure 5.1 Homepage ................................................................................... 19
Figure 5.2 Tag Cloud Generation based on Population of District ..................... 20
Figure 5.3 Tag Cloud Generation based on Population of Castes in a District ........ 20
Figure 5.4 Finding facts with facts finder ......................................................... 21
Figure 5.5 Visualization of data on Map ......................................................... 21
Figure 5.6 Analyzing Feasibility of Federal States ......................................... 22
Figure 5.7 Detailed Information of created Federal States .............................. 22
# LIST OF SYMBOLS AND ABBREVIATIONS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJAX</td>
<td>Asynchronous JavaScript And XML</td>
</tr>
<tr>
<td>ASP</td>
<td>Active Server Page</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>HTML</td>
<td>HyperText Markup Language</td>
</tr>
<tr>
<td>IDE</td>
<td>Integrated Development Environment</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notion</td>
</tr>
<tr>
<td>KML</td>
<td>Keyhole Markup Language</td>
</tr>
<tr>
<td>MS-SQL</td>
<td>Microsoft Structured Query Language</td>
</tr>
<tr>
<td>PDF</td>
<td>Portable Document Format</td>
</tr>
<tr>
<td>XML</td>
<td>eXtended Markup Language</td>
</tr>
<tr>
<td></td>
<td>Language</td>
</tr>
</tbody>
</table>


1. INTRODUCTION

1.1. Background
Data that are presented in numbers can be boring to any one and it is difficult to pull out information from it. If these data can be presented in a more visually attractive way preserving the integrity of the data, maximum benefit can be obtained. The use of computer based tools like animations, graphics, and artificial intelligence will make the tool more interactive. Huge collection of data is much more difficult to deal with and these tools can efficiently mine and present the clear cut theme of the data in a simple way.

Our project aims not just to create a web-based tool but also to test its effectiveness with real world data taken from the census and election of Nepal. The project can be classified into three parts: data extraction, visualization and processing. The data is extracted from different files and saved to the database. The data saved in the database is visualized using maps, tag cloud, charts and graphs. Further, meaningful information can be obtained by analyzing the feasibility of federal stated. Besides, information can be extracted by using the fact finder.

1.2. Objectives
i. To extract data from different files and save it to the database.
ii. To provide a means to visualize the data in an informative way and extract information from it.

1.3. Scope
This project provides a platform to extract data from different files of specified standard and save it to the database. It can be used as a tool to visualize data in maps which represents the distribution of data all over Nepal. Facts finder can be very helpful in extracting information from data stored in the database.
2. LITERATURE REVIEW

Currently, the world has progressed remarkably in the field of data analysis and visualization. New tools and methodologies are being developed so that people can analyze huge amount of data within few seconds. Traditional tools for visualizing data like table have today been old fashioned. Even statistical tools like charts have been replaced by more advanced, computer-based tools.

Our country has not been able to match its pace with the revolution in data analysis. Census data in Nepal is still not informative to most of the people. Relationship between population data and measures of development has not been understood properly due to lack of appropriate techniques for data visualization. Use of electronic method have been initiated but is moving at a tortoise’s pace. The data visualization in websites of government agencies doesn’t utilize the power of modern computational tools and is not in any sense up to the minimum standards. Use of computational methodologies for forecasting of census and election has not even been initiated by government or other sectors.

Modern computer-based tools for data visualization include Maps, Tag clouds, Mind maps and animations. US census website 2010.census.gov uses these tools for showing the data and has obtained a lot of positive feedback. Other countries are in the verge of using these tools. Election and Census forecasting has been done using Artificial Intelligence in Indian states election. The results of forecasting by Artificial Intelligence methodology have been more accurate compared to traditional methodologies.
3. METHODOLOGY

3.1. Collection of Data

- **Collection of Data from Census Bureau and Election commission of Nepal**
  
  We have obtained different files containing the data of Census of Nepal for the year 2001, 1991, 1881, 1971, 1961 as well as 1952-54. We have also obtained the preliminary report of census of Nepal 2011. Besides, we also have obtained files containing the data of Constituent Assembly election as well previous two elections of Nepal.

- **Collection of GIS data for locating places in map of Nepal**
  
  Nepal’s GIS data was obtained from ICIMOD’s website. As the file size was large, we used mapshaper.org to reduce the size and finally converted it to KML file (a variation of XML file).

3.2. Programming Languages

C#: The primary programming language used for the development of this project is C#. It is used in conjunction with ASP .Net for the creation of the website.

JavaScript: For the client side programming, we have used JavaScript. It is used primarily in displaying the map and handling the events like mouse click, mouse hover.

3.3. Developing Environment

ASP .Net: We have used ASP .Net Framework 4.0 for the development of the project.

MS-SQL Server: Ms-SQL Server 2008 is used as the DBMS with MS-SQL as the query language.

Microsoft Visual Studio: Microsoft Visual Studio 2010 is used as our primary IDE for programming in C#.

Microsoft SQL Server Management Studio: Microsoft SQL Server Management Studio is used for efficient management of database.
Firebug/Google Chrome Developer Tool: They are used for debugging the JavaScript codes.

Microsoft Visio: Microsoft Visio 2010 is used for drawing the Entity Relationship diagram as well as database schema.

3.4. Other Technologies/APIs

jQuery: It is a JavaScript library which is used primarily in our project for sending and receiving AJAX requests.

JSON: It stands for JavaScript Object Notation and is used as the data format for retrieving data from the server sent during AJAX request.

AJAX: It is a technology used for communication of browser with the server. With AJAX the data obtained from the server can be displayed on the webpage without completely reloading it. It is used for visualizing data obtained from server in the map.

Google Maps API: We have implemented Google Maps API v3 for displaying the map and region specified by the GIS data.
4. SYSTEM DESCRIPTION

A web page is displayed when a person enters the system. Administrators can enter the admin panel by logging in with an ID and a password. They have privileges to enter and modify the data into the database. On the other hand, normal users can view the data but not modify them. They can also visualize the data in graphical form with charts, tag clouds, maps as well as tabular form based on their selection of data. They can also view the best model for federalism.

Figure 4.1 System Block Diagram
Figure 4.2 Database Schema Diagram of the System
Figure 4.3 Class Diagram of the system
Figure 4.4 Use Case Diagram of the system
4.1. Extraction and Management of Data

4.1.1. Extraction of Data

- **Extraction of data for population of Nepal from HTML files**
  The data was available for different districts for different VDCs. The main idea behind the extraction of html data was implementation of *Regular Expression*. The parser collected data for different ethnicities from different VDCs and grouped them by districts and finally saved to the database.

- **Extraction of data for Election from XML files**
  The data for election of Constitutional Assembly was in xml format. Candidate Details and Party list was provided in two different files. The data from these files was extracted by using System.xml library. The data extracted was then saved to our database.

- **Extraction of data for District Level Indicators from PDF files**
  The data of 28 different district level indicators for all 75 districts was available in pdf format file. *Pdfbox* library was used for converting pdf to text. Then the required data was extracted from txt using regular expressions. And finally it was saved to the database.

- **Extraction of GIS data from KML file**
  We have successfully saved the data to our database via XML parser. The data was then finally saved to our database which runs on top of MS-SQL server 2008.

4.1.2. Admin Section

The admin section of the project is designed for the admin of the system to manage the entire system. Admin has access to modify the data used in the system. Admin section can be partitioned into different portions based on the activities performed by the admin. The activities can be described as:

- **Login:**
  The admin needs to login to enter the admin section. More admin can be created by one or more of the super admin to access the system data. The registration
information of the admin is saved in the database separate from the system database that is maintained by the super admin.

- **Add record:**
  Admin can add record in the existing database inside the add portion. To add the data, admin needs to go to the table where new data is to be added. The data is added through a form containing the fields that correspond to the attributes of the table. The values added in the fields are bound in an object that is the instance of the class corresponding to the table. This object is passed to add corresponding values in the database.

- **Search record:**
  Another important task that admin can perform is search the required data. The database contains lots and lots of data stored in its server. It is quite difficult to search for a particular record looking up in the entire database. Instead, admin can search for the particular record which also helps in editing and deleting the record efficiently.

  The admin can go to a table and the data appears in the grid format as stored in database. There are different criteria provided for the admin to search data effectively. Each table has its own criteria based on the attributes of table and admin can search using either one or all of those criteria. The search function takes up the value in the search text box and matches the criteria with that on database through SQL query and finally displays the matched record in the same grid view.

- **Edit and Delete record:**
  Updating of record is very important to keep the record up-to-date. It also helps in correcting the mistakes in the stored data. Admin has access to update the record stored in the existing database. The edit and delete can be performed on a particular record obtained after searching or record initially displayed.

  Each record has its own unique representation to differ from another record. To edit the record, system first finds the unique identity of the record and the updated value is stored in the database for that record based on the identity. Admin should be
careful not to update the unique identity and in most cases the identity is made not editable.

The deletion of record can be performed similarly. The admin should be careful while deleting any record as it removes the record from the database also. The system removes the record along with its identity when the record is deleted.
4.2. Visualization of Data

4.2.1. Data Visualization Using Google Map

Figure 4.5 Activity diagram for Data Visualization on Map
One of the methods used for visualization is map. Google Maps has been used as the platform for obtaining Maps. Both client-side and server-side programming has been for this purpose. JavaScript along with Google Map API is used as client-side programming language whereas C#.Net is used as server-side programming language.

On client-side, the map is loaded on the web page using Google map API. The coordinates for districts to be plotted are obtained from our own Database. For this AJAX request is sent to the server. Sever then fetches the coordinate data on database and returns them in JSON format to the browser. The districts are plotted by creating layer on Google Map.

After map is displayed on web page, the user selects which values to visualize on map. The values are classified into two categories: population based on caste and numerical value of development indicators like availability of toilet facility, electricity facility and literacy rate. Based on selected value, an AJAX request is sent to the server. The server fetches data on Database and responds to client by sending data in JASON format again. The given data is visualized.

For visualization the numerical data like population has been changed to the strength of color. The following formula has been used to obtain the darkness:

\[
Darkness \ of \ Color = \frac{Given \ Value \ for \ district}{Max \ Value} + 0.05
\]

As a result, the districts with higher value for given domain appear darker while with lower value appear lighter in the map.
4.2.2. Tag Cloud

Tag Cloud is a good way to represent how the data is scattered among various categories. It shows the items with different weights in different font size and styles so that the viewer gets a clear picture instantaneously. The items are randomly scattered each representing one category of corresponding weight which is a number based on which the cloud is generated.

We have implemented tag cloud to represent the distribution of population among various districts of Nepal. Each district is shown as a word in the tag cloud with color and font size varying according to the population of the district. The district with highest population is shown the largest and the one with lowest population is shown the smallest while others lie in between. Further the population of each district is shown on the tooltip of the mouse pointer when the pointer hovers over the district word. This makes easy to get the exact count of population too if needed. Each word which represents a district is clickable i.e. one can click on the word to navigate to another place. If a word, representing a particular district is clicked, then the control will navigate to another page where a tag cloud of various castes within the district is shown. Same as before, here also the population of each caste can be viewed through the tooltip of the pointer hovered over the word.

To implement tag cloud, we have used the asp control `literal` to hold the name of all the words. Each district is assigned a weight according to its population. The weight is calculated by statistical method by calculating the mean and the standard deviation. And the district is shown with the style corresponding to the weight. This is done via CSS. Each weight class has different styles i.e. color and font size which makes the picture of the cloud better.

The weight \( w \) is calculated by using the formula

\[
W = \frac{x - x'}{\sigma}
\]

Where \( x = \) population of district/caste;
\( x' = \) mean population of district/caste;
\( \sigma = \) standard deviation
Further, standard deviation is calculated by using the formula,

$$\sigma = \sqrt{\frac{\sum(x-x')^2}{N}}$$

Where N=total number of district/caste
4.3. Data Processing

4.3.1. Analyzing the Models of Federalism

For finding out best model of Federalism in Nepal, the user is allowed to create different states by grouping the districts. Different statistics for each group including, the top 10 castes and their population, the top 10 parties elected in previous election and the composite development index of the state created. Based on the comparison between the states, users can check if the creation of such state is feasible in country or not.

For this purpose, we have allowed users to add districts to the states via Map. The map is loaded to the web page as described in the portion ‘data visualization on map’. User can create new state by clicking on button. For adding districts to the state, user clicks on district on the given map. After all districts are added, another state is created. Finally, the states with districts are sent to server for finding statistics.
The server calculates the aggregate population, development index and aggregate results in election and displays the data on table as well as bar graph on a new page.

4.3.2. Facts Finder

![Sequence diagram of Facts Finder](image)

It is essential to find the facts from the large volume of data to make sense. Visualization aids in viewing the pattern, comparing data but data processing must be done to yield useful result. For this, we have created facts finder.

It finds different interesting facts from the data as specified by the user. The user can choose which type of fact he wants and can also include some constraints to bind the fact. The fact will then be generated. A fact is generated randomly and displayed in the homepage which gives user some information. User can also quickly view another fact by clicking on the link placed nearby. It displays another fact which is different from the already displayed facts.

To create fact finder, we have given user the choice to input constraint variables which is not mandatory. By inputting the variables of choice, user can view the fact just by clicking the *find fact* button. For this, we have taken all the input values and written a SQL query with these input values as filter conditions. The result of the query is then used to construct
the sentence for fact. While creating sentences, various input values are checked to ensure that correct and meaningful fact is generated.

In case of random fact generation, the values that were taken as input from user are chosen randomly and fed to generate a fact. While choosing another fact from the home page, we have used **AJAX** to ensure that only the fact is reloaded but not the whole page.
5. OUTPUT

Figure 5.1 Homepage
Figure 5.2 Tag Cloud Generation based on Population of District

Figure 5.3 Tag Cloud Generation based on Population of Castes in a District
You can find different interesting facts about the Election of Nepal, Census of 2001.

**Figure 5.4 Finding facts with facts finder**

**Figure 5.5 Visualization of data on Map**
Figure 5.6 Analyzing Feasibility of Federal States

Figure 5.7 Detailed Information of created Federal States
6. CHALLENGES

The project was not as convenient as we expected it to be and we had to face quite a lot of challenges. Most of them were related to data and files we needed for the project. The problems we encountered during our project are listed below:

- Our main challenge was to work with real data. They were distributed randomly and we needed to collect them from all distributed source.
- The records we obtained were in different formats. We needed to make parser for each of the format to extract the required data.
- The data was very inconsistent and there were lots of incorrect data. We had to manage them manually.
- The data was in huge quantity and database management was a tedious task. We had to run complex queries to manage and store the records.
- Query processing took a lot of time so we had to perform query optimization.
7. CONCLUSION

Despite the challenges and problems, we completed our project in specified duration. The program was first tested among the developers and was finally with end users. The project is finally obtained as a webpage with options for data management, visualization and processing.
8. FUTURE ENHANCEMENTS

We have tried our very best to enhance our project with more and more features but still there are a lot that can be added in it. Following enhancement can be done in our project:

- Extraction of data from different file formats automatically when the file is provided by the users via web interface.
- More statistical analysis on the dataset including regression, correlation and hypothesis testing.
- Data forecasting for upcoming and previous election and census.
9. BIBLIOGRAPHY


