Techniques for Consolidation and Exploration of Information on a Webserver

Dissertation in fulfillment of the requirements for the academic degree

Doctor of Technical Sciences in Computer Science

Submitted by

By

Rizwan Mehmood

Matrikelnummer: 1232066

First Reader: O.Univ.Prof. Dr. Hermann Maurer

Second Reader: O.Univ.Prof. Dr. Leopold Wildburger

I hereby certify that the work reported in this dissertation is my own and that work performed by others is appropriately cited.

Hiermit bestätige ich, dass die Arbeit in dieser Arbeit berichtet, ist meine eigene und die Arbeit von anderen durchgeführt wird, entsprechend zitiert.

Rizwan Mehmood

Acknowledgements

First of all, I would like to thank Allah Almighty, for His divine guidance and providence. He blessed me with motivation, passion, and hard work. It was His blessings which made me able to plan, visualise, and execute my dreams into reality. I would like to thank my parents Riaz Mahmood and Shaukat Ara, whose immeasurable sacrifices have led to what I am today.

I would like to thank Professor Hermann Maurer, for his guidance and support throughout my PhD. He always encouraged me when I had tough times. He gave me funding for conferences. I am really grateful to him. His ideas about virtual lab and story generation based on unusual facts helped me in accomplishing the projects in a proper way. He helped me a lot in publications and especially in writing the thesis. Apart from academics he and his wife helped us when my family needed their support during our temporary stay in Graz. I would like to dedicate this thesis to Professor Maurer.

I would like to thank Professor Dr. Leopold Wildburger for being second reviewer of my thesis.

I would like to thank Dr. Tanvir Afzal, for his guidance regarding the scholarship. He introduced me to Professor Maurer and helped me in getting the acceptance letter.

I am extremely grateful to all the colleagues, especially Narayanan Kulathuramaiyer and dear friends (Bruno Bakula, Thomas, Paso, Daud) at IICM who helped me during my course work and PhD. I would like to say thanks to all HEC scholars particularly (Raheel and Hassan) for their continuous support and guidance in different activities. I would like to say thanks to Gerhard Wurzinger for his guidance regarding Austria-Forum.

At this point I am also obliged to Higher Education Commission of Pakistan (HEC) for awarding me the scholarship for studying at Graz University of Technology. I want to say thanks to OEAD for their excellent administrative support during my stay in Graz.

Finally, I would like to say thanks to my wife Mustbeen who really supported me in my education.

Abstract

This thesis deals with techniques of information consolidation and exploration, as applied to collections of pictures and geographic data, the latter including material often not considered in geographic data bases.

The first part of the thesis explores problems and solutions for merging image databases. As initial step, suitable pictures had to be extracted from a large database. They then had to be cleverly integrated into already existing collections of pictures in a server called the Austria-Forum.

The second main part of the thesis deals with the creation of a geography server that differs from existing ones in a variety of important ways.

One aspect is our focus on verifying information using different approaches. We have tried to verify the area of countries, the size of major cities and mountain ranges using various data sources. We also highlight the technique of geographic data verification using spatial methods.

Our efforts culminated in a project "global-geography.org" that covers substantial information about all countries and territories of the world, beyond traditional facts. It provides a global platform for sharing knowledge about different countries of the world. A number of novel techniques have been developed to collect data as needed.

The research in the thesis also allowed to develop a unique "Virtual Lab" built specifically for exploring information about countries in a highly interactive fashion: The number of reports that can be generated by selecting sets of countries and properties is almost uncountable.

We have developed visualisations that allow the comparison of countries using data from Worldbank, UNO and International Energy Statistics. As part of our investigations into visualization we had to also consider aspects of "responsive visualization" to assure that graphics can be viewed in a reasonable way independent of screen size of the device used.

A further section of this thesis presents interactive visualisations and mapping services. Our maps show many important geographical features like boundaries, neighbouring countries, but also districts and capitals of countries. Users can also create markers on the maps and add valuable information about locations in a country. Of course there are many areas for future work, like automatically showing major features like rivers, lakes, mountain ranges, location of raw materials, of sights to visit, etc. Detailed research on this would go much beyond a Ph.D. thesis, yet we have included some aspects (like world heritage sites) in categories like culture, pictures with detailed descriptions and special stories (stories typical for the country at issue). The idea to semi-automatically generate stories typical for a country has provided many interesting problems for which we offer partial solutions. It is well known that the generation of automatic stories is currently a hot topic (e.g. for news channels). We have managed to come up with new heuristics to generate unusual information for our purposes.

As part of our thesis we also did systematic evaluations of user interfaces with cohorts of users, mainly students. We have compared our system with other geographic servers to show the novelty of our approach. We also mention some aspects of potential further improvements.

Summarizing, the thesis is a complex mix of research, state of the art IT technology and making material in many databases useable. The fact that by now members in an international consortium from all continents have agreed to participate in using and further building up what we have started, by easy-to-use ways for feedback is a good indication that our research, our implementation and our ideas have created international attention.

Kurzfassung

Diese Arbeit befasst sich mit Techniken der Datensammlung und Datenkonsolidierung, wie der Sammlungen von Bildern und Geo-Daten, das letztere einschließlich Material, das oft nicht in geographischen Datenbanken berücksichtigt wird.

Der erste Teil der Diplomarbeit untersucht Probleme und Lösungen für das Zusammenführen von Bilddatenbanken. Als ersten Schritt mussten geeignete Bilder aus einer großen Datenbank extrahiert werden. Dann mussten sie passend in bereits bestehende Sammlungen von Bildern in einem Server Austria-Fourm.org integriert werden.

Der zweite Hauptteil der Arbeit befasst sich mit der Erstellung eines Geographie-Servers, der sich von bestehenden in einer Vielzahl von wichtigen Punkten unterscheidet.

Ein Aspekt ist unser Fokus auf die Überprüfung der Informationen mit unterschiedlichen Ansätzen. Wir haben versucht, die Größe der Länder, die Größe der Großstädte und Bergketten mit verschiedenen Datenquellen zu überprüfen. Wir greifen dabei auch auf die Technik der Geo-Daten-Verifizierung mit räumlichen Methoden zurück.

Unserer Bemühungen gipfelten in einem Projekt "global-geography.org", das wesentliche Informationen über alle Länder und Gebiete der Welt, auch jenseits traditioneller Tatsachen umfasst. Es bietet eine globale Plattform für den Austausch von Wissen über verschiedene Länder der Welt. Eine Reihe von neuartigen Techniken um Daten zu sammeln wurde entwickelt.

Die Forschung in der Arbeit hat es auch erlaubt, ein "Virtual Lab" zu entwickeln, das sich dazu eignet Informationen über Länder in interaktiver Weise erkunden: Die Anzahl der Berichte, die generiert werden können, indem man Gruppen von Ländern und Eigenschaften auswählt ist fast unzählbar.

Zudem haben wir Visualisierungen entwickelt, die es ermöglichen, Länder auf Grund von Daten der Weltbank, der UNO und internationaler Energiestatistiken anschaulich darzustellen. Im Rahmen unserer Untersuchungen zur Visualisierung mussten wir auch Aspekte der "reagierenden Visualisierung" berücksichtigen, um sicherzustellen, dass die Grafiken in zumutbarer Weise unabhängig von der Bildschirmgröße des verwendeten Geräts angezeigt werden können.

Ein weiterer Abschnitt dieser Dissertation präsentiert interaktive Visualisierungen und Mapping-Diensten. Unsere Karten zeigen viele wichtige geographische Merkmale wie Grenzen, Nachbarländer, aber auch Bezirke und Hauptstädte der Länder. Benutzer können auch Marker auf den Karten erstellen um Informationen über spezielle Aspekte eines Landes einzubinden. Natürlich gibt es viele Bereiche für zukünftige Arbeit, wie das automatische Anzeigen wichtiger Details wie Flüsse, Seen, Bergketten, die Lage von Rohstoffen, Sehenswürdigkeiten, etc. . Ausführliche Forschung dazu würde den Rahmen einer Doktorarbeit sprengen. Jedoch haben wir einige Aspekte (wie Stätten des Weltkulturerbes) aufgenommen. In Kategorien wie Kultur, Bilder mit detaillierten Beschreibungen und besonderen Geschichten (Geschichten typisch für das Land in Frage) gehen wir weit über die üblichen geographischen Datenbanken hinaus. Die Idee, semi-automatisch Geschichten zu generieren, die typisch für ein Land sind hat viele interessante Fragen aufgeworfen, wofür wir Teillösungen erarbeiten konnten. Die Erzeugung von automatischen Geschichten ist ein wichtiges Thema (z.B. derzeit für Nachrichtenkanäle). Es gelang uns mit einem Bündel von Heuristiken ungewöhnliche Informationen für unsere Zwecke zu generieren.

Als Teil unserer Arbeit haben wir auch eine systematische Auswertung von Benutzeroberflächen mit Gruppen von Benutzern (vor allem Studenten) durchgeführt. Wir haben unser System mit anderen geografischen Servern verglichen, schneiden dabei gut ab, aber erwähnen auch einige weitere Verbesserungsmöglichkeiten.

Zusammenfassend ist diese Dissertation eine komplexe Mischung von Forschung, Stand der Technik der Technologie und Methoden, Material in vielen Datenbanken nutzbar zu machen. Die Tatsache, dass Fachleute als Mitglieder eines internationalen Konsortiums aus allen Teilen der Welt ihre Mitarbeit zugesagt haben ist ein Zeichen dafür, dass unsere Ideen, unsere Forschung und unsere Realisierung internationale Aufmerksamkeit erreicht haben.

Contents

Chapter 1 : Introduction and Motivation

The Web contains some 500 billion pages, more specifically over 1 billion websites and this number keeps growing [1][2]. It is not exaggerated to say that most of human knowledge is now available on the Web. Yet even with search engines constantly improving it is often an exercise in frustration to find substantial information on a topic of interest.

This is due, we believe, to three reasons.

First, the Web often has too much or too little redundancy. Too much in the sense that a query can deliver an immense number of entries, yet many of them contain the same or very similar stuff. On the other hand, in a typical encyclopaedia oriented server like Wikipedia by the very definition of encyclopaedia there will be only one entry for each topic. This is not satisfactory for contentious topics where a number of different views would be helpful, rather than one "compromise", even if the compromise does contain some pros and cons.

Second, material found is often not *coheren*t enough. To find more than superficial information on a particular issue is often a taunting task, requiring to follow hundreds of links and slowly piecing together a satisfying picture using hundreds of information snippets.

Third, material can often not be judged to be *reliable*: Authors may be anonymous, their bias towards certain worldviews is hard to assess, some information may be more advertisement than truth, other material may be influenced by politics, etc.

We believe it is necessary to reduce redundancy by pulling information together, to make sure that coherent material akin books on certain topics is available and that reliability is increased by insisting that information is provided not anonymously but by authors whose qualification can be checked and who can be contacted if necessary.

Above credos have been driving forces for research and experiments at the Institute for Information Systems and Computer Media at Graz University of Technology, the institute at which this Ph.D. thesis was written.

A good part of this thesis is concerned with improving coherency by gathering and checking information from various sources. Pulling information from different data sources with minimal or no redundancy is clearly a challenging task, yet of course an essential task necessary for many data oriented applications. As second effort we have tried to improve on static Web pages by providing a high degree of interactivity. Both ideas have been investigated in this thesis using a particular platform, Austria-Forum [3].

This chapter gives an overview of the website Austria-Forum.org. It highlights research challenges that are the basis of this thesis, allowing to explain the objectives of thesis.

The contribution of the thesis covers fields like database systems, information extraction, data extraction from the semantic web, spatial data handling, information integration, information visualisation, information architecture and web applications development. The thesis also highlights new ways of information verification that are critical for publishing high-quality data on the Web.

Given the huge amount of information we are confronted with, information integration is one of the biggest challenges of this century. This thesis describes methods and techniques for the information integration process. Web data sources are queried in different fashion due to the differences of models

used by different sites and databases. Dynamic web content that is displayed by a browser usually originates from databases and is presented without any structure. Various problems are addressed in this thesis that are encountered when extracting images and factual data. The major contributions are information retrieval of images followed by their meaningful integration in a website, data verification methods, data exploration via a virtual lab, interactive visualisations and intelligent story generation by aggregating multiple scenarios.

The rest of this chapter is divided into four sections. Section 1.1 is about the website Austria-Forum. Section 1.2 is about research motivation. Section 1.3 presents research questions that form the foundation of thesis. Major scientific contributions are listed in Section 1.4. Section 1.5 presents a general overview of the thesis.

1.1 Austria-Forum

Austria-Forum is a substantial web site (technically speaking a JSP wiki with many extensions) trying to improve on some issues encountered e.g. in Wikipedia. In this section we describe what the Austria-Forum is all about and why it is being developed. We believe that it is an important model and an attempt to beat the tyranny of search engines and information servers whose contents are of unknown quality, see [4].

Austria-Forum is based on a JSP-Wiki system, enriched with many plug-ins. It contains Wiki-pages as one would expect, but it also includes so-called Web Books: full books with page-turning facility, full text search, zooming, and other features, particularly the possibility to have links to and from book-pages to ordinary web pages. Further, we have integrated dynamic vector maps where user can place markers along with panning, zooming and other options. Users can be lead from a maker on a vector map of any country to any other page in- or outside the Austria-Forum. Thus, Austria-Forum is tying together ordinary web pages material with books and maps in a rather novel way. We will return to these features in later chapters of the thesis, and specifically on how to weave pages and maps together using markers and popups and how to use data-bases in an interactive way.

The website Austria-Forum is a large site that contains “materials mainly of interest to Austrians”. It started as an electronic encyclopaedia on Austria some 20 years ago and has grown by now to a collection of over 670,000 “objects”, an object being a piece of text or some multimedia object like a picture or a video-clip. The main entry point of Austria-Forum is shown in the Figure 1.1.

It is important to understand that Austria-Forum is different from e.g. Wikipedia in a number of important ways. For completeness sake we briefly mention the main distinctive features as discussed in [4].

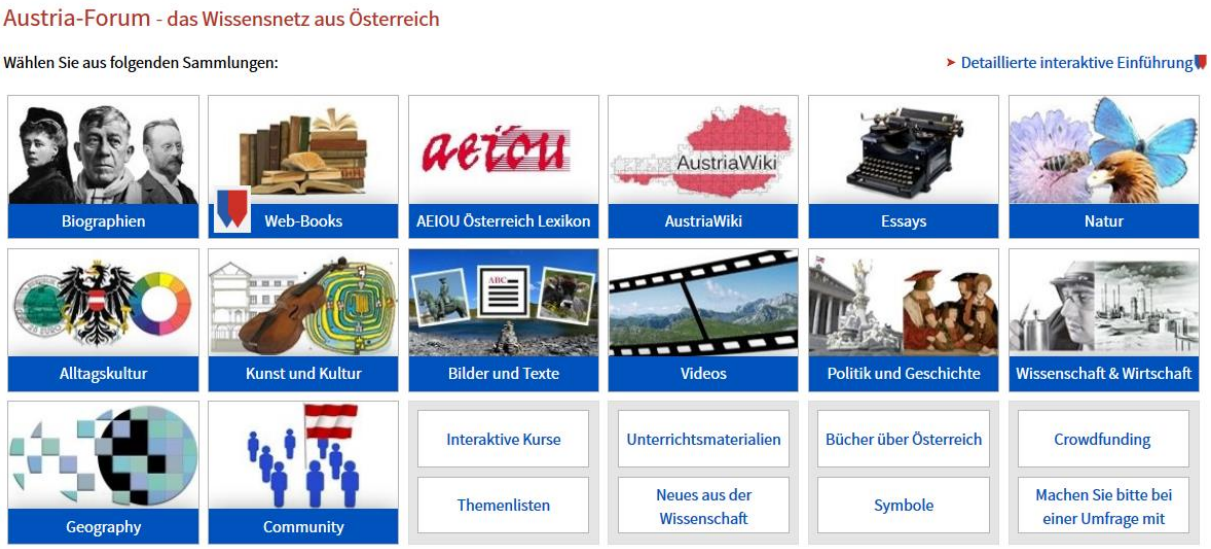


Figure 1.1: Austria-Forum Main Entrance

The Austria-Forum is a special kind of knowledge collection whose unique features include:

- It is only concerned with contributions of some kind of temporal stability (i.e. no information on current weather, sports, events, etc.).
- Most of its contributions are frozen, i.e. do not change any more.
- As a rule, contributions must have known authors or known reliable sources to make them citable.
- Multiple contributions on the same topic are welcome as one cannot understand a complex issue unless one examines it from different points of view.
- Books are part of the Austria-Forum as Web-Books: when opened, they look like a printed book and pages can be turned, yet full-text searching is possible and- most importantly-links to other book pages or Austria-Forum or arbitrary Web pages (and conversely) are possible.
- Austria-Forum also integrates contribution from other sources (data bases like from the UN, the Worldbank and including Wikipedia) with an attempt to verify imported data or adding further substance to it.

One of the most important features of the Austria-Forum is the increase of interactivity. Thus, Austria-Forum turns more and more into a repository of information that can be experimented with in countless ways as is becoming more and more common with the use of server side Apps. In order to make Apps universally useful, it makes sense to have them running on the server, rather than as an add-on to the device (as is often the case in proprietary software for e.g. the I-Phone). The best examples in Austria-Forum for interaction are found in the projects “sparkling science” and “global-geography” exploring information related to countries that will be discussed in chapters 4 and 9 respectively.

For the sparkling science project approximately 300 students of 5 schools wrote essays on a range of topics. It is possible for instance to compare the choice of topics by school, by gender, by age, by grade, etc. It is this kind of experimenting with data that can also be done for geographic data of l93 UN countries. Austria-Forum, like large parts of the WWW, is starting to turn from a search or menu based information server into a data-base of information open to uncountable experiments.

1.2 Research Motivation

Due to the proliferation of data sources on the Web, information integration has emerged as a research topic of great interest. The crucial challenge is to stitch heterogeneous data together. The situation becomes more challenging when applied to the integration of images. For instance, consider images that we can extract from some websites or data bases, and- by knowing their context- automatically associate them with suitable documents homogeneously. This thesis explains new ways of consolidation of images into one of the largest Austrian online repositories, the Austria-Forum. The process involved a large set of heuristics. The tools developed made sure that only a moderate effort from a human expert was required to check whether the images are suitable as objects in the context proposed by the system, see [5]. This corresponds to one of our credos: as long as AI (Artificial Intelligence) cannot do everything, AI techniques can and should be used to help humans.

The web has significant impact on the lives of people and its importance is further growing. The culture of printed books has changed with the invention of web. The problem in the Web is not, however, the lack of information, but the reliability of information and its fragmentation as claimed and highlighted in [6]. The author showed the unreliability and difference of facts in different data sources found on the web. One cannot imagine a web without data. But data has to be of good quality and reliability to be really useful. The quality of data is a critical factor for all kinds of decision-making and transaction processing and its importance is also clear when used for learning. We will present ways to measure the reliability of data.

Geographic web servers and geospatial databases are emerging on the web, offering information about countries and places in the world. Digital content is increasing at a staggering rate due to community collaboration and the integration of information from webcams and sensors. Like in case of Wikipedia, some geospatial databases allow everyone to edit the content. We cannot ignore the role of wikis and geospatial databases particularly Wikipedia, Wikicommons, GeoNames, etc., as they have replaced the traditional encyclopaedias and they are empowering information seekers by providing information easily. However, there is no guarantee of validity and authenticity of the information provided. The reason is that very little attention has been given to verify information before publishing it on the Web. Also, to find particular information about countries, web users mainly rely on search engines such as Google which often points to Wikipedia. We will identify some inconsistencies in online facts such as area of countries, size of cities and mountains rankings using multiple data sources in Chapter 5 and 7. Our investigations reveal that there is a need for a reliable geographic web server which can be used for learning and other purpose. We will explain how we managed to devise a mechanism for collecting and verifying different facts. Our attempt to provide a reliable geographic web server has resulted in a comprehensive collection covering a wide range of information aspects such as culture, geography, economy etc. that are associated with a country. We will also describe our approach to measure the reliability of geographic facts such as area, cities and mountains rankings for all countries in Chapters 5 and 7.

The rapid expansion of Web content has raised numerous concerns related to reliable and meaningful information delivery. What users actually need is information that suits the task and context at hand, from a reliable knowledge source, with an assurance of information objectivity and integrity. In this respect, most of the numerous web sites that offer geographical information have, serious drawbacks: Often, advertisements are intermixed. It is hard to tell if information is slanted or censored. Facts presented are usually taken from one single source and not checked against others. Also, most information is of numeric nature. This is particularly true of databases like the one provided by the UN, the Worldbank, Wolfram Alpha, DBpedia, etc. A modest mixture of textual information and a selection of pictures is found in Factbook, Wikipedia, Britannica or Factfish, and maps are much employed in systems such as Open Street View, Google Street View, or Geonames. However, cultural or country-

specific issues are rarely addressed. We have started what we believe is the first serious attempt to provide a site that tries to overcome some of those weaknesses in global-geography.org. Without reinventing the wheel, i.e. using information from various sources as far as copyright issues permit, we are building a site that beyond presenting consolidated and verified numerical data we try to also report on cultural and other important aspects of countries.

Concerning Austria-Forum, global-geography is the first time a serious attempt to integrate and publish information in English, beyond the limited material on the Austria-Forum available in languages other than German before. Interactive support was missing that is critical to access information hidden in huge data repositories, see [7]. Interactive visualizations are the core of the sub-server global-geography.org. It offers static views of facts along with dynamic interactivity for exploring information about countries. We dig deep into the hearts of countries by mentioning cultural aspects, stories typical for a certain region and a range of pictures with solid description showing the uniqueness of parts of the world.

1.3 Research Questions

This section describes research questions that are addressed in this thesis. Wherever applicable, we will break down these questions into more specific ones. The research questions 1 and 2 are addressed in Chapter 2, 3 and 4. The research questions 3, 4 and 5 are addressed in Chapter 5, 6 and 7. The research question 6 and 7 are investigated in Chapter 8, while research questions 8 and 9 are addressed in Chapter 9. The solution of research question 10 can be found in Chapter 10. Research questions 11 and 12 are solved using visualisation approaches and are discussed in Chapter 11. Research questions 13 and 14 addressed in Chapters 12 and 13 respectively. Finally, the research questions 15 and 16 are addressed in Chapter 14 and 15 respectively.

RQ 1. How can we integrate images into our existing collection of biographies, villages, lakes etc. from other data sources? More specifically, we have to choose only relevant ones, and avoid duplicates, or very similar ones.

RQ 2. How can we clearly visualise and compare statistical data on students by gender, age, schools and other properties, based on essays they have composed?

RQ 3. How can we verify information using online sources on the Web?

RQ 4. How can we verify geographic information using maps?

RQ 5. How can we integrate information from various data sources and build a new geographic server that is part of Austria-Forum, yet to the outside world acts like a separate server?

RQ 6. How can each country be represented in Austria-Forum consistently? This question is mainly a design aspect that deals with the presentation of entry page of each country in the geographic server.

RQ 7. How can we extend country data with information related to its culture?

RQ 8. How can we present a dynamic interactive web interface that allows users of the system to explore countries meaningfully?

RQ 9. How can we compare two or more countries?

RQ 10. What is the ranking of a country among 193 UN countries regarding different parameters?

RQ 11. How can we visualise time oriented data?

RQ 12. How can we visualise trade among countries?

RQ 13. How can we extend the map section of each country and provide interactive dynamic maps?

RQ 14. How can we semi-automatically generate a story about countries that are of interest to users?

RQ 15. How can we create visualisation that work on different devices, i.e. use responsive visualisation?

RQ 16. How can we evaluate different modules of system?

1.4 Scientific Contributions

This thesis deals with techniques related to information consolidation and exploration. The thesis makes the following scientific contributions:

1.4.1 Merging Image Databases

The thesis presents innovative ideas about how to merge image databases. Merging two image databases is not a trivial problem. We have investigated a set of heuristics to achieve our goals. We have automatically prepared approval and disapproval lists of pictures which have reduced human efforts concerning association of images with documents. Carefully designed regular expressions, when applied on "image descriptions", confirmed that most images are automatically associated with a correct document. A special picture arrangement tool solved the problem of arranging pictures in a meaningful way. Further details are discussed in Chapters 2 and 3.

1.4.2 Interactive Bar Chart Tool

We have succeeded in developing an interactive bar chart tool to monitor activities related to a "Sparkling Science" project, by choosing multiple series in a bar chart and appropriate selection controls. The tool helps in exploring facts about student's essays. Further details can be found in Chapter 4.

1.4.3 Verification methods

We have looked at different data sources to verify facts. Even if many sources provide the same information it still may be wrong: Often, sources copy from one particularly prestigious source and errors in such a source are then propagated. One typical example is "Water area" of Ethiopia. According to Factbook the water area of Ethiopia is at 104,300 sq km more than 10% of the area of the country. This is clearly wrong as one look at a map of the country shows. When we looked at different data sources for verification we found that several data sources e.g. "Index mundi" is just copying facts from Fact book. We only found the correct entry (7,444 sq. km) in Wolfram Alpha and of course use this figure. More general, we have conducted a case study to check the consistency of facts (area of countries, city and mountain rakings based on population size and height respectively). We have also used geospatial methods to verify some of the geographic facts. Further details can be found in Chapter 5, 6.

1.4.4 Geographic Server

Over the past few years many geographic servers have been developed but they often cover limited types of information. Our geographic web server (global-geography.org) is fairly unique since it also contains much non-numeric data, some not found in any data base known to us. Also users have to use search engines to find information related to different aspects of a country. We have solved this problem by consolidating substantial amounts of information related to a country. The comparison with leading geographic servers have shown that our system is leading in cultural and media (maps, images) aspects. Further details can be found in Chapters 8 and 15.

1.4.5 Virtual Lab

We have introduced a virtual lab that helps in exploring facts about countries. We have designed different types of reports that have made it much easier for users to explore countries. The virtual lab is fully interactive. Users can compare multiple countries with different parameters. The lab provides meaningful reports conveying key information and engage users in the learning process. Further details can be found in Chapter 9. We have also used data from the virtual lab to expand the story section of countries by providing unusual rankings.

1.4.6 Interactive Visualisations of Facts

Information visualisation techniques have been used for a long time to visualise large chunks of data. Our visualisations allow to compare countries using parameters like (population, energy productions, CO2 emissions etc.). Our ideas of visualising trade and different facts are fairly unique. We have developed an interactive globe and intelligently bind some statistical data to the globe. Further details about innovative visualisations can be found in Chapter 11. We have also used choropleth/thematic maps to visualise some of the statistical data of Europe.

1.4.7 Semi-automatic story generation

This thesis also explains how we have generated stories of about countries. We have employed clever heuristics for story generation. Some are based on statistical facts (exports, imports, labour force by occupation etc.). Some are chosen intelligently (statistical distribution of different parameters associated with Nobel Prize winners) that appeal to the curiosity of users. The users' feedback showed that the stories are quite interesting. The idea of stories is fairly new and appreciated by the users of Austria-Forum. They are also contributing and writing stories, see [8], [9], [10]. Further details can be found in Chapter 13.

1.4.8 Survey

The thesis also presents two detailed surveys. The first is about multiple data sources that contain facts about countries. It highlights the ways of data extraction from these data sources. The second survey covers approaches to responsive visualisations. The idea of responsive visualisations is similar to responsive web design which is an emerging trend in web development due to variety of screen sizes. Further details can be found in Chapter 7 and 14 respectively.

1.4.9 Objective Truth

The thesis also deals with some of issues related to objective truth. Often questions asked are not specific enough, e.g. the question "How many Nobel Prize winners are in a country?" This is not specific enough: Does one mean the number of winners born in that country, or the number of those that lived in the country when they got the prize, or the number of winners who got the prize for something they did while living in that country, etc., etc. We will discuss this in Chapter 5 that without clear definitions questions about the number of hospitals, number of airports, biggest cave in a country, longest river, etc. do not make sense. To avoid ambiguity, we have a web page in our server named as "Fields Description" [11] where users can see the definition of a parameter that is used either in virtual lab or in visualisations.

1.5 Thesis Overview

The current chapter serves as an introduction to the thesis explaining the history of the website Austria- Forum, research challenges accompanied with research questions. The remaining part of thesis is organized as follows. Chapters 2 and 3 are about information integration. The specific example of merging sets of images is discussed. The interactive visualisation in sparkling science (a project initiated for managing and keeping track of students' essays in different fields) is discussed in Chapter

4, concentrating on visualisation aspects of sparkling science project. Chapter 5 will highlight some of the inconsistencies found in online data sources when we started collecting geographic data for our server. Chapter 6 will highlight some of the geographic verification methods that can be used to verify online information using Maps. Chapter 7 presents a concrete view of facts collection and verification methods that are used. Chapter 8 is about the core project "global-geography.org". We will elaborate the key features. The "Virtual Lab" for exploring countries is discussed in Chapter 9. In Chapter 10 we will explain underlying hidden statistic of countries which are not obvious otherwise e.g. (Top ten and bottom rankings of 193 UN countries, similar countries, etc.). Chapter 11 focuses on information visualisation. Chapter 12 is mainly concerned with geo visualisations. Chapter 13 is about semi-automatic story generation. We end with a conclusion and mention future research work that will become part of the geographic server. "global-geography" is an international project: Different institutions are starting to participate, world-wide. Note, readers might find some repetition of text, as text is taken from papers published in different journals.

References

[1] "Total number of Websites" [Online]. Available: http://www.internetlivestats.com/total-number-of-websites/. [Accessed: 09-Nov-2015].

[2] "After 18 years, Internet archive site amasses 410 billion web pages" [Online]. Available: http://www.gmanetwork.com/news/story/360568/scitech/technology/after-18-years-internet-archive-site-amasses-410-billion-web-pages. [Accessed: 09-Nov-2015].

[3] "Austria-Forum." [Online]. Available: http://austria-forum.org. [Accessed: 02-Nov-2015].

[4] H. Maurer, "Austria-Forum and Beyond," Trans. Internet Res. TIR, vol. 10, no. 2, pp. 1–6, 2014.

[5] P. Korica-Pehserl and H. Maurer, "Semi-automatic information retrieval and consolidation with a sample application," 2012 Int. Conf. Emerg. Technol., pp. 1–6, 2012.

[6] G. Wurzinger, "Information Consolidation in Large Bodies of Information," J. Univers. Comput. Sci., vol. 16, no. 21, pp. 3314–3323, 2010.

[7] S. Murray, Interactive Data Visualization for the Web. OREILLY Media Inc., 2013.

[8] "The_ Waitomao_Caves." [Online]. Available: http://austria-forum.org/af/Geography/Australia/New_Zealand/Special_Information/The_Waitomao_Caves. [Accessed: 19-Oct-2015].

[9] "Shahara - The incredible city." [Online]. Available: http://austria-forum.org/af/Geography/Asia/Yemen/Special_Information/Shahara-_the_incredible_city. [Accessed: 19-Oct-2015].

[10] "From Bungabiddy Rockhole to Mount Leisler via the Sandy Blight Junction Road." [Online]. Available: http://austria-forum.org/af/Geography/Australia/Australia/Special_Information/. From_Bungabiddy_Rockhole_to_Mount_Leisler. [Accessed: 19-Oct-2015].

[11] "Fields Description". [Online]. Available: http://austria-forum.org/af/Geography/Lab/Fields_Description. [Accessed: 19-Oct-2015].

Chapter 2 : Information Integration

By integrating information we mean pulling information pieces from various sources together with as little loss of information and as little redundancy as possible. Given the huge amount of information we are confronted with, information integration is one of the biggest challenges of this century. This chapter describes methods and techniques for the information integration process. Information integration is an important topic but, in general, no really convincing approaches have been discovered, so far. By limiting the domain to the integration of image databases the problem becomes tractable, by using the techniques and methods ranging from image processing, knowledge management and natural language processing to feature engineering**.**

2.1 Introduction

A short definition on information integration is: "Integrating information means pulling information pieces together that belong to one specific domain or topic, yet come from a variety of sources and in this process, as little information as is possible is lost and redundancy is reduced to a minimum".

Information integration is sometimes also called information consolidation, sometimes information fusion.

The definition of information integration in Wikipedia reads a bit different: *"*Information integration (also called de-duplication and referential integrity) is the merging of information from disparate sources with differing conceptual, contextual and typographical representations. It is used in data mining and consolidation of data from unstructured or semi-structured resources. Typically, information integration refers to textual representation of knowledge but is sometimes applied to rich-media content."

Information fusion is a related term that means "the combination of information into a new set of information towards reducing uncertainty".

Examples of technologies to integrate textual information include string metrics which allow the detection of similar text in different data sources by fuzzy matching and similarity recognition of pictures when images are concerned. Many methods in these research areas have been developed and will be briefly mentioned below. They include some techniques known from Operation Research such as heuristic algorithms and meta-heuristics, similarity measures, use of constraints, combinatorial algorithms and collaborative efforts, as discussed in many papers in a rather different context, see [1]. In this chapter we concentrate on an important sub-area of information integration, the integration of pictures from various sources into a single collection, where pictures are grouped according to which documents they should be associated with.

Suppose we have a large collection C of pictures in a certain domain, each picture or subset of pictures associated with some "document". We want to merge an additional set of pictures P in a clever way with the sets in C.

"Clever" means using heuristics and constraints for situations such as: (i) Pictures in P outside of the domain of C are ignored; (ii) Pictures in P that already occur in C or are very similar to pictures in C are also ignored; (iii) Each remaining picture has to be associated with a suitable document in C subject to three restraints: (a) If a suitable document does not exist, a new document is created in C; (b) If a document gets "too large" some additional action (moving images or creating a new document) is needed; (c) An image may be associated with more than one document.

In what follows we discuss problems and solutions arising that may well be useful in other contexts, but then concentrate on specific data bases C and P and concrete constraints. However, it is worth to consider the scope of the general problem faced before we dive into details.

Knowledge is doubling in many fields (biology, medicine, computer science, etc.) within 6 years, data is doubling every 8 months, see [2]. Our overall objective is to be able to extract this information and integrate it with other sources and then summarize the most important parts for users to consume.

An information integration process starts after data sources are located by using techniques related to information retrieval. According to various definitions, see [3] [4], information retrieval is the science of searching for documents, for information within documents, and for metadata about documents, including searching relational databases and the World Wide Web. There are various methods of extracting information from the Web starting from parsing HTML pages with heuristics, or by using extraction rules, see [5]. Regarding general extraction of information from images, the attributes “legend”, “title” or the “alt” is very useful. The “legend” or “title” attribute contains text and usually carries valuable meta-information describing the image. Furthermore, information extraction methods can include computer vision methods like image recognition, see [5]. Machine learning approaches have been shown to be very useful for information extraction from text, including approaches that learn to extract various categories of entities (e.g., biographies, cities) and relations (e.g., company produces product) from structured and unstructured text, see [6], [7]. In order to extract the information from images we use text extraction methods combined with intelligent heuristics to identify terms needed.

Once extracted, data needs to be integrated. The result of automatic information integration should be a small collection of relevant information and documents on a particular topic in reliable quality. Methods like Natural Language Processing, String Matching and other AI methods are used. Another possibility of information integration is semantic integration of data, see [8].

This kind of integration is currently not yet possible on the Web, in general. However, given a restricted domain, there have been advances in integration of information - for example in domains of scientific papers such as [2].

We also have to consider that some relevant information will get lost or will be mismatched when a fully automatic integration process is used. To achieve better results humans might help by using sophisticated tools to look at what would be hopefully a comparatively small set of data of doubtful cases that are identified as such by the integration process.

One aspect seems certain: With the growing avalanche of information, information integration will be one of the most important topics confronting us in this century. Much of what has been called knowledge management for some 15 years, see [9], is closely connected to information integration. It pops up in many situations, of which we just mention a few as samples:

(a) In companies, pulling information together from various teams or other sources (like the experience reports of eight deep-sea oil-drilling teams to help the next team)

(b) In research, pulling together all papers dealing with topic x (but if z is an improvement of y, forget y, it would just increase redundancy)

(c) In news media (pulling thousands of reports on an earthquake together into a few coherent reports: this requires text merging, a very difficult issue as such)

(d) Creating a CV from a number of CVs of the same person.

(e) Pulling information on a person together. Academia Europaea is very much interested in applying this to members and to persons who are not, but should be members

(f) Combining image databases.

The last point (f) is indeed what this paper concentrates on. However, we restrict our attention to a situation that occurs when trying to enrich an existing Web site that already contains some images with further images from a large database of picture.

We will define the situation we want to deal with and look at the main issues in the next section 2.2. We will narrow our focus in section 2.3 and discuss the problems we were faced with and the new tools and algorithms developed as the technical heart of this research. We end with a short conclusion and a representative list of references.

Some of this work is based on previous research, see [10].

2.2 The problem in a general setting

In this section we deal with the following kind of situation. We have a large server which we call AF (Austria-Forum is abbreviated in this way) with material in a certain domain with the named documents. Each document may or may not have already pictures associated with it. Observe that the collection of all pictures in AF is what was called C above, and the term "document" corresponds to the documents in AF. The name n(d) of a document d in AF, each such name consisting of one or a few words, is descriptive of its content: if the document describes a person, the name of the document will be the name of that person; if the document describes a city, the document name will be the name of the city, etc. Further information on the document might possibly be gained by analysing not just the name of the document, but its contents.

In addition, we have a database P of pictures at our disposal. Each picture p of P has a bit of textual metadata associated with it. This database allows simple queries, each query returning a set of images, each image with its metadata. Using the names n(d) of the documents in AF to query P, we want to identify images in P that we can add as further pictorial material to one or more of the documents in AF.

We have to deal with a number of potential problems. Approaches to partially solve them depend much on the server AF and the picture database P at issue. For this reason we are formulating the types of problems arising when trying to identify new pictures in P and associate them with documents in AF in what follows, but indicate how we deal with them specifically only in the context of a particular application in section 2.3. We also accept the fact that a certain percentage of situations arising cannot be solved by any program, but will require human intervention. The challenge is to keep the percentage of such problems small, and provide tools to deal with them efficiently.

The solution of many of the problems encountered will be based on cleverly exploiting the textual data and meta-data associated with a document d of AF on one hand, and the meta-data associated with a picture in P on the other. Metadata describes various attributes of information objects and gives them meaning, context, and organization. In the process of information integration of images the above mentioned attributes are used for extracting information from image descriptions in the first phase, possibly supported by similarity recognition of pictures. In some cases image recognition techniques may indeed be essential: after all, if textual and meta-data information will not be enough to determine if a picture in the new database is indeed an additional picture, then it might be necessary to compare the pictures as such, using techniques from image processing, particularly using tools like openCv, PIL or imgseek which can perform this kind of similarity recognition. Image processing in general is used to carry out some operations on (digitized) pictures in order to get an enhanced image, or to extract some useful information from it, or to find similar pictures.

Suppose a query based on the name n (d) of a document d of AF yields a set of pictures X.

For each picture p in X we have to determine:

(a) Is the picture p already part of d or is it already associated with some other document? In this case we will ignore the picture.

(b) Is the picture p fitting into the domain of AF? If AF e.g. deals with cars, then surely a picture of an animal does not fit.

(c) If the picture fits the domain of AF, does it fit the document d? We e.g. do not want to associate the picture of some castle with the document "Robert Castle" describing the director of the International Theatre New York: after all, we queried for "castle" intending to find information on a person with name "Castle" but ended up with pictures of buildings that are castles! If the picture does not fit d, could it be that it fits another document? May be the picture of the castle mentioned fits a document describing the structure of medieval castles! This problem is not easy to solve in general, but as we will see in section 2.3 it can be handled fairly easily using human intervention with high accuracy in the particular application we are interested in.

(d) Is the picture p we found based on n(d) similar to a picture already associated with d? Of course we have to somehow define what we mean by similarity and then deal with such pictures appropriately.

(e) If d has already many pictures associated with it we have to handle the new picture (or even ones already attached to d) in a different fashion: remember, the pictures associated are part of a document, and even if we attach all pictures to the document as thumbnails to be enlarged when desired, anything approaching 100 thumbnails is creating a document that is too large. Thus, in this case we have to either attach the new picture to another document or create a completely new document. The latter is probably only possible with human intervention (choosing a proper name, etc.) but the first alternative often does yield a solution and can be done automatically in many cases. What we can do is this: we determine all documents u, v, w, in addition to "d" to which the picture can be associated with and attach it to the one with the smallest number of (so far) attached images. If all of them have already many images attached, we still need not necessarily create a new document, but we can try to move another image attached to "d" to some other document to "create space" for the new image. This idea is similar but an extension of the one used in [10].

To resolve all the aforementioned issues we have applied some heuristics. Experimental evidence has shown that the heuristics worked quite well, see section 2.4. They did fail occasionally but we consider this part of "bearable noise". An open issue is whether it may not be useful to have the same picture in more than one document. If so, we would suggest not using a physical copy but Nelson's transclusion idea that we have successfully implemented in our concrete application to which we are now turning.

2.3 A concrete application

The server AF our application is concerned with is the Austria-Forum [11] as discussed in Chapter 1.

Documents in the Austria-Forum have a name that consists of one or more words, like "Graz" or "Graz University of Technology". Special characters are not allowed, except for a few. Like "Kreisky, Bruno" is the name of a famous former Austrian chancellor, "Kreisky" as family-name, "Bruno" as first name. Documents consist mostly of text or pictures, or often both.

The database of pictures P from which we want to import additional material is the IMAGNO database, see [12]. It has some 100,000 images, some already incorporated into the Austria-Forum (three years ago), some not (since they have been added to the database more recently, or where overlooked the first time around). The aim of our project is to determine suitable additional pictures from IMAGNO and to attach them to suitable documents in AF.

Before proceeding let us make clear that in what follows, we take the liberty to translate some parts of documents in Austria-Forum and in the IMAGNO database into English, or choose English examples that are analogous to German ones for better readability.

Pictures in IMAGNO have a unique numeric identifier (which will turn out to be critical to our efforts) and some textual description (metadata) associated with them. The IMAGNO database can be basically queried by inputting a phrase or a number of words for which IMAGNO searches with an "AND" connective in the metadata. Unfortunately, the structure of the metadata is not consistent; thus much has to be based on experiments and textual search. To stick to the example above, both the query "Kreisky, Bruno" and "Bruno Kreisky" yield all pictures of Bruno Kreisky in the database. However, if there is a picture whose metadata contains "Peter Kreisky, son of Bruno Kreisky" this image will also be found, since both words "Bruno" and "Kreisky" occur in the meta-data. There are even more serious problems: some queries produce entries that one would not expect. The query "Graz" yields among many pictures of the city Graz also pictures of some other villages or even historic events that are from an archive in Graz, but have little else to do with Graz. Further, the IMAGNO database also contains pictures completely unrelated to Austria, like the launch of the first submarine in UK. Such pictures have to be excluded. There are also border-line cases. Like, would one include a picture of a musical performance in London simply because an Austrian was the conductor of the musical?

We will discuss in what follows how we handle this and other problems following the steps (a) through (f) described in section 2.2.

However, it is convenient to deal with (a) and (b) together, i.e. with the questions:

(a) Is a picture x already in the Austria-Forum?

(b) If not, is it fitting into the domain of the Austria-Forum?

To make sure that we get pictures that are not complete misfits we use as queries only such terms that are based on names of documents in the Austria-Forum. Actually, we are restricting attention to the main general purpose encyclopaedia within Austria-Forum, AEIOU, but we will use AEIOU and Austria-Forum synonymously in what follows for convenience.

AF Data Classification

We apply some heuristics to further break down the set of documents we consider. This helps in the querying process and during manual intervention of humans. Specifically we consider:

(i) Long Biographies:

They can be determined easily due to a special mark-up. They are the first ones to be used for querying IMAGNO.

(ii) Short Biographies:

For a name to be in this category we exploit the fact that all biographies contain at least a date of birth preceded by * like *14.1.1975 or date of death which is preceded by † as shown in Figure 2.1.

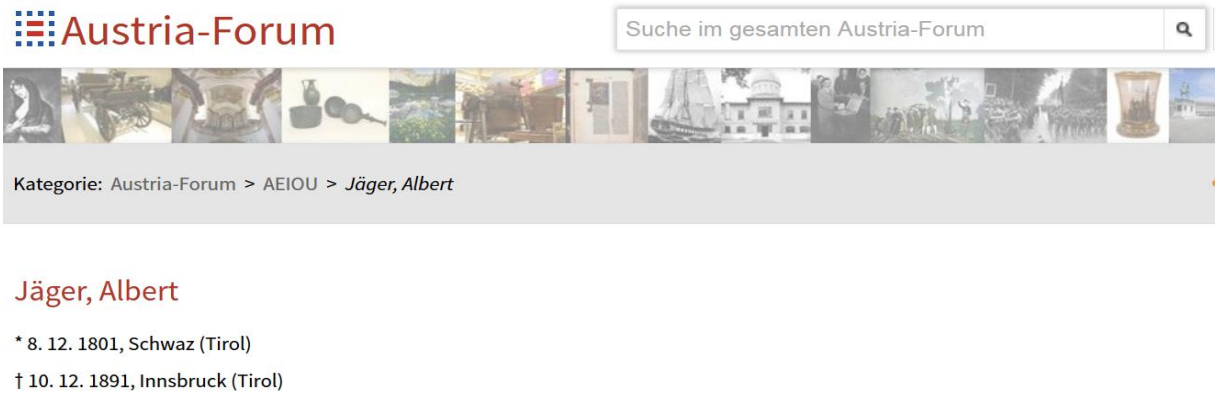


Figure 2.1: Austria-Forum document of Jäger, Albert

(iii) Towns/Villages:

The document text contains special mark-ups that indicate that the document is dealing with a town or village as shown in the Figure 2.2.

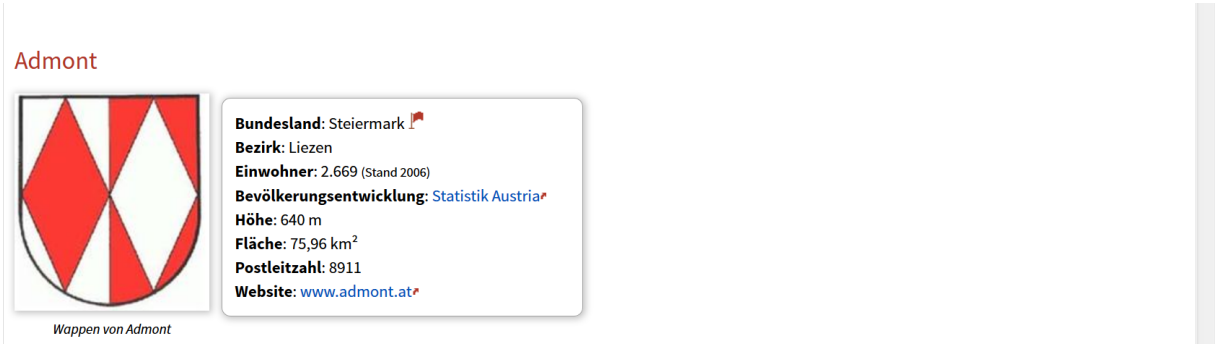


Figure 2.2: AEIOU representing document of village Admont

(iv) Other entries (single word, multiple words):

The remaining AF documents are then classified as either “one word entry” or “multiple words entry” based on the number of words in their names.

We deal with the above mentioned categories in the order indicated. Figure 2.3 shows the number of entries in each category.

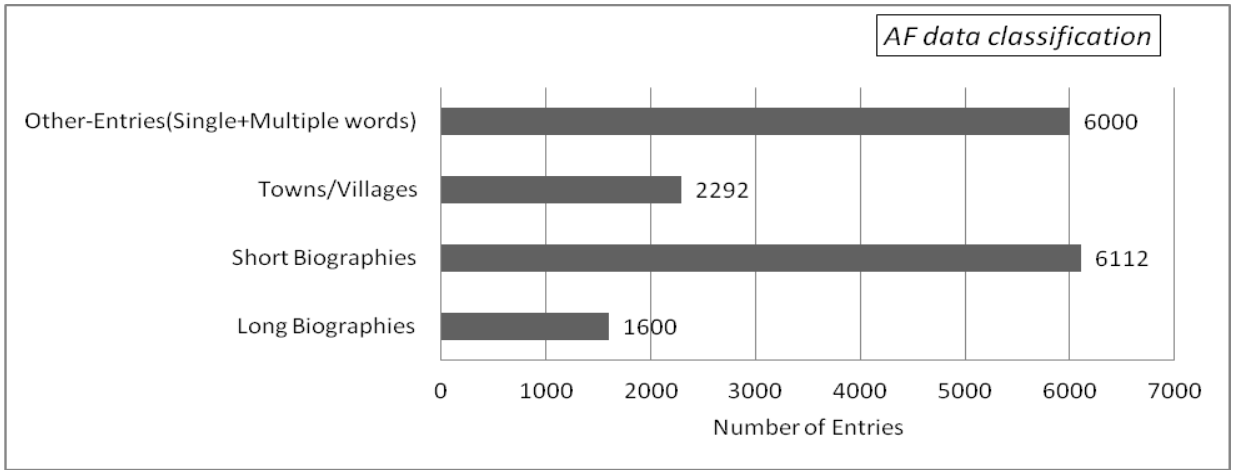


Figure 2.3: AF Data Classification

Let us consider the set of long biographies mentioned in some detail, since many of ideas presented also apply to the other sets. Austria-Forum contains roughly 1600 long biographies and

had, to start with, already 1500 pictures from the database P (IMAGNO) associated with them. To find further images we used the names of persons in the biographies.

In German, most persons have a surname (like “Roessel”) and a first name (like “Franz”). Sometimes there are further middle names. Sometimes names contain titles that belong to the name (like Richard, Baron of Martinage: here Martinage is a kind of the surname, Richard the first name and “Baron of” a filler that should be ignored). This can be further complicated due to diacritical characters (like the infamous “Umlaute” in German, or the slash in the letter O when the name comes from Danish, or some special diacritical signs like in Andrè). There are further situations to be considered: surnames sometimes come hyphenated; some are followed by an attribute like “the younger”; and some actually consist of two entirely different names like “Hartmann Hermann, really Arthur Nestranowitsch” or “Mertine Susan, nee Martonati” (indicating that her name before marriage was actually Martonati.) This list is by no means exhaustive but shows the difficulty one is confronted with, even if we forget about names like Sherbakova and Sherbakov, just indicating that the first is a woman, the second a man if surnames are of Slavic origin.

One first measure before searching is to resolve hyphenation (by either contraction or replacing the hyphen by a blank: the second version turned out to work better), to replace diacritical symbols by their “base” character, e.g. “Umlaut a” by “a”, “Umlaut u” by “u”, etc. It is also clear that it does not make sense to search in the picture database for part of a name: In case of a person like “Roessel Franz” both “Roessel” and “Franz” (the latter, a very common German name) would give pictures of any one in whose metadata either of the two names occur. To search for “Roessel Franz” will not work either, since the search is an AND search, so if the description contains a part like “Franz Willerin kissing Susan Roessel” would give a picture that certainly does not show Roessel Franz.

As a result, it turned out that the search for a person with just surname and first name has to consist of two phases: one, in which all pictures whose description contains both words is carried out, but then the description is checked by using a regular expression that more or less says: “either Roessel has to occur directly before Franz or the other way around”. Regular expressions therefore play an important role in the second phase.

The “more of less” comes from the fact that still more is needed: in a case like “Richard, Baron of Martinage” the word “of” can be deleted: the deletion of any word with a maximum of 3 letters turned out to be quite a good strategy, particularly if the word starts with a small letter which in German indicates it is just a preposition, an article, or such. Also, certain words like “Baron” can be collected in a “stop words list” and can also be ignored. Thus, the first phase would do a search for Richard AND Martinage, the second phase would check that both words Richards and Martinage occur “reasonably” close together, since in the description it may well say “Richard, Baron of Martinage”, i.e. the two critical words do not occur directly behind each other.

Sometimes the title consists of more than two words like Abraham_a_Santa_Clara. To handle such cases we apply regular-expression which checks the occurrence of all the words of title, in the image description.

Clearly, the situation gets still more complicated with hyphenated surnames, multiple first names/ middle names, special characters like “King Karl Eduard III” or pen names, etc. These are special cases that can be handled by considering the approaches as Common Authority File, GND, see [13]. However, we are pulling out pictures from the IMAGNO database that does not allow Roman numerals such as I, II, III; therefore we are forced to treat Roman numerals as stop words, requiring manual checks afterwards.

We do not want to discuss all details here, but just want to emphasize that our methods do not work 100% automatically as some artificial intelligence addicts would want, but we accept the fact that heuristic algorithms, observation of constraints etc. can help to solve the majority of

problems, leaving some cases to be resolved by human intelligence. Above is no attempt to describe all details that had to be handled, but rather to make sure that readers understand the variety of cases and hence the large number of techniques that had to be applied to reach our goals.

We ended up developing fairly sophisticated methods to find for each entry in the biography (hopefully) all pictures, eliminating those already in our collection of pictures (using the picture identifiers of IMAGNO mentioned above) and checking the rest for “approval” or “disapproval”.

Our heuristics found a list of additional pictures fitting (“approved”), and a list not fitting (“disapproved”). For the 1600 biographies at issue we found 4101 pictures. This list is further cut down by removing duplicate entries that is, by ignoring images that are already present in AF by comparing image identifier. This can be extended by eliminating pictures that are very similar using appropriate measures. Our approach resulted in approximately 2000 images as approved images.

At this point we had to develop tools that would allow to check easily how the heuristics performed. Would the heuristics be good enough that no manual checking would be necessary? Let us briefly describe the tool developed for the “approved” list: the tool for the “disapproved list” is fairly analogous.

The tool in Figure 2.4 shows the name of the document in AF, pictures and descriptions of the images from the IMAGNO database and gives the option to either do nothing (heuristic worked perfectly) or to disapprove a picture by clicking “Disapprove” button.

Approve
Disapprove

Achleitner,_Friedrich
6:00610578
Description: Das 2. literarische cabaret der Wiener Gruppe. Gerhard Rühm, Friedrich Achleitne und Oswald Wiener. Wien. Photographie. 1959. Author: Franz Hubmann.

Approve
Disapprove

Achleitner,_Friedrich
8:00577597
Description: Das 2. literarische cabaret der Wiener Gruppe. In der Nummer "2 Welten" zertrümmern Friedrich Achleitner und Gerhard Rühm ein Klavier auf der Bühne. Wien. Photographie. 1959. Author: Franz Hubmann.

Approve
Disapprove

Achleitner,_Friedrich
9:00577596
Description: Das 2. literarische cabaret der Wiener Gruppe. In der Nummer "2 welten" zertrümmern Friedrich Achleitner und Gerhard Rühm ein Klavier auf der Bühne. Wien. Photographie. 1959. Author: Franz Hubmann.

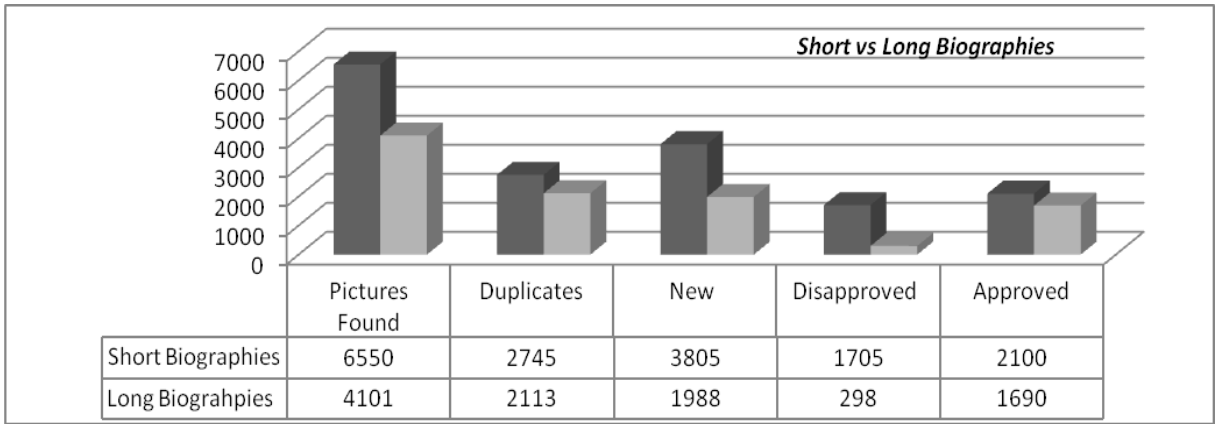
Figure 2.4: Tool for checking approved/disapproved pictures. User can disapprove pictures using this interface as it shows all approved pictures.

The time invested in developing complex heuristics and combinatorial techniques paid off: the approval worked very well: we found only 2% spurious errors in the 60% cases that were manually checked. Thus, the pictures approved by our heuristics need not be manually checked, quite a triumph, with one small sad footnote: Some photographers shoot 50 pictures of the same scene/person within a few seconds hence with only trivial variations: left hand in pocket, cigarette

in hand, cigarette in mouth, etc.: if one wants to cut down on such a collection of very similar images the (multiple) disapprove option of the tool comes in handy. Clearly, some sophisticated picture similarity recognition algorithm could cut down on this if it really is a big problem in other applications.

The picture disapproval heuristics did not do quite as well for obvious reasons. Suppose we look for pictures of Ludwig van Beethoven. We find many pictures with the proper search, but the checking disapproves some since only Beethoven occurs in the text (metadata) with no Ludwig at all. Here, the (multiple) approval tool came in handy. Of 100 disapproved cases we approved 7, i.e. the disapproved cases do require manual checks. However, since the situation only has occurred in a few cases (there are not that many very famous persons where this phenomenon happens) we found this combination of clever programs and human intervention quite tolerable.

Let us now turn to (ii), short biographies. They are also comparatively easy to extract from AF due to special mark-ups. We deal with them exactly like with the long ones, with no big surprises occurring, so we present only numbers for comparison, see Figure 2.5.



	Pictures Found	Duplicates	New	Disapproved	Approved
Short Biographies	6550	2745	3805	1705	2100
Long Biograhpies	4101	2113	1988	298	1690

Figure 2.5

Now we turn to (iii), i.e. villages and towns in AF. Again, due to special mark-ups we can easily extract them, and we find 4415 of them, with already 1593 pictures attached to them.

Like in the previous cases (i) and (ii) the idea is to use two phases: the first consists of extracting all pictures from P, the second examining them, providing two lists of “approved” and “disapproved” images that then can be further manipulated using appropriate tools.

Names of villages or towns in Austria-Forum often consist of one word, like “Linz” or “Graz”. Other names are more complex like “Bruck an der Mur”: This distinguishes that city as situated on the river Mur from Bruck an der Leitha (Leitha is another river) or Bruck an der Glocknerstrasse (Glocknerstrasse is a famous alpine road). The rule to eliminate diacritical symbols and to neglect words of length three or less and to ignore words with an initial small letter works here fine, also, like for “Kirchberg am Wechsel”, “Übelbach bei Graz” (note Umlaut!) , etc. In all those cases we do a search for the critical words (Bruck – Mur, Bruck- Glocknerstrasse, Kirchberg- Wechsel, Ubelbach-Graz) and in the second phase try to determine if what we found is applicable.

The list of stop words is now rather different: important are “Bad” (a prefix added to names if the village is a recreational area), or “St.” and “Sankt” (a prefix somehow associating the village name with a saint), the prefix “Markt” (signifying a certain status of the village), “Burg”, “Festung”, “Schloss” indicating that the village was built near a fortification or castle, etc.

The approval/ disapproval algorithm becomes tricky, since e.g. “Traunstein” is both a village and the name of a mountain, so metadata and text analysis (or picture recognition) has to try to eliminate the mountain, by detecting the word “Berg” (mountain) or such. Figure 2.6 shows the overall results regarding images of villages and towns.

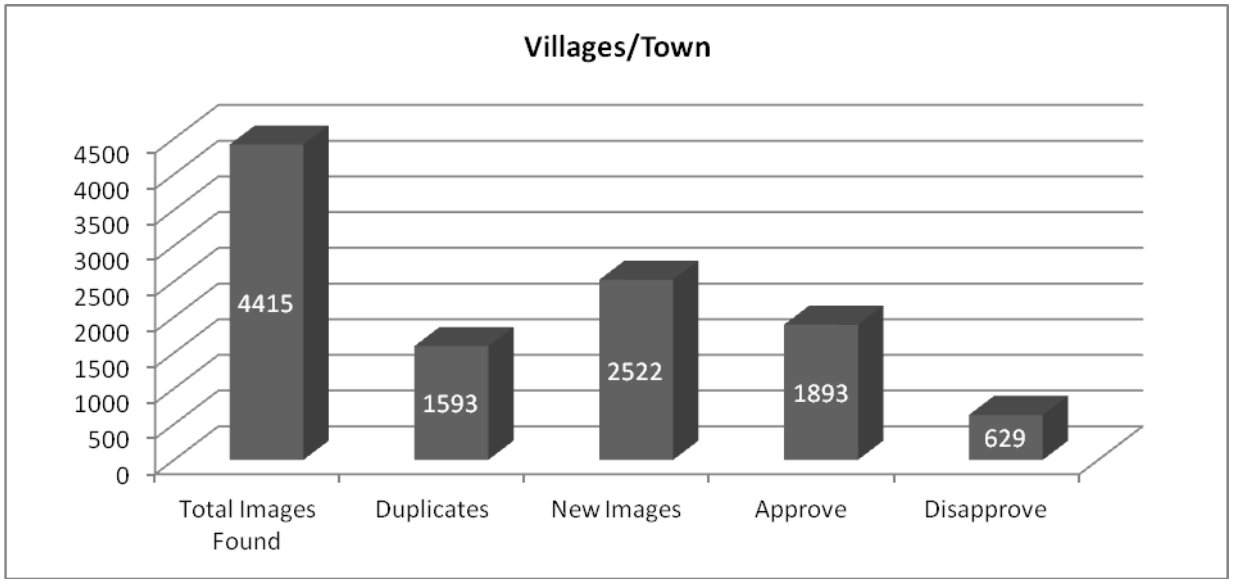


Figure 2.6: Village images breakdown

However, the biggest difficulty is the large number of images found for e.g. cities like Salzburg or Graz. Images there should often be associated with other documents (like a building or festival in that city rather than the main entry on the city). This has been described before and in the paper [10] and has been partly taken care of as pictures of persons, where Salzburg is mentioned in the description, have already been dealt with in (i) and (ii).

(iv) Other AF-Entries (Single word and multiple words).

In this category some manual work was necessary, since the entries in the AF were often generic entries like describing what a steamboat “Steamboat” is, yet the picture would be much too specific to fit, like “Wreck of a steamboat after being hit by a bomb in 1944”. And sometimes it was the other way around: a very specific entry on e.g. the seed “poppy” and a generic picture “meadow in Northern Austria with a mixture of flowers, including corn flowers and poppies”.

2.4 System Evaluation

We searched images for approximately 10,000 entries (document names) in Austria-Forum. This yielded approximately 20,000 images from IMAGNO. First we removed all the duplicate images using IMAGNO image identifier and then from the remaining set of new pictures, the system associated around 12,000 images as approved image which are then used by Austria-Forum to populate its documents. The failure rate of algorithm in case of approved pictures was less than 5% where as in disapproved cases the failure rate of algorithm was less than 10%.

Table 2.1: Summarized Image Integration Analysis against Each AF Category

S#	Description	Results				
1	Total Images Found	18,766				
2	Long Biographies	*Total Images*	*Duplicates*	*New Pictures*	*Approved*	*Disapproved*
		4101	2113	1988	1690	298
3	Short Biographies	*Total Images*	*Duplicates*	*New Pictures*	*Approved*	*Disapproved*
		6550	2745	3805	2100	1705
4	Villages/Towns	*Total Images*	*Duplicates*	*New Pictures*	*Approved*	*Disapproved*
		4115	1593	2522	1893	629
6	Other AF-Entries	*Total Images*	*Duplicates*	*New Pictures*	*Approved*	*Disapproved*
		4000	*1300*	*2700*	*2200*	*500*

Accuracy rate Calculation

We have calculated the accuracy of our classifier through investigating approved and disapproved lists of pictures. The results shown in Table 2.2 reveal the accuracy of classifier which is found to be approximately 97% in case of approved pictures for "other AF-Entries" and calculated using the equation below.

% Accuracy= $Correct\ Approved/Approved * 100$

% Accuracy =2150/2200*100→97%

Table 2.2: Classifier Accuracy

S#	***Category***	***Approved***	***Correct Approved***	***Disapproved***	***Correct Disapproved***	***Accuracy***
1	Long Biographies	1690	1680	298	280	**99**
2	Short Biographies	2100	2000	1705	1600	**95**
3	Villages/Towns	1893	1800	629	513	**95**
4	Other AF-Entries	2200	2150	500	405	**97**

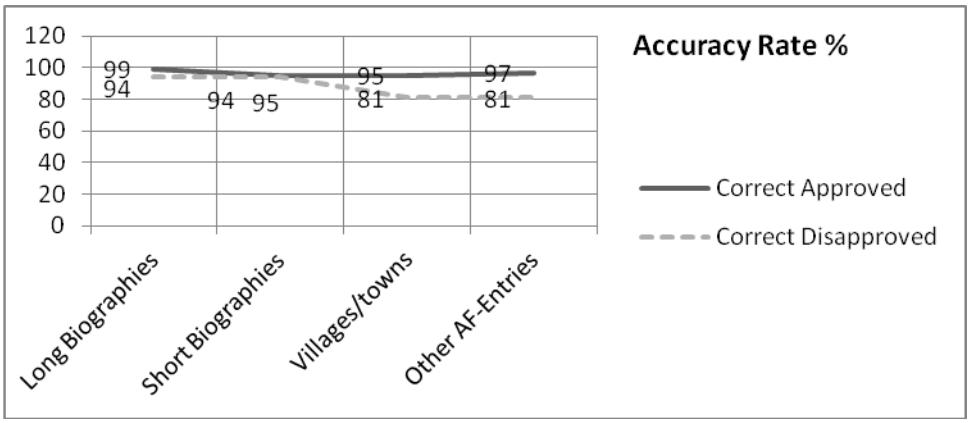


Figure 2.7

2.5 Conclusion and future work

Before this project was started, Austria-Forum contained 8,000 pictures of IMAGNO. After this project, this number increased significantly to 20,000. This is certainly a success, yet it is clear that some pictures in IMAGNO were not found, since only entries in Austria-Forum were used as query terms. Thus, if a person x had no entry in Austria-Forum but IMAGNO did contain some pictures of x, such pictures of x would only be found if x would be present with other persons (or objects) that do occur in AF.

This raises another issue: Although a picture was sometime associated with more than one entry, this was not done systematically. For instance, a group photo showing 8 persons would usually be only associated with the "main" person and the choice was often not clear. A first impression might be that this means that pictures of persons in a group might "hide" in AF without a way of being found. This is, fortunately, not the case: since the text ("caption") of a group photo would usually mention all persons, a full text search for any of the persons on the picture would find that picture.

We also have to state here that our approach was only successful because of the consistency of the picture ids. Thus, merging pictures from a database where one cannot rely on picture ids, would require still deeper manipulation of the textual and metadata description associated with the pictures.

Overall, it is clear that merging databases or integrating various information sources is an extremely important task, but at the moment all techniques that can be used are typically heuristic in nature and quite domain dependent.

There are many other large image databases that can be used for AF. Examples are image databases of Austrian National Library, the one from the Austrian Press Agency, or the one of the Austrian Tourism Board etc. However, the challenge is to cope with the difference of metadata associated with the pictures as their structure is different in each of the databases.

Observe further that we have used only entries in AF as a starting point for the queries. Clearly other sources are also possible, thus expanding the scope of the future work still very much.

References

[1] Z. Blázsik, C. Imreh, and Z. Kovács, “Heuristic algorithms for a complex parallel machine scheduling problem,” Cent. Eur. J. Oper. Res., vol. 16, no. 4, pp. 379–390, 2008.

[2] G. Wurzinger, “Information Consolidation in Large Bodies of Information,” J. Univers. Comput. Sci., vol. 16, no. 21, pp. 3314–3323, 2010.

[3] C. D. Manning, P. Raghavan, and H. Schüze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[4] D. A. Grossman and O. Frieder, Information Retrieval - Algorithms and Heuristics. Springer, 2004.

[5] W. Nakapan, G. Halin, J.-C. Bignon, and M. Wagner, “Extraction of building product image from the Web,” Int. J. Intell. Syst., vol. 19, no. 1–2, pp. 65–78, 2004.

[6] A. Carlson, J. Betteridge, R. C. Wang, J. Estevam R. Hruschka, and T. M. Mitchell, “Coupled semi-supervised learning for information extraction,” Proc. Third ACM Int. Conf. Web Search Data Min., pp. 101–110, 2010.

[7] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002, pp. 71–78.

[8] M. R. Genesereth, A. M. Keller, and O. M. Duschka, “Infomaster: An Information Integration System,” SIGMOD, Tucson, AZ, pp. 539–542, 1997.

[9] H. Maurer and K. Tochtermann, “On a New Powerful Model for Knowledge Management and its Applications,” JUCS, vol. 8, no. 1, pp. 85–96, 2002.

[10] P. Korica-Pehserl and H. Maurer, “Semi-automatic information retrieval and consolidation with a sample application,” 2012 Int. Conf. Emerg. Technol., pp. 1–6, 2012.

[11] “Austria-Forum - das Wissensnetz aus Österreich.” [Online]. Available: http://austria-forum.org/. [Accessed: 26-Oct-2015].

[12] “IMAGNO.” [Online]. Available: http://www.imagno.at/. [Accessed: 22-Nov-2015].

[13] “Katalog der Deutschen Nationalbibliothek.” [Online]. Available: https://portal.dnb.de/opac.htm?method=simpleSearch&cqlMode=true&query=idn%3D118601121. [Accessed: 22-Nov-2015].

Chapter 3 : Implementation Aspects of Image Integration System

Web data sources are queried in different fashion due to the modelling differences. Dynamic web content that is displayed in a browser usually originates from the databases and is presented without any structure. Different problems are encountered during extraction of images and their description from Austria-Forum and IMAGNO website. We will highlight those problems in this chapter along with their solutions. Further, the mechanism of proper arrangement of pictures before uploading them in Austria-Forum will be discussed. The procedure of meaningful association of images with multiple documents of Austria-Forum will also be explained.

With the growth of the Web different novel fields have emerged, information integration is one of them. There is a wide range of data sources that include structured data like relational databases or object oriented databases, semi-structured data like in XML files, and unstructured data like text on the Web, data residing in spread sheets or in text files. Even more data exists in other formats such as pictures in our particular application. Prior to integrating data sources, the first step is to locate these data sources on the Web. Information retrieval is the science of searching for relevant documents from a large collection. Relevant data has to be extracted prior to the integration process.

The task of information integration is also summarized by a quote of a well-known researcher [1]:

“There is abundant amount of data available on the Internet and there are many opportunities to combine this information to build new applications and tools." (Craig A Knoblock).

This chapter will address the following research questions and mainly discuss the implementation details:

RQ1. How do we extract images from IMAGNO? Most specifically, we will explain regular expressions for extracting images automatically.

RQ2. How to search for images in IMAGNO using document names from Austria-Forum? This problem is related with Meta data associated with images (image number, description). There are two types of Meta data i) Descriptive: This informs about the creation of documents/images ii) Semantic: This informs about semantics of documents/images.

RQ3. How to remove redundant pictures?

RQ4. How to extract existing images from Austria-Forum about biographies, villages etc. (It is essential to extract existing images to keep track of duplicates).

RQ5. How to associate images with other documents of Austria-Forum suggested by a special tool (provision of auto-suggestion service).

RQ6. How to deal with roman numbers that are often found in document names of Austria-Forum.

RQ7. How to arrange the images to be integrated (sorted meaningfully) before uploading them in Austria-Forum.

RQ8. How to upload the finally selected pictures in Austria-Forum?

3.1 Introduction

To fully comprehend an information integration system, it is essential to have a look at some of the definitions found in the literature, see [2] [3] [4].

1. Information integration is the process of combining data residing in different data sources and providing users a unified view of these data sources.
2. Information integration systems offer a uniform interface to a multitude of data sources, whether structured like databases, semi structured like XML or unstructured like text documents.
3. It is the process of providing uniform access to a set of autonomous and heterogeneous data sources.
4. Enterprise information integration (EII) is the ability to support a unified view of data and information for the entire organization.
5. Information integration can be viewed as the problem of modelling information sources in a coherent manner so that they talk to each other.
6. Information integration is the problem of providing a unified and transparent view to a collection of data stored in multiple, autonomous and heterogeneous data sources. The unified view is achieved through a global schema which is linked to the data sources by means of mappings.
7. Combining information from multiple autonomous information sources and answering queries using the combined information.
8. The aim of information integration is to support query processing over structured and semi-structured sources as well as services.
9. To make disparate data sources work together to better reply user queries.

3.2 Image Extraction Procedure

In this particular section we explain the image extraction procedure using two Web data sources: one is Austria-Forum and the other is IMAGNO. The reason for choosing IMAGNO is that it has historical images which are not commonly found in other data sources. Also, the owner of the database, Dr. Christian Brandstätter allowed Austria-Forum to use the pictures as follows:

- Upto a certain size without watermarks.
- Somewhat larger pictures with watermarks.
- High resolution pictures have to be down-sized.

3.2.1 IMAGNO Images Extraction Procedure

IMAGNO offers an interface with a search textbox in which users type the query term as shown in the Figure 3.1. For instance, if users want to see the pictures about Graz they enter the term “Graz” in the textbox as shown in the Figure 3.2. In order to automate the process of query search with document names of Austria-Forum, we have used Html Unit [5], a GUI less browser which provides an API for filling HTML forms automatically.

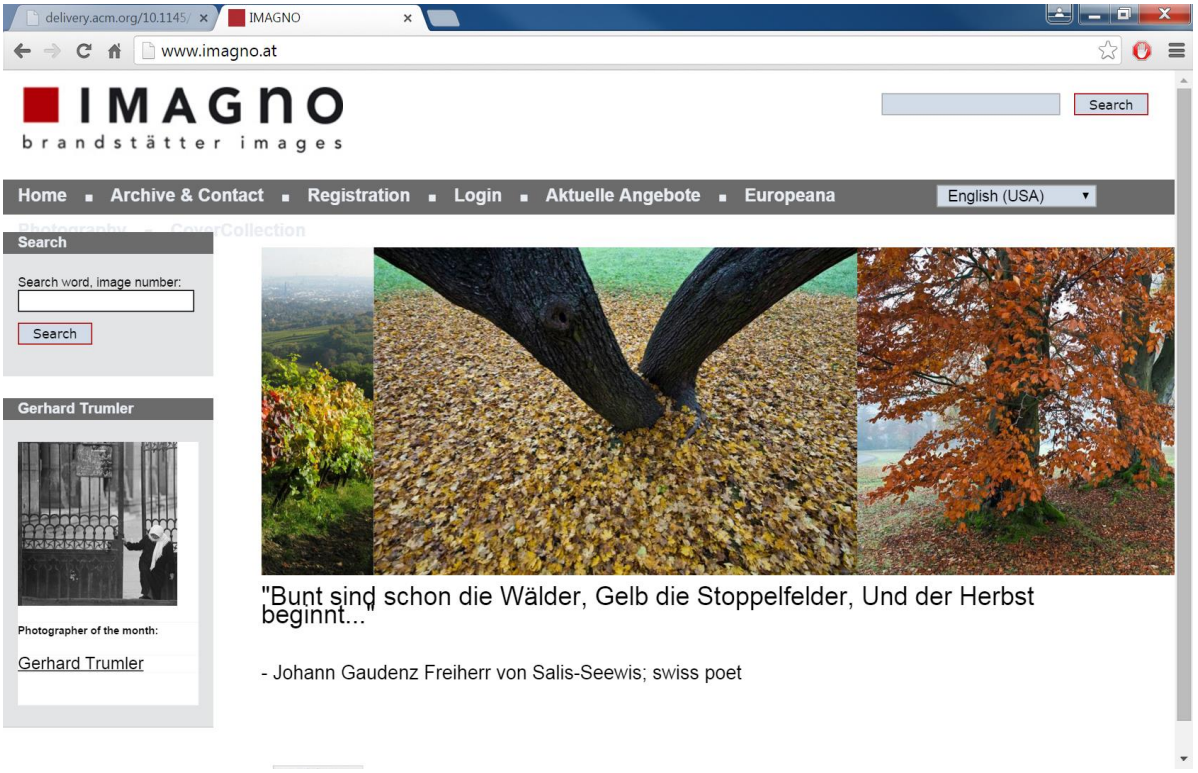


Figure 3.1: IMAGNO main interface: a huge data repository of more than 50,000 historic images.

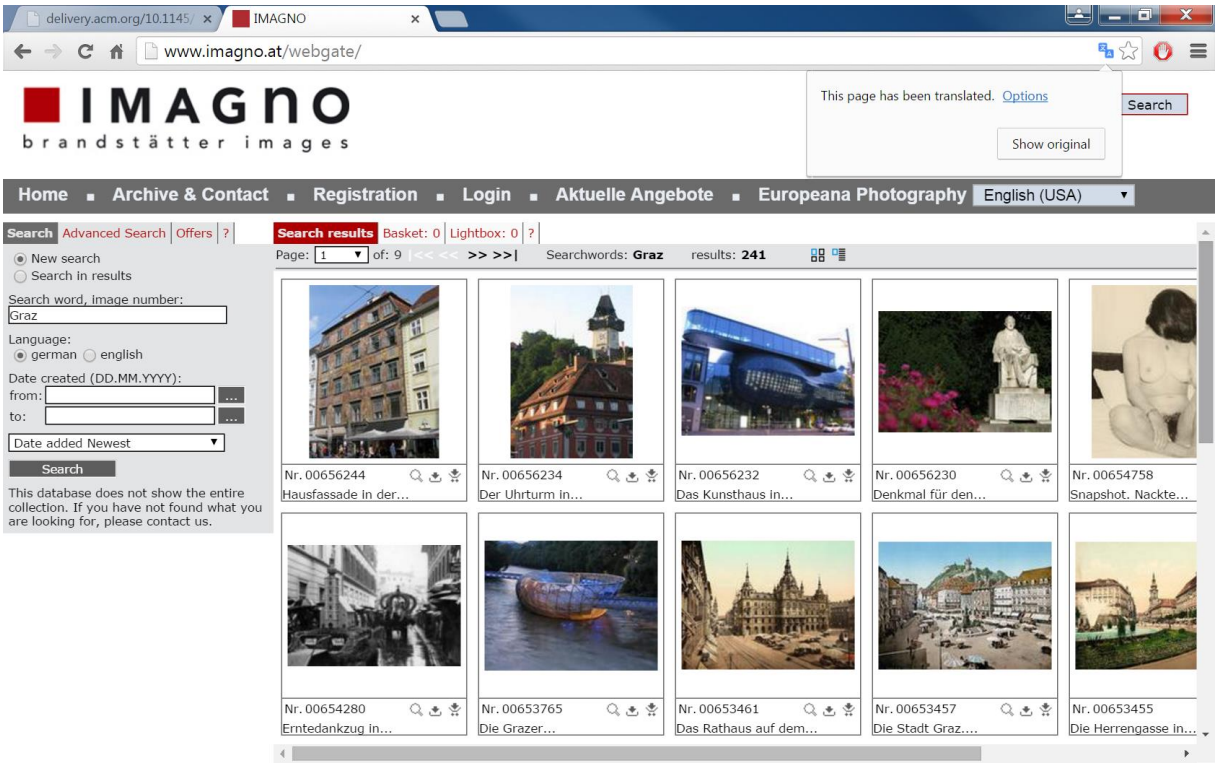


Figure 3.2: 241 images are listed against search term “Graz” from IMAGNO database.

The images shown in Figure 3.2 are just thumbnails. Once a thumbnail is clicked a larger image is displayed along with Meta data and description as shown in the Figure 3.3.

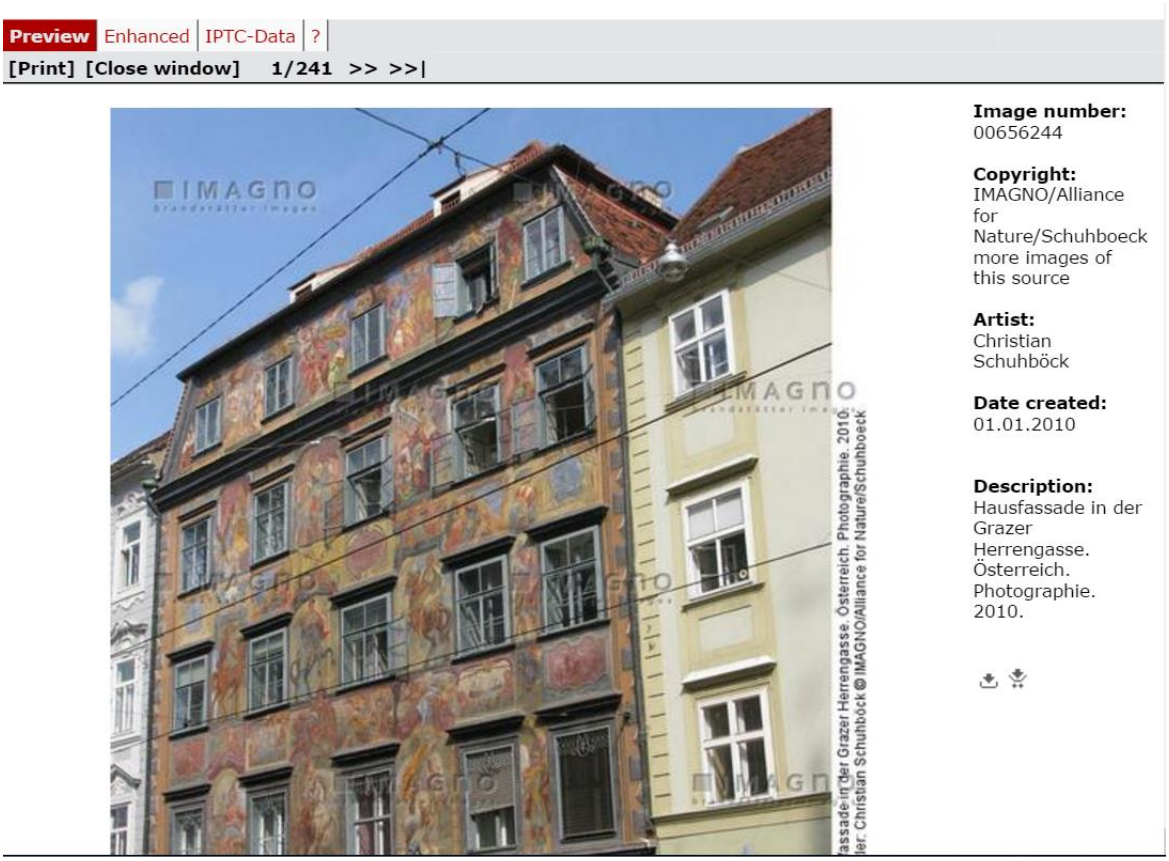


Figure 3.3

We have used XPath[1] which is the suggested way to navigate through elements and attributes in an XML document. XPath uses path expressions to select nodes or node-sets in an XML document [6]. Once we have a reference to the HTML page in XML format, we search for a <tr> and <td> elements of HTML table by XPath which contain images and textual description as shown in the Figure 3.2. Another problem is to extract images from multiple HTML pages as IMAGNO displays only 30 images per page. We have to click on the "next" button to gather further results. To navigate through all these pages automatically we dynamically bind the value of parameter PAGECURRENT as this value is used for discrimination between pages that show images for the same query term. The value is calculated by dividing the value "total number of images" by the constant 30 (as 30 images are displayed per page). The value "total number of images" shows total search results and is found on the HTML page displaying images. We use the XPath expression */b[@class=searchresultnum]* to get this value, as it is stored in HTML tag <b> with class "searchresulnum" shown in the Figure 3.4.

[1] http://www.w3schools.com/xpath/

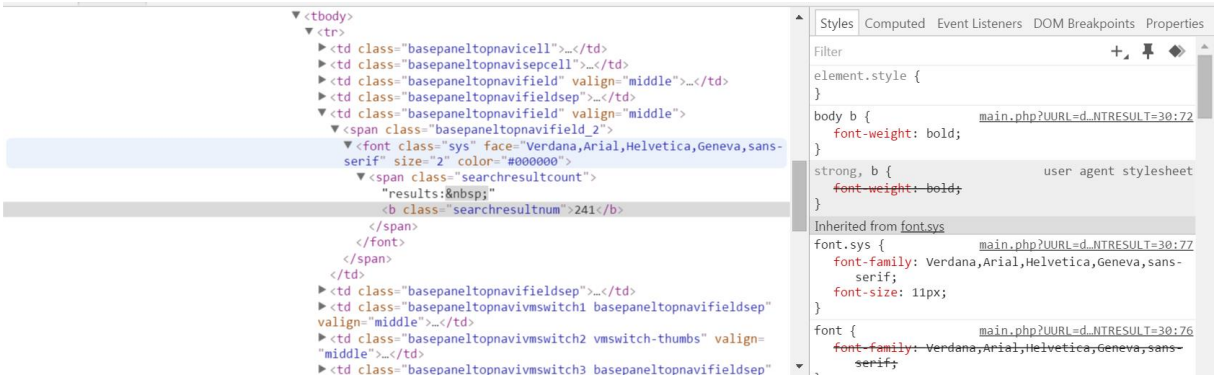


Figure 3.4: Firefox console elements tab representing html of web page of IMAGNO.

The IMAGNO website is built using HTML frames, hence exact source code is placed in the HTML pages enclosed in frames. We therefore first parse the frame and extract the URL of HTML page located in the src attribute of the frame.

A typical HTML source code for IMAGNO image is listed below.

```
<img class="thumbimage" src="thumb.php/00629384.jpg" title="Description: Graz. Steiermark. Bismarckplatz. Farbige Bildpostkarte . 1905. Photographie." />
```

IMAGNO images are displayed using <img> tags with class thumbimage, therefore we use XPath expression `img[@class=thumbimage]` to locate them. We extract src attribute values for getting image URLs and extract title attribute values for image description. Every image in IMAGNO has an identifier that is 8 bit digit so we use regex `[0-9]{8}` to extract image number from HTML pages and we populate the relational table, More details can be found in the papers [7][8].

3.2.2 AUSTRIA-FORUM image Extraction Procedure

It is necessary to extract existing images from both Austria-Forum and IMAGNO. One obvious reason is to keep track of duplicates. The second reason is we want to replace the old version of images with new enlarged images as IMAGNO has larger version of images in its repository since 2013 (with or without watermarks in some cases).

Austria-Forum does not provide form like interfaces. It has sets of categories.

Figure 3.5: Biographies in Austria-Forum listed alphabetically in lexical order

Each category has a number of terms associated with it. For example, persons associated with biographies are listed in Figure 3.5. To see the pictures concerning a particular term one has to click on that term which takes us to an HTML page containing images. These HTML pages are referenced by URLs enclosed in HTML table with class wikitable. Therefore we apply XPath expression *{//table[@class=wikitable]}* to extract URLs shown below.

```
http://austria-forum.org/af/AEIOU/Bilder_Admont
http://austria-forum.org/af/AEIOU/Bilder_Graz
```

A source code for images stored in HTML pages corresponding to above URLs is displayed here

```
<div class="image_image"><img src="AEIOU_Graz_Bilder_Graz_00201145.jpg" height="375"
width="250" title="Arkadenhof des Landhauses in Graz" /></div>
```

We use XPath expression *//div[@class=image image]* to locate the div element containing img tags and then we extract image URLs and populate a relational data source “AustriaForum” for further comparison with IMAGNO images.

3.3 Smart Search

Querying IMAGNO using document names often produces an unnecessary set of images. For instance, biographies in the Austria-Forum have documents with names such as:

1: Lossl Friedrich Ritter von
2: Lassnig, Maria

These names, when used as query terms may either produce large sets of images or no result. Therefore we remove some characters {underscores, commas, hyphens} from document names before querying IMAGNO. Sometimes a set of words like {von, admiral} affect the query results, therefore we first generate a list of query terms that can produce a reasonable set of images.

3.4 Redundancy Removal

Information integration may produce redundant data that has to be removed. As indicated earlier, Austria-Forum contains some IMAGNO images already, most of them fortunately uniquely identified by the image numbers e.g. 00414866 or 00623029 etc. This makes it easy to remove duplicate images from the set of images retrieved from IMAGNO.

3.5 Image association with Multiple Documents in Austria-Forum

To associate images with documents in Austria-Forum we first identify a suitable document based on the image description. The classification of images based on their textual description is a challenging and important problem in automatic image retrieval and integration [9],[10]. Here are a few simple heuristic rules which proved to be very useful in the document identification process from textual description associated with each image. The terms which are commonly found in image description but do not represent document names belong to the set

{month, year, date, preposition, conjunctions, roman characters, less than 3 character words}.

However, it can be the case that a term starting with a capital letter represents a document name and terms which are classified as nouns are serious contenders for document names.

Our tool discussed in Chapter 2 also allows to associate images with multiple documents using the proposed document names besides approval and disapproval. User can associate an image with multiple documents using the tool shown in the Figure 3.6.

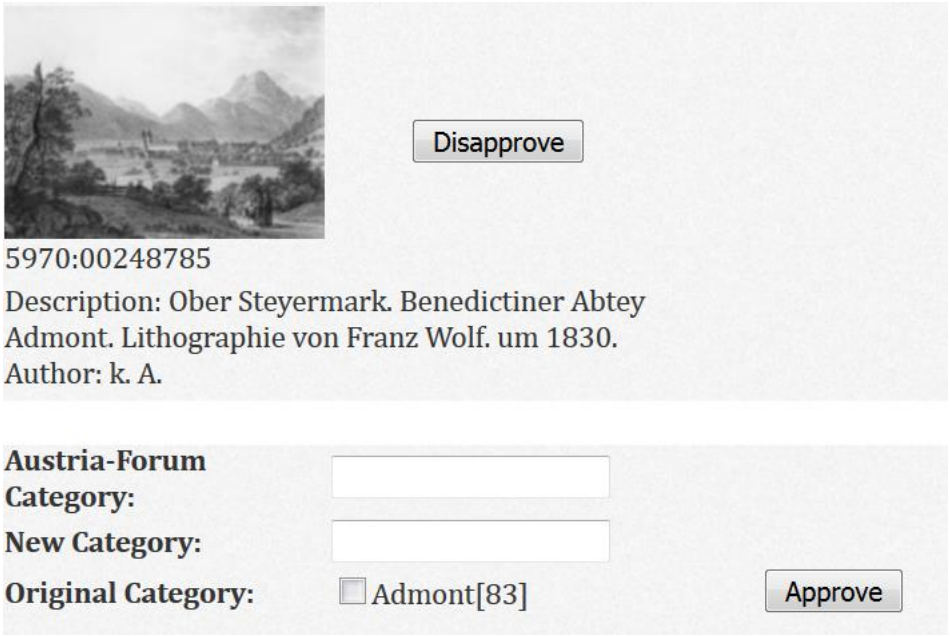


Figure 3.6: Multiple categories approval

3.5.1 Dynamic documents suggestion

Dynamic document suggestion service (often referred as autocomplete) enrich the text box with suitable options. As users enter characters in the search box one at a time, document suggestions are offered to them in drop down form as shown in the Figure 3.7.

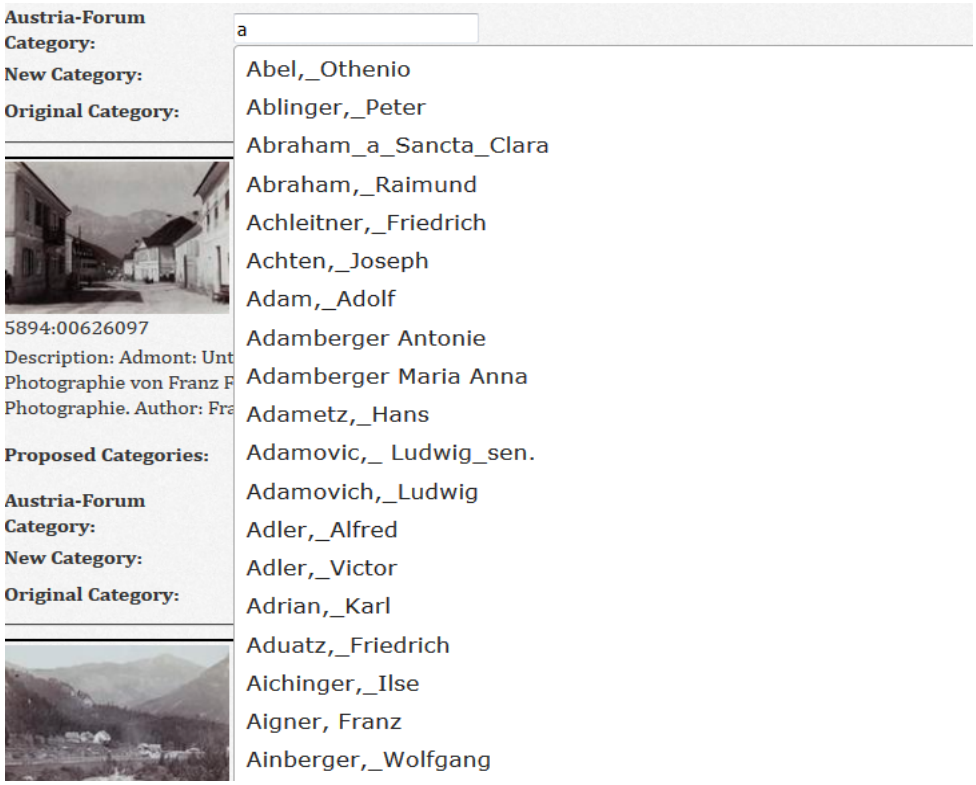


Figure 3.7: Users are facilitated by an auto suggestions service that displays list of documents in Austria-Forum that starts with letter “a”.

3.6 Handling the special case of Roman Numerals

There are a few documents about emperors or other prominent persons with the same name in Austria-Forum. We cannot search IMAGNO with full emperor name as there are Roman numerals attached to some names. Roman numerals in the search term often produce null results. The only way is to first search IMAGNO with search term without Roman numerals and then associate pictures with emperors by examining their birth date; as pictures in IMAGNO usually have year in their meta tags or in their description.

3.7 Pictures Arrangement tool

After merging the two sets of pictures, the next step is the arrangement of the merged pictures in some reasonable order.

Dragging and Dropping in browser cannot be handled by ordinary tools, therefore a drag and drop plugin "TableDnD" was used in the development of this picture arrangement tool, see [11]. This plugin allows reordering of rows in table. The plugin requires JQuery[2] for proper execution [12]. We have tweaked this plugin for the pictures arrangement in browser.

Let us dig deeper and explore the tool with some examples. To illustrate what we mean by "arrangement", take an example of "Altenberg". For instance the pictures of Altenberg,_Peter are placed in an unordered layout without conveying much sense as shown in Figure 3.8. Pictures having different context and meaning are mixed with each other. The picture with id "00115601" is representing a painting painted by Altenberg. Picture with id 00621843 shows a pose of Altenberg (in thinking mood). To present the pictures in a meaningful way we have developed a tool which is used for arrangement of pictures.

Figure 3.9 shows the pictures of Altenberg,_Peter which are now in a more logical sequence. The first group of pictures shows his face, hence the user who is viewing the document quickly recognizes him. The second group of pictures is related to his activities followed by his paintings and so on. It makes it much easier to arrange the pictures, as pictures can be easily dragged and dropped at the desired location to obtain a logical sequence. We have used the same sequence in storing those pictures in the Austria-Forum.

Figure 3.8: Unordered layout of pictures

[2] https://jquery.com/

Figure 3.9: Ordered layout of pictures

The “picture arrangement tool” also provides an additional option of deleting unwanted and surplus pictures, thus making it easy to select the most suitable picture among similar pictures and discarding the rest. To illustrate this, consider the set of pictures of “Christian Ludwig” as displayed in Figure 3.10.

Figure 3.10: Similar group of pictures

The first five pictures with ids 00180673, 00238067, 00238069, 00620936 and 00238068 are quite similar in nature and just showing different poses of “Christian Ludwig” with a monkey. The pictures with ids 00620905 and 00238022 also show the same posture of the people sitting on chairs. After removing similar and unwanted pictures using the “Delete” option of the picture arrangement tool, the final set of pictures hence selected is shown in the Figure 3.11.

Figure 3.11: Final set of selected pictures

3.8 Pictures in Austria-Forum

Our ultimate goal was to enhance the Austria-Forum documents with new and enlarged pictures. The final set of pictures of village Admont is merged into Austria-Forum document that represents village Admont, as shown in the Figure 3.12. The small pictures in IMAGNO have no watermarks whereas large pictures have watermarks except picture postcards. All downloading of images in bulk is done with special permission from IMAGNO administration using proper user name and password. There were some issues regarding size of thumbnails which were handled by using image API “imgscalr” in JAVA [13].

Figure 3.12: Final set of selected pictures of Admont in Austria-Forum.

References

[1] "Craig Knoblock." [Online]. Available: http://usc-isi-i2.github.io/knoblock/. [Accessed: 25-Nov-2015].

[2] M. R. Genesereth, A. M. Keller, and O. M. Duschka, "Infomaster: An Information Integration System," SIGMOD, Tucson, AZ, pp. 539–542, 1997.

[3] G. Wurzinger, "Information Consolidation in Large Bodies of Information," J. Univers. Comput. Sci., vol. 16, no. 21, pp. 3314–3323, 2010.

[4] R. Baeza Yates, Modern information retrieval the concepts and technology behind search, 2nd ed. ACM Press Books, 1999.

[5] "HtmlUnit – Welcome to HtmlUnit." [Online]. Available: http://htmlunit.sourceforge.net/. [Accessed: 21-Oct-2015].

[6] "XPath Syntax." [Online]. Available: http://www.w3schools.com/xsl/xpath_syntax.asp. [Accessed: 25-Nov-2015].

[7] R. Mehmood and H. Maurer, "Towards the integration of images on the Web," in Proceedings of iiWAS, 2013, doi: 10.1145/2539150.2539250

[8] R. Mehmood and H. Maurer, "Aspects of Information Integration in a Wiki," in Proceedings of iiWAS, 2013, doi: 10.1145/2539150.2539230

[9] W. Nakapan, G. Halin, J.-C. Bignon, and M. Wagner, "Extraction of building product image from the Web," Int. J. Intell. Syst., vol. 19, no. 1–2, pp. 65–78, 2004.

[10] L. Dey, M. Abulaish, J. Sharma, and G. Sharma, "Text Mining through Entity-Relationship Based Information Extraction," 2007 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. - Work., pp. 177–180, 2007.

[11] "Table Drag and Drop JQuery plugin" [Online]. Available: http://isocra.com/2008/02/table-drag-and-drop-jquery-plugin/. [Accessed: 19-Oct-2015].

[12] "jQuery." [Online]. Available: https://jquery.com/. [Accessed: 25-Nov-2015].

[13] "imgscalr – Java Image Scaling Library" [Online]. Available: http://www.thebuzzmedia.com/software/imgscalr-java-image-scaling-library/. [Accessed: 25-Nov-2015].

Chapter 4 : Sparkling Science

Information visualisation is a discipline that deals with visualising abstract structures, as discussed in [1] . However, visualisation of data is often only part of what one wants, the other part is interaction with data. Such interactive visualisations allow users to work and experiment with data. Interactivity becomes even more important when dealing with complex data covering multiple dimensions. In this project we have provided interaction support along with visualisation to explore facts related to essays written by students on different topics, in different schools, of different ages (12-17) and involving both male and female students.

The project under discussion was called WWWW ("Was Wir Wissen Wollen", an acronym not easily translated into English, so we will just call it W4). It was part of a "Sparkling Science" initiative funded by the (then) "The Federal Ministry for Science and Research (BMWF)". It was an attempt to actively involve students in the research process. The project was aimed at students from grade 6 to grade 11 (12 to 17 years old) and was coordinated by the Institute for Information Systems and Computer Media (IICM) TU Graz. The idea was to allow students to compose a multimedia essay on a topic of their choice but only after researching the topic in a multitude of sources. One reason was the concern that plagiarism is on the rise in the sense that writing an essay or seminar paper for a school or university assignment is now often done by using a search engine to locate relevant literature and use copy and paste [2]. To tackle this, students were allowed to use the material from the Web that was classified as "reliable" with the constraint that also books, typical from school libraries or the provincial library have to be consulted.

This chapter addresses the following research questions.

RQ1. How to visualise data based on essays produced.

RQ2. How to program an interactive interface that allows a number of different applications.

RQ3. How to summarise information visually.

4.1 Project Description

This project addressed the copy-past dilemma. The major concern was that copy-paste is used by students to such an extent that they sometimes do not even read pieces that are pasted together. They often compose documents/essays by gluing together snippets from various servers (mind you, some clever plagiarism detection algorithm may notice the change of style between parts of the essay on the Web with a minimum of "filling" in between to yield a "new" essay [2]).

In the project W4 we tried to force students (on their own or in small teams) to work in a different fashion. Students could choose some topic they were interested in to compose a multimedia essay. However, before doing so they had to consult the school library, search the Web for relevant material and pictures or other media clips they could use or link to (without violating copyrights). They had to document what they have looked at, adding to their essay a list of "proper citations". They were only allowed to compose their document after the topic and the references found were deemed to be satisfactory by the teachers and members of the IICM team. Figure 4.1 shows the main entry to Sparkling Science in Austria-Forum.

Figure 4.1: Sparkling Science main entry in Austria-Forum, taken from [3]

It was interesting to note the choice of topics, revealing much about the interest of students, in some cases even revealing potential psychological problems. Like when one student compiled how “effective weapons were during WWII to kill people”, etc.

One source of reliable information was the Austria-Forum. Since the final documents would reside on Austria-Forum, students were allowed to include parts of material on that server, particularly pictures, by using the “Insert Page” functionality, that allows to transclude (to use Ted Nelson’s formulation) pictures not by copying, but by reference. A good example is the essay on the Alps, see [4]. This essay contains a large number of pictures, a few of them shown in Figure. 4.2. Part of the source code for the picture section is shown below:

```
[{InsertPage page='AEIOU/Glockner' start='1' end='2'}]
[{InsertPage
page='Wissenssammlungen/Bibliothek/Österreich_aus_der_Vogelperspektive/Tirol/Wildsp
itze' start='1' end='2'}]
[{[InsertPage page='Bilder_und_Videos/Herrliche_Natur/Großvenediger' start='1'
end='2'}]
[{InsertPage page='AEIOU/Dachstein' start='3' end='4' }]
```

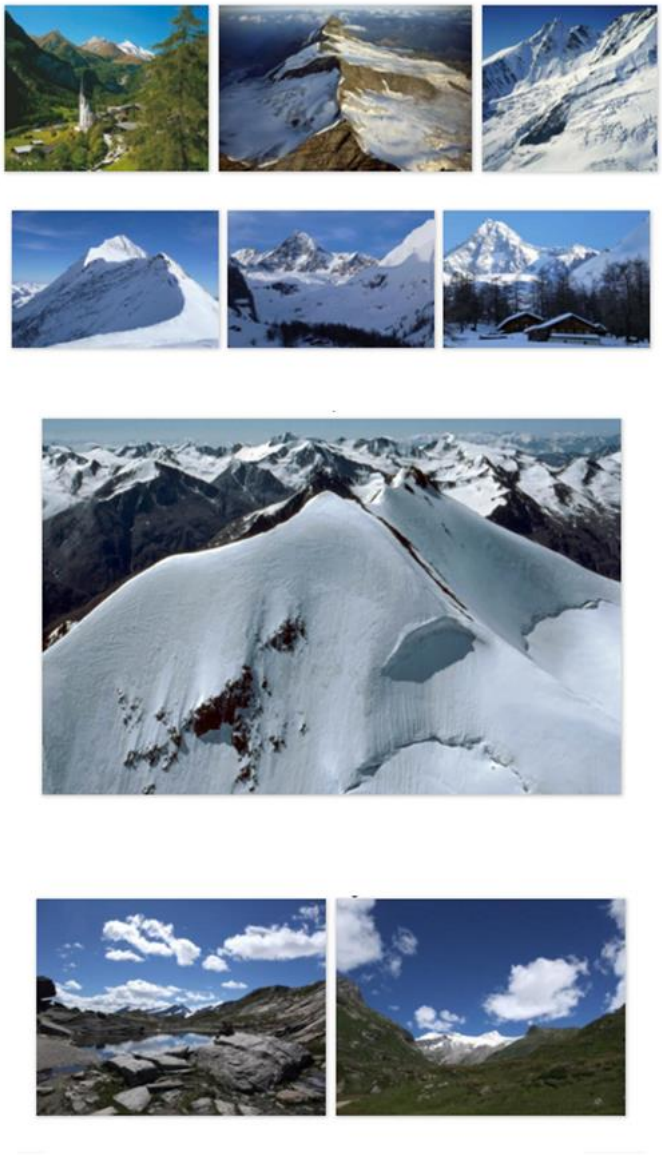


Figure 4.2

This shows that quite a few pictures were not copied into the document but transcluded, see [5],[6],[7]. The main aim of the project was to collect essays in an optimal searchable fashion using meta data and such. Eventually, a pool of approximately 400 essays was consolidated under the project Sparkling Science[3].

As information analysts our concern was to build a tool that unveils underlying statistics e.g. how many male or female students participated from which school in which area, and such.

4.2 Interactive Bar Chart Tool

The interactive bar chart tool, shown in the Figure 4.3 shows the distribution of essays that were written by group of students in different fields. A bar chart is simple but elegant way to visualise data [8]. We have used d3 (JavaScript Visualisation Library) for this visualisation. The code snippet to generate bar chart is shown below.

[3] http://austria-forum.org/af/Sparkling_Science

```
1: bargroup.append("rect").
2: .attr("height",function(d,i)
3:{return barHeight-yScale(gebiet[Object.keys(glabel).sort()[i]]);})
4: .attr("width","20px").
5: .attr("fill","black")
```

The text labels that represent essays topics e.g. physics, biology etc. are rotated vertically so that visualisation fits fell within the screen. The following code snippet is used for text rotation.

```
1: group.append("text").text(function(d){return d})
2: .attr("transform", function(d) {
3: return  "rotate(-90)"
4: });
```

With the interactive bar chart tool shown in Figure 4.3, one can do interesting investigations. Some of examples are given below. First, we observe that a total of 493 essays were produced.

4.2.1 Comparison by gender

When selecting the gender in the first selection bar as female (weiblich) shown blue, and in the second as a male (männlich), shown black, one can see the total distribution by gender: it involved 260 male and 233 female students as shown in the Figure 4.4.

Although male students contributed, not surprisingly, more to “physics”, “technology” and “sports”, yet there are some surprises: male students were more interested in “gastronomy” (cooking!) than girls, but girls were more interested in e.g. mathematics, by a factor of 12!

4.2.2 School-wise comparison

The interface shown in the Figure 4.5 presents school wise comparison. The comparison is shown between schools Pestalozzi and Körösi. There are more contributions from Pestalozzi in almost all fields. The teachers motivated their students well to participate. However, in Psychology and Medicine there are more contributions from students of Körösi.

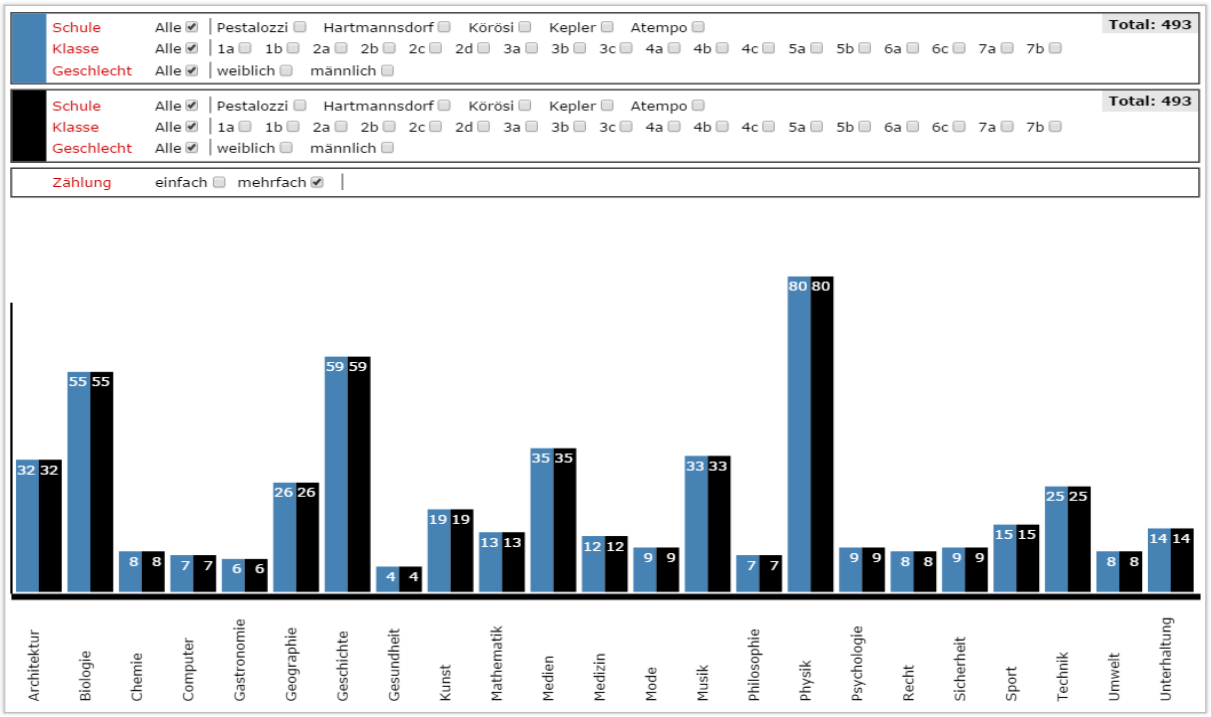


Figure 4.3: Interactive Bar Chart tool

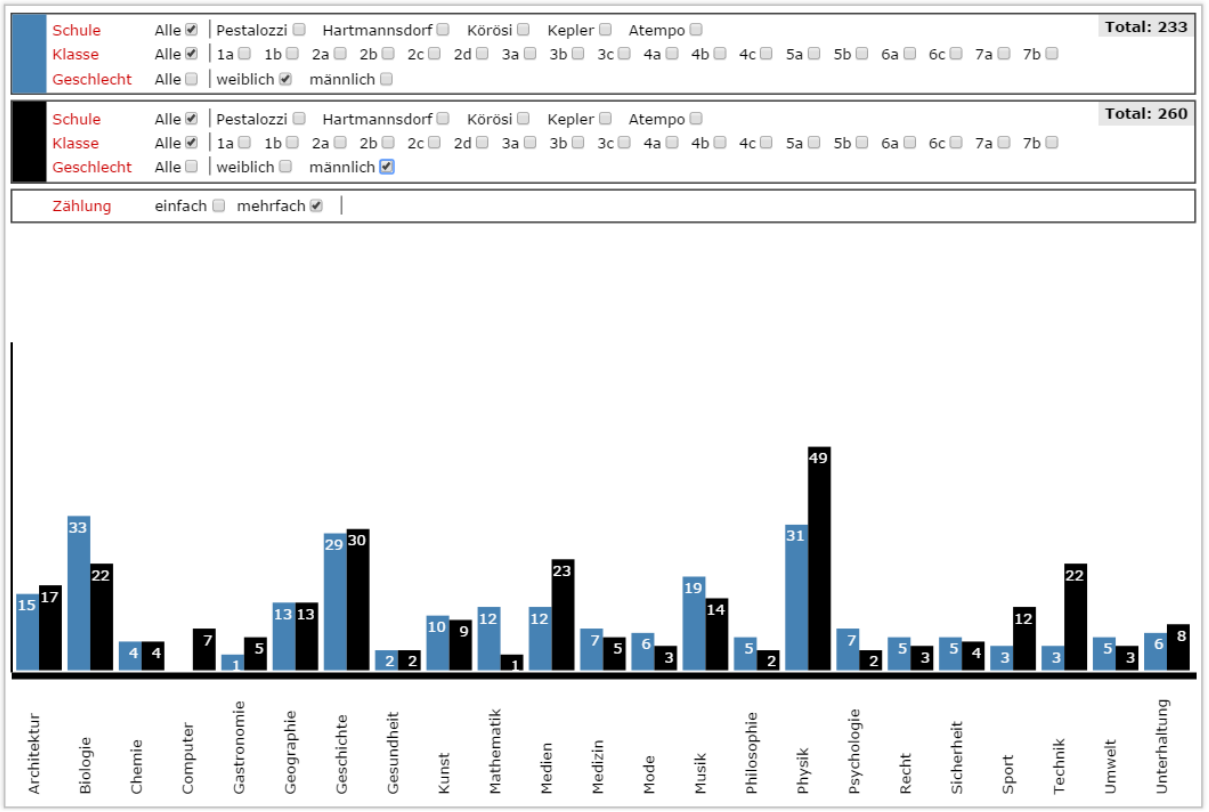


Figure 4.4: Interactive comparison by gender.

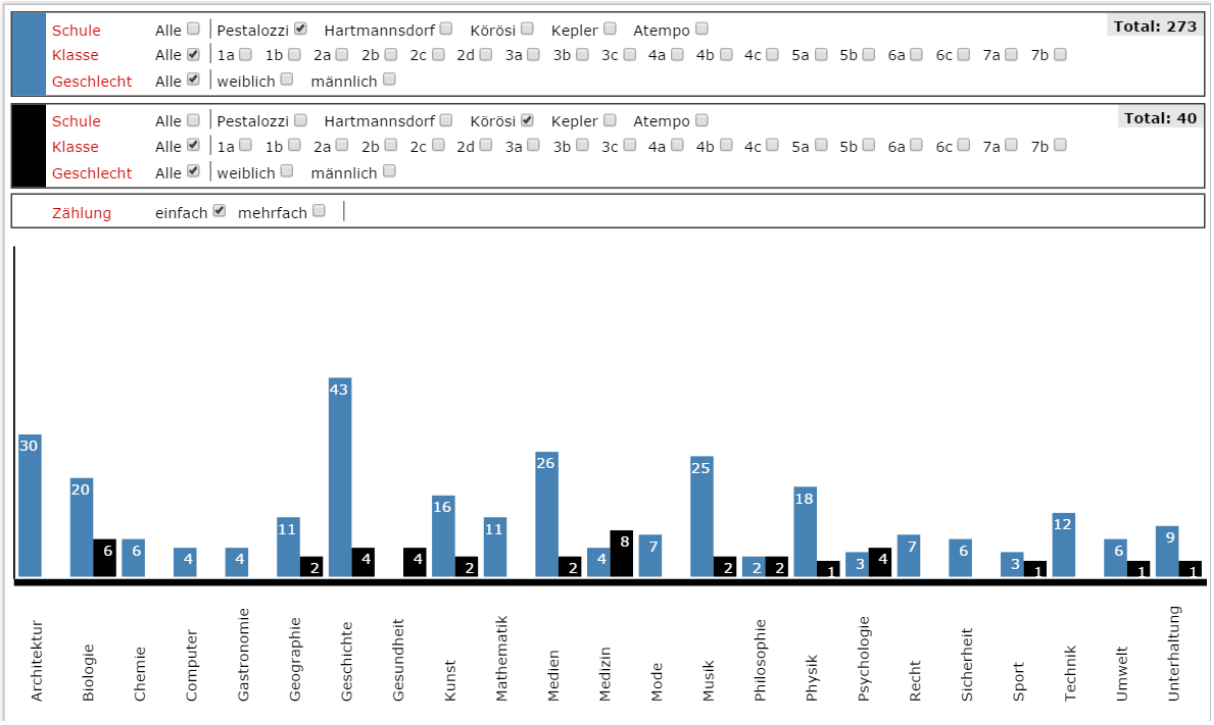


Figure 4.5: School wise comparison

4.2.3 Comparison by age

Age wise comparison is shown in the Figure 4.6. The chart is showing selection of lower classes (blue) versus upper classes (black).

Note that there were only about half as many students age 10-14 compared to those age 15-17. Thus, to get the real distribution of interest one basically has to double the blue column. This leads to interesting observations: Interest in “biology” or particularly “physics” is decreasing dramatically with age, while interest in e.g. “media” and “art” is increasing, etc. The results lead us to interesting discussion with the teachers involved, e.g. concerning the startling drop in e.g. “physics”. It turns out that physics in lower grades is dominated by fun experiments, in higher grades by calculations that are considered more boring! Deeper investigations reveal that the drop of interest in “physics” is smaller for girls than for boys. This is probably explained by the earlier fact pointed out that girls are more interested in mathematics than boys are!

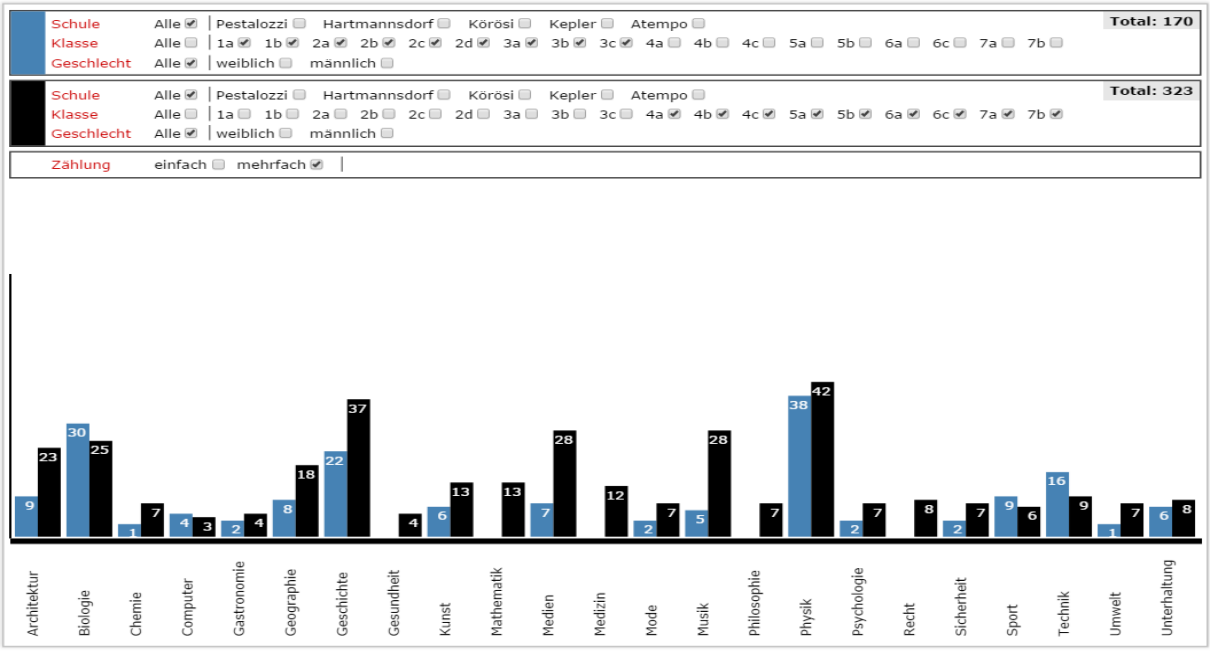


Figure 4.6: Chart comparing ages 10-14 against 15-17.

4.3 Overall Analytical Analysis

The charts shown in Figure 4.7 and 4.8 presents the overall summary. It shows the diversity of contributions made by students. One can see at a glance that most contributions are allocated to the theme/essay “History” (53), followed by “Biology” (51) and “Physics” (46). There are slightly less contributions in the fields of “Media” (32), “Music” (31), “Architecture” (30); followed by the other topics. The numbers displayed correspond to those when one does not take into account multiple authors. They correspond to those in the underlying interactive bar chart tool, if the box "Single" is checked as shown in the Figure 4.3. By clicking there on "Multiple" (this option considers authors who are writing an essays. E.g. an essay is counted as 2 if is written by two authors, 3 if it is written by three authors), then “Physics” comes on top with 80 and dominated “History” with number 59 which comes at place two. The “Biology” with 55 moves to rank 3. In other areas, the figures do not change significantly. This expresses that many posts were written by a team of 2-3 people in physics, but in the other areas most contributions are written and send by the only one student.

Figure 4.7: Visualisations using D3 (A JavaScript library). The Tree map showing summary of student posts/essays.

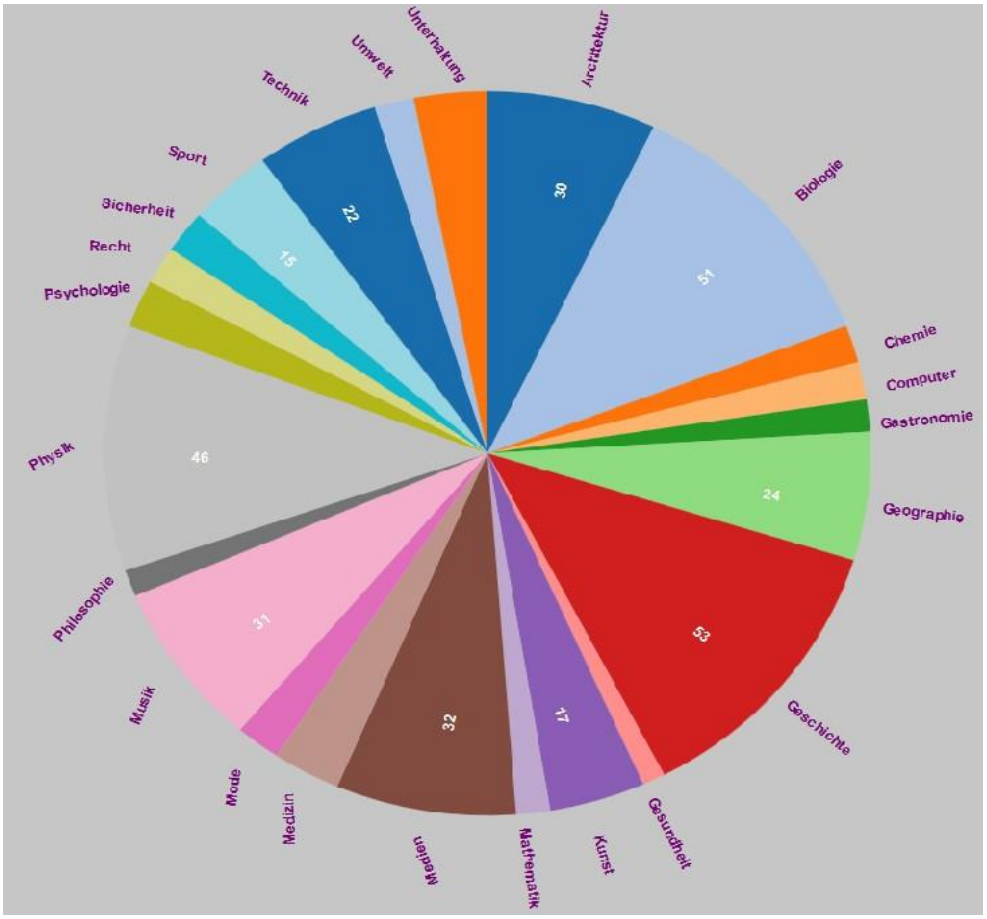


Figure 4.8: Pie Chart using D3 (A JavaScript library) showing summary of student posts/essays. Essay Topics are arranged along sectors.

References

[1] "Information Visualisation." [Online]. Available: http://courses.iicm.tugraz.at/ivis/ivis.pdf. [Accessed: 21-Oct-2015].

[2] H. Maurer, R. Mehmood, and P. Korica-Pehserl, "How Dangerous is the Web for Creative Work?," J. Comput. Technol., vol. 21, no. 2, pp. 59–69, 2013.

[3] "Sparkling Science" [Online]. Available: http://austria-forum.org/af/Sparkling_Science. [Accessed: 06-Dec-2015].

[4] "Alpen" [Online]. Available: http://austria-forum.org/af/Sparkling_Science/Aufsatzsammlung/Alpen. [Accessed: 06-Dec-2015].

[5] "Ted Nelson: Literary Machines" [Online]. Available: http://u-tx.net/ccritics/literary-machines.html. [Accessed: 06-Dec-2015].

[6] T. H. Nelson, Computer Lib: You Can and Must Understand Computers Now:, 2nd ed. Microsoft Press, 1974.

[7] "Wikipedia:Transclusion - Wikipedia, the free encyclopedia." [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Transclusion. [Accessed: 13-Dec-2015].

[8] "Let's Make a Bar Chart." [Online]. Available: http://bost.ocks.org/mike/bar/. [Accessed: 27-Oct-2015].

Chapter 5 : Some Aspects of Reliability of Information on the Web

The main aim of this chapter is to consider how true (or reliable) information is that we find using some search engines, some special services, Wikipedia, or other databases.

When we look up information in the WWW we hope to find information that is correct, fitting in quantity for our purposes and written at a level that we can understand. Unfortunately, very often one of the above criteria will not be met. A young person looking for information on some aspect of physics may well be frustrated when finding a complex formula whose understanding requires higher mathematics. In other cases, information may be much too voluminous or too short. This seems to indicate that what we need is presentation of material at various levels of detail and complexity. But most important of all, and this is what we are going to discuss in this chapter, is: how do we know that what we read is actually true? We will analyse this problem in the introductory section. We will show that it is impossible to expect too much. We will argue that some improvements can be made, particularly if the domain is restricted. We will then examine certain types of geographical information. Detailed research shows that some quantitative measurements like the area of a country or the highest mountains of a country, even if different sources disagree, can be verified by explaining why the discrepancies occur and by trusting numbers if they are identical in very different databases.

Following are the research questions that will be addressed in this chapter.

RQ1: Is Web data reliable? A case study is presented that examines reliability of web data.

RQ2: How to verify the area of countries. We propose a novel verification mechanism.

RQ3: What are the explanations in case of significant difference in area figures of countries?

RQ4: What is the situation with the reliability of facts related to mountains and cities rankings? (A few cases are examined.)

This is essentially a paper co-authored with Professor Hermann Maurer [1].

5.1 Introduction

First of all it is necessary to understand that in many cases an objective truth does not exist. After all, many things like historical events or actions of persons can be interpreted in one way or another. Has Lenin's work on communism helped mankind or caused catastrophic events? Has the discovery of nuclear energy made the world a better or a more dangerous place? Is the internet turning us into dummies [2] or is it helping us to achieve new levels of knowledge?

Clearly this list can be expanded arbitrarily: in many cases there is no objective truth but there are just different views of seeing the same event, person, phenomenon etc. The second author in [1] has already tried to explain many years ago [3] that if we want to understand something well we have to look at it from different angles. While this fact is well accepted for physical objects (how can we know what a coin looks like when we don't look at both sides?) it is unfortunately less clear to most persons that it is also necessary to look at abstract things like ideas from different points of view to fully comprehend them. Thus, in many cases whenever we find some piece of information it will only represent a single point of view or an opinion. The situation is worse, since the opinion may be from someone not knowledgeable in the area at issue, or presents only a partial view either intentionally or not, or even a lie or a distorted view of whatever is being described.

Hence we believe that it is fair to say: If we want to understand any moderately complex issue we have to examine more than one source.

In the past, a number of top universities like Cornell and U.Texas [4] have started to list some criteria to allow to judge the reliability of a Web page. It is interesting to note that in all cases the first of the criteria is that the source should be known, telling us much about the expected quality or potential tainting of facts. Better still, we should be able to contact whoever is behind a statement and to engage the author(s) in a discussion. After all the information is on the Web that provides easy ways to discuss and communicate. Hence, the communicational facilities should be used; maybe we should even discourage anonymous postings completely? Well, the situation may not be that easy. After all, there is a reason why some organisations still allow the use of a box where one can drop suggestions or complaints anonymously. Thus we feel that anonymous postings are ok in order to protect the reputation or safety of persons but should be banned completely from serious information sites. This argument has been presented to major information providers such as Wikipedia by some of us repeatedly. Note in passing that anonymity makes it easier to be sloppy, since no one has to take responsibility for whatever is being spread.

Thus, complex issues need to be presented by knowledgeable persons from different angles. However, are there not many issues that can be settled with a definite no or yes answer or a figure, hence should we not be able to get correct and trustworthy information in such cases? Unfortunately, even for answers to questions whose answer is a yes, a no, or a quantitative measurement we found that WWW servers turn out to be unreliable in the sense of giving differing answers. Again, there can be a variety of reasons for why the answer may be either deliberately (in particular to gain an advantage for a product, just to mention one “application”) or by some coincidence to be wrong. It may be eye-opening to mention some real cases: On checking with the most widely used search engine for the “boiling point of Radium” at the time of writing we found a range of answers (in degrees Kelvin): 1413, 1809, 1973, and 2010. Of course one may argue that the “boiling point of Radium” is not such an essential quantity for most of us. However, when checking the edibility of a special wild mushroom we found three entries describing it as “delicate edible fungus” but also two entries stating it as “deadly poisonous”. Since that mushroom has a long tradition as a delicacy in Europe we really wondered about the truth in such remarks and did some serious research. It turns out that in all sources that were written before 2005 this mushroom was considered good for eating. After this, statements became more negative. The reason is this: after a meal involving the mushrooms two persons died. It has been suspected that this was due to the mushrooms. Hence they are now labelled as “deadly poisonous”. However, when thousands of other people had eaten that mushroom without any ill effects before, should one not at most state: “It seems that in rare cases some potentially life threatening allergic reaction is possible” (as is done for peanuts, milk-products, etc.)

To put it more mildly: truth keeps changing. At some stage the world was considered flat and the sun circling the world; someone was considered a witch for curious reason and had to be burned at a stake; swans were considered as prime example for “white” for ages [5]; Pluto was once considered a planet, atoms indivisible, etc., etc. In considering the notion of truth which can change dynamically with times and circumstances, there is a need to watch with caution the manipulative projection of truth as a means to gain symbolic power. Algorithmic approaches are now capable of emphasizing subjective relationships that can automatically be determined based on usage patterns. Baker & Potts [6] have described situations where degrading auto-complete suggestions were listed whenever the name of particular user names were typed in a search engine. These situations are common to all search tools that offer the auto-complete suggestion service. Linking a user name to derogatory terms such as *conman and fraud* [6] and other terms that result from mere rumors that are easily spread on

the web can thus become a serious concern. There have even been cases where companies such as Google have been sued for defamatory auto-suggestions.

Such a service can be further exploited for performing a character assassination of individuals. Google has been known to remove auto-suggestions, but it does not vigilantly watch out for emerging associations. The web has thus produced a new form of meaning creation [6]. The traces of character-strings left by users in performing searches, are in themselves meaning-creating; sequences of letters in a trace can become leads to the selected search directions for other users. By trying to provide a useful service, this approach enables search engines to influence directly the search process of users.

Baker & Potts [6] also point out that negative stereotyping of vulnerable groups is an unavoidable consequence of such common actions that arises from collective consciousness of a large enough number of users. In efforts to predict things that a user may find interesting or is likely to search for, search engines are presenting suggestions that may include product-oriented links (see Figure. 5.1), cultural biases or even unexpected connections. Curious users may select unexpected links, mainly as an exploration of its validity or its source and therefore inadvertently reinforce their importance. At the same time it is also highly likely that users become distracted and drift away from the original search intent. Search engines are thus indirectly re-shaping the reality for many users. The original small group of users who establish a correlation between terms are thus able to shape the search experience of millions of users.

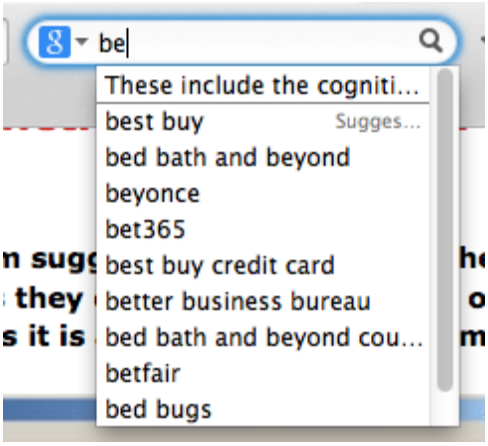


Figure 5.1

Baker & Potts also demonstrate the stereotypical framing of questions asked by users which reflect biased opinions on social groups relating to factors such as race, gender, religion. Even if only a small number of users asked an initial question on some social groups, interesting connections are attractive enough to catch on and become a buzz on the Web.

Information providers play an important role in shaping the reality for millions of users. The billions of daily searches tend to have a severe implication on the lives of people. Many new users start to fully trust he rapid answers that the Web is able to provide. Auto-suggestions as described above becomes a powerful medium that intervenes strategically in the intention-specification stage of searching. In

this process this service guides the articulation of intentions leading to fast and easy predicted (and orchestrated) 'relevant answers'.

By providing a standard shallow answer that could even be directed by business motives this can kill the spirit of inquiry and leave users in a worse situation than before the search was performed. Distorting reality by restricting and manipulating user perception [7] and in a way indirectly altering the recording of history [8] is irresponsible and can be extremely dangerous [9]. The internet revolution has thus failed miserably in its promise of "*bringing more truth to more people, more depth of information, more global perspective and more unbiased opinion from dispassionate observer"* as described in [10].

The question of whose responsibility it is to protect the interest of vulnerable groups or to protect users from the spread of dangerous sentiments or beliefs is yet to be resolved. Issues of reliability of information and notions and truth, the accountability for providing them and the resulting consequences have to be studied and carefully addressed.

Though no solution exists at this moment, why can't the power of massive collaboration systems be used to address this in a meaningful way? We believe that indeed collaboration and involving "the crowd" should be able to play an important role here. A discussion on this idea is presented in [11].

Acquiring reliable knowledge and inquisitive searching for truth are pre-requisite core research skills for learning to understand the world or specific issues. Information providing sources should explore ways of engaging communities in scholarly pursuits that helps to nurture these traits. Responsible information sources should be distinguished and recognized for providing alternative measures of reliability or by providing exploratory directions. In any case the source of information has to be duly considered, scrutinized and made known to users. This has to be done for all forms of information, including facts considered as 'taken for granted'.

Being particularly motivated about the perception of truth is a characteristic that cannot be compromised, particularly when it comes to scholarly activities. Such motivation which closely relates to inquisitiveness and passion for knowledge, is now being replaced by an acceptance of stereotypical ideas and the focusing on the trivialities as propagated by social media through repeated posts and reposts. As users are being left in a distracted state of not even able to able acquire the intended information directly, we do have a daunting task ahead.

Another aspect related to the notion of truth is seen in rating systems. In these systems collective human judgment is used as a basis for ranking of sites, services and information clips or media units. These rating services that are capable of providing truth ratings are being subject to malicious activities, multiple identity attacks and the orchestrated ratings by buying in users, see [12]. As progress into these research areas continue, we have to take steps in engaging the users and motivating them to consider more seriously the accuracy of perceived truth notions.

Ways to prevent users from consuming information that constitute partial truths of non-diligently controlled sources need to be explored. The notion of truth needs to be re-emphasized and 'discovery of truth' needs to once again become a core activity.

Information providing sites should take into consideration different needs of users, rather to propagate popular directions only. As an illustration, we note that even for those searching for information on tallest mountains the desired outcome may be either:

- An approximate answer is all that this needed
- A single reliable answer is sufficient but it has to be within a context

- A comparative analysis of all possible facets of answers with an indication of sources is required for research purpose
- An overview of discourses within a trusted community to help validate and verify the best answer, as best as possible.

On the search for more clearly answerable questions we considered the following: What is the largest cave? What is the biggest river? How high is the highest mountain on moon? Should there not be clear answers to such "quantitative" questions? No, in each case!

After all, what does "largest" in connection with "cave" mean: the longest from entrance to where it ends (even if this is accepted: is a single straight cave 10 km in length considered longer than one that branches into 6 caves with each being over 5 km long?), the highest (in what sense? How far a stone could drop vertically or rather the difference of the elevation of one point and another point?) Or do we mean the cave with the largest volume, etc. The "largest river" … do we mean length (and what if parts of the river have different names), or do we mean the water flow (Maximum flow? Average flow?) The question about the highest mountain on the moon is particularly funny: Since we measure height usually from the level of oceans we are in trouble on the moon where there are no water oceans. So do we measure the height relative to the lowest point "nearby" (is "nearby" 2, 5 or 150 km?), or do we convert the moon into an abstract sphere (by "filling lower parts with material from higher parts") and measure from the surface of that sphere?

The important point to note is that we are often asking questions that are ill-defined and are based on measurements as if it is possible to describe important facts by just one figure: how can we quantify the most beautiful woman, or the most intelligent human, or the best athlete (are you as lost as we are when we have to compare a top golfer with a top mountaineer?).

As a consequence, if we want to talk about reliability of a WWW page or WWW sites we have to be very modest. As a key step in this direction, we check the reliability of geographic information of basic quantitative information. Our findings reveal that these quantitative measures do not exist independently of historical developments, evolving geographical boundaries and current state of affairs and the dominant forces of change. These associated developments tend to become ignored or overlooked, in a number of cases either inadvertently or deliberately. We will present in what follows a project whose aim is to set up a server for geographic information that is reliable to some extent. More specifically, we collect information (or links to information) from a number of reliable sources. For detailed description of data sources see Chapter 7.

We can only do this for selected types of information but even so we ran into unexpected difficulties. If we find an agreement between all or almost all sources we assume that the information is correct and we display it. Otherwise we list the different results while, trying to guess why the discrepancies occur, but are dependent on the community to complete our job in many cases.

In Section 5.2 we explain our main approach a bit more carefully, then present in Sections 5.3-5.5 three different topics that we checked. Section 5.6 is a short conclusion.

5.2 Testing Reliability of Information in a Special Case

We started a project to set up a server "global-geography.org" [13]. We initially imported the information on all countries listed in [14] covered by the topics "Introduction", "Geography", "People and Society", "Government", "Economy" , "Energy", "Communications" and "Transportation", but omitted more contentious and rather time dependent parts like "Military" and "Transnational issues". However, we introduced additional categories "Culture", "Pictures", "Special items" and are

encouraging the community with a category "Community Contributions" Under "Culture" we imported for each country data (or links) to all UNESCO heritage sites from the server [15], and similarly information on Nobel Prize laureates from [16]. The details of data extraction from different data sources are provided in Chapter 7. Awardees of the Wolf prize from the [17] and Field medallists are also included. We hope to also add further information to other areas in the future, partially by appealing to the community.

The entry "Pictures" is supposed to give a pictorial overview of each country by providing a selection of an average of 100 pictures per country: some of them are taken from [14] when available, with many others coming from elsewhere. The main point is not to just offer pictures (zillions of them can be found on the web already) but to offer pictures with substantial descriptions, and even pictures of the past, of buildings, monuments or nature, as destroyed by time, industrial developments or wars. The pictures of Syria can be seen at [18].

The entry "Special items" is reserved to report on potentially interesting and unusual facts about each country, to present maps, important links, etc.

The entry "Community Contributions" is a plea to the community to improve and add information.

However, most emphasis (in this chapter) is placed by us on checking the data under "Geography" for all countries or expanding the data to some extent. The first major three steps are described in the next three sections: We decided to check the information on the area for each country (Section 5.3), and we were in for some surprises. We added major cities to each country (Section 5.4) and major mountains (Section 5.5), provided that mountains above 2000 meters do occur in that country. In each case more than one source was used to extract the information, and other sources were used to verify them as has been alluded to in Section 5.1.

If all this sounds simple it is not. Remember, we are concentrating on information concerning countries, so we need a list all countries for consideration. Factbook [14] lists 263 items under "countries". Yet a first look shows that many of them are not countries: "Antarctica" does not qualify, nor do the "Ashmore and Cartier islands" (a group of uninhabited small islands and reefs belonging to Australia, located North of the continent and South of Timor), or "Jersey", one of the British channel islands, etc. etc. Another problem is in the naming of countries. While Italia (official name) might be easily distinguished from its German version Italian or its English version Italy, some cases become very complicated due to the different transcription from other languages and alphabets: Azerbaijan and Aserbaidschan do at least sound similar, yet that this country was at some stage called Albania (!) can be quite confusing (since Albania now, for many years, denotes a small country on the SW Balkan). Some transcriptions of Russian or Arabic names are hardly recognizable, and countries may have changed their name, like Burma to Myanmar (a problem still more severe when it comes to cities). We find in [Factbook 2014] North Korea and South Korea, although their official names are "Democratic Republic of Korea" and "Republic of Korea", respectively, (the attribute "democratically" is misused for "communistic" as it is in connection with a number of names of countries). Thus, we need a more solid list of country names. So why not use the list of 193 of UN countries [19]? Unfortunately, this is not satisfactory either. Although Taiwan (officially the "Republic of China on Taiwan") has all the trimmings of a country like passports, visas, government, flag, national anthem etc., it is not a UN country due to the opposition of mainland China (the "one China policy"). On the other hand, Sudan is a UN country but has not existed as a single country for some time, since it has been divided into North and South Sudan with continuing border disputes. Many countries are recognized as such at some stage, but not uniformly enough to be accepted by the UN. The republic of Cyprus, a UN country, has de jure control of the whole island, yet 41% of it is claimed and occupied by Turkey as Turkish Republic of Cyprus, recognized as state only by Turkey. There are many similar situations.

Despite the problems mentioned with the UN classification we have decided to use this one as the best, yet not an ideal alternative. In our project "country" therefore means country by UN definition; we call all other 71 entries in Factbook [14] "territories". Note that the entry "Special items" allows us to include also special cases as just mentioned.

Let us now turn to consider the reliability of area specifications (in square kilometres) of all countries.

5.3 Area Verification

Before looking at details let us point out that even the definition of area of a country is done in a rather arbitrary way (even definition of country is not clear): what is not counted is not the actual area that one obtains by counting every square meter that can be viewed, but what is counted is the area of the projection of the surface of the country on a plane. Putting it differently, because of slopes or mountains, the viewable (and useable) area may be considerably larger than the area of projection. To be concrete, consider an area that shows on a map as 40 times 40 square meters (i.e. 1.600 square meters in projection) that is on a slope rising 30 meters high, then the actual area for e.g. a meadow on this part of land is (by Pythagoras theorem) 2.000 square meters! Clearly, the steeper the incline, the more the projection will differ from the viewable area of the object: With high and almost vertical cliffs the difference can clearly be dramatic. Of course it also critical whether the area of inland lakes or even parts of the ocean inlets are included. If the area of rivers is included, if the river flows into the sea in a large delta, at what degree of salinity do we still count the water as part of the river rather than part of the sea? What if the tide makes a big difference in shorelines or where a river "ends?

And different servers use different criteria! Let us add another curiosity: Countries may have very large cave systems whose area is never taken into account when talking about the area of a country. Extreme cases may be the Carlsbad Caverns in new Mexico which are not only comparable in size to the gigantic Mulu caves system of Borneo, but some of its area, hundreds of meters below the ground is used commercially, e.g. for a veritable super market… with an elevator directly back to the surface!

We have discussed this issue at some length to reiterate that we are all the time using terminology without giving much thought to what terms really mean and that statements of areas that are exact to the last digit do not make sense. It may be surprising: but there is (as has been explained) no internationally agreed way to measure the size of a country in square kilometres. Therefore we consider the figures for the area of a country correct, if they differ by at most 1/10 of one percent.

Above sounds reasonable (does it?). However, we have to be more careful with our definition. We use sources, call them s(1), s(2),…, s(6). We choose in the tables to follow the "average approach": we find the average of the 6 numbers, call it av, i.e. av= (s(1)+s(2)+…+s(6))/6 ; we then take 1/10 of one percent of av, let us call it y, i.e y = av/1000 and calculate for each source s(i) (i=1, 2, …, 6) the value z(i) = abs (s(i) –av)/y. We define the Difference as the maximum of the 6 values s(i), rounded to an integer. When the Difference is not more than 1 we consider that the six measurements agree.

It might be useful to given an example. Suppose we find for a country the following square kilometre measurements: 100.000, 100.300, 100.300, 99.700, 100.000, and 99.700. Then av= 100.000, hence y= 100 and hence Difference= max (0, 3, 3, 3, 0, 3) = 3. Putting it differently, if the Difference is 10 or less i.e. discrepancies do not exceed 1% we do not need to worry too much. Still, what is the reason that so many UN country measurements differ at all?

Using Difference=1 as upper threshold we find that all sources 1-6 agree for all countries except 94 of the 193 UN countries. It might be quite a surprise that there is no agreement concerning the area of a

country in over 45% of the cases! If we consider Difference=10 as threshold we still find 40 countries whose area in square kilometres differs according to the sources by more than 1%. For a list of all UN countries see UN [19], we list in Table 5.1 for brevity only those in Europe ones with a Difference>1. For full results see Appendix B.

Table 5.1: Area (sq km) of countries

Country	Factbook	Dbpedia	Geoname	Infoplease	Britannica	Wolfram	Difference
Finland	338145	338242	337030	338145	390903	338145	127
France	643801	674843	547030	547030	543965	551500	154
Ireland	70273	84421	70280	70280	70273	70273	162
Macedonia	25713	-	25333	25333	25713	25713	6
Malta	316	316	316	321	315	316	14
Montenegro	13812	12999	14026	14026	13812	13812	54
Netherlands	41543	41541	41526	41526	41850	41543	6
Norway	323802	385183	324220	324220	385186	323802	118
Serbia	77474	88360	88361	77474	77498	77474	89
United Kingdom	243610	243610	244820	244820	243073	243610	4

Now let us consider some of the above cases with a Difference of 10 or more. We have managed to find out why the differences occur in some cases, but do hope for further results from experts we are consulting with and from the community. The numbers in brackets show the difference.

In Europe we find Finland (127), France (154), Ireland (162), Malta (14), Montenegro (54), Norway (118) and Serbia (89). We explain why we get discrepancies in square kilometers for a few of above cases. On our server we will clarify of course many more cases than the few samples we present in this section.

Finland	338145	338242	337030	338145	390903	338145	127

Finland is listed with about 338.150 km² in Sources 1, 2, 4 and 6 and is close to the 338.420 km² as the most reliable reference source in German Brockhaus [20] and Countrycode [21] are quoting. The figures include the area of lakes (roughly 35.500 km²) which seems reasonable. Worldbank [22] gives only around 305.000, i.e. is not counting the lakes. The scattering of islands (like Aland-island 40 km off the Swedish coast that belong to Finland) and lakes is likely to explain the difference of 1.200 km² with source 3. However, the large figure in entry 5 is obtained by adding in some 52.000 km² of ocean also claimed by Finland! Note that this poses a new problem: should the parts of the ocean claimed by a country added to its area? It seems that most sources do not, but Britannica seems to do it in some cases!

France	643801	674843	547030	547030	543965	551500	154

For France, source 5 (Britannica) gives the smallest area. This agrees exactly with the area in Brockhaus [20]: It follows the French Land register data that excludes lakes, ponds and glaciers larger than 1 km² and the estuaries of rivers. This is very much in contrast to how the figures are arrived at for Finland. Including bodies of water the French National Geographic Institute arrives at the figure in source 6 above. Those figures do include the area of the island of Corsica but do not include overseas departments and oversea territories. If one counts them in, a figure higher than 640.000 is obtained. None of the figures include the 320.000 km² of Antarctica where sovereignty has been suspended since the signing of the Antarctic Treaty in 1959. The main overseas departments (by all laws equivalent to parts of continental France listed by their rough size in brackets) are French Guiana (83.0000), Réunion (2.500), Guadeloupe (1.600), Martinique (1.100), Mayotte (370), Saint Pierre and Miquelon (240). The overseas territories (with a different legal status) are New Caledonia (19.000), French Southern and Antarctic Lands (7.600, the largest Kerguelen Island 7.200), French Polynesia (over 100 island with a total of 3.500), Wallis and Futuna (140). With a few very small islands (like St. Martin and St. Barth) still missing and various places having different political status it becomes clear why there are discrepancies in the figures. For completeness let us mention that the figure in Countrycode [21] agrees with Source 1, in Worldbank [22] with Source 3.

The situations concerning other European countries with large Difference are easy to explain. In case of Ireland the two essentially different figures come from whether the whole island or only the Republic of Ireland (without the British North) is counted. The disagreement of over 60.000 km² in some figures for Norway just comes from whether Spitzbergen and nearby islands are counted as part of Norway or not. The figures in Serbia differ by 10.800 km² depending on whether one counts Kosovo as part of Serbia or as separate country. A particular curious case (that is not even listed above) is Denmark. All 6 Sources see UN [19] assign some 43.000 km² to it, but ignore the over 2 million square kilometers of Greenland that is an "autonomous country of the Kingdom of Denmark": It does belong to Denmark in a strong sense yet is not mentioned as part of Denmark nor accepted as a UN country! An interesting side-remark: in Factbook [14] Greenland is listed as a territory of North America, something most Europeans would certainly not agree on!

Having looked at European countries in some detail it has become clear that differences in areas of countries are due to three reasons: political aspects (like is Kosovo part of Serbia or not), are remote regions (often with slightly different status considered or not: Spitzbergen for Norway, Falkland for UK, oversea territories for France, Greenland for Denmark, etc.) and finally, are inland waters and glaciers or even parts of the Ocean (like in Finland or the waterways at the tip of South America or the deep bays in Vavau, the northernmost island group of Tonga) counted or not. Consequently, when looking up even something as simple as the area of a country or territory the figures you should not use them without further investigation.

We now look briefly at some other parts of the world starting with Africa, and again only looking at some cases with big discrepancies.

Comoros	2235	2235	2170	2170	1862	2235	134

Comoros is geographical an archipelago and set of reefs that originally was a colony of France. The largest part (shown by the figure of source 5) became independent in 1975 but with much political unrest afterwards. One island, Mayotte, remained with France and actually became an overseas department on 31 March 2011 and an "Outermost region of the European Union" on 1 January 2015. Add its 373 km² to the figure of source 5 and you get exactly the figures of Sources 1, 2 and 6 which do

not reflect the political reality. The figures of Sources 3 and 4 probably come from adding to 1862 km² the area of small islets like the Banc du Geyser, a reef claimed by Comoros, France and Madagascar, of Glorioso Islands and others with political unclear status. Countrycode [21] agrees with Source 1, Worldbank [22] with 5.

Gambia, The	11295	10689	11300	11300	11632	11295	50

Except for Source 2 there is agreement that Gambia has about 11.300 km², a figure also supported by Brockhaus [20], Countrycode [21]. That Source 2 lists 10689 (like the English Wikipedia) despite the fact that the German Wikipedia and the French Wikipedia also list what the others say again shows that not only general reports but also quantitative facts vary between versions of Wikipedia! That the usually reliable Worldbank [22] also reports a figure significantly below 11.000 explains the lower figures: They ignore the area of inland water (Worldbank [22] always does), but is a particularly tricky affair in case of Gambia, since most of the inland water is the long and wide mouth of the river. Hence there is really no clear distinction between the fresh water river and the deep ocean bay!

Let us finish this section by discussion three countries of Asia with very large differences:

Georgia	69700	0	69700	69700	69700	153900	778
Bhutan	38394	38394	47000	47000	38394	38394	139
Oman	309500	309498	212460	212460	309500	309500	233
Pakistan	796095	796095	803940	803940	881889	796095	85

The case Georgia is easy to explain: The size is definitely 69.700. The figure 153.900 was given by Wolfram as the size of the USA state Georgia! By correctly specifying not just Georgia but Country Georgia Wolfram also yields 69.700.

Bhutan has been found in 8 additional sources also with 38.394 km². Using Natural Earth[4], Daftlogic[5] 2014 and Freemaptools[6] (one of which will be incorporated into our project) we also found roughly 39.000 km². Only in the French Wikipedia the mysterious figure 47.000 appears. The explanation may be that there are some 6.500 km² that are contentious between China and Bhutan: counting those to Bhutan would make the difference.

Oman is listed by 4 of the 6 sources with an area of 309.500. This agrees with other geography books we have checked. We have used area measurement tools mentioned above and have obtained also around 310.000 km². Hence the significantly lower figures in Source 3 and 4 can just be considered as wrong: there is a bit of trouble with Yemen in the South but certainly no area has been seized by Yemen at the time of writing.

For Pakistan the dominating figure (Sources 1, 2 and 6) is 796.095 km², yet two sources are about 7.000 km² higher, and Source 5 even an astounding 85.000 km². The reason for this is how much of "Jamnu

[4] http://www.naturalearthdata.com/
[5] https://www.daftlogic.com/
[6] https://www.freemaptools.com/

and Kashmir" belongs to Pakistan, how much to India, see Figure 5.2. Additionally, the boundary in the North to China is not clear at all, but there is almost continuous fighting at altitudes over 5.000 m, actually more between India and China than Pakistan and China!

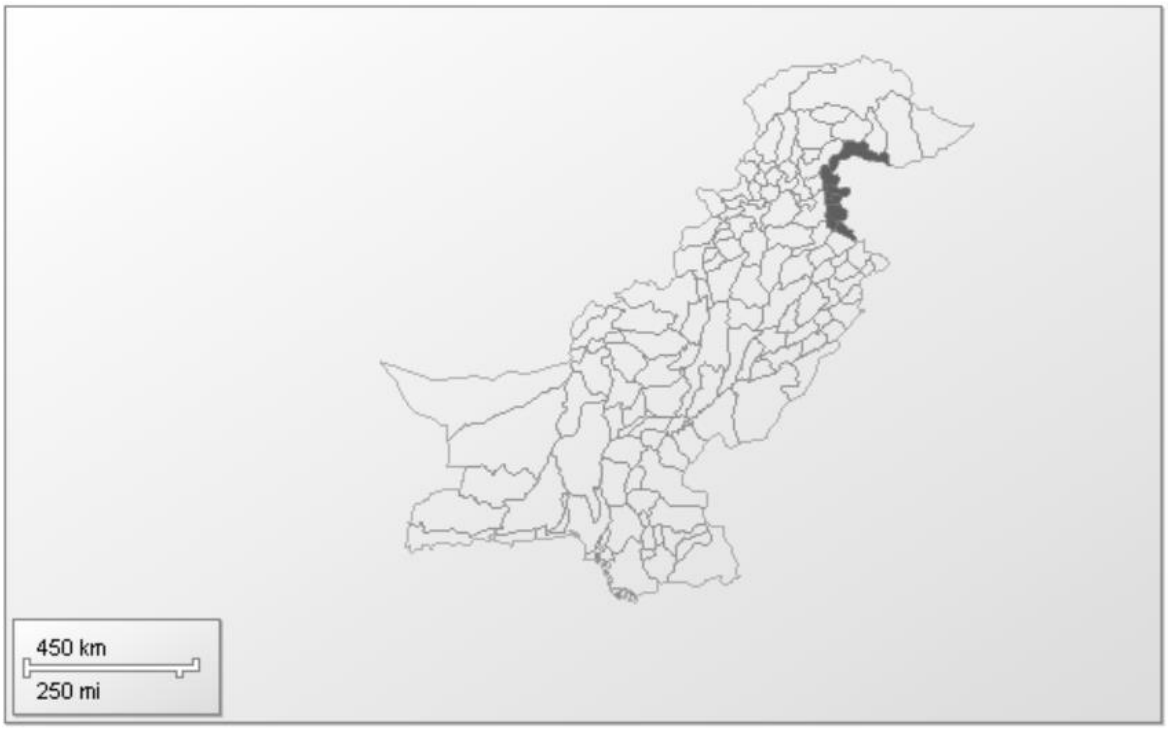


Figure 5.2: Pakistan (Map taken from Natural Earth, the dark part is the contentious area of Kashmir)

Let us just briefly mention some points concerning the territories which are not in UN [19]:

Jan Mayen	377	373	-	-	-	380	10
Svalbard	62045	61022	-	-	-	62045	11

Jan Mayen is a small (about 375 km²) island that belongs to Norway. It is situated North of Iceland, East of Greenland and Southwest of Spitzbergen. For the latter reason it is sometimes mentioned together with Spitzbergen (whose Norwegian name is Svalbard), despite the fact that Spitzbergen is more loosely connected to Norway than Jan Mayen is. See the discussion of Spitzbergen and size of Norway under European countries.

Greenland	2166086	2165512	2166086	341701	2166086	2166000	816

The area of Greenland is indeed what all Sources above except one indicate. Greenland is an autonomous region of Denmark as explained earlier. The much larger figure of over 3 million km² can only be obtained if part of Northern Canada (Ellesmeere Island) is included, an area indeed not far West of Greenland. Note that Europa considers Greenland definitely part of Europe, but (according to Factbook [14]) it is part of North America. It is also worth noting that Worldbank [22] gives only an area of 410.450 km², i.e. discounts all areas covered by deep ice.

We will not discuss further the area of countries in this section, but our server will of course have many more comments, explanations and pleas for help!

The Continent-wise Verified and Not Verified cases of 193 UN member states are shown in the Table 5.2 and 5.3 using upper threshold 1 and 10 respectively.

Table 5.2: Continent-wise Verified and Not Verified countries (When % average difference is greater than 1 we call it Not Verified)

Continents	Verified	Not Verified
Europe	26	17
Africa	30	24
Australia	8	5
America	15	20
Asia	20	28
Total	99	94

Table 5.3: Continent-wise Verified and Not Verified countries (When % average difference is greater than 10 we call it Not Verified)

Continents	Verified	Not Verified
Europe	36	7
Africa	42	12
Australia	11	2
America	30	5
Asia	34	14
Total	153	40

5.4 Some cities of countries ranked by population size

In this section we use three as primary sources for information and checking: Wolfram, Geonames and Infoplease, but will also again involve manual checks against e.g. Brockhaus [20] or other language Wikipedias.

In ranking cities in a country we run into two major problems:

First, the names of cities may be quite different. “Wroclaw” and “Breslau” are the same city in Poland and both names are still in international use. Both “Wien” and “Vienna” as names for the capital of Austria are acceptable. We have (new) Mumbai and (former) Bombay, the old name having almost disappeared, Louangpraban und Lunang Prabang as second largest city of the state officially called “Sathalanalat Paxathipatai Paxaxon Lao of the Democratic People’s Republic of Laos”, etc.

Second, population counts often reflect the number of people at different moments in time which can possibly lead to a change of ranking.

Third, and often most serious, sometimes a name stands for a part of the community or the whole community. A typical case is Auckland in New Zealand. Up to 2010, the “City of Auckland” with some 400.000 inhabitants was the biggest city of New Zealand, yet the Metropolitan area Auckland (to which people usually referred to) was well over one million at that time. A decision in 2010 combined a number of areas including the “City of Auckland” into “Auckland Council” with now over 1,4 million people. And nowadays this is what is usually meant when talking about Auckland.

The following Tables 5.4, 5.5, 5.6 show a few samples of countries of cities arranged by size and problems encountered.

Table 5.4:

Ranking	Wolfram	Geonames	InfoPlease	Wikipedia
1	Karachi	Karachi	Karachi	Karachi
2	Lahore	Lahore	Lahore	Lahore
3	Faisalabad	Faisalābād	Faisalabad	Faisalabad
4	Rawalpindi	Rawalpindi	Rawalpindi	Rawalpindi
5	Multan	Multān	-	Multan
6	Hyderabad	Hyderabad	-	Gujranwala
7	Gujranwala	Gujranwala	-	Hyderabad
8	Peshawar	Peshawar	-	Peshawar

As shown in Table 5.4, in Pakistan the rankings agree except for rank 6 and 7, sometimes 6 given to Hyderabad, sometimes to Gujaranwala. According to most recent records from Brockhaus [20], Gujaranwala wins with a small margin, so both cities can be seen as more or less the same size (1.4 million).

Table 5.5: Top Cities of Saudi Arabia

Ranking	Wolfram	Geonames	InfoPlease	Wikipedia
1	Riyadh	Riyadh	Riyadh	Riyadh
2	Jiddah	Jeddah	Makkah	Jeddah
3	Makkah	Mecca	Jeddah	Mecca
4	al-Madinah	Medina	-	Medina
5	ad-Dammam	Sultanah	-	Al-Ahsa
6	Taif	Dammam	-	Ta'if
7	Tabuk	Taif	-	Dammam
8	Buraydah	Tabuk	-	Khamis Mushait

In Table 5.5, Jiddah (Jeddah) is considerably larger than Makkah, so we consider the ranking in Infoplease as plain wrong. The mentioning of a comparatively small village of Sultanah in Geoname in rank 5 is very surprising. Dammam is a city of some 900.000, Taif a bit more than that. Al-Ahsa is an old city now quite small (but was at rank 10 in the world 1000 years ago!), but listed here since the region around it is close to one million and has an international airport with its name. It is clear to the authors that only specialist familiar with the region can do a proper ranking.

Table 5.6: Top Cities of Canada

Ranking	Wolfram	Geonames	InfoPlease	Wikipedia
1	Toronto	Toronto	Toronto	Toronto
2	Montreal	Montréal	Montreal	Montreal
3	Calgary	Vancouver	Vancouver	Calgary
4	Ottawa	Calgary	Calgary	Ottawa
5	Edmonton	Ottawa	Edmonton	Edmonton
6	Mississauga	Edmonton	Quebec	Mississauga
7	Winnipeg	Mississauga	Hamilton	Winnipeg
8	Vancouver	NorthYork	Winnipeg	Vancouver

In Table 5.6, Ranks 1 and 2 are undisputed: the core of both cities is close to (Montreal) or above (Toronto) 2 million, the metropolitan area in both cases more than twice as much. The rest becomes murky, since core cities and their metropolitan areas are very different. Ranking by core cities we have from rank 3 onward: Calgary, Otttawa, Edmonton, Winnipeg, Vancouver; by metropolitan population however: Vancouver, Ottawa, Calgary, Edmonton, Quebec (a real curiosity, since the city is only ¼ of the metropolitan population) and Hamilton. That Mississauga shows up twice, a more or less artificial union of suburbs of Toronto, is only due to the summed up population of a large area but should really not appear: Even its metropolitan area does not allow it to rank it under the first 8.

Table 5.7: Top cities of Malaysia

Ranking	Wolfram	Geonames	InfoPlease	Wikipedia
1	Kuala Lumpur	Kota Bharu	Kuala Lumpur	Kuala Lumpur
2	Klang	Kuala Lumpur	Kelang	Johor Bahru
3	Subang Jaya	Klang	Johor Bharu	Ipoh
4	Johor Bahru	Kampung Baru Subang		Shah Alam
5	Ipoh	Johor Bahru		Petaling Jaya
6	Ampang Jaya	Subang Jaya		Kuching
7	Kuching	Ipoh		Kota Kinabalu
8	Petaling Jaya	Kuching		Kuala Terengganu

Table 5.7 shows top cities of Malaysia. The correct list of towns and cities according to the population within the local government areas are specified in the document on statistics for local authority areas, by the department of statistics [23] and is reflected correctly in Wikipedia. Wolfram referred to another document by the statistics department that does not distinguish between cities and municipality areas: This explain the reasons for the discrepancy. The list by Geonames is a list compiled from users without considering the documents from the statistics department.

To see overall results of city verification of 193 UN countries, see Chapter 7.

5.5 Some mountains of countries ranked by height

In this section we are considering only countries with mountains higher than 2.000 meters. In trying to rank them the major difficulty is that borders often go on top mountains, so the country they "belong to" is not clear. Then there is also a petty difficulty: countries want to have a high mountain, so a mountain 3993 m will often turn into a just above 4000 m for PR reasons.

Like with all quantities if they are "too exact" they are misleading. After all, we measure the altitude of mountains as "above sea level", yet the sea level is not the same all over the world (and changes with the tide), so there is some curious definition of "mean sea level" that is usually used. Further, if we believe some climatologists, the level of oceans is going to rise. Does this mean we will have to adjust the height of all mountains accordingly?

The SQL query to extract top 5 mountains of Austria from GeoNames is shown below.

```
Select top 5 name, elevation from GeoNames
Where country_code=AT and  feature_class=T
And feature_code=MT and
elevation > 0 order by elevation desc.
```

Table 5.8: Top 5 mountains of Austria

Ranking	Wolfram	Elevation(m)	Geonames	Elevation(m)
1	Grossglockner	3798	Großglockner	3798
2	Wildspitze	3772	Wildspitze	3774
3	Weisskugel	3739	Palla Bianca	3738
4	Grossvenediger	3674	Großvenediger	3662
5	Similaun	3606	Ramolkogel	3550

As shown in Table 5.8, the mountains ranked 1 and 2 are undisputed. Both have secondary peaks (Kleinglockner, Southern Wildspitze) with both 3770 m but are usually no considered separate mountains. Hence Weisskugel (whose Italian name is Palla Bianca and can be counted to Austria or Italy, since the peak is at the border) and Großvenediger are rank 3 and 4. Rank 5 is wrong in both lists, since Hinterer Brochkogel (3628) and Hintere Schwärze (3624) are a bit higher than Similaun, the lowest Austrian peak above 3600m. Both Wiesbachhorn (3564) and Rainerhorn (3560) are however still higher than Ramolkogel, so his rank is "far off", even if we are talking only of a range of 50 meters.

Table 5.9: Top 5 mountains of Nepal, Pakistan, and India

Ranking	Wolfram	Elevation(m)	Geonames	Elevation(m)
Country : Nepal				
1	Mount Everest	8848	Mount Everest	8848
2	Kangchenjunga	8586	Kānchenjunga	8586
3	Kangchenjunga West	8505	Makālu	8463
4	Lhotse	8501	Dhaulāgiri	8167
5	Makalu	8462	Manāslu	8163
Country : Pakistan				
1	K2	8612	K2	8611
2	Nanga Parbat	8125	Nanga Parbat	8125
3	Gasherbrum	8068	Gasherbrum Shan	8080
4	Broad Peak	8047	Broad Feng	8051
5	Gasherbrum II	8035	Gasherbrum II Feng	8034
Country : India				
1	Kangchenjunga	8586	Nanda Devi	7816
2	Kangchenjunga West	8505	Kāmet	7756
3	Kangchenjunga South	8494	Saser Kangri	7672
4	Kangchenjunga Central	8482	Kabru	7412
5	Distaghil Sar	7885	Badrīnāth	7138

The problems with many of the mountains in Table 5.9 is because of borders on or near the peak, hence the mountains can be claimed by more than one country, and some listed as mountains can also be seen as secondary peaks just separated by a saddle or such from the higher cousin.

Concerning Nepal and the list according to Wolfram, Kangchenjunga West can be considered a side peak of Kangchenjunga, and even more so Lhotse (side peak of Mount Everest). Hence listing Makalu at rank 3 makes sense. Dhaulāgiri and Manāslu (located fully in Nepal) then come next, if one does not consider Cho Oyu (8.201) whose peak is at the border between Nepal and China. It may well be that

Geonames does not list it, since till 1984 its height was considered to be 8153, just below Dhaulāgiri and Manāslu. More recent measurements have yielded 8201.

Concerning Pakistan, the lists agree and further checks have confirmed their correctness.

India is complicated, since Kangchenjunga and its side peaks are shared with Nepal. Distaghill Sar is often considered the 7th highest mountain of Pakistan! Kangchenjunga (on the border of Pakistan) can certainly be also be counted to India, and then would be ranked 1, of course. Of the 5 listed by Geonames Kabru is contentious, since it is also claimed by Nepal. Overall, the boundaries in the Himalayas are not well defined and often are defined by peaks, so mountains are often claimed to belong to more than one country.

One major problem when talking of the highest mountains of some region is the question when a peak that is connected to another mountain is counted as separate mountain or not. One attempted (but not universally accepted definition) is that the "prominence" of the peak (the height above the highest saddle connecting it to a higher summit) is at least 300 m. Thus, although Lhotse is only some 2 km from Mt. Everest in distance, its prominence is 610 m, hence is usually seen as fourth highest mountain of the world (and one of the 14 mountains higher than 8.000 m). However, Lhotse is a borderline case in a sense: if we use 1500m prominence instead of 300m there are still 13 mountains higher than 8.000 m (but Lothse is then only a side peak of Mt. Everest). Thus, those 13 peaks can be truly called (separate) mountains. For the list of the 14 highest mountains (choosing prominence 300) see Table 5.10.

Table 5.10

Rank	Mountain	Height (m)	Prominence	Parent Mountain
1	Mount Everest	8848	8848	None
2	K2	8611	4017	Mount Everest
3	Kangchenjunga	8586	3922	Mount Everest
4	Lhotse	8516	610	Mount Everest
5	Makalu	8485	2386	Mount Everest
6	Cho Oyu	8188	2340	Mount Everest
7	Dhaulagiri 1	8167	3357	K2
8	Manaslu	8163	3092	Cho Oyo
9	Nanga Parbat	8126	4608	Dhaulagiri
10	Annapurna 1	8091	2984	Cho Oyu
11	Gasherbrum 1	8080	2155	K2
12	Broad Peak/K3	8051	1701	Gasherbrum 1
13	Gasherbrum 11	8035	1524	Gasherbrum 1
14	Shishapangma	8027	2897	Cho Oyu

5.6 Conclusion

In this chapter we have tried to show that even when using multiple Web sources and even simple quantitative questions they are often not easy to resolve without the help of specialists.

The main reason is that questions are very often too vague, hence allow different interpretations and hence different answers. In any reliable source the answer to a "vague" question has to be accompanied by the assumptions the answer is based on. It is our aim that global-geography.org will do so.

References

[1] N. Kulathuramaiyer,H. Maurer and R. Mehmood “Some Aspects of the Reliability of Information on the Web,” JUCS, vol. 20, no. 9, pp. 1284–1303, 2014.

[2] H. Maurer, R. Mehmood, and P. Korica-Pehserl, “How Dangerous is the Web for Creative Work?,” J. Comput. Technol., vol. 21, no. 2, pp. 59–69, 2013.

[3] H. Maurer, Xperten Bd.1: Der Telekinet., 1st ed. Freya, 2002.

[4] “How Can I tell if a Website is Reliable.” [Online]. Available: https://www.edb.utexas.edu/petrosino/Legacy_Cycle/mf_jm/Challenge 1/website reliable.pdf. [Accessed: 14-Dec-2015].

[5] T. Nassim Nicholas, The Black Swan: The Impact of the Highly Improbable. Random House (US), 2007.

[6] P. Baker and A. Potts, “Why do white people have thin lips?’ Google and the perpetuation of stereotypes via auto-complete search forms,” Crit. Discourse Stud., vol. 10, no. 2, pp. 187–204, 2013.

[7] N. Couldry, “Extending the Range of Bourdieu’s Field Theory,” Theory Soc., vol. 32, no. 5, pp. 653–677, 2003.

[8] I. H. W. and M. G. and T. Numerico, WEB Dragons: Inside the Myths of Search Engine Technology. CA: Morgan Kaufmann, 2007.

[9] S. Weber, Das Google-Copy-Paste-Syndrom (TELEPOLIS): Wie Netzplagiate Ausbildung und Wissen gefährden. 2008.

[10] A. Keen, The Cult of the Amateur: How blogs, MySpace, YouTube, and the rest of today’s user-generated media are destroying our economy, our culture, and our values. 2008.

[11] N. Kulathuramaiyer and H. Maurer, “A Survey of Communications and Collaborative Web Technologies,” J. Comput. Inf. Technol., vol. 23, no. 1, pp. 1–18, 2015.

[12] A. Molavi Kakhki, C. Kliman-Silver, and A. Mislove, “Iolaus: securing online content rating systems,” in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 919–930.

[13] “Geography im Austria-Forum.” [Online]. Available: http://austria-forum.org/af/Geography. [Accessed: 02-Nov-2015].

[14] “FactbookXML - All textual information of the CIA World Factbook in one XML file.” [Online]. Available: http://jmatchparser.sourceforge.net/factbook/. [Accessed: 21-Oct-2015].

[15] “Österreichische UNESCO-Kommission.” [Online]. Available: http://www.unesco.at/. [Accessed: 14-Dec-2015].

[16] “All Nobel Prizes.” [Online]. Available: http://www.nobelprize.org/nobel_prizes/lists/all/. [Accessed: 14-Dec-2015].

[17] “Wolf Foundation.” [Online]. Available: http://www.wolffund.org.il/index.php?dir=site&page=content&cs=3044&langpage=eng. [Accessed: 14-Dec-2015].

[18] “Pictures” [Online]. Available: http://austria-forum.org/af/Geography/Asia/Syria/Pictures. [Accessed: 19-Oct-2015].

[19] “Member States of the United Nations.” [Online]. Available: http://www.un.org/en/members/. [Accessed: 14-Dec-2015].

[20] “BROCKHAUS.” [Online]. Available: http://www.brockhaus.de/. [Accessed: 21-Oct-2015].

[21] “Country Codes, Phone Codes, Dialing Codes, Telephone Codes, ISO Country Codes.” [Online]. Available: https://countrycode.org/. [Accessed: 14-Dec-2015].

[22] “Indicators Data.” [Online]. Available: http://data.worldbank.org/indicator/. [Accessed: 14-Dec-2015].

[23] “Department of Statistics Malaysia Official Portal.” [Online]. Available: https://www.statistics.gov.my/. [Accessed: 02-Jan-2016].

Chapter 6 : Geographic Data Verification

The dramatic increase of data sources on the web does not make it easier to find reliable information, since often data sets found do not agree. Therefore a mechanism is needed to verify online information.

In this chapter we start by defining the term geography and identifying geographic data covering different branches of geography. The prime objective of our efforts is to automatically verify parts of geographic information which is found on the web using different methods and approaches. Our contribution focuses on two issues; (i) addressing the reliability of information and (ii) verifying geographic data found in different online data sources by applying spatial queries on maps of the world.

This chapter explains the following spatial methods using real life examples also discussed in [1]:

- STArea ()
- STLength()
- STIntersect()
- STTouches()
- STCrosses ()
- STTouches()
- STBoundary()
- STIntersection()
- STExteriorRing()

6.1 Introduction

There are innumerable definitions of the word geography. Geography is a branch of science that deals with the study of earth and the factors that affect the living things on earth. Very broadly it can be seen as dealing with all aspects if they are location oriented. There are different types of geography. Each type covers a different type of data. We present a classification of geographic data that covers different domains of geography as shown in Figure 6.1. The major emphasis in this chapter is on verifying information related to physical geographic features as highlighted in Figure 6.1. For instance, the examples of data that relate to physical geography of any country; such as area of the country, neighbours of the country, total length of political boundary of the country etc. One of the key aspects related to geographic information on the web is information quality. The dubious nature of geographic data needs special attention in terms of reliability. For instance, the political boundaries of some countries often remain an unsolved issue. Pakistan, India and China are well known examples, but even Russia and Japan cannot agree who should own a number of islands, nor is it clear at this point whether the peninsula Krim belongs to Russia or Ukraine, etc. Where a river starts and where it ends is not simple to answer, since e.g. the tide does turn around some slow flowing rivers like the Sarawack in Borneo for dozens of kilometres!

There are two obvious methods for ascertaining the correctness of physical geographic data (i) textual verification (ii) geometric verification particularly in case of physical geography using maps.

Textual verification of geographic data can be done using online encyclopedias such as Britannica[7] and Brockhaus[8] (the largest German encyclopedias). A few examples are already discussed in Chatper 5.

For geometric verification, we have conducted an analytical study that consists of two steps; (i) calculation of physical features of countries of the world using spatial queries and (ii) comparison of results of geometric calculations with the geographic data that is found in different online data sources such as Wikipedia or Factbook.

The rest of the chapter is structured as follows. Section 6.2 presents a literature review that explores the factors contributing towards the reliability of web page. It presents glimpses of research work conducted in past, to verify geographic information. Spatial data handling tools and techniques that play a vital role in calculation of physical geographic features are discussed in Section 6.3. Section 6.4 is based on the actual verification of geographic features using examples. This is followed by a short conclusion.

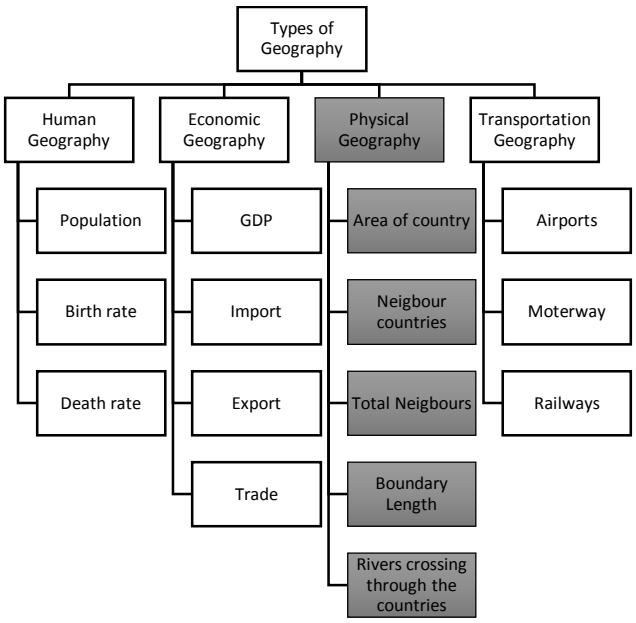


Figure 6.1: Examples of geographic data covering different branches of geography

6.2 Literature Review

This section presents an overview of current approaches and methods that have been proposed for judging the reliability of online information sources. The features contributing towards reliability of an online data source are shown in the Figure 6.2. Often, information is found on the websites that give no indication of a specific source from which the information is taken. Most of the time, outdated information is presented. Time and date are very important dimensions that make information reliable. The valid indication of the reference, from which information is taken, makes it more trustworthy. The data is usually considered more reliable if it is found in a website having ".gov" domain. The reason behind this lies in the fact that it is produced by the government of a particular state, city or country.

[7] http://www.britannica.com/
[8] https://tugraz.brockhaus-wissensservice.com/

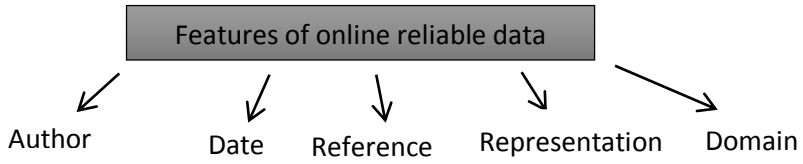


Figure 6.2: Features contributing towards reliability of an online source of information

Wikipedia ensures verifiability of articles by the presence of cited reliable sources so that readers of the article can verify the information using a cited source. The difficulty is that cited sources are usually not accessible electronically and are often available in printed form only in rather specialized libraries.

The textual verification is a well-known research area in terms of verifying information existing in Wikipedia. Calzada and Dekhtzar presented an approach to check the quality of Wikipedia articles using encyclopaedias [2]. Weather verification using geographic systems is done by Ken Waters[3]. In contrast to textual verification our efforts are based on geometric verification.

This chapter is a step towards verifying online data using maps. In order to verify geographic data we have used spatial queries. A spatial query means that we want to select certain data based on its location. Ujaval Gandhi in [4] has highlighted the role of spatial queries to verify geographic data. He has presented the verification of geographic data using a case study about populated places located near rivers. Two shape files that contain data of rivers and populated places in the world were taken for spatial analysis. He used buffer query to calculate the populated places near the selected river crossing through a particular region on the earth.

To verify information we take data from online data sources, particularly from CIA world fact book and Wikipedia. The CIA world fact book is the most widely used online information source for geography. It covers many aspects of geography. Much geographic information is also found in Wikipedia. It presents data in the form of tables and info boxes. The info boxes in Wikipedia contain factual information about countries of the world and other geographic entities such as cities, mountains, rivers, etc.

6.3 Spatial Data Analysis

This section is a brief introduction to spatial data and serves as a background to understand the verification of geographic data in Section 6.4. Geographic features of earth can be represented by three main types of geometry such as point, line, and polygon as shown in the Figure 6.3.

Point is used to represent geographic location such as cities, parks or such. Line is used to represent

rivers, roads, railways etc., and polygon is used to represent features like countries, islands etc. There is also the possibility of combining geometries in a 'GeometryCollection' data type. It may contain any type of geometries. An example of 'Geometric Collection' is the country Japan which consists of many islands or "the great wall of china"[5][6]. There is clearly no exact type of geometry that can be used to represent all geographic features. The choice of choosing a particular type to represent data depends on the requirement of the application[6].

Most database management systems are providing spatial data handling facilities. This technology enhancement has made it easier for geographers and researchers to analyse spatial entities. Spatial databases play a vital role in a number of applications such as GIS, AutoCAD etc. as highlighted in [7]. Spatial operators come handy during spatial analysis.

For verifying physical features we have used two main spatial data types namely "geography" and "geometry". Geography data type is used to represent geodetic data. Geometry data type is used to represent planar data i.e. data that defines an object on a flat 2D surface.

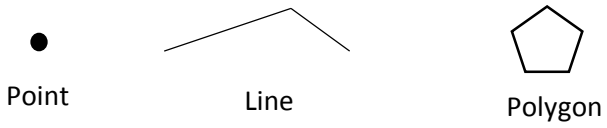


Figure 6.3: Examples of objects that are used to conceptualize geographical features like river, roads, and countries

To represent a geographical entity, a coordinate system is needed. The most widely used reference system is the World Geodetic System (WGS 84). It is a mathematical system used to plot locations globally. It is a standard coordinate system. This is the default system used by GPS devices to represent geodetic data. In the process of analysing spatial data, the spatial reference system affects the calculation of physical features as different reference system provides different units of measurement.

Geospatial topology explores the rules concerning the relationships between points, lines, and polygons that represent the features of a geographic region. Topological concepts are also used for analysing spatial relationships in many situations and the importance of topological relations is widely recognized [12] [13].

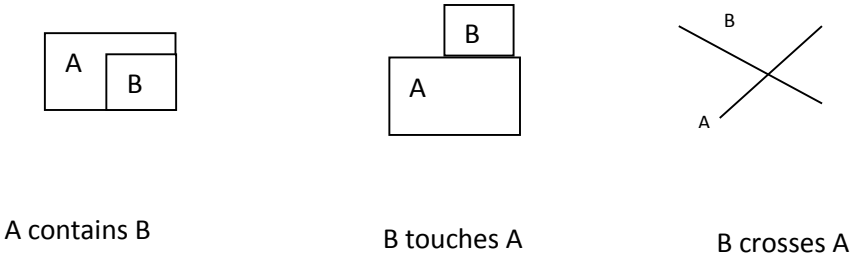


Figure 6.4: Examples of topological relationships

There are various topological relations such as (Equals, Disjoint, Intersects, Touches, Contains, Within, Covers) that can exist between two geometric items. The conceptual description of few of them is shown in the Figure 6.4. In our efforts we have applied these relationships to calculate the physical features of the world countries like: What are the neighbours of a particular country?

6.4 Geographic Data Verification

This section demonstrates the research challenges that we have encountered during verification of data using maps. In order to verify geographic data, the first step was to calculate geographic features of selected countries from maps followed by their comparison with the corresponding geographic facts found in the CIA world fact book and Wikipedia. For this purpose we have downloaded the shape file (.shp) of the world countries from Natural Earth[9] which turned out to be very helpful resource for geographic shape files. The SQL server spatial tool "shape to sql" was used to upload the shape files in a database. Afterwards spatial queries were applied to find out the topological relationships between countries on the map. The map of countries of the world is shown in the Figure 6.5.

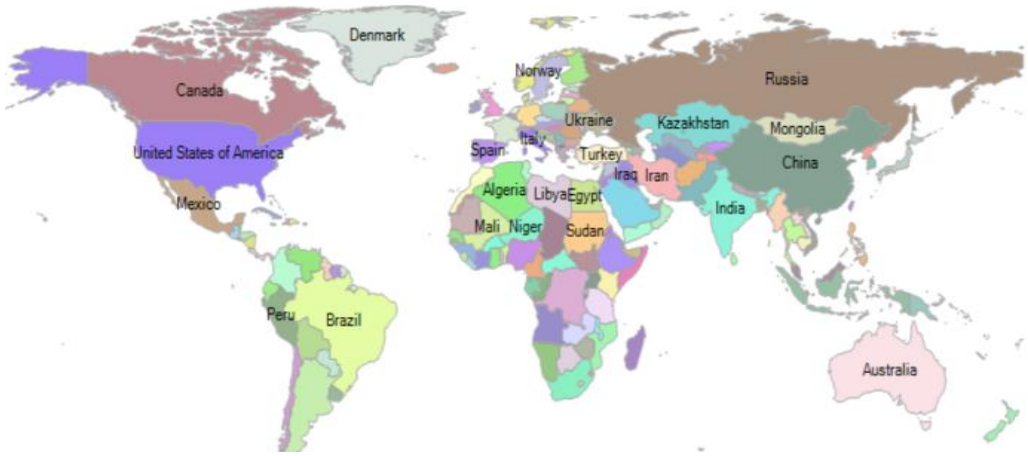


Figure 6.5: World map showing countries

6.4.1 Research Question 1: How to calculate distance between cities/locations.

STDistance () gives the shortest distance between two points. When used on geographic instances it gives the shortest great elliptic arc distance between two locations in meters. We can calculate the shortest distance between two cities or any two locations with a spatial query using STDistance() method as also discussed in [8].

```
1: Declare @Graz geography=geography::Point(47.066,15.45,4326)
2: Declare @Vienna geography=geography::Point(48.2,16.366,4326)
3: select @Vienna.STDistance(@Graz)/1000
```

Code Listing 6.1 a: Distance calculation between Vienna and Graz

Prior to SQL Server 2008, geography locations such as bank branches, restaurant locations and household locations, were usually stored as latitude and longitude coordinates in two separate columns [9]. With the introduction of spatial datatypes particularly Geography data type (latitude and longitude can be stored as point in single column). Point accepts first latitude and then longitude as shown in above example.

To create a geographic instance e.g. Graz, we can use method STPointFromText () or STGeomFromText () but the longitude is specified first before the latitude as shown in the code Listing 6.1 b.

[9]http://www.naturalearthdata.com/

```
Declare @Vienna geography
Declare @Graz geography
SET @Graz=geography::STGeomFromText('POINT(15.45 47.066)', 4326)
set @Vienna=geography::STGeomFromText('POINT(16.366 48.2)', 4326)
select @Vienna.STDistance(@Graz)/1000
```

Code Listing 6.1b: Distance calculation between Vienna and Graz

The above query returns the distance between two cities in kilo meters as 143.648 km.
The STDistance () method, see Line 3 in code Listing 6.1 a, gives the distance between two points/locations in meters; therefore we are dividing distance by 1000 to get distance in km. One can see the unit of measure against reference ids using the query in the Listing 6.2.

```
1: SELECT unit_of_measure from
2: Sys.spatial_reference_systems
3: where authorized_spatial_reference_id=4326
```

Code Listing 6.2: Checking units of measurement

The output of query is in meters. When we perform a calculation on spatial data, it returns results according to units relevant to the spatial reference system. In above distance calculation example "4326" id is used.

6.4.2 Research Question 2: How to verify neighbouring countries of a particular country?

To find out the neighbour countries we have used an adjacency query. Figure 6.6 shows the field entry "Land boundaries" from CIA world fact book. The sub entry "border countries" tells that there are 6 countries that are sharing a border with Afghanistan. Figure 6.7a verifies this information as the map shows that there are 6 neighbouring countries of Afghanistan. This map is drawn using adjacency spatial query which extracts all objects in the map that are adjacent to a particular object in the map as shown in Code Listing 6.3. The map showing neighbour countries of Austria is also shown in Figure 6.7 b. The STTouches () method can also be used. This function is used to find out geometric objects that touch other instance.

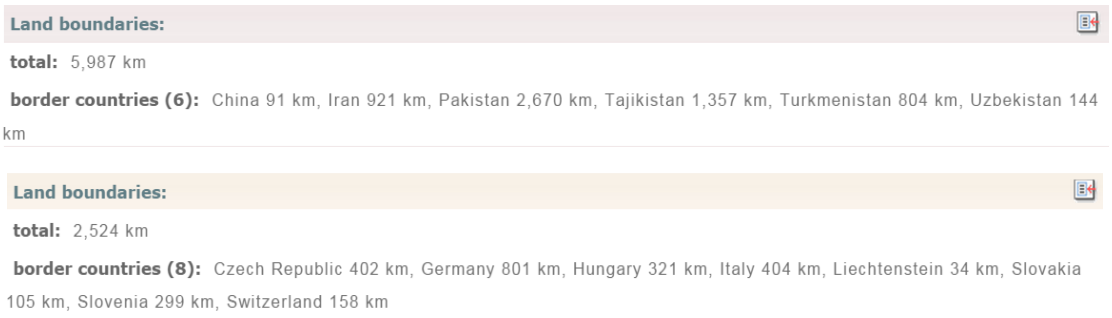
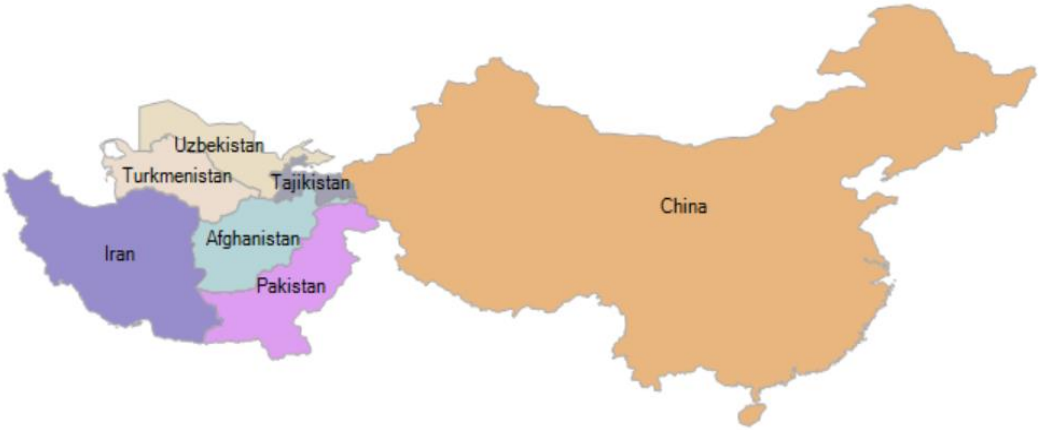


Figure 6.6: World Fact book Field Entry showing border countries of Afghanistan and Austria.



6.7 a: Neighbouring countries of Aghanistan

6.7 b: Neighbour countries of Austria (Liechtenstein is not shown in the map as too small. It lies between Switzerland and Austria)

Figure 6.7: Map showing neighbour countries

```
1: Declare @g geometry
2: Select @g=geom from World_Countries where
3: country='Afghanistan'
4: select country,geom from World_Countries
5: where geom.STIntersects(@g)=1
```

Code Listing 6.3: Spatial query to find neighbouring countries of Afghanistan

6.4.3 Research Question 3: How to verify the boundary length of neighbouring countries?

A boundary is a real or imaginary line that separates two things. The most obvious type of boundary is a physical boundary. A physical boundary is a naturally occurring barrier between two areas. A political boundary is a boundary line drawn between two countries. Political boundaries are created by humans, unlike natural boundaries which are created by natural processes leading to formation of rivers, watersheds, mountain ranges or coastal lines. Natural boundaries have a pronounced obstacle impeding movement which may differ from one species to another. Humans usually ignore the natural boundaries when setting political boundaries. We are verifying the length of political boundaries in this particular geographic verification. The STIntersects () method finds those spatial objects which touch each other. It works with both Geometry and Geography objects.

Which countries border Austria?

According to our research engine:

The following countries directly border Austria (border lengths included):

- Czech Republic 362 km
- Germany 784 km
- Hungary 366 km
- Italy 430 km
- Liechtenstein 35 km
- Slovakia 91 km
- Slovenia 330 km
- Switzerland 164 km

Figure 6.8: Facts taken from Country Facts [10]; showing boundary length of countries sharing border with Austria

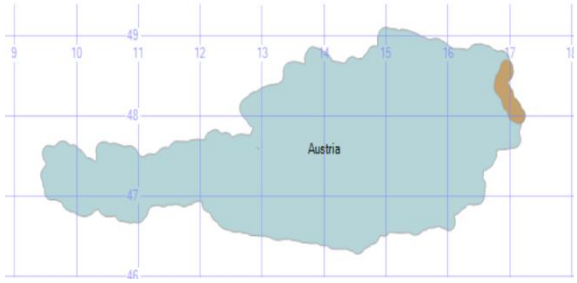


Figure 6.9: Map of Austria showing border with Slovakia.

The country facts [10] page shown in the Figure 6.8 represents that Slovakia has 91 km long border with Austria. This fact is verified by the map shown in the Figure 6.9. The map is drawn using queries listed in Code Listing 6.4. The boundary is calculated using STLength() function. The trick here is to calculate the length of shared boundaries between Austria and Slovakia using Intersection () function, as it returns the shared geometric region that is common between two geographic objects which are countries in our case. The unit of length is calculated according to the reference system. To calculate the length in meters the SRID 4326 is used.

```
declare @g1 geography
declare @g2 geography
select @g1=Geography::STGeomFromText(geom.STAsText(),4326) from countries where
SOVEREIGNT='Austria'
select @g2=Geography::STGeomFromText(geom.STAsText(),4326) from countries where
SOVEREIGNT='Slovakia'
select @g1.STIntersection(@g2).STLength()/1000
```

Code Listing 6.4: Spatial query to find the boundary length of Slovakia sharing border with Austria

STIntersection () method returns an object that represents the points where a **geometry** instance intersects another **geometry** instance.

6.4.4 Research Question 4: How to verify area of a country?

In this section we will explain the power of STArea () method. There are some online tutorials[11] [12] that are about using this method to calculate the area of places like city of “Chicago” and “Berkshare”. We are following the same examples and apply this method on polygons of countries. The area of a country is one of the main geographic features. Russia is the largest country in the world with respect

to area. To verify this feature we have taken Austria as an example. The area of Austria is found to be approximately 83,871 sq. km as shown in Figure 6.10.

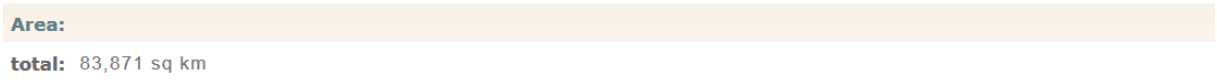


Figure 6.10: World Fact book Field Entry showing area of Austria

The total area of Austria in square kilometres can be calculated using Listing 6.5. The STArea () returns the results in sq meter; therefore we are dividing this by 1000,000 to get the results in sq km.

```
1:DECLARE @g geography
2:select
3:@g=geography::STGeomFromText(geom.STAsText(),4326) from World_Countries where
4: Country='Austria'
5:select 'Area '+@g.STArea())/1000000  + 'sq km'
```

Code Listing 6.5: Spatial query to find the area of Austria

Table 6.1 shows countries with total area (sq km) taken from Factbook and STArea () calculation and difference

Table 6.1: Results of calculated and observed area of countries (sq km) taken from Factbook

S#	Country	Fact book Total Area (sq km)	Calculated Area (STArea()) Sq km	Difference (sq km)
1	Austria	83871	83993	122
2	Aruba	180	169	11
3	Andorra	468	452	16
4	Belgium	30528	30665	137
5	Belarus	207600	207499	101

6.4.5 Research Question 5: How to verify the boundary of a country?

A polygon is defined by a closed line string called exterior ring [1]. As countries in the map are represented by polygons, boundary calculation of countries becomes easier. The length of the outer boundary of the country can be calculated using the spatial query listed in Code Listing 6.6 a. The STExteriorRing() method is used. It returns outer closed Line String of the Polygon as shown in the Figure 6.11, also discussed in [13]. The STLength () function is then used to calculate the length of the border of Belgium. The boundary of Belgium is approximately found to be 1145 km; which is very close to the entry in World Factbook. The STBoundary() method can also be used to extract the outer ring of polygon. It returns the boundary of a **geometry** instance.

Note, the boundary length and area can be inaccurate due to the fact that the polygons used to approximate the country's map are often not detailed enough. Natural earth[10] data comes in different scales of 1:10m, 1:50m, and 1:110 million scales. This factor also affects the calculations.

```
1: select geography::STGeomFromText((geom.STExteriorRing().MakeValid())
2:.STAsText(),4326).
3: STLength()/1000
4: SOVEREIGNT  from World_Countries where NAME='Belgium'
```

Code Listing 6.6 a: Spatial query to find the boundary length of Belgium

[10] http://www.naturalearthdata.com/

```
select geography::STGeomFromText(geom.STBoundary().STAsText(),4326).STLength()/1000
from countries where SOVEREIGNT=' Liechtenstein'
```

Code Listing 6.6 b: Spatial query to find the boundary length of Liechtenstein

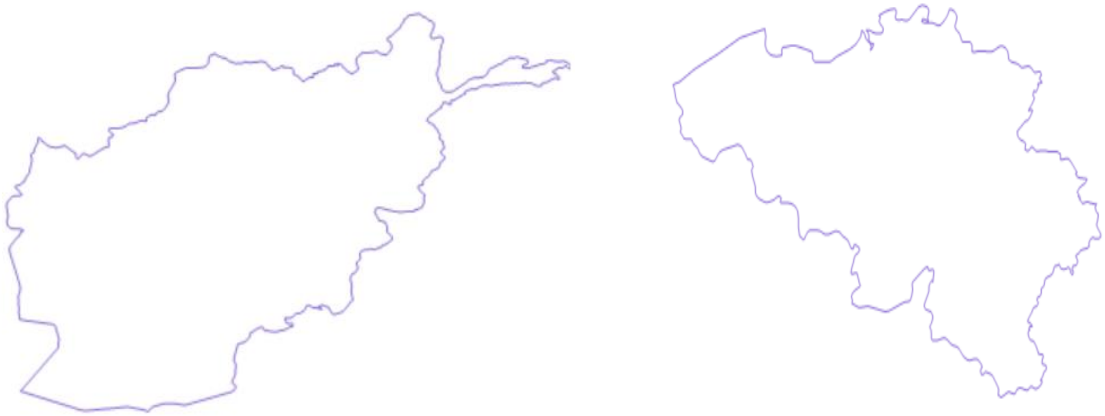


Figure 6.11: Map of Afghanistan and Belgium. STExteriorRing () method returns Line String geometry shown in purple colour

Table 6.2: Results of calculated and observed land boundaries of countries (km) taken from Factbook

S#	Country	Fact book Total boundary (km)	Boundary Calculation km	Difference (km)
1	Afghanistan	5529	5225	304
2	Moldova	1390	1324	66
3	Andorra	120.3	96.6	23.7
4	Benin	1989	1996	7
5	Bhutan	1075	1020	55
6	Rwanda	892	822	70

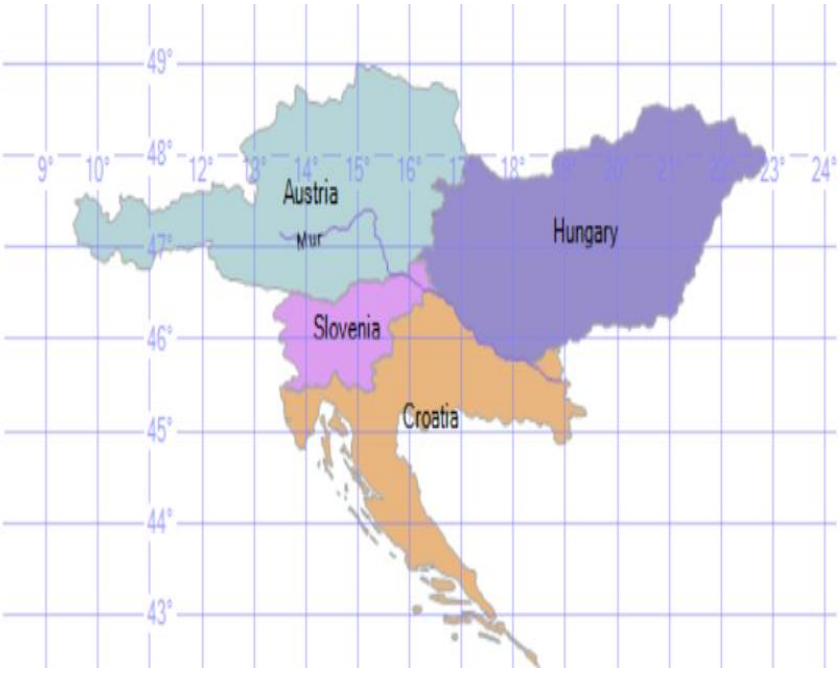


Figure 6.12: Map showing river Mur crossing through countries Austria, Slovenia, Hungary, and Croatia

Origin	Hohe Tauern at Muhr, Austria 47°7'48"N 13°20'49"E
Mouth	Drava River near Legrad, Croatia 46°17'54"N 16°53'9"E
Basin countries	Austria, Slovenia, Croatia, Hungary
Length	480 km (300 mi)
Source elevation	1,898 m (6,227 ft)
Mouth elevation	130 m (430 ft)
Avg. discharge	166 m^3/s (5,900 cu ft/s) at mouth
Basin area	13,824 km^2 (5,337 sq mi)

Figure 6.13: Wikipedia Article of River Mur

6.4.6 Research Question 5: How to verify countries through which a river crosses?

To answer this question let us take an example. The river Mur which is a big river flowing in Austria. We have drawn a map shown in the Figure 6.12 using Code Listing 6.7. This verifies the fact about the countries displayed in the infobox of Wikipedia article as shown in the Figure 6.13. It is clearly shown in the map of Figure 6.12 that river Mur, shown by a line string, originates in Austria flows through Austria, Slovenia, Croatia and Hungary represented by the coloured geometric elements in the map.

```
1:Select name,geography::STGeomFromText
2:(geom.STAsText(),4326) from rivers
3:where featurecla='River' and name='Mur'
4:union all
5:select
6:country,geography::STGeomFromText(geom.STA
7:sText(),4326)from World_Countries where
8:country='Austria' or
9: country ='Croatia' or country ='Slovenia'
10:or country ='Hungary'
```

Code Listing 6.7: Spatial query to draw map to show countries through which river Mur crosses

One way of doing automatic verification is using STCrosses () method as shown in Code Listing 6.8. It returns true if a geometry instance crosses another geometry instance and returns false if it does not. This function is of extreme importance, as it finds out the countries through which a river crosses. Special care must be needed before applying this method, as it cannot be directly applied to geographic instances: they have to be converted to geometric data types [1]. For performing automatic verification of rivers crossing through countries we have taken two shape files one describing rivers and the other representing world countries. After converting them to geometric data type, we applied the STCrosses() method to verify the countries through which each river crosses, this leads to automatic verification of parameter "basin countries of rivers". We can pass any river name (Indus in this case) in Code Listing 6.8 and the query draws the map of basin countries for that particular river as shown in the Figure 6.14. River Indus flows mostly in Pakistan (93%) but also covers India (5%) and China (2%).

```
declare @g geometry
select @g=geom from rivers where name='Indus'
select name,geom.MakeValid() from countries where
@g.STCrosses(geom.MakeValid())=1
union all
select 'Indus',@g
```

Code Listing 6.8: Spatial query to automatically verify countries through which a river crosses

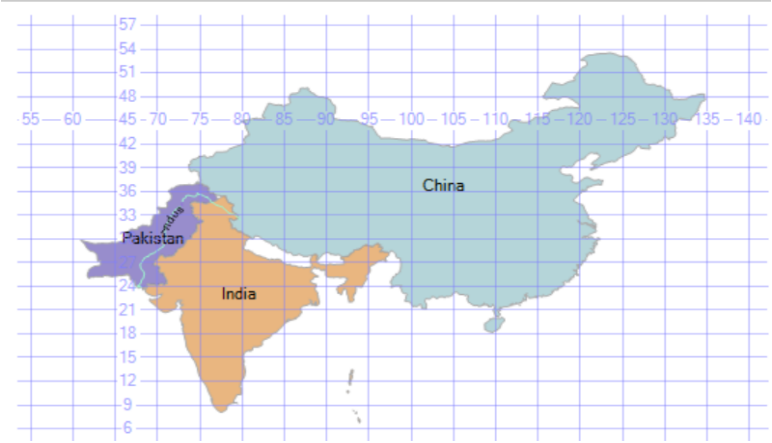


Figure 6.14: Countries of River Indus.

6.5 Conclusion

In this chapter we have explained the significance and use of spatial queries. We have applied a unique way of verifying geographic data. Our efforts are mostly based on querying maps using spatial methods. This is a step towards automatic verification of data sources. We have examined the role of spatial operators and particularly linked their significance to the verification of geographic information with real world examples. The calculated results however of course depends on the quality of shape files. The numbers can vary to some extent as shown here. How we handle small differences in area in global-geography.org (1/10 of average formula) is already discussed in Chapter 5.

References

[1] "Geometry Data Type Methods in SQL Server." [Online]. Available: http://www.techytube.com/geometry-data-types-methods-in-sql-server/. [Accessed: 20-Oct-2015].

[2] G. De la Calzada and A. Dekhtyar, "On measuring the quality of Wikipedia articles," Proc. 4th Work. Inf. Credibil. - WICOW '10, p. 11, 2010.

[3] "Weather Verification Using Geographic Information Systems" [Online]. Available: http://www.nws.noaa.gov/regsci/gis/presentations/ESRI_Fall06.pdf. [Accessed: 27-Jan-2016].

[4] "Performing Spatial Queries — QGIS Tutorials and Tips." [Online]. Available: http://www.qgistutorials.com/en/docs/performing_spatial_queries.html. [Accessed: 27-Jan-2016].

[5] E. Clementini and P. Di Felice, "Approximate topological relations," Int. J. Approx. Reason., vol. 16, no. 2, pp. 173–204, 1997.

[6] A. Aitchison, Beginning Spatial with SQL Server 2008. Apress, 2009.

[7] A. Silberschatz; and H. Korth, Database System Concepts. Mcgraw Hill Book, 2010.

[8] "Calculating a distance between two points of interest using SQL Server 2008 geography data type." [Online]. Available: http://blogs.msdn.com/b/pekorica/archive/2009/03/16/calculating-a-distance-between-two-point-of-interest-using-sql-server-2008-geography-data-type.aspx. [Accessed: 24-Jan-2016].

[9] "SQL Server 2008 - Convert Lat/Long to Geography Point." [Online]. Available: http://www.sql-server-helper.com/sql-server-2008/convert-latitude-longitude-to-geography-point.aspx. [Accessed: 17-Jan-2016].

[10] "Which countries border Austria?" [Online]. Available: http://country-facts.findthedata.com/q/96/14319/Which-countries-border-Austria. [Accessed: 20-Oct-2015].

[11] "Working with Spatial Data in SQL Server 2008." [Online]. Available: http://www.dotnetsolutions.co.uk/working-with-spatial-data-in-sql-server-2008/. [Accessed: 17-Jan-2016].

[12] "SQL Server Geospatial Part 3 – Visualizing Spatial Data in SSRS." [Online]. Available: http://community.rightpoint.com/blogs/viewpoint/archive/2010/07/24/sql-server-geospatial-part-3-visualizing-spatial-data-in-ssrs.aspx. [Accessed: 17-Jan-2016].

[13] "How do I add polyline length to a Bing Maps polygon - Stack Overflow." [Online]. Available: http://stackoverflow.com/questions/25169679/how-do-i-add-polyline-length-to-a-bing-maps-polygon. [Accessed: 18-Jan-2016].

Chapter 7 : Data Sources used for Facts collection and Verification

Geographic data sources have been around for a long time but have been extended by much pictorial material due to advances in remote sensing, aerial photography, satellite communication and crowdsourcing. The aim of this chapter is to identify data sources visible on the web having different aspects of information about countries, as also discussed in [1]. In this chapter we will examine particularly existing geographic data sources that contain vast amount of geographic information. The goal of this chapter is to present ways of data extraction from existing data sources that we are using either for data verification or else as facts in the appropriate sections of our geographic server. We will also summarize ways of information verification and present city rankings verification results of 193 UN countries.

This chapter will address the following research questions

RQ1. Search and categorize data sources on Web that contain different aspects of information about the countries.

RQ2. How to extract data from online sources (free to use or require some license to overcome copyright issues).

RQ3. What would be a convenient way of relational data representation that helps in automatic generation of Wiki pages in Austria-Forum?

RQ4. How to verify information using different methods?

RQ5. What is the situation of city rankings in data sources particularly when examining Wolfram Alpha vs Geonames? (Note that in contrast to Chapter 5, we will cover all 193 countries in this chapter)

RQ6. How to create a meaningful Wiki interface that coveys proper messages when data sources agree or disagree about a fact?

7.1 Data Sources

This section looks at sources of information that provide information related to countries. We group data sources according to their type, as shown in Figure 7.1. We will briefly look at each type. We will explain what kinds of facts are stored in a particular data source. We talk about the data storage mechanism followed by extraction of facts in this section.

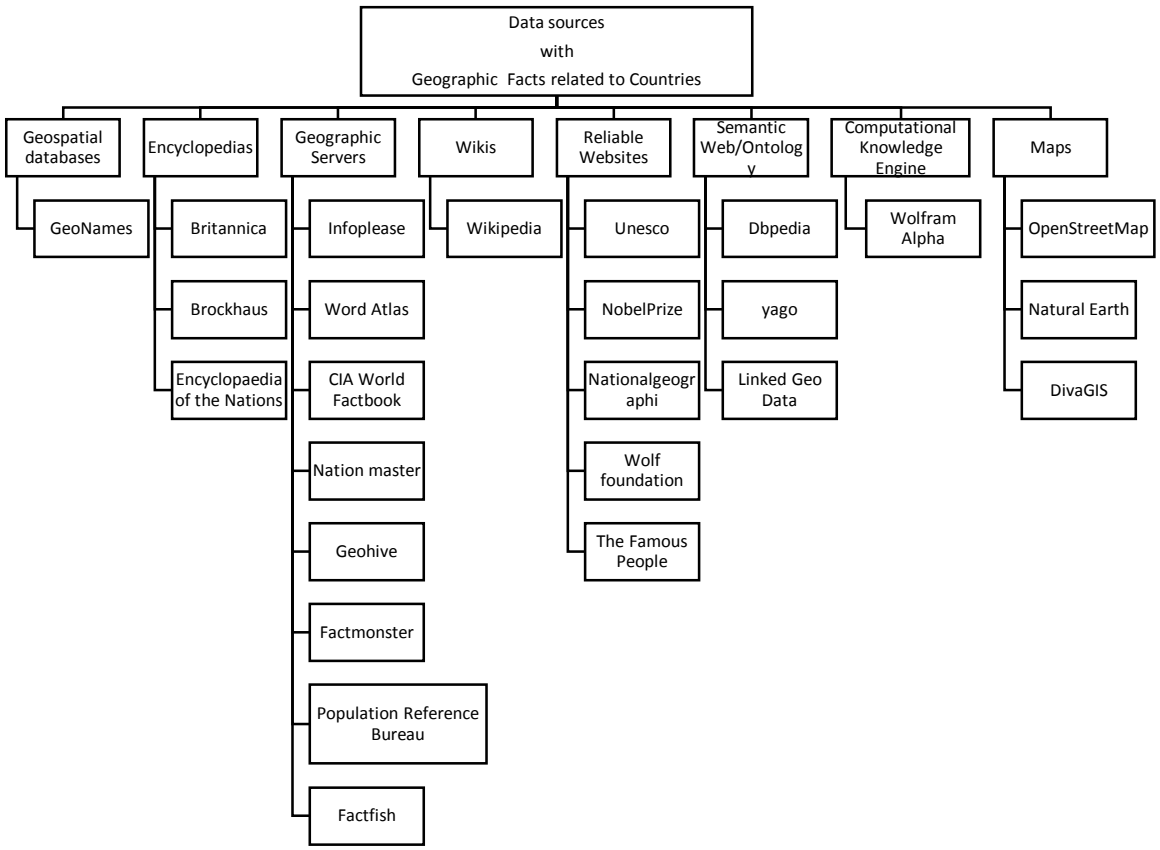


Figure 7.1: Data sources

7.1.1 Geospatial databases

7.1.1.1 GeoNames

It contains over 8 million place names that are available for download free of charge. The GeoNames data dump is available in the form of a text file which requires a little bit of tweaking before importing it into SQL Server DBMS. Listing 7.1 shows a query to extract 5 highest mountains of Austria from GeoNames.

```
1: select top 5 name,elevation
2: from GeoNames
3: where country_code=AT
4: and feature_class=T
5: and feature_code=MT
6: and elevation > 0
7: order by elevation desc.
```

Listing 7.1

GeoNames is a geographical database; it covers all countries. Further, it provides web services to access data.

GeoNames is an open source geographic database. It provides very useful "REST[11] (Representational State Transfer)" web services for extracting structured data in XML or JSON format. A login is required to access the data using web services. These web services return different types of information such as countries and their capitals.

```
Websservice: Country Info (Bounding Box, Capital, Area in square km, Population)
Web service Type: REST
URL: http://api.GeoNames.org/countryInfo?username=demo
```

[11] http://www.ibm.com/developerworks/library/ws-restful/

The ***CountryInfo*** web service provides information about the countries and it returns data in JSON and XML format as shown in Figure 7.2. The <geonameId> of Pakistan is 1168579. This geonameId is passed as reference to other web services to get more facts about the Pakistan.


```
▼<country>
  <countryCode>PK</countryCode>
  <countryName>Pakistan</countryName>
  <isoNumeric>586</isoNumeric>
  <isoAlpha3>PAK</isoAlpha3>
  <fipsCode>PK</fipsCode>
  <continent>AS</continent>
  <continentName>Asia</continentName>
  <capital>Islamabad</capital>
  <areaInSqKm>803940.0</areaInSqKm>
  <population>184404791</population>
  <currencyCode>PKR</currencyCode>
  <languages>ur-PK,en-PK,pa,sd,ps,brh</languages>
  <geonameId>1168579</geonameId>
  <west>60.878613</west>
  <north>37.097</north>
  <east>77.840919</east>
  <south>23.786722</south>
 </country>
```


Figure 7.2: XML File displaying facts of Pakistan in GeoNames.

The ***children*** web service returns the childrens (admin divisions and populated places) for a given geonameId. The children web service is invoked using URL as shown in the box below. The geonameId is the id of Pakistan where username is IICM. It is a username that is created for accessing data from Geonames. The web service returns the data of provinces of Pakistan in XML format as shown in Code Figure 7.3.

```
http://api.GeoNames.org/children?geonameId=1168579&username=iicm
```


```
▼<geoname>
  <toponymName>Balochistān</toponymName>
  <name>Balochistān</name>
  <lat>28.47849</lat>
  <lng>65.64355</lng>
  <geonameId>1183606</geonameId>
  <countryCode>PK</countryCode>
  <countryName>Pakistan</countryName>
  <fcl>A</fcl>
  <fcode>ADM1</fcode>
  <numberOfChildren>34</numberOfChildren>
 </geoname>
▼<geoname>
  <toponymName>Federally Administered Tribal Areas</toponymName>
  <name>Federally Administered Tribal Areas</name>
  <lat>33.01455</lat>
  <lng>69.99925</lng>
  <geonameId>1179245</geonameId>
  <countryCode>PK</countryCode>
  <countryName>Pakistan</countryName>
  <fcl>A</fcl>
  <fcode>ADM1</fcode>
  <numberOfChildren>14</numberOfChildren>
 </geoname>
▼<geoname>
```


Figure 7.3: Children web service displaying facts about provinces of Pakistan in XML.

Geonames provide different country codes for countries. A brief list is given in Table 7.1. We have used them in extracting data from other websites. For example the Nobel Prize winner API requires ISO-3166 alpha 2 code for data extraction. DiVA GIS web site which stores countries boundaries (shape files) require ISO-3166 alpha 3 for data extraction.

Table 7.1: Country codes

Country	ISO-3166 alpha 2	Iso-3166 Alpha 3	ISO-3166 Numeric	Fips
Andorra	AD	AND	020	AN
Austria	AT	AUT	040	AU
Armenia	AM	ARM	051	AM
United Arab Emirates	AE	ARE	784	AE
Afghanistan	AF	AFG	004	AF
Albania	AL	ALB	008	AV
Australia	AU	AUS	036	AS

7.1.2 Semantic Web

7.1.2.1 DBpedia

DBpedia is one of the online structured information resources. The idea behind constructing DBpedia[12] was to extract structured content from the information which is created as part of the Wikipedia project. This structured information is then made available on the World Wide Web and extracted using SPARQL [2] queries. It contains geographic facts about the countries in the form of triples that were originally stored in the form of tables and info boxes in Wikipedia.

The RDF triples describing facts of Pakistan are shown in Figure 7.4. Each RDF triple represents a subject, a predicate and an object. Pakistan is a subject; predicates are “capital”, “currency” and “area Total”. The objects are dbpedia:Islamabad, dbpedia:Pakistani_rupee and 796095.

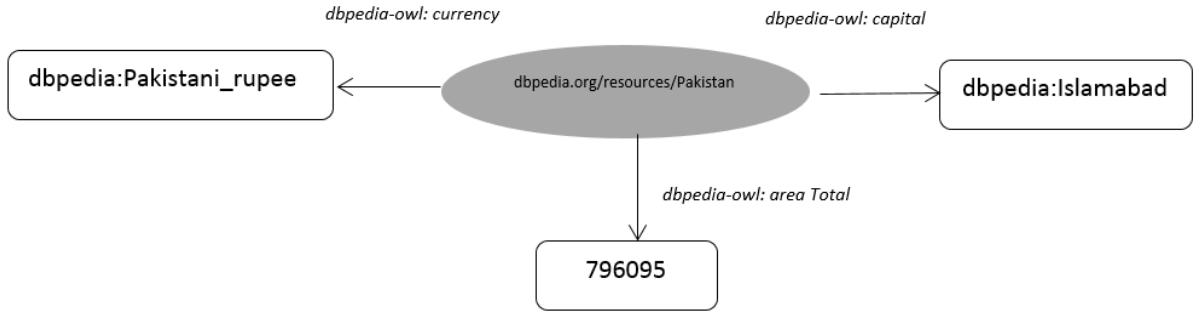


Figure 7.4: RDF triples representing information about Pakistan.

Listing 7.2 shows a SPARQL query to extract facts of Asia from DBpedia.

```
PREFIX dbpprop : <http://dbpedia.org/property/>
PREFIX db:<http://dbpedia.org/resource/>
PREFIX dbpedia-owl:<http://dbpedia.org/resource/>
SELECT ?pop ?countries ?area WHERE {
db:Asia dbpprop:population ?pop.
db:Asia dbpprop:countries ?countries.
db:Asia dbpedia-owl:areaTotal ?area. }
```

Listing 7.2

[12] http://dbpedia.org/About

7.1.3 Geographic Servers

7.1.3.1 CIA world fact book

The CIA world fact book[13] is the most widely used online information source for geography. It covers different aspects of geography. The data can be downloaded from CIA world fact book freely. It provides an archive of geographic information for the last 10 decades which can be used for statistical inference. It is updated biannually. It is itself often referenced by other information sources. The data is in structured form that is easy to extract and use. The different menus displaying information about particular geographic facts about Albania is shown in Figure 7.5. CIA world fact book is up-to-date geographic data source.

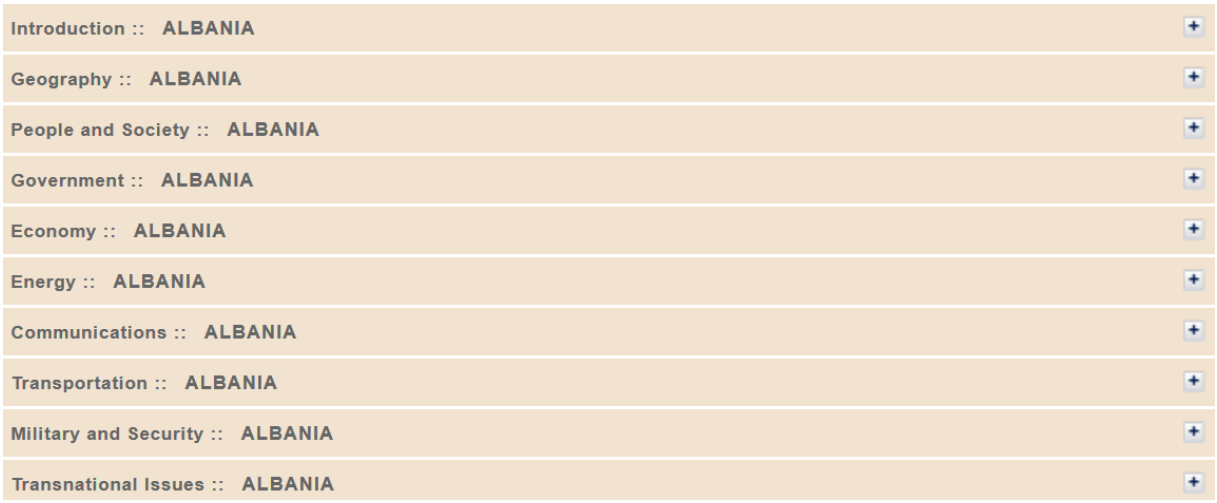


Figure 7.5: World Fact book page of Albania

7.1.3.1.1 Data Extraction from Fact book

The project named "FactbookXML" started by Michael Schierl is commonly used, as it allows to download all fact book data in one xml file. We have downloaded files from FactbookXML [3]. It contains two main files factbook schema and actual data. We have uploaded them in our data store in MySQL (a database management system) database.

1. factbook-mysql-tables.sql (schema)
2. factbook-mysql.sql (data)

7.1.4 Reliable Web sites

This category contains those web sites which contain fairly reliable data. These web sites include UNESCO heritage sites, National geographic, Nobel Prize winners.org etc. The culture section of global-geography.org contains pointers to valuable information on these websites. We are using both links and facts from these web sites. We redirect users to these web sites if they want to learn more about facts.

Nobel prize.org provides API to extract data. An example of extracting Nobel Prize winners from Austria is shown below. Here the code of Austria is "AT" and assigned to "bornCountryCode" in queryString.

```
http://api.nobelprize.org/v1/laureate.json?bornCountryCode=AT
```

[13] https://www.cia.gov/library/publications/the-world-factbook/

7.1.5 Computational Knowledge Engine

7.1.5.1 Wolfram Alpha:

It is different from search engines like Google. It is an online service that answers factual queries directly by computing the answer, rather than providing a list of documents or web pages. WolframAlpha [4] provides an API for extracting data. It needs a subscription to allow open access to data. The wolfram API returns data in XML format. We are using python library xml.etree.ElementTree for parsing xml response, see Listing 7.3, which makes it easy to parse and extract cities from WolframAlpha. We have taken city and mountain rankings from Wolfram Alpha to verify GeoNames rankings.

```
1: _input=Large+five+cities+of+"+country
2:Url="http://api.wolframalpha.com/v2/query?input=_input+"&podtitle=Result"
3: response = urllib2.urlopen(url)
4: tree=ET.parse(response)
5: root=tree.getroot()
6: dict=root.attrib
7: status=dict.values()
8: if status[1]!='false':
9: cities=root[0][0][0].text
```

Listing 7.3: Python code to extract large cities of a country from Wolfram Alpha

7.1.6 Encyclopaedias

7.1.6.1 Britannica:

It provides a general article about a country which includes historical and cultural information. We have used QuickFacts[14] page of each country to verify the area of countries taken from Fact book. The Web page of Finland in Britannica is shown in the Figure 7.6

Official name[1]	Suomen Tasavalta (Finnish); Republiken Finland (Swedish) (Republic of Finland)
Form of government	multiparty republic with one legislative house (Parliament [200])
Head of state	President: Sauli Niinistö
Head of government	Prime Minister: Juha Sipilä
Capital	Helsinki
Official languages	none[1]
Official religion	none
Monetary unit	euro (€)
Population	(2014 est.) 5,461,000
Total area (sq mi)	150,928
Total area (sq km)	390,903

Figure 7.6: QuickFacts Page of Finland in Britannica

Let's consider the code which is used to extract area of countries from Britannica as shown in Listing 7.4. We are using Htmlunit [5], a window-less browser API. It allows to query a web page using a

[14] britannica.com/EBchecked/topic/438805/Pakistan

WebClient object, followed by information extraction using XPath expressions. Once we have the access to table rows, using the XPath expression (shown in line 4, 5 of Listing 7.4), we iterate through the rows and extract a specific table row that contains area of a particular country. Afterwards, we populate our data store Britannica which is latter used for verification.

```
1: String URI="britannica.com/topic/"+value1+"/"+value2+"-quick-facts";
2: final WebClient client = new WebClient();
3: final HtmlPage mainPage = client.getPage(URI);
4: links1=(List<?>)mainPage.getByXPath("//tr[@class='eb-profile-table-even']");
5: HtmlTableRow tr=(HtmlTableRow)(links1.get(j));
6: links2=tr.getChildNodes();
7: while(k<links2.size()) { if(k==0) {
8: th=(HtmlTableHeaderCell)links2.get(k);} if(k==1
9: &&th.asText().contains("sq km"))
10: {td=(HtmlTableDataCell)links2.get(k);
11:    td.asText());
12:  found=1;
13: }k++; }
```

Listing 7.4: Java code to extract total area in sq km from Britannica

7.1.7 Maps

7.1.7.1 OpenStreetMap (OSM)

It is a freely editable map of the world. It was started by Steve Coast in 2004. Osmosis is a command line Java application for processing OSM data. An osmosis command which is used to extract schools in Austria from OpenStreetMap is shown below, for more examples see [1] .

```
C:\ >osmosis --rbf austria.osm.pbf – nkv keyValueList="amenity.schools" --wx school.osm
```

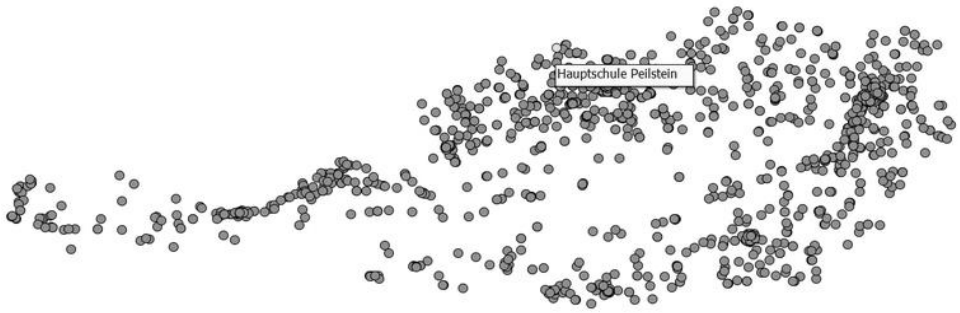


Figure 7.7: Map of Austria taken from OpenStreetMap.

7.1.7.2 Natural Earth

It is an online resource for free geographic dataset, maps and shape files. Natural Earth provides both vector and raster graphics that can be used to draw map visualizations. It allows to download vector map files representing each country.

7.1.7.3 CartoDB

We have used basemaps and tiles from CartoDB. The CartoDB basemaps can be used in any CartoDB.js and Leaflet maps. Adding them is simple and takes just a few lines of code.

```
var layer =
L.tileLayer('http://{s}.basemaps.cartocdn.com/light_all/{z}/{x}/{y}.png', {
attribution: '&copy; <a
href="http://www.openstreetmap.org/copyright">OpenStreetMap</a> contributors,
&copy; <a href="http://cartodb.com/attributions">CartoDB</a>' });
```

We will explain the mapping services of global-geography.org in detail in Chapter 12.

7.2 Relational Schema

Figure 7.8 shows the relational schema behind global-geography.org. The table "factbook_countries" contains information about all the countries of the world. After extracting facts from different data source mentioned in Section 7.1 we store them in MySQL data store.

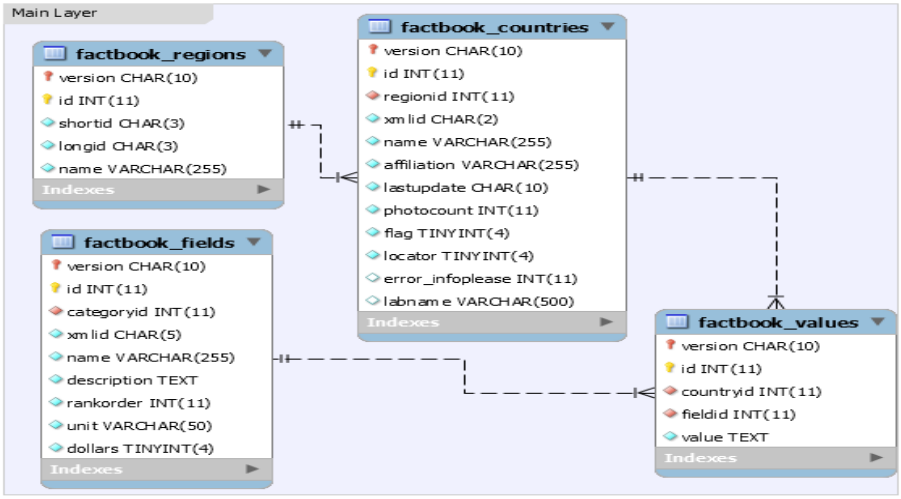


Figure 7.8: Relational schema of Geographic Server; few relevant tables are shown.

The tables "nobelwinners", "wolfprize", "fieldmedals" contain the extracted facts from the respective websites as shown in the Figure 7.9.

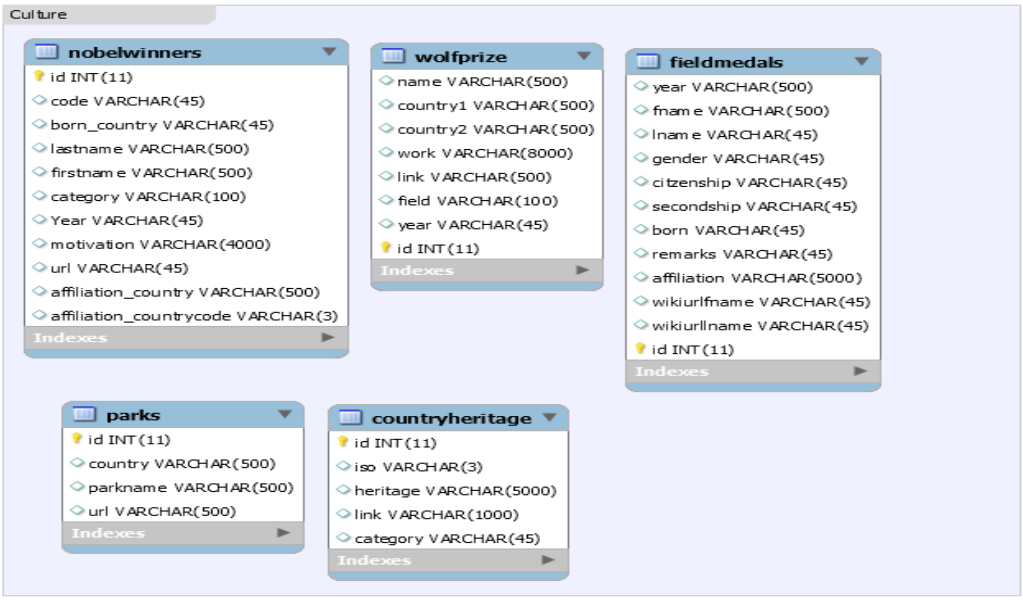


Figure 7.9: Relational schema related to culture section

7.3 Summary of Verification Procedures

7.3.1 Verification using Maps

Physical geographic data such as area of country, land boundaries, neighbour countries of a particular country can be verified using maps. The details of data verification using maps is highlighted in our previous paper, see Chapter 6. A "Spatial" operator allows automatic calculation of geographic features. Using spatial methods such as STArea (), STLength () we can verify facts such as area of country, boundary length of countries etc. The STArea () method returns the total surface area of a geographic entity such a polygon representing a country. More specifically, it returns area in the square of the unit of measure, used by the spatial reference identifier of the geography instance; for example, if the SRID of the instance is 4326, STArea () returns results in square meters, see Chapter 6 for more details. Therefore we can verify geographic facts of countries found in different data sources using calculations based on geometries of vector maps representing countries.

7.3.2 Verification using Encyclopaedias

Traditional encyclopaedias rely on domain experts to generate content and use formal methods to validate it before allowing public access to the information. In contrast, Wikipedia uses the power of online user's community. It allows them to edit the existing entries and therefore depends on its contributors which are sometimes not experts. Several studies in the past have shown, the use of encyclopaedia Britannica and Brockhaus, as examples of an accurate reference, see [6], that Wikipedia is not much worse than a good reliable encyclopaedia provided (!) an entry has been visited by many persons and has been around without change for an extended period. We use Britannica [7] and Brockhaus [8] for verification of facts.

7.3.3 Verification using other online Geographic Sources

We also decided to go one step further and look at other data sources discussed in Section 7.1. We want to ensure the reliability by comparing facts with other online geographic servers/portals. Therefore we choose infoplease.com which is a very reliable web server. The code to extract area of different countries from Infoplease[15] is shown in Listing 7.5. In Infoplease, the area facts are stored in HTML pages as shown in the Figure 7.10. These facts are embedded in <p> tag with class area,

[15] http://www.infoplease.com/country/

therefore we are capturing those <p> elements which contain area of country using XPath expression shown at line 5 of Listing 7.5.

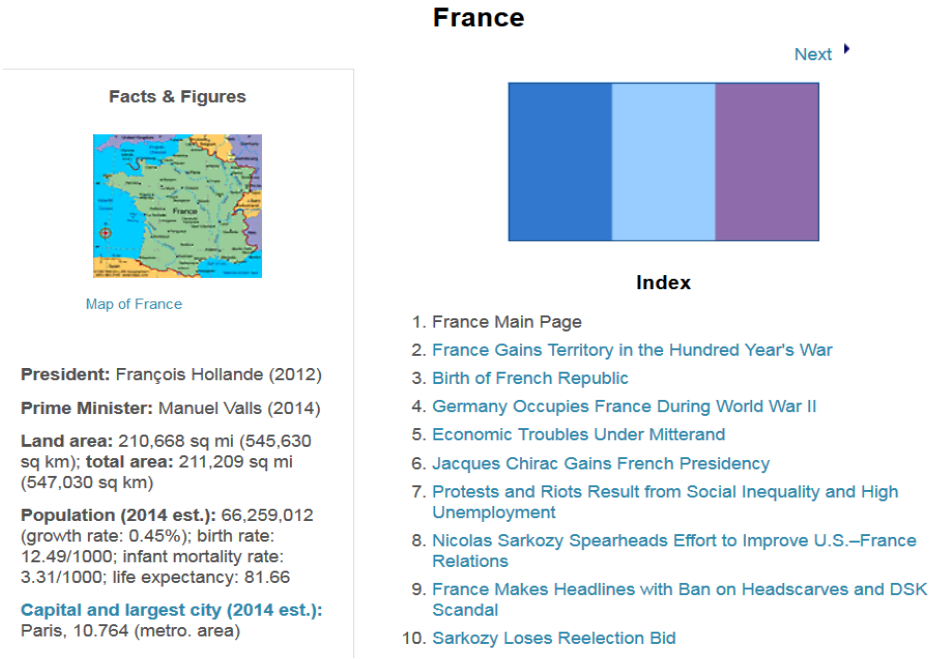


Figure 7.10: Web page of France in Infoplease[16].

```
1: URI="http://www.infoplease.com/country/"+country+".html";
2: final WebClient client = new WebClient();
3: final HtmlPage mainPage = client.getPage(URI);
4: links1 = (List<?>) mainPage.getByXPath
5: ("//p[@class='area']/text()");
```

Listing 7.5: Java code to extract area in sq km from Infoplease

7.4 System Architecture

Figure 7.11 summarizes the overall process of data collection and verification. We extract facts from a number of data sources (a sample is shown). After extracting facts we verify them using Britannica, WolframAlpha etc. Moreover we compare facts with existing web portals such as infoplease and word atlas for additional check.

Our geographic server has a static collection of facts in the form of wiki pages which is based on JSP wiki syntax, defined in [9]. Therefore we generate them automatically using our verified and reliable collection of facts. Users, domain experts can interact with our geographic portal and perform different activities; for example users can explore a country and find information related to different aspects such as geography, economy, Nobel Prize winners, parks etc. Domain experts and editors can give their comments and update the existing wiki pages after log in. Interactive data visualisations are also provided for meaningful understanding of geography. They will be covered in Chapter 11.

[16] http://www.infoplease.com/country/france.html

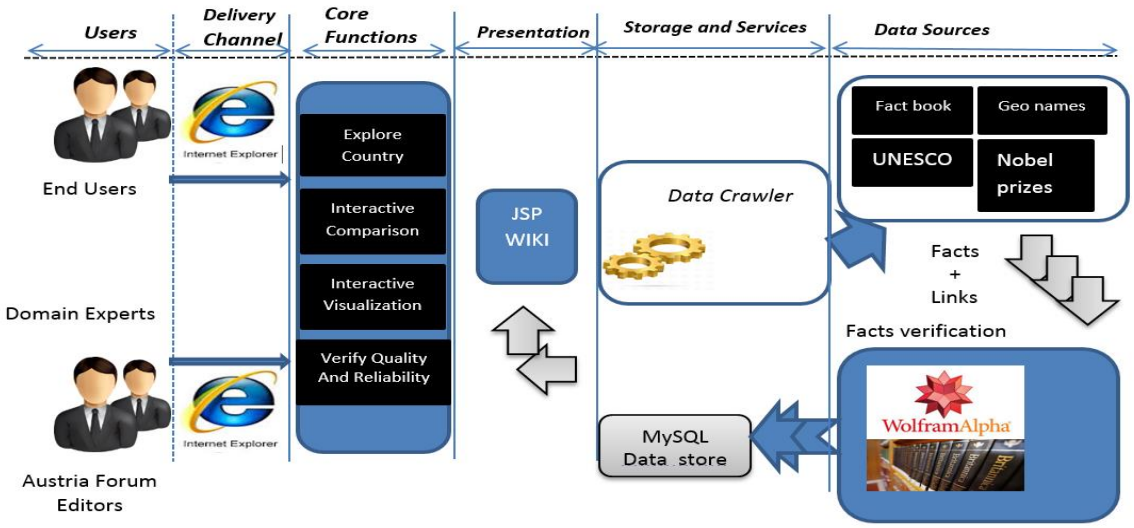


Figure 7.11: Overall System Architecture

7.5 City Verification

This section presents the summarized results of the verification process for 193 UN countries. The continent-wise city ranking results based on population, using two data sources (WolframAlpha, GeoNames) are shown in Table 7.2.

Table 7.2: Results of City Verification

Continents	Verified	Partially Verified	Verified+ Explanation	Not Verified
Europe	17	2	3	21
Africa	18	2	2	32
Australia	2	0	0	11
America	8	7	0	20
Asia	20	1	5	22
Total	65	12	10	106

For this particular study we are presenting top 5 city rankings. City rankings are found correct in case of 65 out of 193 countries (UN member states); whereas in case of 106 countries, city rankings did not match. We classified our verification results as:

i) Verified
ii) Partially Verified
iii) Verified + Explanation
iv) Not Verified.

Verified rankings are rankings of those countries where two lists representing cities and their ranks exactly match. In the area Australia we have found two countries “Australia” and “Vanuatu” where city rankings perfectly match as shown in the Table 7.3 and 7.4. Whereas in case of 11 countries, rankings did not match.

Table 7.3: City lists of Australia

Rank	Wolfram	GeoNames
1	Sydney	Sydney
2	Melbourne	Melbourne
3	Brisbane	Brisbane
4	Perth	Perth
5	Adelaide	Adelaide

Table 7.4: City lists of Vanuatu

Rank	Wolfram	GeoNames
1	Port-Vila	Port-Vila
2	Luganville	Luganville
3	Norsup	Norsup
4	Port Olry	Port Olry
5	Isangel	Isangel

Partially Verified rankings are rankings of those countries where two lists partially match. For instance, the city list of Sao Tome and Principe shown below is representing “Partially Verified” case. The Sao Tome (capital city) comes on top in both data sources.

Wolfram: Sao Tome; Santo Amaro; Santana; Neves; Trindade
GeoNames: Sao Tome; Santo Amaro

Verified+Explanation category points to verified rankings but they require some explanation. For instance let us look at the list of large 5 cities of Moldova in Wolfram and GeoNames.

Wolfram: Chisinau; Tiraspol; BalTi; Tighina; Ribnita
GeoNames: Chisina; Tiraspol; Balti; Bender; Rîbnita

From the above two lists the 4th entry is showing two different city names; the Wikipedia page says: Tighina and Bender are same city name and Tighina is also known as Bender. Similarly, below is the list of Tajikistan and represents Verified + Explanation case. The city ranked 5th (Istaravshan) is called (Uroteppa) in Tajik.

Wolfram: Dushanbe; Khujand; Kulob; Qurgonteppa; Uroteppa
GeoNames: Dushanbe; Khujand; Kulob; Qurgonteppa; Istaravshan

The problem of different names for the same entity of course has an impact on verification results (e.g. Mounteverest and Chomolungma refer to the same mountain). Therefore we are investigating these cases more carefully in our server.

For example the lists of Nepal are shown below. The official name of Patan is Lalitpur.

Wolfram: Kathmandu; Pokhara;Lalitpur; Biratnagar; Birganj
GeoNames: Kathmandu; Pokhara;Patan; Biratnagar; Birganj

The “Not Verified” represents those rankings which are different in two data sources. Let us look at a concrete case.

Listing 7.6 is used to generate Table 7.5 which presents a Not Verified case. Brasilia is ranked 4th by WolframAlpha whereas Fortaleza is ranked 4th in GeoNames. According to Wikipedia Belo Horizonte is the 6th largest city in Brazil, but it is ranked 5th in GeoNames (We find it surprising: One city, three different rankings in three different data sources). Similarly if we look at city rankings of Bosnia in Table 7.6, we find that Banja Luka is ranked 2nd in GeoNames whereas Zenika is ranked 2nd in Wolfram. According to Wikipedia Zenica is the 4th largest city in Bosnia and Herzegovina. Likewise, Table 7.7 shows the difference of rankings in city lists of Venezuela.

Much also depends on whether town, village, atoll, capital city is excluded or included in largest city list e.g. Vaiaku is a village that is included in city list of Tuvalu (a country associated with Australia) in Wolfram.

```
1: select w.city as Wolfram_city, g.city as Geonames_city, g.rank as Rank
2: from Wolfram w join isocodes iso on
3: iso.country=w.country join
4: Geonames g on g.country_code=iso.iso
5: and g.rank=w.rank where iso.iso=BR
```

Listing 7.6: Using joins to get city ranks from both tables Wolfram and GeoNames based on ISO codes.

Table 7.5: City lists of Brazil

Rank	Wolfram	GeoNames
1	Sao Paulo	Sao Paulo
2	Rio de Janeiro	Rio de Janeiro
3	Salvador	Salvador
4	Brasilia	Fortaleza
5	Fortaleza	Belo Horizonte

Table 7.6: City lists of Bonsnia

Rank	Wolfram	GeoNames
1	Sarajevo;	Sarajevo
2	Zenica	Banja Luka
3	Banja Luka	Zenica
4	Tuzla	Tuzla
5	Mostar	Mostar

Table 7.7: City lists of Venezuela

Rank	Wolfram	GeoNames
1	Caracas	Caracas
2	Maracaibo	Maracaibo
3	Valencia	Maracay
4	Baraquisimeto	Valencia
5	Ciudad Guayana	Baraquisimeto

7.6 Proper Message Handling in Geographic Server.

This sections presents the web interface for displaying meaningful messages in our geographic server in case of verified and not verified cases. We try to make it easy for domain experts to participate and give their opinion in case of contradictory facts. It also helps users and people to better understand verified and contradictory facts.

7.6.1 Verified Case

For instance Figure 7.12 and 7.13 show a Verified case; when information is found similar in multiple databases, for example, all data sources agree about the area of Albania. Also, the city list of Equatorial Guinea is found identical in the data sources considered.

Albania: Geography

Location	Southeastern Europe, bordering the Adriatic Sea and Ionian Sea, between Greece in the south and Montenegro and Kosovo to the north
Geographic Coordinates	41 00 N, 20 00 E
Area	***total:*** 28,748 sq km ***land:*** 27,398 sq km ***water:*** 1,350 sq km [***Verified in 8 databases***]
Land boundaries	***total:*** 717 km ***border countries:*** Greece 282 km, Macedonia 151 km, Montenegro 172 km, Kosovo 112 km
Coastline	362 km
Elevation Extremes	highest point: Maja e Korabit (Golem Korab) 2,764 m lowest point: Adriatic Sea 0 m

Figure 7.12: Verified Case

Large Cities	Bata ; Malabo ; Ebebiyin ; Aconibe ; Anisoc **[List Verified]**

Figure 7.13: Large cities of Equatorial Guinea

7.6.2 Not Verified Case

Figure 7.14 shows a “Not Verified Case” when information is not found similar in multiple databases. For example France has different area figures in multiple databases; therefore we are listing all the sources and also provide an attempted explanation for the discrepancies.

Area	***Total:*** 643,801 sq km; 551,500 sq km (metropolitan France) ***land:*** 640,427 sq km; 549,970 sq km (metropolitan France) ***water:*** 3,374 sq km; 1,530 sq km (metropolitan France) Note: The first numbers include the overseas regions of French Guiana, Guadeloupe, Martinique, Mayotte, and Reunion ***We are unable to verify those figures from Factbook, since the 6 main sources we used for checking give different figures as follows:*** ***Factbook: 643801*** ***DBpedia: 674843*** ***Geoname: 547030*** ***Infoplease: 547030*** ***Britannica: 543965*** ***Wolfram: 551500*** **Attempted Explanation**: For France, Britannica gives the smallest area. This agrees exactly with the area in the largest German encyclopaedia Brockhaus (2014): It follows the French Land register data that excludes lakes, ponds and glaciers larger than 1 km² and the estuaries of rivers. (This is, by the way, very much in contrast to how the figures are arrived e.g. for Finland that not only includes freshwater lakes like many databases but alos ocean channels!). Including bodies of water the French National Geographic Institute arrives at the figure shown by Wolfram. Those figures do include the area of the island of Corsica but do not include overseas departments and oversea territories. If one counts them in, a figure higher than 640.000 is obtained. None of the figures include the 320.000 km² of Antarctica where sovereignty has been suspended since the signing of the Antarctic Treaty in 1959.

Figure 7.14: Not Verified Case

7.6.3 Verified + Explanation Case

Figure 7.15 shows Verified + Explanation case. The city list of Moldova is verified, since Tighina and Bender are same city names. Similarly Figure 7.16 shows that lists of Nepal are verified and explanation is also given about names (Patan and Lalitpur).

Moldova: Geography

Large Cities	Chisinau ; Tiraspol ; BalTi ; Tighina ; RibniTa Chişinău; Tiraspol; Bălţi; Bender; Rîbniţa **[List Verified]** **Explanation:** Tighina and Bender are same city name

Figure 7.15: Verified + Explanation Case

Large Cities	**Kathmandu ; Pokhara ; Lalitpur ; Biratnagar ; Birganj** **Kathmandu; Pokhara; Pātan; Birātnagar; Bīrganj** **[List Verified]** **Explanation:** The official name of Patan is Lalitpur

Figure 7.16: Verified + Explanation Case

7.6.4 Not Verified Case (City Rankings)

Figure 7.17 shows a Not Verified case of the city list of Venezuela. The list is not verified since Valencia is ranked at 3rd position according to Wolfram where as in GeoNames, it is ranked 4th, therefore we are listing both lists.

Large Cities	**Due to difference in city rankings taken from two data sources we are listing here both lists :** According to Wolfram: Caracas ; Maracaibo ; Valencia ; Barquisimeto ; Ciudad Guayana According to Geonames: Caracas; Maracaibo; Maracay; Valencia; Barquisimeto **Attempted Explanation:** Please help us to try to explain the discrepancies by sending us helpful information to office@austria-forum.org
Geography-note	on major sea and air routes linking North and South America; Angel Falls in the Guiana Highlands is the world's highest waterfall

Figure 7.17: Not Verified Case

Reference

[1] R. Mehmood and P. Korica-pehserl, “The Web is Big Business,” J. Comput. Inf. Technol., vol. 23, no. 1, pp. 19–27, 2015.

[2] “SPARQL Query Language for RDF.” [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/. [Accessed: 21-Oct-2015].

[3] “FactbookXML - All textual information of the CIA World Factbook in one XML file.” [Online]. Available: http://jmatchparser.sourceforge.net/factbook/. [Accessed: 21-Oct-2015].

[4] “Wolfram Alpha: Computational Knowledge Engine.” [Online]. Available: http://www.wolframalpha.com/. [Accessed: 21-Oct-2015].

[5] “HtmlUnit – Welcome to HtmlUnit.” [Online]. Available: http://htmlunit.sourceforge.net/. [Accessed: 21-Oct-2015].

[6] G. De la Calzada and A. Dekhtyar, “On measuring the quality of Wikipedia articles,” Proc. 4th Work. Inf. Credibil. - WICOW ’10, p. 11, 2010.

[7] “Britannica.com.” [Online]. Available: http://www.britannica.com/. [Accessed: 21-Oct-2015].

[8] “BROCKHAUS.” [Online]. Available: http://www.brockhaus.de/. [Accessed: 21-Oct-2015].

[9] “Apache JSPWiki.” [Online]. Available: https://jspwiki.apache.org/. [Accessed: 21-Oct-2015].

Chapter 8 : Geographic Server

Geography of the world is a fairly unique project where data from multiple data sources (as mentioned in Chapter 7) are merged together. The major emphasis of this project is to provide reliable information to people, going beyond numerical facts. Thus, the main directions of the research involved are twofold. One is to collect data from different sources that relates to a country, but also to assure correctness as much as possible by comparing data from different geographic data bases and geographic search engines. And two, much information beyond numeric facts is provided. In this chapter we will explore some of these features. We will concentrate in this chapter on the design of Web pages that allow to intuitively easily access various types of information.

This chapter will address the following research questions.

RQ1. What type of data can we extract and store for each country?

RQ2. How can we classify and arrange countries and territories?

RQ3. How can we represent each country and territory in our Geographic Server?

RQ4. How can we generate and enhance the culture section of each country? (E.g. by using facts and hyperlinks that allow users to visit reliable websites from which facts are taken)

RQ5. How can we link different subsections of each country with the main entry page? (Using suitable icons and internal wiki links)

RQ6. What is the significance of the picture section of global-geography.org compared to pictures on other geographic server?

8.1. Motivational Factor

We want to give some background and motivation for the special type of Geographic Server in this section, as discussed by Professor Maurer (The Editor-in-Chief of Austria-Forum) in his book [1].

“As a member of the Executive Board of the Academia Europaea I had contacts with prominent researchers from all disciplines in Europe. One such person was Anne Buttimer, professor of geography in Dublin, Ireland, She received the highest prize that the International Geographical Association awards in 2014.

In a demonstration of the Austria-Forum she was quite impressed but regretted that an international component was missing. She suggested to extend the Austria-Forum by an appropriate geography component through which global cooperation could be possible.

In fact, there are a number of geography servers, so to build just another from scratch seemed pointless. Buttimer suggested to use material from different servers but where possible to emphasize particularly those aspects that are hardly found in other geography servers: information on the culture of countries, specific stories that are characteristic of countries and image collections with detailed descriptions missing in most image collections.

Buttimer agreed to establish contacts with other institutions that would join the project once it would be a separate unit with data and English interface. Thus, as part of the Austria-Forum is a category

"Geography of the World"[2], which is accessed via URL global-geography.org was established and is scheduled for release in April 2016. The theme of geography seems good at a time when many are interested in remote countries, yet travel is restricted due to political reasons. It is hoped that many people will find it useful, even for English lessons.

8.2 Types of Data

In this section we will look at different types of facts that we are offering to users via global-geography. Users of the system can find various kinds of information about the countries. Let us look at some of the facts related to Geography and Government as shown in the Figure 8.1 a and 8.1 b respectively. For Geographic data we have geographic coordinates, area, land boundaries, coastlines etc. about all countries of the world. Government data is about country name, capital, administrative divisions etc.

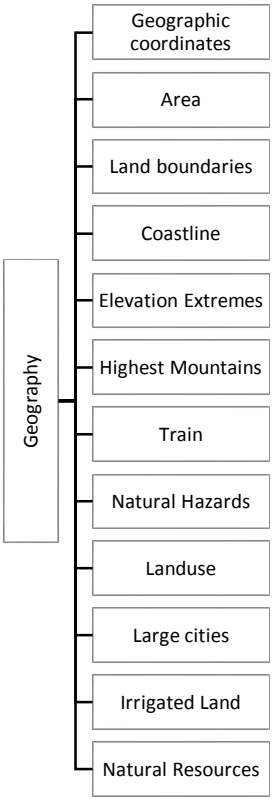


Figure 8.1 a

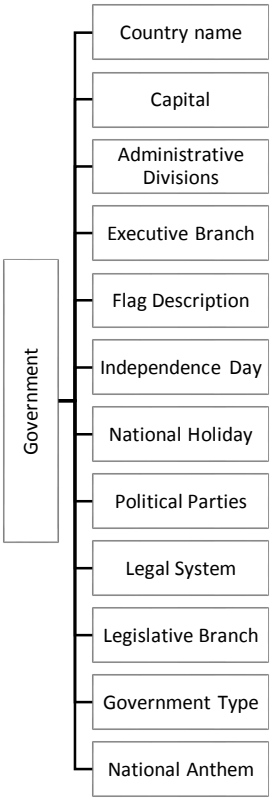


Figure 8.1 b

Figure 8.1: Types of data related to Geography and Government in Geographic Server

Under the tab “People and Society” Users can find information related to population, population growth rate, birth rate etc. The economic data is about GDP, exports, imports, industrial production growth rate etc. For full details see Figure 8.2. Figure 8.3 shows Transportation and Communication data.

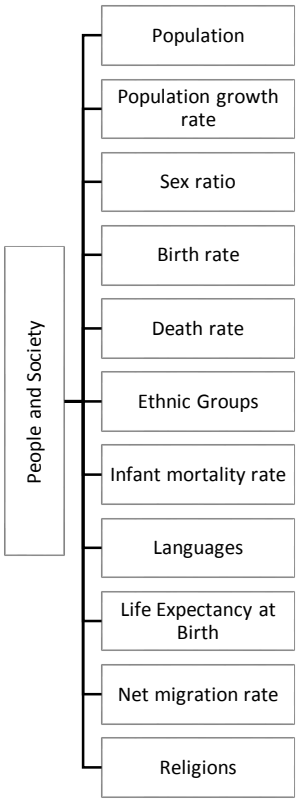


Figure 8.2 a

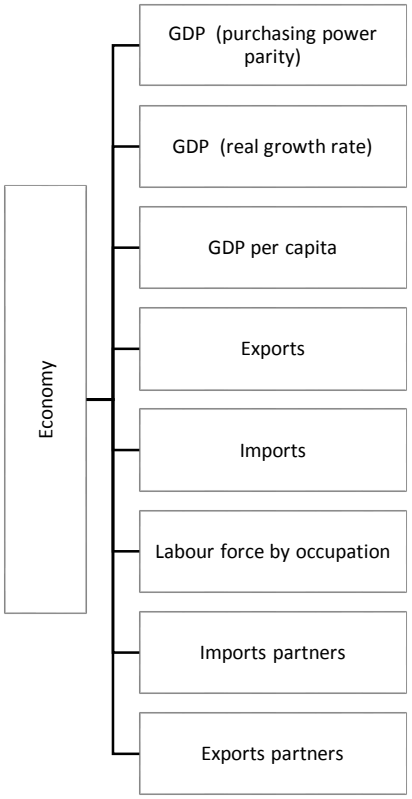
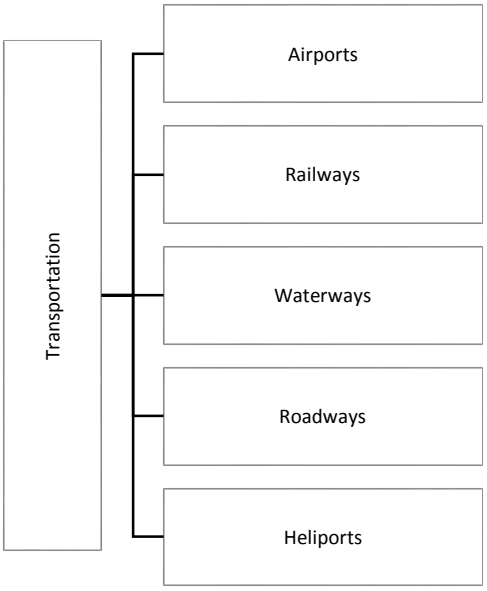


Figure 8.2 b

Figure 8.2: Type of Data related to "People and Society" and Economy in Geographic Server



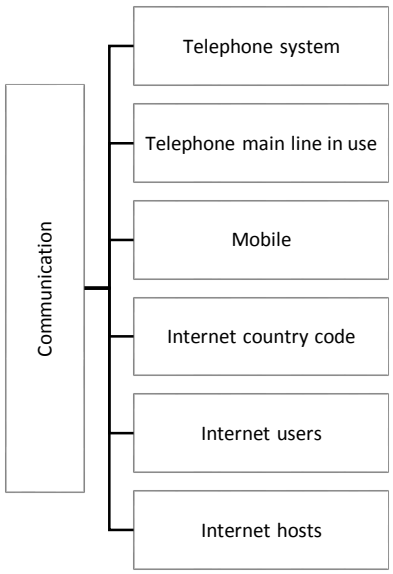


Figure 8.3: Facts related to "Transportation" and Communication in Geographic Server

We finish this section by showing the energy facts in Figure 8.4. It covers Electricity, Crude oil, Natural gas, Petroleum products, Carbon dioxide emission etc.

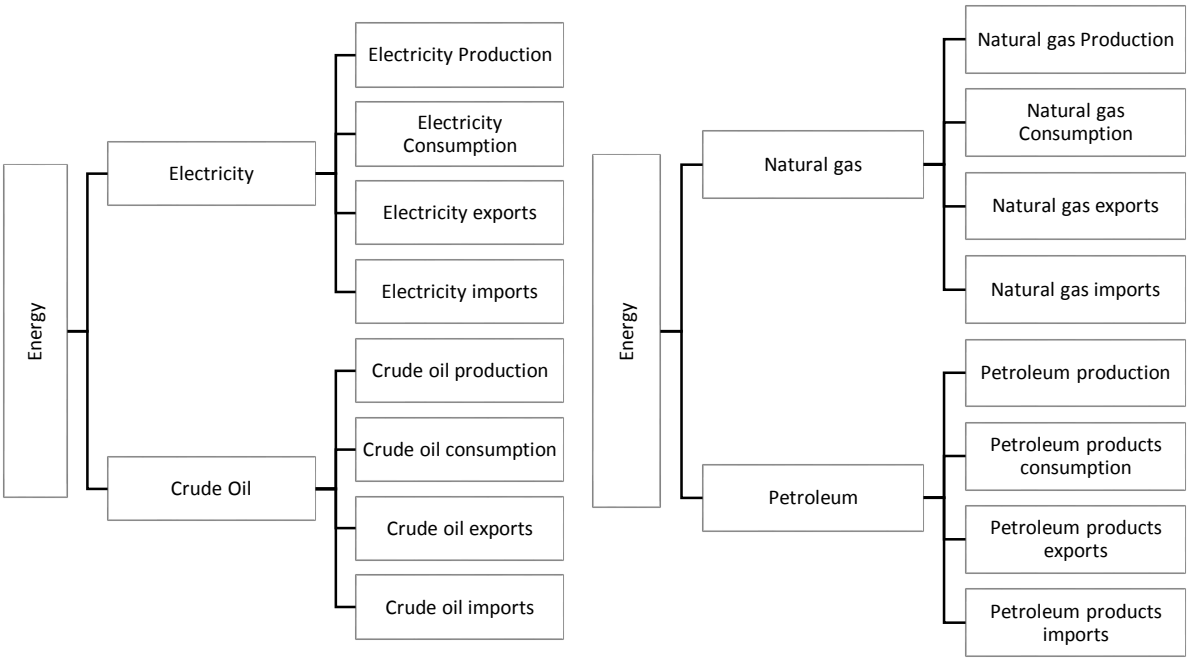


Figure 8.4: Facts related to “Energy”

8.3 Main Interface of Global-Geography

The main interface of geographic server is shown in the Figure 8.5. Users can navigate through different services of Geographic server using this interface. We will focus on the part “Find. Information on any country you are interested in“ in this section.

Figure 8.5: Main interface of global-geography.org

8.3.1 Index Page of Countries and Territories

We have divided the index page into two sections. One represents UN member states. In this category we are listing those countries that are recognised by UN. We make use of a classification by countries within continents even if this does create some problems: Is e.g. Armenia in Europe or Asia? Or how about Russia? Does Greenland belong to Europe or America? There are 43 countries that are listed in Europe, that are recognised by UN as shown in the Figure 8.6. We place Russia rather arbitrarily into Asia in our categorization.

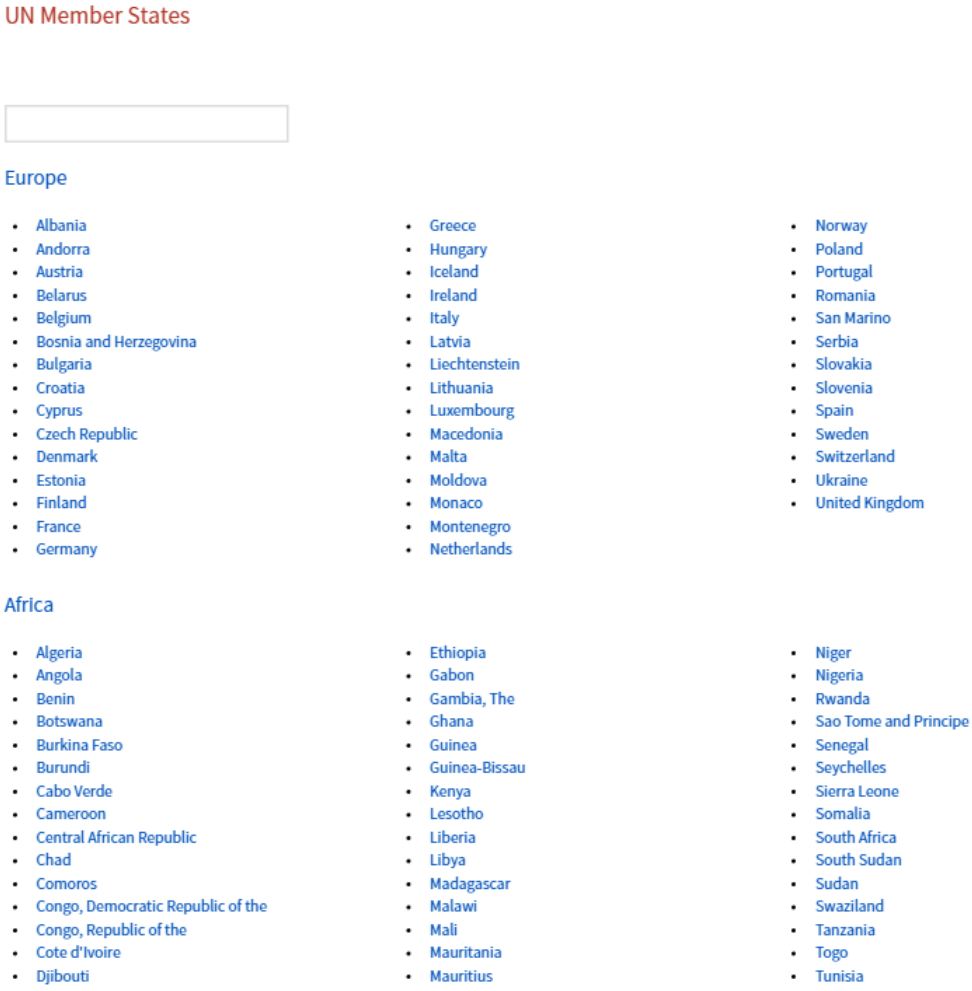


Figure 8.6: UN member states of Africa and Europe

8.3.2 Index Page of Territories, Oceans and Islands

The territories oceans and islands that are not recognised by UN are listed under the heading “Territories Oceans and Islands” as shown in the Figure 8.7. For instance Akrotiri and Western Sahara are listed in this category as they are not recognised by UN. We are presenting facts about Oceans of the world. The Geography Web page of Arctic Ocean is shown in the Figure 8.8, showing different geographic facts about Arctic Ocean.

Territories, Oceans and Islands

- Akrotiri
- American Samoa
- Anguilla
- Antarctica
- Arctic Ocean
- Aruba
- Ashmore and Cartier Islands
- Atlantic Ocean
- Bermuda
- Bouvet Island
- British Indian Ocean Territory
- British Virgin Islands
- Cayman Islands
- Christmas Island
- Clipperton Island
- Cocos (Keeling) Islands
- Cook Islands
- Coral Sea Islands
- Curacao
- Dhekelia
- European Union
- Falkland Islands (Islas Malvinas)
- Faroe Islands
- French Polynesia
- French Southern and Antarctic Lands
- Gaza Strip
- Gibraltar
- Greenland
- Guam
- Guernsey
- Heard Island and McDonald Islands
- Holy See (Vatican City)
- Hong Kong
- Indian Ocean
- Isle of Man
- Jan Mayen
- Jersey
- Kosovo
- Macau
- Montserrat
- Navassa Island
- New Caledonia
- Niue
- Norfolk Island
- Northern Mariana Islands
- Pacific Ocean
- Paracel Islands
- Pitcairn Islands
- Puerto Rico
- Saint Barthelemy
- Saint Helena, Ascension, and Tristan da Cunha
- Saint Martin
- Saint Pierre and Miquelon
- Sint Maarten
- South Georgia and South Sandwich Islands
- Southern Ocean
- Spratly Islands
- Svalbard
- Taiwan
- Tokelau
- Turks and Caicos Islands
- United States Pacific Island Wildlife Refuges
- Virgin Islands
- Wake Island
- Wallis and Futuna
- West Bank
- Western Sahara

Figure 8.7: Territories Oceans and Islands

Arctic Ocean: Geography

Location	body of water between Europe, Asia, and North America, mostly north of the Arctic Circle
Geographic Coordinates	90 00 N, 0 00 E
Area	*total:* 14.056 million sq km *note: includes Baffin Bay, Barents Sea, Beaufort Sea, Chukchi Sea, East Siberian Sea, Greenland Sea, Hudson Bay, Hudson Strait, Kara Sea, Laptev Sea, Northwest Passage, and other tributary water bodies*
Land boundaries	
Coastline	45,389 km
Highest Mountains	
Terrain	central surface covered by a perennial drifting polar icepack that, on average, is about 3 meters thick, although pressure ridges may be three times that thickness; clockwise drift pattern in the Beaufort Gyral Stream, but nearly straight-line movement from the New Siberian Islands (Russia) to Denmark Strait (between Greenland and Iceland); the icepack is surrounded by open seas during the summer, but more than doubles in size during the winter and extends to the encircling landmasses; the ocean floor is about 50% continental shelf (highest percentage of any ocean) with the remainder a central basin interrupted by three submarine ridges (Alpha Cordillera, Nansen Cordillera, and Lomonosov Ridge)
Natural Hazards	ice islands occasionally break away from northern Ellesmere Island; icebergs calved from glaciers in western Greenland and extreme northeastern Canada; permafrost in islands; virtually ice locked from October to June; ships subject to superstructure icing from October to May
Natural Resource	sand and gravel aggregates, placer deposits, polymetallic nodules, oil and gas fields, fish, marine mammals (seals and whales)
Land Use	
Climate	polar climate characterized by persistent cold and relatively narrow annual temperature ranges; winters characterized by continuous darkness, cold and stable weather conditions, and clear skies; summers characterized by continuous daylight, damp and foggy weather, and weak cyclones with rain or snow
Irrigated Land	
Renewable Water Resources	
Environment_CurrentIssues	endangered marine species include walruses and whales; fragile ecosystem slow to change and slow to recover from disruptions or damage; thinning polar icepack
Environment - international agreements	

Figure 8.8: Geography section of Arctic Ocean

8.3.3 Automatic Search

In addition to navigation through menus discussed earlier, users can type in desired countries in the search text box as shown in the Figure 8.9 either in English or German. (This might be extended by other language version if demanded).

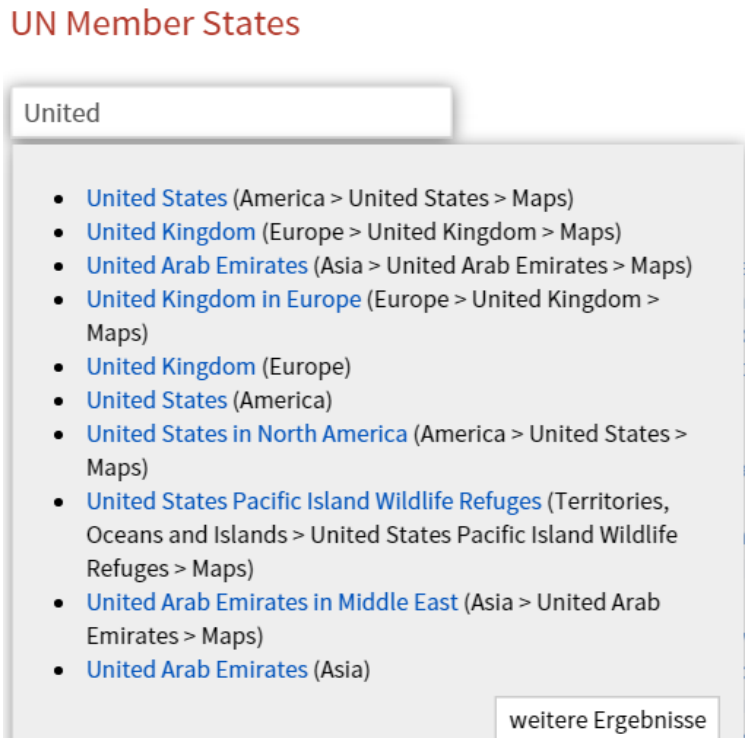


Figure 8.9: Search facility to find a particular country

8.4 Main Entry Page of Each Country

The main entry page of Austria is shown in Figure 8.10. Ranging from geography, government, culture, maps, up to special stories about countries. Users will hopefully find much of what they want to know about a particular country. We take care of standard features associated with every country. For instance, both flag and emblems of each country are displayed on the respective entry page of that country. Meaningful icons[17] supported by English phrases lead to sub-pages where users can find corresponding information.

[17] Katharina Ziegler who is graphic designer in Austria-Forum created these icons

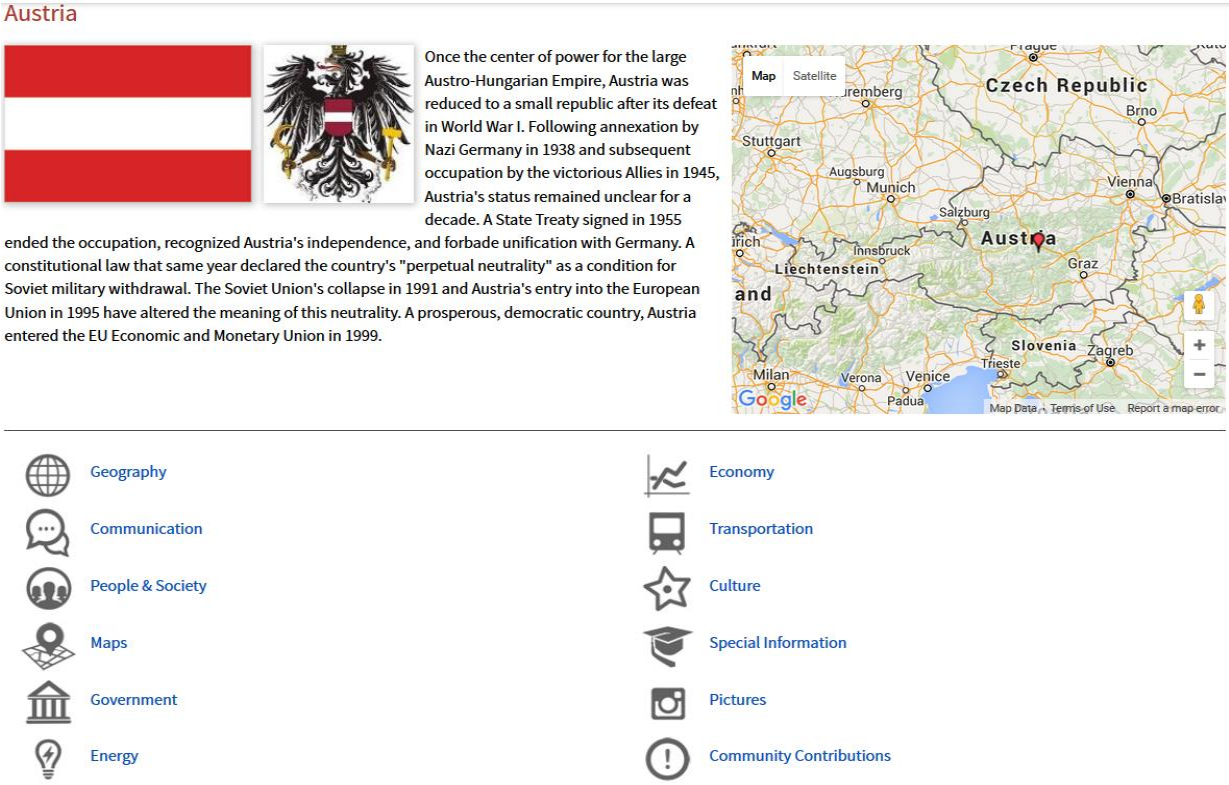


Figure 8.10: Main page of Austria. Icons lead to sub-pages.

8.5 Sub Pages of Each Country

In this section we will explain features of subpages that relate to different aspects of a particular country.

8.5.1 Geography

As an example, geographic features of Albania are shown in the Figure 8.11. This starts by defining the location of country. The important geographic facts such as area, elevation extremes, and climatic conditions are listed afterwards. The flag symbol shown in the Figure 8.11 in front of "Geographic Coordinates" and elevation extremes leads to Google map where users can see the location of place on the map.

Albania: Geography

Location	Southeastern Europe, bordering the Adriatic Sea and Ionian Sea, between Greece in the south and Montenegro and Kosovo to the north
Geographic Coordinates	41 00 N, 20 00 E
Area	***total:*** 28,748 sq km ***land:*** 27,398 sq km ***water:*** 1,350 sq km [***Verified in 8 databases***]
Land boundaries	***total:*** 717 km ***border countries:*** Greece 282 km, Macedonia 151 km, Montenegro 172 km, Kosovo 112 km
Coastline	362 km
Elevation Extremes	highest point: Maja e Korabit (Golem Korab) 2,764 m lowest point: Adriatic Sea 0 m
Highest Mountains	**Due to difference in mountain rankings taken from two data sources we are listing here both lists :** According to Wolfram: Golemi Korab 2764 m ; Maja Jezercë 2694 m ; Gjalices 2537 m According to Geonames:Maja e Jezercës 2694 m ; Grámos Óros 2523 m ; Mali i Mesit 2480 m ; Mali i Gramës 2345 m ; Pllaja e Pusit 2288 m **Attempted Explanation:** Please help us to try to explain the discrepancies by sending us helpful information to office@austria-forum.org

Figure 8.11: Geographic Section of Albania

8.5.2 Culture

The culture section of Austria is shown in Figure 8.12. Concerning culture, we have used information from the UNESCO heritage site [3], information on National Parks from various sites including National Geographic [4]. For famous person, we have used various sites of prestigious awards, like Nobel Prize [5], Wolff Prize[6].

World Heritage

- Historic Centre of the City of Salzburg
- Palace and Gardens of Schönbrunn
- Hallstatt-Dachstein / Salzkammergut Cultural Landscape
- Semmering Railway
- City of Graz – Historic Centre and Schloss Eggenberg
- Wachau Cultural Landscape
- Fertö / Neusiedlersee Cultural Landscape
- Historic Centre of Vienna
- Prehistoric Pile dwellings around the Alps

Nobel Prize Winners (19)

Here persons are counted for the country they born in or if their affiliation is with this country.

Name	Category	Year	Motivation
Robert Bárány	Medicine	1914	"for his work on the physiology and pathology of the vestibular apparatus"
Alfred Hermann Fried	Peace	1911	
Victor Franz Hess	Physics	1936	"for his discovery of cosmic radiation"
Elfriede Jelinek	Literature	2004	"for her musical flow of voices and counter-voices in novels and plays that with extraordinary linguistic zeal reveal the absurdity of society's clichés and their subjugating power"
Eric R. Kandel	Medicine	2000	"for their discoveries concerning signal transduction in the nervous system"
Martin Karplus	Chemistry	2013	"for the development of multiscale models for complex chemical systems"
Walter Kohn	Chemistry	1998	"for his development of the density-functional theory"
Richard Kuhn	Chemistry	1938	"for his work on carotenoids and vitamins"
Karl Landsteiner	Medicine	1930	"for his discovery of human blood groups"
Otto Loewi	Medicine	1936	"for their discoveries relating to chemical transmission of nerve impulses"
Konrad Lorenz	Medicine	1973	"for their discoveries concerning organization and elicitation of individual and social behaviour patterns"
Wolfgang Pauli	Physics	1945	"for the discovery of the Exclusion Principle, also called the Pauli Principle"

Figure 8.12: Culture section of Austria

8.5.2.1 Ordering by Year

Nobel Prize winners are shown in tables. They can be arranged by year, category or by their name. For instance, in Figure 8.13 Nobel Prize Winners list of Pakistan in displayed in descending order, based on year. Malala Yousafzai who got Nobel Prize in 2014 is listed on top. Figure 8.14 shows another view of Noble Prize Winners list which is sorted in ascending order based on year. Dr. Abdul Salam who got Noble Prize in 1979 is listed on top in the list.

Pakistan: Culture

Nobel Prize Winner

Name	Category	Year	Motivation
Malala Yousafzai	Peace	2014	"for their struggle against the suppression of children and young people and for the right of all children to education"
Subramanyan Chandrasekhar	Physics	1983	"for his theoretical studies of the physical processes of importance to the structure and evolution of the stars"
Abdus Salam	Physics	1979	"for their contributions to the theory of the unified weak and electromagnetic interaction between elementary particles, including, inter alia, the prediction of the weak neutral current"

Figure 8.13: Nobel Prize Winners list in descending order based on year

Nobel Prize Winner

Name	Category	Year	Motivation
Abdus Salam	Physics	1979	"for their contributions to the theory of the unified weak and electromagnetic interaction between elementary particles, including, inter alia, the prediction of the weak neutral current"
Subramanyan Chandrasekhar	Physics	1983	"for his theoretical studies of the physical processes of importance to the structure and evolution of the stars"
Malala Yousafzai	Peace	2014	"for their struggle against the suppression of children and young people and for the right of all children to education"

Figure 8.14: Nobel Prize Winners list in ascending order based on year.

We always try to explain our choices, if they are not obvious. Let us take an example.

Consider Nobel Laureates: It is not clear how to associate them with a particular country due to the following reasons:

Is our criterion to list a person for a country:

- In the country where the person was born?
- Where the person obtained the award?
- Where the person did the work for which the award was given?
- Also, the country where e.g. a person was born may not exist anymore!

Overall, we consider the country of birth most important, but also try to list persons under the country they live in when the award was given. In case of country changes, we add them to potentially more than one country.

For example Mother Teresa was born in Ottoman Empire which is now Macedonia (a country in Europe). She was living in India when the Nobel Prize was given and she also died in India.

All this shows: Questions we often pose are ill-posed. There is no clear answer to what is the largest cave, or how many Nobel Laureates belong to a country etc. unless the question is formulated more specifically, as we have discussed in Chapter 5.

Figure 8.15 shows the Nobel Prize winners list of Macedonia.

Macedonia: Culture

World Heritage

- Natural and Cultural Heritage of the Ohrid region

Nobel Prize Winners (1)

Here persons are counted for the country they born in or if their affiliation is with this country.

Name	Category	Year	Motivation
Mother Teresa	Peace	1979	

Figure 8.15: Culture Section of Macedonia.

Figure 8.16 shows the Nobel Prize winners list of United States. In the culture section we are listing Nobel Prize winners for the country they born in or if their affiliation is with this country.

Nobel Prize Winners (362)

Here persons are counted for the country they born in or if their affiliation is with this country.

Name	Category	Year	Motivation
Alexei A. Abrikosov	Physics	2003	"for pioneering contributions to the theory of superconductors and superfluids"
Jane Addams	Peace	1931	
Peter Agre	Chemistry	2003	"for the discovery of water channels"

Figure 8.16: Three of 362 Nobel Prize winners of the United States. The message over the table is displayed to avoid any ambiguity.

We have also downloaded famous people from the website "The Famous People" [7]. A partial list of famous people of Sweden is shown in the Figure 8.17.

Partial List of Famous People

- Agnetha Faltskog
- Alfred Nobel
- Alva Myrdal
- Anders Celsius
- Anders Zorn
- August Strindberg
- Bertil Gotthard Ohlin
- Birgit Nilsson
- Carl Linnaeus
- Dag Hammarskjöld
- Greta Garbo
- Gunnar Myrdal
- Hugo Theorell
- Ingmar Bergman
- Ingrid Bergman
- Ingvar Kamprad
- Izabella Scorupco
- Johannes Rydberg
- Laila Freivalds
- Markus Persson
- Princess Madeleine, Duchess of Hälsingland and Gästrikland
- Svante Arrhenius
- Swedenborg
- Tomas Tranströmer

Figure 8.17: Partial List of Famous People of Sweden

When we extract facts for the culture section we always extract both facts and links. This makes it easy and gives freedom to users for navigation. By clicking on the facts they can easily visit the websites to learn more about the facts. These facts are stored in Austria-Forum as "External links". A few examples are shown below

```
[Swedenborg (Scientist, Philosopher, Christian Mystic and
Theologian)|http://www.thefamouspeople.com/profiles/swedenborg-85.php]

[Torsten N.
Wiesel|http://www.nobelprize.org/nobel_prizes/medicine/laureates/1981/wiese
l.html]|Medicine|1981|"for their discoveries concerning information
processing in the visual system"
```

8.5.3 Pictures

We have obtained many interesting pictures from the community and are particularly grateful for access to the picture archives of Gerhard Huber [8], Ewald Judt[14], Hasso Hohmann [9] and Jontes, Günther[13]. Some pictures of Greece (provided by Gerhard Huber) are shown in Figure 8.18. A set of photos of Iranian architecture [11] taken over 40 years ago by Hasso Hohmann are shown in the Figure 8.19. We have also used pictures from Factbook [12], satellite images from NASA, and some from Wikipedia under Creative Common license, or from Pixabay where applicable.

We believe that pictures can provide valuable information about a country, including its culture, folklore, people, and more. The Web is overflowing with zillions of pictures. However, we believe that most pictures one can find lack good descriptions of what they show and usually are rather recent. By way of example, a description “Sagrada Familia built by Gaudi” is not enough, if not more is mentioned about this building in Barcelona and its architect Gaudi. 700 pictures of Syria[10] taken in 2009 have now historic value, since many of the famous parts of the six UNESCO heritage sites have been virtually destroyed by a conflict that is still ongoing at the time of writing. But the photos not only show the sight still intact but provide ample textual explanation (even bilingual in this case) as shown in the Figure 8.20.

On the other hand, hundreds of pictures of various islands in the Philippines show the mixture of different ethnic groups and allow to better understand why frictions may occur and allow us to better understand actions of governments involved.

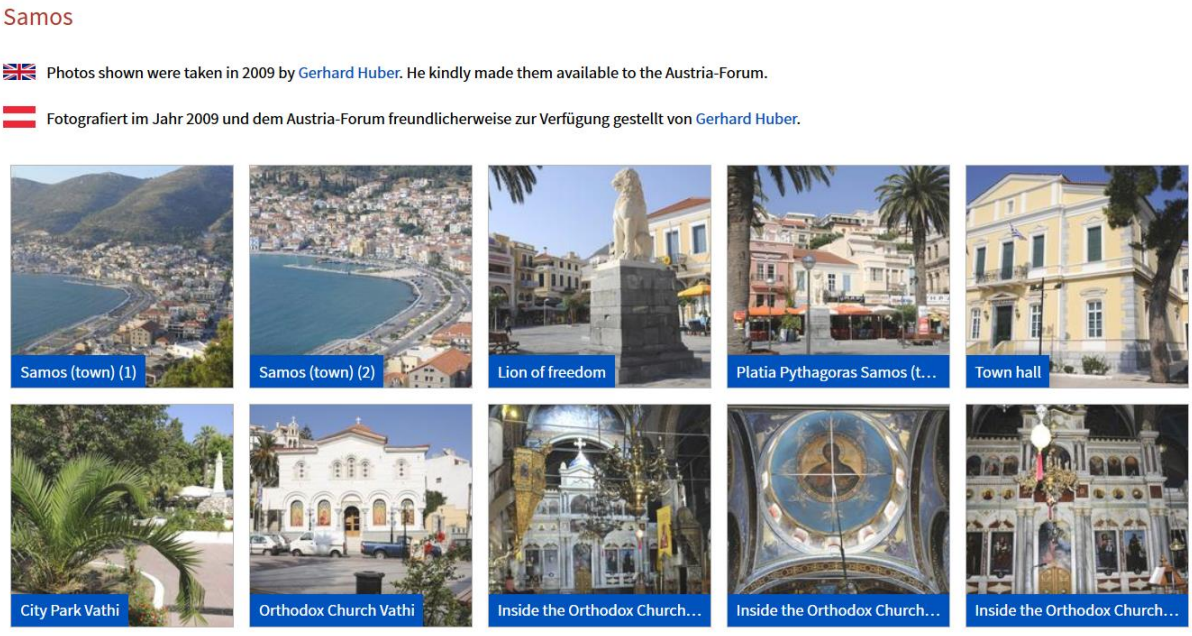


Figure 8.18: Pictures of the island Samos in Greece

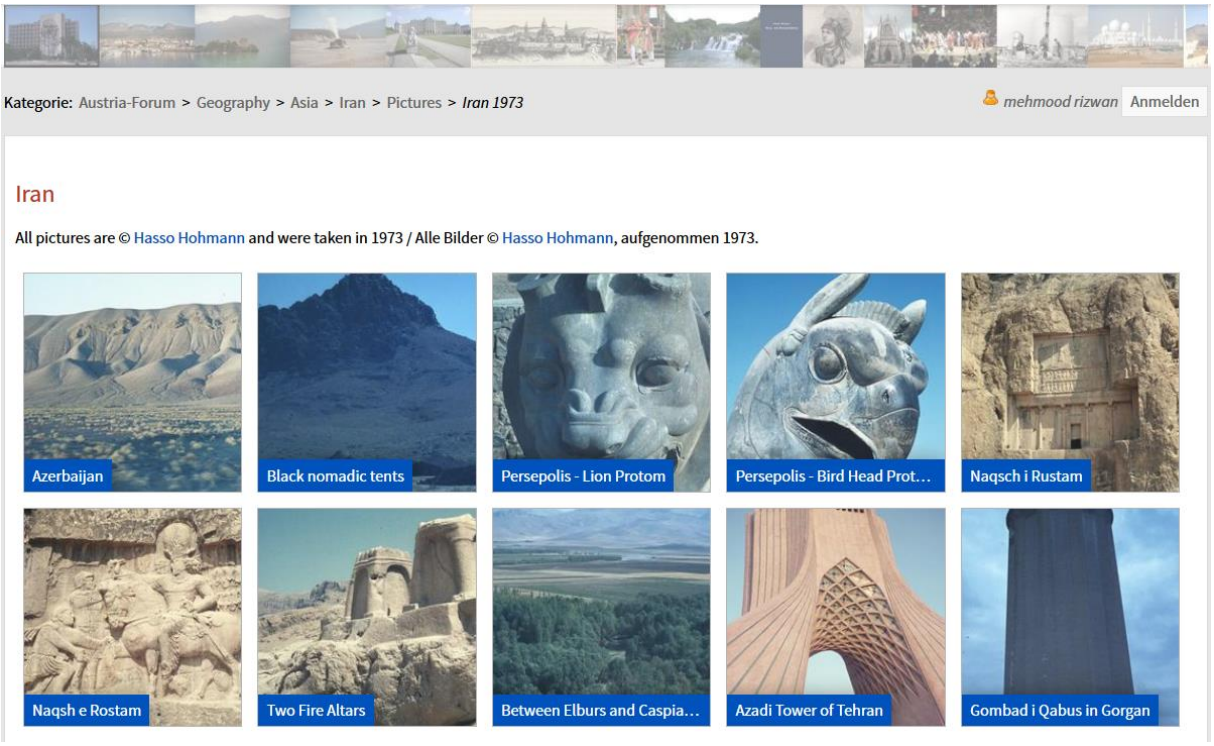


Figure 8.19: Picture section of Iran

Figure 8.20: Funerary Temple

Taiwan, United States, Austria have lots of pictures that are taken from different sources and partners of global-geography, see Taiwan and United States Picture section in Figure 8.21 and 8.22.

Figure 8.21: Picture Section of Taiwan

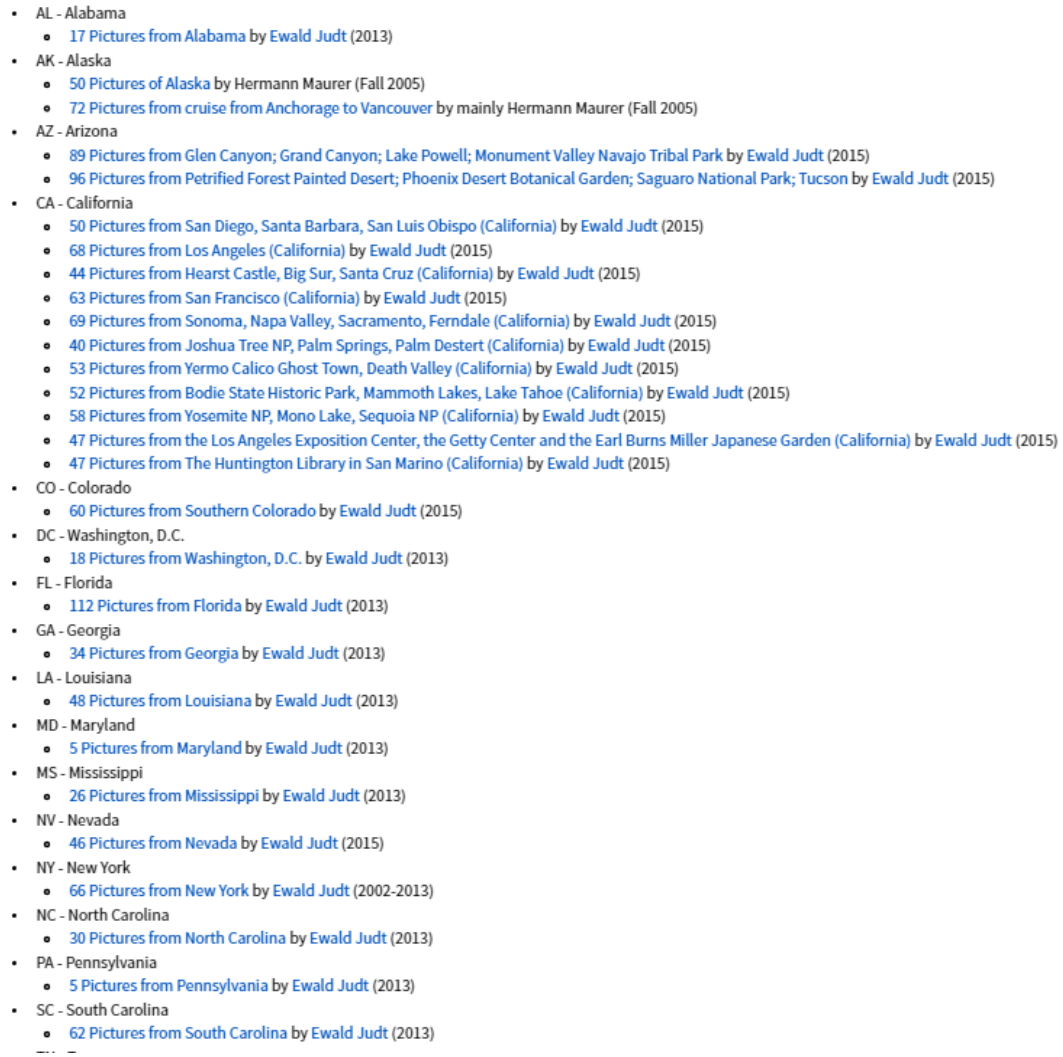
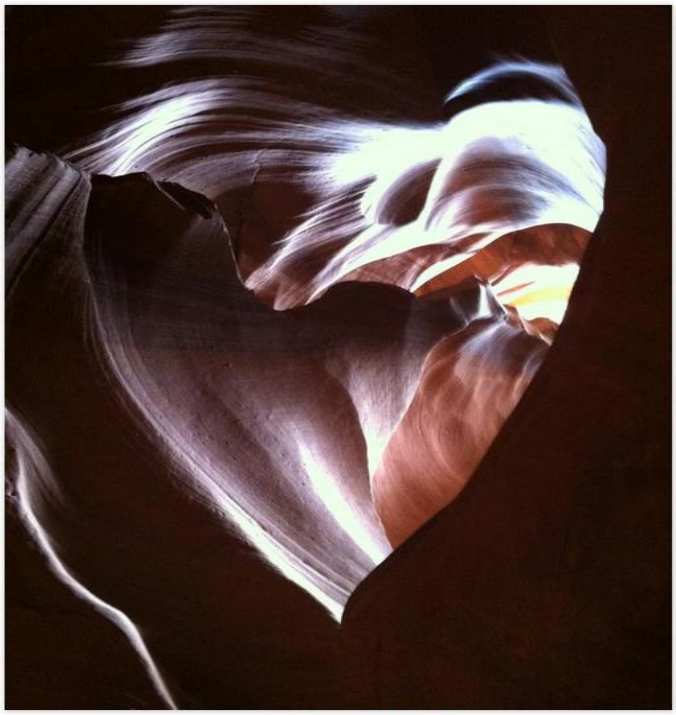


Figure 8.22: Picture Section of United States.

Let us look at the example of picture about “Antelope Canyon” taken from “Picture section” of United States from Arizona, part 1. The picture has a detailed explanation as shown in the Figure 8.23. A different view of canyon is shown in the Figure 8.24 where no description is given. In our server we give detailed description only for the first view of an object. The remaining views of the same object are shown without description, unless there is something special to be mentioned.

Antelope Canyon



Antelope Canyon is a narrow canyon ("slot canyon") on Navajo land east of Page , Arizona.

Antelope Canyon includes two separate picturesque sections, the Antelope Canyon as such and the Upper Antelope Canyon.

The canyons have been eroded mainly by water which in heavy rushes through the canyon. The sand carried by the water adds to erosive force.

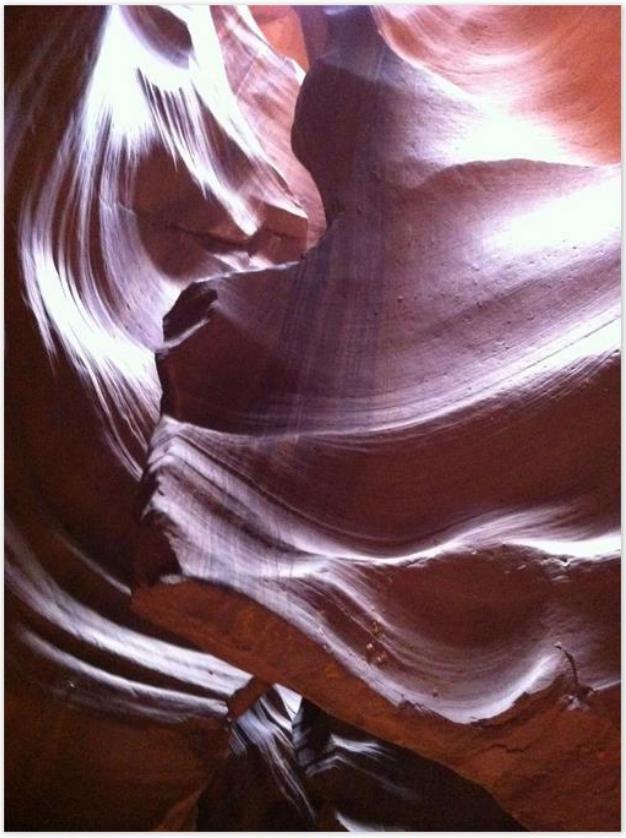
Like in some other canyons, flash floods can and have killed people. They can occur with no rain or clouds visible, rather caused by some thunderstorm even 15 miles upstream that dumps lots of water. Such flash floods can therefore come with no warning at all.

Antelope canyon can only be visited in guided tours.

Antelope Canyon, May 2015, © Ewald Judt

Figure 8.23: Picture of "Antelope Canyon" taken from pictures of United States

Antelope Canyon



Antelope Canyon, May 2015, © Ewald Judt

Figure 8.24: Picture of "Antelope Canyon" taken from pictures of United States

References

[1] “Web-Books im Austria-Forum.” [Online]. Available: http://austria-forum.org/web-books/austriaforum01de2015iicm. [Accessed: 02-Nov-2015].

[2] “Geography im Austria-Forum.” [Online]. Available: http://austria-forum.org/af/Geography. [Accessed: 02-Nov-2015].

[3] “UNESCO World Heritage Centre - World Heritage List.” [Online]. Available: http://whc.unesco.org/en/list/. [Accessed: 19-Oct-2015].

[4] “United Kingdom National Parks -- National Geographic.” [Online]. Available: http://travel.nationalgeographic.com/travel/parks/united-kingdom-national-parks/. [Accessed: 19-Oct-2015].

[5] “Nobelprize.org.” [Online]. Available: http://www.nobelprize.org/. [Accessed: 19-Oct-2015].

[6] “Wolf Foundation.” [Online]. Available: http://www.wolffund.org.il/. [Accessed: 19-Oct-2015].

[7] “Famous People.” [Online]. Available: http://www.thefamouspeople.com/. [Accessed: 23-Feb-2016].

[8] “Gerhard Huber.” [Online]. Available: http://austria-forum.org/af/User/Huber%20Gerhard. [Accessed: 19-Oct-2015].

[9] “Hohmann, Hasso, Univ. Doz. (Architektur, Baukunst)” [Online]. Available: http://austria-forum.org/af/Infos_zum_AF/Editorial_Board/Hohmann, Hasso, Univ. Doz. (Architektur, Baukunst). [Accessed: 19-Oct-2015].

[10] “Pictures” [Online]. Available: http://austria-forum.org/af/Geography/Asia/Syria/Pictures. [Accessed: 19-Oct-2015].

[11] “Iran 1973.” [Online]. Available: http://austria-forum.org/af/Geography/Asia/Iran/Pictures/Iran_1973. [Accessed: 19-Oct-2015].

[12] “The World Factbook.” [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2147rank.html. [Accessed: 19-Oct-2015].

[13] “Jontes, Günther .“ [Online]. Available: http://austria-forum.org/af/Wissenssammlungen/Biographien/Jontes%2C_G%C3%BCnther. [Accessed: 19-Oct-2015].

[14] “Ewald Judt.“ [Online]. http://austria-forum.org/af/Infos_zum_AF/Editorial_Board/Judt%2C_Dr._Ewald_%28Wirtschaftswissenschaft%29. . [Accessed: 19-Oct-2015].

Chapter 9 : Virtual Lab for Exploring Information about Countries

Most existing geographic servers provide static web pages with geographic information. There is no dynamic interface that allow users to experiment with data, like explicitly answering queries that require exploratory analysis e.g. what is the area of a country and its rank among 193 UN member states? Which of the countries have longer land boundaries compared with others? This list is quite long and covers the spectrum of our attempts.

In this chapter we present a "Virtual Lab" that is specifically designed to explore geographic facts. It helps in understanding rankings of countries in different properties. One of our goals in designing a virtual lab is to provide a convenient user interface that helps in exploring geographic facts. This lab also presents comparisons of countries based on different types of facts. It is a tool that can simplify and accelerate geographic investigations, and hence foster geographic learning.

This lab is built by using techniques from different fields that includes geography, data extraction, database systems and AJAX (Asynchronous JavaScript and XML) technology [1]. The AJAX technology is used for faster processing of queries in a client server environment.

The following research questions are dealt with in this chapter:

RQ1: How can we arrange countries and properties on a web page in virtual lab?

RQ2: What is the selection procedure of properties and countries?

RQ3: What is the significance of different reports that will be presented to user? Case studies are presented to illustrate this.

9.1 Introduction

Having data from all countries of the world (UN member states) it clearly makes sense to provide a "Virtual Lab" for exploring countries. In this lab one can check for each country an arbitrarily large subset of some 100 properties, or one can compare properties of countries. It is probably best to explain this by means of some examples.

Rank	Country	Population density
1	India	376.1 (14)
2	Chile	22.97 (149)

Rank	Country	Median age (years)
1	Chile	33.3 (63)
2	India	27.0 (105)

Figure 9.1: A comparison of Chile and India concerning population density and median age

Figure 9.1 shows that among others, the India is fairly densely populated (only 13 other countries are more densely populated). The mean age of 27.0 as compared to the rest of the UN countries ranks 105, i.e. fairly in the middle.

Of course we could now be interested to find out which countries have population density still higher than India. The report is shown in Figure 9.2.

This gives rise to another obvious questions: Are there countries very similar, i.e. agreeing in many properties? This question is particularly relevant if one normalizes figures (per capita, per square kilometre, etc.), for comparison of countries, see chapter 10.

Rank	Country	Population density
1	Singapore	7,988 (1)
2	Bahrain	1,729 (2)
3	Maldives	1,321 (3)
4	Malta	1,306 (4)
5	Bangladesh	1,155 (5)
6	Barbados	673.67 (6)
7	Mauritius	652.53 (7)
8	Lebanon	565.63 (8)
9	Korea, South	491.78 (9)
10	Rwanda	468.42 (10)
11	Netherlands	406.26 (11)
12	Marshall Islands	392.17 (12)
13	Israel	376.59 (13)
14	India	376.1 (14)

Figure 9.2

Rank	Country	Total renewable water resources	Rank	Country	Total Area
1	Germany	154.00 cu km	1	Germany	357,022 sq km
2	Austria	77.70 cu km	2	Austria	83,871 sq km
3	Switzerland	53.50 cu km	3	Switzerland	41,277 sq km

Figure 9.3

Consider another case as shown in the Figure 9.3. Note that both in Total Area and Total renewable water resources Germany ranks before Austria and Switzerland. However, Germany is almost 4 times larger than Austria but has only twice as much renewable water resources; and Switzerland is half the size of Austria, yet has about 3/4 of renewable water resources when compared to Austria.

9.2 Virtual Lab Main Interface

We have tried to design the virtual Lab according to rules and principles of information architecture and web usability. The main interface of the lab is shown in the Figure 9.4. The countries are arranged alphabetically on the left side in 4 columns whereas on right side, the properties are grouped according

to their categories (in 3 columns). For example, properties such as land boundaries, total area, and land area are grouped under Geography. The properties like airports, heliports, railways, roadways are grouped under Transportation.

9.2.1 Flexible Selection of Countries and Properties

Although there can be innumerable combinations that can be used while exploring, we will highlight few concrete cases below.

9.2.1.1 Countries Selection

Users can select countries in three ways:

1. Select all countries.
2. Select continent-wise countries as shown in the Figure 9.5 (European countries are selected in this case).
3. Individual countries can be selected as well.

9.2.1.2 Properties Selection

Like for countries, users can select all properties at once or select category-wise properties. A set of Individual properties can also be selected. The properties listed under "Transportation" can be selected at once by selecting "Transportation" checkbox from the top menu, see Figure 9.5.

Figure 9.4: Virtual lab showing countries and properties

Figure 9.5: European countries selection via selection of Europe checkbox. Transportation properties selection via selection of Transportation checkbox.

9.3 Report Display

There are four different types of reports that we are providing to users.

9.3.1 Case 1: One Property and Many Countries

This report helps users to see the rankings of countries for the selected property. Figure 9.6 shows the scenario when a user selects all the countries and one property, the system will redirect user to the page shown in the Figure 9.7. The report shows the ranking of countries for the selected property "Land boundaries".

Figure 9.6: Virtual lab showing selection of all countries and one property.

Virtual Laboratory for Exploring Countries

Explanation
Fields Description

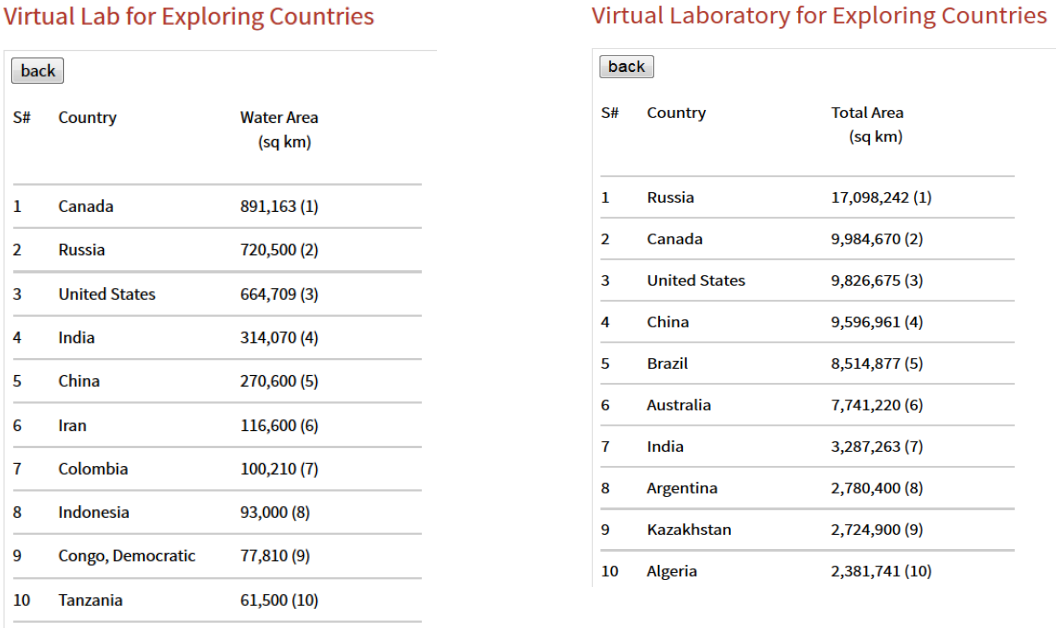
back

The number in brackets shows the rank of a country among 193 UN countries.

S#	Country	Land boundaries (km)
1	China	22,117 (1)
2	Russia	20,242 (2)
3	Brazil	16,885 (3)
4	India	14,103 (4)
5	Kazakhstan	12,185 (5)
6	United States	12,034 (6)
7	Congo, Democratic	10,730 (7)
8	Argentina	9,861 (8)
9	Canada	8,893 (9)
10	Mongolia	8,220 (10)

Figure 9.7: Above report is shown when user selects all countries and only one property. China is listed on top in length of land boundaries.

The report in Figure 9.8 a shows the countries with the largest water area in the world. The water area is the sum of the surfaces of all inland water bodies, such as lakes, reservoirs, or rivers. The list in Figure 9.8 b on the right side shows "Total Area" in sq km of countries. Russia is leading the list in "Total Area".



Virtual Lab for Exploring Countries

back

S#	Country	Water Area (sq km)
1	Canada	891,163 (1)
2	Russia	720,500 (2)
3	United States	664,709 (3)
4	India	314,070 (4)
5	China	270,600 (5)
6	Iran	116,600 (6)
7	Colombia	100,210 (7)
8	Indonesia	93,000 (8)
9	Congo, Democratic	77,810 (9)
10	Tanzania	61,500 (10)

Virtual Laboratory for Exploring Countries

back

S#	Country	Total Area (sq km)
1	Russia	17,098,242 (1)
2	Canada	9,984,670 (2)
3	United States	9,826,675 (3)
4	China	9,596,961 (4)
5	Brazil	8,514,877 (5)
6	Australia	7,741,220 (6)
7	India	3,287,263 (7)
8	Argentina	2,780,400 (8)
9	Kazakhstan	2,724,900 (9)
10	Algeria	2,381,741 (10)

9.8 a 9.8 b

Figure 9.8: Above report is shown when user selects all countries and only one property. "Water Area" is selected in Figure 9.8 a, (we have corrected the entry about Ethiopia as discussed in Section 1.4.3). Canada is listed on top in Water Area. In the right side in Figure 9.8 b "Total Area "is selected by the user.

9.3.2 Case 2: All Properties and one Country

This report helps users to understand a particular country which is selected. For example, a user selects “Afghanistan” with all properties as shown in the Figure 9.9.

Figure 9.9: User selects all properties and one country. Afghanistan is selected in this case with all properties.

The user can see the top rankings of Afghanistan as shown in the Figure 9.10. Afghanistan comes at top in infant mortality rate (which is indication of a health crises). Likewise, the death rate is also high. Therefore one can easily find country rankings that are higher or lower in properties using this report. The properties are listed in order of decreasing rank for a particular country.

Virtual Laboratory for Exploring Countries

back

S#	Property	Afghanistan
1	Infant mortality rate - deaths/1,000 live births	117.23 (1)
2	Age Structure: 15-24 years - %	22.2 (6)
3	Death rate - deaths/1,000 population	14.12 (7)
4	Highest Point - meter	7,485 (7)
5	Birth rate - births/1,000 population	38.84 (10)
6	Total fertility rate - children born/woman	5.43 (11)
7	Administrative divisions	34.0 (19)
8	Heliports	9.0 (19)
9	Irrigated Land - sq km	32,080 (19)
10	Land boundaries - km	5,529 (24)
11	Maternal mortality rate - deaths/100,000 live births	460.00 (24)
12	Age Structure: 0-14 years - %	42.0 (27)
13	Health expenditures - % of GDP	9.60 (30)
14	Freshwater withdrawal per capita - cu km/yr	823.1 (32)
15	Freshwater withdrawal - cu km/yr	20.28 (33)
16	Population growth rate - %	2.29 (35)
17	Land Area - sq km	652,230 (39)
18	Total Area - sq km	652,230 (40)
19	Population	31,822,848 (40)
20	% Irrigated land	4.92 (42)
21	Ethnic groups	8.0 (48)

Figure 9.10: Report displaying ranking of Afghanistan in different properties

9.3.3 Case 3: Many Properties and four Countries

Users can select many properties and up to four countries. For instance a user selects Afghanistan, Albania, Algeria and Austria as shown in the Figure 9.11. Suppose now the user wants to compare these countries with respect to Geography. The report shown in the Figure 9.12 presents the comparison. Properties are listed on the left side and the values are shown in the respective columns of countries. Afghanistan ranks high in highest point. Noshak is the highest point in Afghanistan.

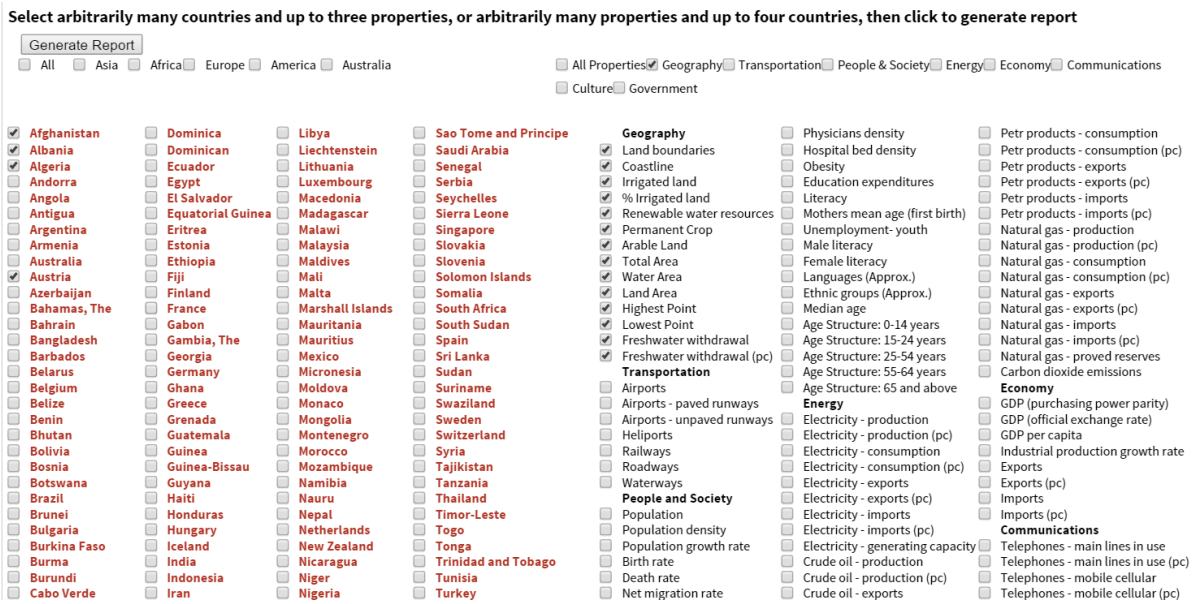
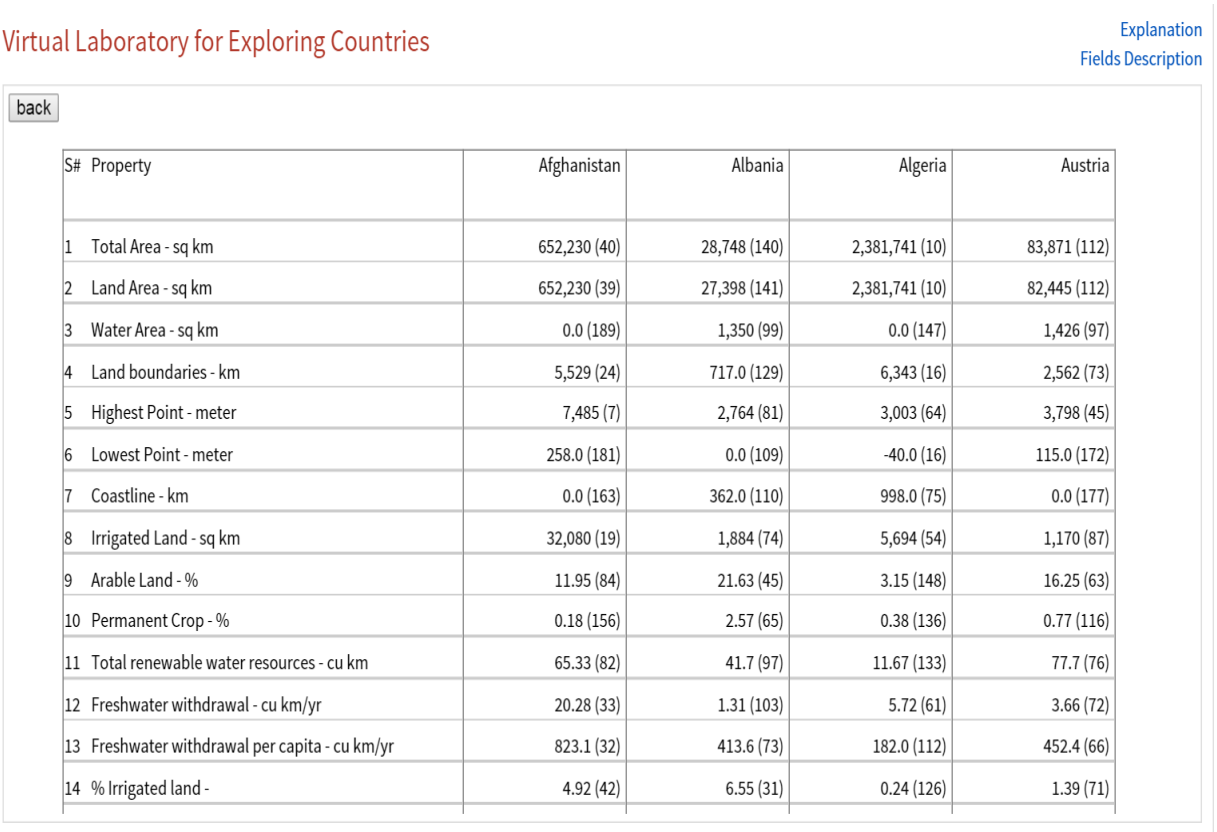


Figure 9.11: Selection of 4 countries and different geographic properties.



S#	Property	Afghanistan	Albania	Algeria	Austria
1	Total Area - sq km	652,230 (40)	28,748 (140)	2,381,741 (10)	83,871 (112)
2	Land Area - sq km	652,230 (39)	27,398 (141)	2,381,741 (10)	82,445 (112)
3	Water Area - sq km	0.0 (189)	1,350 (99)	0.0 (147)	1,426 (97)
4	Land boundaries - km	5,529 (24)	717.0 (129)	6,343 (16)	2,562 (73)
5	Highest Point - meter	7,485 (7)	2,764 (81)	3,003 (64)	3,798 (45)
6	Lowest Point - meter	258.0 (181)	0.0 (109)	-40.0 (16)	115.0 (172)
7	Coastline - km	0.0 (163)	362.0 (110)	998.0 (75)	0.0 (177)
8	Irrigated Land - sq km	32,080 (19)	1,884 (74)	5,694 (54)	1,170 (87)
9	Arable Land - %	11.95 (84)	21.63 (45)	3.15 (148)	16.25 (63)
10	Permanent Crop - %	0.18 (156)	2.57 (65)	0.38 (136)	0.77 (116)
11	Total renewable water resources - cu km	65.33 (82)	41.7 (97)	11.67 (133)	77.7 (76)
12	Freshwater withdrawal - cu km/yr	20.28 (33)	1.31 (103)	5.72 (61)	3.66 (72)
13	Freshwater withdrawal per capita - cu km/yr	823.1 (32)	413.6 (73)	182.0 (112)	452.4 (66)
14	% Irrigated land -	4.92 (42)	6.55 (31)	0.24 (126)	1.39 (71)

Figure 9.12: Report showing comparison of four countries Afghanistan, Albania, Algeria and Austria with respect to different geographic properties.

To understand, how this report helps in comparison of countries we are showing another example. Figure 9.13 shows a comparison of three countries concerning certain health parameters

S#	Property	Hungary	Netherlands	New Zealand
1	Physicians density - physicians/1,000 population	3.41 (30)	3.92 (11)	2.74 (50)
2	Hospital bed density - beds/1,000 population	7.2 (11)	4.7 (41)	2.3 (93)
3	Population -	9,919,128 (88)	16,877,351 (64)	4,401,916 (124)
4	Mothers mean age at first birth - years	28.2 (19)	28.9 (13)	27.7 (22)
5	Population growth rate - %	-0.21 (179)	0.42 (141)	0.83 (119)
6	Birth rate - births/1,000 population	9.26 (179)	10.83 (157)	13.4 (137)
7	Death rate - deaths/1,000 population	12.72 (23)	8.57 (71)	7.3 (108)
8	Life expectancy at birth - years	75.46 (67)	81.12 (17)	80.93 (18)
9	Maternal mortality rate - deaths/100,000 live births	21.0 (134)	6.0 (168)	15.0 (142)
10	Infant mortality rate - deaths/1,000 live births	5.09 (158)	3.66 (178)	4.59 (162)
11	Obesity - adult prevalence rate - %	27.6 (34)	18.8 (99)	28.3 (31)
12	Population density -	106.63 (67)	406.26 (11)	16.44 (159)

Figure 9.13: Comparison of three countries involving twelve properties

Figure 9.13 seems to indicate that the health system in Hungary is much worse than the one in The Netherlands and New Zealand: Infant mortality rate and Maternal mortality rate are both high, Life expectancy is low, Death rate is high, all a bit surprising when noticing the High hospital bed density. The hospital bed density in Hungary is high but it appears that many lack important facilities. Again a typical case for our contention, that both questions and data are often not specific enough.

When comparing Hungary using the same parameters with Cambodia and Nigeria in Figure 9.14. Hungary comes out well ahead in health parameters, however: it has higher Physicians density, higher Hospital bed density, higher life expectancy, much lower maternal mortality and Infant mortality rate, etc.

S#	Property	Cambodia	Hungary	Nigeria
1	Physicians density - physicians/1,000 population	0.23 (142)	3.41 (30)	0.4 (128)
2	Hospital bed density - beds/1,000 population	0.7 (168)	7.2 (11)	0.53 (175)
3	Population -	15,458,332 (67)	9,919,128 (88)	177,155,754 (7)
4	Mothers mean age at first birth - years	22.8 (67)	28.2 (19)	20.9 (89)
5	Population growth rate - %	1.63 (68)	-0.21 (179)	2.47 (30)
6	Birth rate - births/1,000 population	24.4 (58)	9.26 (179)	38.03 (12)
7	Death rate - deaths/1,000 population	7.78 (94)	12.72 (23)	13.16 (19)
8	Life expectancy at birth - years	63.78 (148)	75.46 (67)	52.62 (180)
9	Maternal mortality rate - deaths/100,000 live births	250.0 (45)	21.0 (134)	630.0 (11)
10	Infant mortality rate - deaths/1,000 live births	51.36 (35)	5.09 (158)	74.09 (10)
11	Obesity - adult prevalence rate - %	2.1 (180)	27.6 (34)	6.5 (143)
12	Population density -	85.39 (87)	106.63 (67)	191.78 (43)

Figure 9.14: Hungary compared to less developed countries

9.3.4 Case 4: Five Properties and Many Countries

Users can select up to five properties and many countries. For instance the user selects all countries as shown in the Figure 9.15. He wants to compare these countries with respect to five properties e.g. Land boundaries, Coastline, Irrigated land, %Irrigated land and Renewable water resources. The report shown in the Figure 9.16 presents the results. Only the first column "Land boundaries" is sorted in ascending order based on rank. The country list is sorted according to "Land boundaries". For instance China comes on top in Land boundaries. Other columns (showing different properties) are not sorted. They show the values corresponding to country list (obtained earlier using Land boundaries). China is ranked 10 in coastline. It is ranked 2nd in irrigated land. This interface is helpful for finding the correlations between different properties.

Figure 9.15: Selection of all countries and five geographic properties.

9.3.4.1 Report Layout based on Order of Property Selection

The columns arranged in the report are based on the order of selection of properties. For instance when users first select coastline then the report layout changes as shown in the Figure 9.17. This layout is different compared to Figure 9.16 where land boundaries is selected first and then coastline. Countries are sorted according to coastline. Canada has the longest coastline.

Virtual Laboratory for Exploring Countries

Explanation
Fields Description

back

The number in brackets shows the rank of a country among 193 UN countries.

S#	Country	Land boundaries (km)	Coastline (km)	Irrigated land (sq km)	% Irrigated land	Renewable water resources
1	China	22,117 (1)	14,500 (10)	629,380 (2)	6.56 (30)	2,840 (5)
2	Russia	20,242 (2)	37,653 (3)	43,460 (13)	0.25 (125)	4,508 (2)
3	Brazil	16,885 (3)	7,491 (15)	54,000 (9)	0.63 (99)	8,233 (1)
4	India	14,103 (4)	7,000 (18)	663,340 (1)	20.18 (3)	1,911 (9)
5	Kazakhstan	12,185 (5)	0.0 (154)	20,660 (25)	0.76 (92)	107.5 (65)
6	United States	12,034 (6)	19,924 (8)	266,440 (3)	2.71 (57)	3,069 (3)
7	Congo, Democratic	10,730 (7)	37.0 (145)	105.0 (134)	0.0 (169)	1,283 (10)
8	Argentina	9,861 (8)	4,989 (24)	15,500 (32)	0.56 (104)	814.0 (17)
9	Canada	8,893 (9)	202,080 (1)	8,699 (46)	0.09 (145)	2,902 (4)
10	Mongolia	8,220 (10)	0.0 (155)	843.0 (102)	0.05 (156)	34.8 (102)

Figure 9.16: Report showing comparison of all countries with respect to five properties.

S#	Country	Coastline (km)	Land boundaries (km)	Irrigated land (sq km)	Renewable water resources
1	Canada	202,080 (1)	8,893 (9)	8,699 (46)	2,902 (4)
2	Indonesia	54,716 (2)	2,830 (65)	67,220 (6)	2,019 (7)
3	Russia	37,653 (3)	20,242 (2)	43,460 (13)	4,508 (2)
4	Philippines	36,289 (4)	0.0 (188)	18,790 (28)	479.0 (22)
5	Japan	29,751 (5)	0.0 (171)	25,000 (22)	430.0 (26)
6	Australia	25,760 (6)	0.0 (179)	25,460 (21)	492.0 (21)
7	Norway	25,148 (7)	2,542 (75)	1,149 (90)	382.0 (28)
8	United States	19,924 (8)	12,034 (6)	266,440 (3)	3,069 (3)
9	New Zealand	15,134 (9)	0.0 (158)	6,193 (50)	327.0 (32)
10	China	14,500 (10)	22,117 (1)	629,380 (2)	2,840 (5)
11	Greece	13,676 (11)	1,228 (114)	15,550 (31)	74.25 (78)
12	United Kingdom	12,429 (12)	360.0 (144)	2,280 (71)	147.0 (54)
13	Mexico	9,330 (13)	4,353 (41)	64,600 (7)	457.2 (24)
14	Italy	7,600 (14)	1,899 (90)	39,510 (15)	191.3 (45)
15	Brazil	7,491 (15)	16,885 (3)	54,000 (9)	8,233 (1)

Figure 9.17: Report layout when user first selects coastline and then remaining properties from main interface.

9.4 Normalization

To compare statistical figures it was necessary to normalize some of them. It may be interesting to have e.g. the number of medical doctors or Nobel Prize winners in a country, but it might be more interesting to have them per 1.000 or 1.000.000 people in the country for comparison. While USA has the largest number of Nobel Prize winners (254), per 1 million persons Saint Lucia has more (12.24) compared to 0.8 in the USA. Saint Lucia is a sovereign island country and has a population of around 174,000 people but 2 Nobel Prize winners.

9.5 Non ranked data Display

Our lab also allows to explore parameters that cannot be ranked numerically like capital, Independence day of a country. However we have sorted countries alphabetically in such lists as shown in Figure 9.18. The Independence column is sorted in descending order using parameter independence date. South Sudan is listed on top among countries since it became independent in 2011.

S#	Country	Capital	S#	Country	Independence
1	Afghanistan	Kabul	1	South Sudan	2011-07-09
2	Albania	Tirana	2	Serbia	2006-06-05
3	Algeria	Algiers	3	Montenegro	2006-06-03
4	Andorra	Andorra la Vella	4	Palau	1994-10-01
5	Angola	Luanda	5	Eritrea	1993-05-24
6	Antigua and Barbuda	Saint John's	6	Slovakia	1993-01-01
7	Argentina	Buenos Aires	7	Czech Republic	1993-01-01
8	Armenia	Yerevan	8	Bosnia and Herzegovina	1992-03-01
9	Australia	Canberra	9	Kazakhstan	1991-12-16
10	Austria	Vienna	10	Turkmenistan	1991-10-27
11	Azerbaijan	Baku	11	Armenia	1991-09-21
12	Bahamas, The	Nassau	12	Tajikistan	1991-09-09
13	Bahrain	Manama	13	Macedonia	1991-09-08
14	Bangladesh	Dhaka	14	Uzbekistan	1991-09-01
15	Barbados	Bridgetown	15	Kyrgyzstan	1991-08-31
16	Belarus	Minsk	16	Azerbaijan	1991-08-30
17	Belgium	Brussels	17	Moldova	1991-08-27
18	Belize	Belmopan	18	Belarus	1991-08-25
19	Benin	Porto-Novo	19	Ukraine	1991-08-24
20	Bhutan	Thimphu	20	Russia	1991-08-24
21	Bolivia	La Paz	21	Estonia	1991-08-20
22	Bosnia and Herzegovina	Sarajevo	22	Slovenia	1991-06-25
23	Botswana	Gaborone	23	Croatia	1991-06-25
24	Brazil	Brasilia	24	Georgia	1991-04-09
25	Brunei	Bandar Seri Begawan	25	Yemen	1990-05-22
26	Bulgaria	Sofia	26	Latvia	1990-05-04
27	Burkina Faso	Ouagadougou	27	Namibia	1990-03-21

Figure 9.18: Report showing list of capitals and independence days of countries.

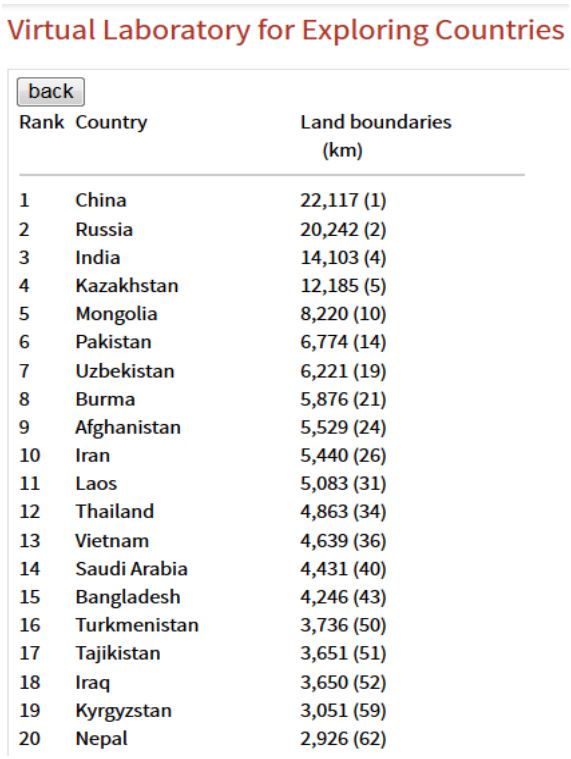
9.6 Relative vs Absolute Ranking

The laboratory provides multiple views of countries and rankings. For instance, the users want to see the ranking of only Asian countries with respect to land boundaries as shown in the Figure 9.19. The corresponding report is shown in the Figure 9.20 where users can view absolute and relative rankings. Concerning Land boundaries Iran is ranked 10 in Asian countries whereas it is ranked 26 among 193 UN countries. India is ranked 3 in Asian countries whereas it is ranked 4 among 193 UN countries. Figure 9.21 shows ranking of European countries with respect of "Airports". Germany is at the top position in the list, however it ranks only 13 among UN countries. The "Airports" is one of the many cases when without knowing the exact definition numbers do not make sense.

According to Factbook "Airports" is:

"This entry gives the total number of airports or airfields recognizable from the air. The runway(s) may be paved (concrete or asphalt surfaces) or unpaved (grass, earth, sand, or gravel surfaces) and may include closed or abandoned installations. Airports or airfields that are no longer recognizable (overgrown, no facilities, etc.) are not included. Note that not all airports have accommodations for refueling, maintenance, or air traffic control."

Figure 9.19: Selection of Asian countries and Land boundaries.



Virtual Laboratory for Exploring Countries

back

Rank	Country	Land boundaries (km)
1	China	22,117 (1)
2	Russia	20,242 (2)
3	India	14,103 (4)
4	Kazakhstan	12,185 (5)
5	Mongolia	8,220 (10)
6	Pakistan	6,774 (14)
7	Uzbekistan	6,221 (19)
8	Burma	5,876 (21)
9	Afghanistan	5,529 (24)
10	Iran	5,440 (26)
11	Laos	5,083 (31)
12	Thailand	4,863 (34)
13	Vietnam	4,639 (36)
14	Saudi Arabia	4,431 (40)
15	Bangladesh	4,246 (43)
16	Turkmenistan	3,736 (50)
17	Tajikistan	3,651 (51)
18	Iraq	3,650 (52)
19	Kyrgyzstan	3,051 (59)
20	Nepal	2,926 (62)

Figure 9.20: Ranking of Asian countries regarding “Land boundaries”.

Virtual Laboratory for Exploring Countries

back

Rank	Country	Airports
1	Germany	539.0 (13)
2	France	464.0 (17)
3	United Kingdom	460.0 (18)
4	Sweden	231.0 (25)
5	Ukraine	187.0 (31)
6	Spain	150.0 (38)
7	Finland	148.0 (39)
8	Italy	129.0 (45)
9	Czech Republic	128.0 (46)
10	Poland	126.0 (47)

Figure 9.21: Ranking of European countries regarding Airports.

9.7 Technical Aspects of Laboratory Development

The first stage behind development of the virtual lab was the data preparation. We first parse the data from Factbook data using functions like “substring_index” to get the actual values from the textual description of properties.

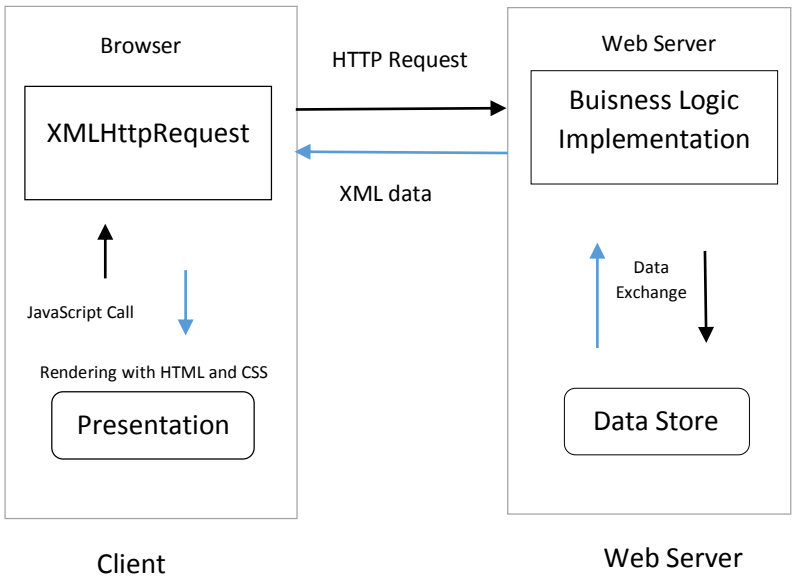
Afterwards, we apply two major operations

- Filtering
- Sorting

As we want to generate rankings for only UN member states, we have excluded the non UN member states from the list of countries. In the next stage we apply sorting and generate the rankings of countries based on their values.

We have used AJAX technology for report processing in the virtual lab. AJAX is a client-side script that communicates to and from a server/database without the need of post backs or completely refresh a page [1] [2].

There are many benefits of AJAX one of which is speed. The process flow of AJAX is shown in the diagram below, for more details see [3]



Process flow diagram describing AJAX concept

The AJAX request using JQuery is shown below.

```
$.post( "ajax/Report1.jsp", function ( data ){$( ".result" ).html( data );});
```

We have used following tools, libraries and components for building lab:

- JavaScript library JQuery for handling business logic.
- HTML elements (checkboxes) in the user interface along with labels for showing countries and properties.
- CSS3 styling for arranging countries and properties in 4 and 3 columns respectively.

References

[1] "AJAX Technologies." [Online]. Available: http://www.tutorialspoint.com/ajax/ajax_technology.htm. [Accessed: 10-Feb-2016].

[2] "What is Ajax and Where is it Used in Technology?" [Online]. Available: http://www.seguetech.com/blog/2013/03/12/what-is-ajax-and-where-is-it-used-in-technology. [Accessed: 10-Feb-2016].

[3] "Introduction to Ajax for Java Web Applications - NetBeans IDE Tutorial." [Online]. Available: https://netbeans.org/kb/docs/web/ajax-quickstart.html. [Accessed: 11-Feb-2016].

Chapter 10 : Mining Countries Statistics

Researchers doing research in data mining always try to uncover hidden facts that are not obvious, otherwise. This chapter reports on our attempts to find the rank of countries with respect to various properties. We have prepared two types of rankings for UN countries.

We call them “Top Ten” and “bottom” rankings. The “Top Ten rankings” of a country means that country has a high rank in some properties, see section 10.1 for better understanding.

We also present the countries comparison by showing which countries have common properties. Specifically, we list countries which are similar in more than 6 properties. In this chapter, we will focus on how we have calculated rankings of countries. We explain our results by means of a few examples.

In addition to stories that are written by people (present in “Special Information” section of each country), there are also those which are compiled based on the statistics of data (approximately 100 properties). For example, for each country there is a list of the properties in which the country is among the top ten UN countries (Top Ten rankings), and among the last (bottom rankings).

MySQL has introduced number of new features (stored procedures being one of the most significant) [1][2]. A cursor allows to iterate through set of rows returned by a query and process each row accordingly. This is particularly useful in calculating similarities of countries.

This chapter tries to answer the following research questions.

RQ1. How can we find countries that are in the top 10 rankings concerning some properties?

RQ2. How can we present rankings in a meaningful way that is easy for the people to understand?

RQ3. How can we find bottom rankings of a country?

RQ4. How can we find similarities between countries with respect to different properties?

RQ5. How many countries come in top 10 rankings among UN member states in more than 20 properties?

RQ6. How many countries come in top 10 rankings among UN member states in only 1 property?

RQ7. How many countries come in bottom rankings among UN member states in more than 10 properties? (Countries like Tuvalu, Nauru have low rankings in many properties)

Note: The values of three properties “Ethnic groups”, “Languages”, “Administrative divisions” are not exact, they are just the approximate values.

10.1 Top Ten Rankings

For each country we have a list called “Top Ten Rankings” showing where the country ranks high in some properties. High ranking is of course not necessarily good in some properties, like High Mortality of New borns etc. We will discuss a few cases in this section. We start our stories with specifically designed template shown below.

Did you know that <Country> is among the top 10 of the 193 UN countries in the following <No of Properties> respects?

In the template <Country> is replaced by country name and <No of Properties> is replaced by total number of properties in which country ranks in top ten. Examples will be presented in the sub-sections.

10.1.1 Top Ten Rankings of Austria

Figure 10.1 shows the situation for Austria.

Did you know that Austria is among the top 10 of the 193 UN countries in the following 13 respects.

1. Natural gas - imports per capita - cu m: 5,176 (Rank 2).
2. Physicians density - physicians/1,000 population: 4.86 (Rank 5).
3. Electricity - imports per capita - KWh: 2,829 (Rank 5).
4. Median age - years: 44.3 (Rank 5).
5. Electricity - imports - KWh: 23,260,000,000 (Rank 7).
6. Nobel Prize Winners per 1000000 people: 2.07 (Rank 7).
7. Age Structure: 65 and above - %: 18.9 (Rank 8).
8. Age Structure: 55-64 years - %: 19.2 (Rank 9).
9. Electricity - exports - KWh: 20,460,000,000 (Rank 9).
10. Natural gas - exports per capita - cu m: 4,226 (Rank 9).
11. Hospital bed density - beds/1,000 population: 7.6 (Rank 10).
12. Electricity - exports per capita - KWh: 2,488 (Rank 10).
13. Imports per capita - $: 20,418 (Rank 10).

Figure 10.1: Top ten rankings of Austria.

Note that Austria is doing well in physician's density (Rank 5 among UN member states), yet this may even be weakly connected with the dismal age structure: Austria is ranked no.8 when it comes to the percentage of persons above age 65.

10.1.2 Top Ten Rankings of India

India is a fairly large country in south Asia. Top ten rankings of India are shown in Figure 10.2. India is doing well in irrigated land, roadways etc. India ranks 2nd among UN member states after China in size of population, hence some high ranks are not at all surprising.

Did you know that India is among the top 10 of the 193 UN countries in the following 30 respects.

1. Irrigated Land - sq km: 663,340 (Rank 1).
2. Freshwater withdrawal - cu km/yr: 761.0 (Rank 1).
3. Roadways - km: 4,689,842 (Rank 2).
4. Population: 1,236,344,631 (Rank 2).
5. Telephones - mobile cellular: 893,862,000 (Rank 2).
6. Languages: 14.0 (Rank 3).
7. GDP (purchasing power parity) - $: 4,960,000,000,000 (Rank 3).
8. % Irrigated land: 20.18 (Rank 3).
9. Land boundaries - km: 14,103 (Rank 4).
10. Highest Point - meter: 8,598 (Rank 4).
11. Railways - km: 63,974 (Rank 4).
12. Electricity - production - KWh: 985,400,000,000 (Rank 4).
13. Crude oil - imports - bbl/day: 3,270,000 (Rank 4).
14. Refined petroleum products - production - bbl/day: 4,220,000 (Rank 4).
15. Refined petroleum products - consumption - bbl/day: 3,290,000 (Rank 4).
16. Carbon dioxide emissions from consumption of energy - MT: 1,730,000,000 (Rank 4).
17. Water Area - sq km: 314,070 (Rank 4).
18. Arable Land - %: 47.87 (Rank 5).
19. Electricity - consumption - KWh: 698,800,000,000 (Rank 5).
20. Electricity - installed generating capacity - KW: 208,100,000 (Rank 5).
21. Refined petroleum products - exports - bbl/day: 1,250,000 (Rank 5).
22. Heliports: 45.0 (Rank 6).
23. Internet users: 61,338,000 (Rank 6).
24. Total Area - sq km: 3,287,263 (Rank 7).
25. Land Area - sq km: 2,973,193 (Rank 7).
26. Imports - $: 515,100,000,000 (Rank 8).
27. Total renewable water resources - cu km: 1,911 (Rank 9).
28. Waterways - km: 14,500 (Rank 9).
29. Airports - with paved runways: 253.0 (Rank 10).
30. Telephones - main lines in use: 31,080,000 (Rank 10).

Figure 10.2: Top ten rankings of India. India comes in top ten rankings in 30 properties

10.1.3 Top Ten Rankings of Bosnia and Herzegovina

All former Yugoslavian countries have high unemployment among young people, but how catastrophic the situation is can be seen in the Top Ten Rankings of Bosnia and Herzegovina shown in the Figure 10.3, where 62.8% of the population between 15 and 24 years of age are unemployed!

Did you know that Bosnia and Herzegovina is among the top 10 of the 193 UN countries in the following 1 respect.

1. Unemployment- youth ages 15-24 - %: 62.8 (Rank 1).

Figure 10.3: Top ten rankings of Bosnia and Herzegovina

10.1.4 Top Ten Rankings of Denmark

In Denmark, the Top Ten rankings are, for example:

Did you know that Denmark is among the top 10 of the 193 UN countries in the following 5 respects?

1. Arable Land - %: 57.99 (Rank 1).
2. Electricity - imports per capita - KWh: 2,859 (Rank 4).
3. Nobel Prize Winners per 1000000 people: 2.15 (Rank 6).
4. % Internet users: 85.29 (Rank 6).
5. Education expenditures - % of GDP: 8.7 (Rank 8).

10.1.5 Top Ten Rankings of Nauru

The top ten rankings of Nauru is shown below. The obesity is the major health concern in Nauru.

Did you know that Nauru is among the top 10 of the 193 UN countries in the following 5 respects?

1. Urbanization - %: 100.0 (Rank 1).
2. Obesity - adult prevalence rate - %: 71.1 (Rank 1).
3. Refined petroleum products - consumption per capita - bbl/day: 0.11 (Rank 5).
4. Refined petroleum products - imports per capita - bbl/day: 0.11 (Rank 6).
5. Permanent Crop - %: 20.0 (Rank 9).

Nauru is a small country (phosphate rock island) having population of about 9,488 people. The fact that Nauru is 100% urbanized is rather counter-intuitive. Nauru is mostly empty but all people are administrated by one or a few villages it comes out as 100% urbanized. The intuitive picture of a highly urbanized country is Singapore, where almost all area is covered with houses. Iceland is 93% urbanized, again because of the strange definition used by Factbook.

"Urbanization: Urban population, describes the percentage of the total population living in urban areas, as defined by the country"

10.1.6 Top Ten Rankings of South Sudan

South Sudan is a country that became independent in 2011. The top ten rankings of South Sudan are:

Did you know that South Sudan is among the top 10 of the 193 UN countries in the following 6 respects?

1. Maternal mortality rate - deaths/100,000 live births: 2,054 (Rank 1).
2. Population growth rate - %: 4.12 (Rank 3).
3. Ethnic groups: 18.0 (Rank 4).
4. Age Structure: 0-14 years - %: 45.8 (Rank 6).
5. Net migration rate - migrant(s)/1,000 population: 11.94 (Rank 9).
6. Total fertility rate - children born/woman: 5.43 (Rank 10).

10.1.7 Top Ten Rankings of Germany

The top ten rankings of Germany are of course more than most other European countries. It ranks high in many properties. The list of top ten rankings is given below.

Did you know this about Germany? Did you know that Germany is among the top 10 of the 193 UN countries in the following 29 respects?

1. Electricity - exports - KWh: 66,810,000,000 (Rank 1).
2. Electricity - imports - KWh: 46,270,000,000 (Rank 2).
3. Telephones - main lines in use per capita: 0.63 (Rank 2).
4. Age Structure: 55-64 years - %: 21.1 (Rank 3).
5. Age Structure: 65 and above - %: 20.9 (Rank 3).
6. Natural gas - imports - cu m: 87,960,000,000 (Rank 3).
7. Exports - $: 1,490,000,000,000 (Rank 3).
8. Imports - $: 1,230,000,000,000 (Rank 3).
9. Nobel Prize Winners - : 80.0 (Rank 3).
10. Median age - years: 46.1 (Rank 3).
11. GDP (official exchange rate) - $: 3,593,000,000,000 (Rank 4).
12. Telephones - main lines in use: 50,700,000 (Rank 4).

13. GDP (purchasing power parity) - $: 3,230,000,000,000 (Rank 5).
14. Internet users: 65,125,000 (Rank 5).
15. Railways - km: 41,981 (Rank 6).
16. Electricity - consumption - KWh: 582,500,000,000 (Rank 6).
17. Electricity - installed generating capacity - KW: 178,400,000 (Rank 6).
18. Crude oil - imports - bbl/day: 1,880,000 (Rank 6).
19. Carbon dioxide emissions from consumption of energy - MT: 814,000,000 (Rank 6).
20. Internet hosts: 20,043,000 (Rank 6).
21. Airports - with paved runways: 318.0 (Rank 7).
22. Refined petroleum products - production - bbl/day: 2,200,000 (Rank 7).
23. Hospital bed density - beds/1,000 population: 8.3 (Rank 8).
24. Refined petroleum products - consumption - bbl/day: 2,400,000 (Rank 8).
25. Refined petroleum products - imports - bbl/day: 758,100 (Rank 8).
26. Heliports: 23.0 (Rank 9).
27. Electricity - production - KWh: 526,600,000,000 (Rank 9).
28. Natural gas - consumption - cu m: 75,200,000,000 (Rank 9).
29. % Internet users: 80.4 (Rank 10).

10.1.8 Procedure to find Top Ten Rankings

The SQL query used to select top 10 rankings of a particular country is given below. We are excluding Independence day, Capital and Internet country code from the list of properties.

```
1: select property, value, unit, rank from reportdata
2: where id=cid and rank <=10
3: and property !='Independence' and
4: property!='Internet country code'
5: and property!='Capital'
6: order by cast(rank as unsigned) asc;
```

Table 10.1 shows the list of countries which are in top ten rankings in 20 or more than 20 properties. Japan has 29 properties where it comes in top rankings among UN member states in different properties. France has 20 properties where it comes in top 10 rankings among UN member states.

Table 10.1: Countries which come in top ten rankings in 20 or more properties

Country	Property count where the country ranks in Top 10
United States	48
Russia	39
China	37
Canada	30
India	30
Japan	29
Germany	29
Brazil	24
Netherlands	21
Norway	21
France	20

Table 10.2 shows countries that occupy a place in top ten rankings among UN member states in only 1 property.

Table 10.2: Countries which come in top ten rankings in only 1 property

Country	Property	Rank
Bosnia and Herzegovina	Unemployment-youth ages 15-24	1
Tonga	Obesity - adult prevalence rate	2
Solomon Islands	Industrial production growth rate	2
Syria	Lowest Point	3
Samoa	Obesity - adult prevalence rate	3
Uzbekistan	Freshwater withdrawal per capita	4
Nicaragua	Age Structure: 15-24 years	4
Ireland	Mothers mean age at first birth	4
Swaziland	Age Structure: 15-24 years	5
Botswana	Education expenditures	5
Panama	Telephones - mobile cellular per capita	5
Egypt	Lowest Point	6
Guatemala	Age Structure: 15-24 years	7
Timor-Leste	Education expenditures	7
Georgia	Physicians density	7
Mauritius	Population density	7
Lithuania	Crude oil - imports per capita	7
Dominica	Permanent Crop	7
Peru	Total renewable water resources	8
Togo	Arable Land	8
Grenada	Permanent Crop	8
Saint Kitts and Nevis	Obesity - adult prevalence rate	8
Laos	Industrial production growth rate	8
Costa Rica	Freshwater withdrawal per capita	9
New Zealand	Coastline	9
Mozambique	Age Structure: 0-14 years	9
Tunisia	Unemployment- youth ages 15-24	9
Tanzania	Water Area	10
Cameroon	Maternal mortality rate	10
Papua New Guinea	Airports - with unpaved runways	10
Jamaica	Age Structure: 15-24 years	10
Suriname	Telephones - mobile cellular per capita	10

10.2 Bottom Rankings

In addition to the Top Ten Rankings, also parameters for which a country is ranked among the Ten Lowest in the UN countries are shown, revealing many interesting aspects, see [3] . Here is one example: In the following 5 countries only 2.1 % or less of the population is older than 65: Uganda, Kuwait, Nauru, United Arab Emirates, and Qatar. United Arab Emirates, Kuwait and Qatar spend a very small percentage of their GNP on health, so these 3 countries do not come as surprise, but Nauru and Uganda are ranked around 30 in their expenditure on health, so why are there not more old people? Well, the island Nauru with just 21 square kilometers (most of it destroyed through phosphate mining, and no natural fresh water resources except rain and an aging desalination plant) and less than 10.000

inhabitants cannot be a serious candidate for statistics. Uganda with about 30 million people, yet 70% 24 years or younger, destructive civil wars, less than 50% of the population having access to clean water and the country where AIDS originated explain why there are few old persons. And since it is considered the main battle ground against AIDS the high medical expenditure (mostly through international aid) explains why the expenditure on health is fairly high.

10.2.1 Procedure to find Bottom Rankings

The SQL query to select bottom rankings is shown below. First, max rank of a property is selected. The second step is to compare rank of particular country with max rank. If difference is less than 10 then property is returned.

```
1: select max (cast (rank as unsigned)) into _max from reportdata
2: where property=_property ;
3: set _diff=_max-_rank;
4: if (_diff>-1 and _diff<10) then
5: begin
6: select id,value,unit,property,rank from reportdata rt
7: where id=cid and property=_property;
8: end;
```

Table 10.3 shows countries with total number of properties in which they rank very low. We are listing countries when property count >=10

Table 10.3: Countries which has low rankings in more than 10 properties

Country	Number of Properties in which country ranks very low
Nauru	22
Tuvalu	15
Kiribati	15
Central African Republic	15
Comoros	15
Niger	14
Tonga	13
Monaco	13
Somalia	13
Marshall Islands	12
Chad	12
Sao Tome and Principe	12
Liechtenstein	11
Burundi	11
South Sudan	11
Ethiopia	10
Qatar	10
Afghanistan	10

Note, this is perhaps not that much surprising that the countries listed in Table 10.3, have low rankings in many properties, because they are very small countries (having area of few square kilometres). We are listing low rankings of Tuvalu mentioned in the list, as such case.

10.2.2 Bottom rankings of Tuvalu

Did you know that Tuvalu is very low in ranking among 193 countries in the following 15 respects?

1. Total Area - sq km: 26.0 (Rank 191).
2. Land Area - sq km: 26.0 (Rank 190).
3. Highest Point - meter: 5.0 (Rank 192).
4. Coastline - km: 24.0 (Rank 148).
5. Airports - : 1.0 (Rank 184).
6. Roadways - km: 8.0 (Rank 192).
7. Airports - with unpaved runways - : 1.0 (Rank 170).
8. Population - : 10,782 (Rank 192).
9. Imports - $: 16,500,000 (Rank 192).
10. GDP (purchasing power parity) - $: 40,000,000 (Rank 193).
11. GDP (official exchange rate) - $: 38,000,000 (Rank 192).
12. Telephones - main lines in use - : 1,450 (Rank 193).
13. Telephones - mobile cellular - : 2,800 (Rank 193).
14. Internet users - : 4,200 (Rank 187).
15. Exports - $: 600,000 (Rank 191).

10.2.3 Bottom rankings of Italy

The bottom rankings of Italy is shown in the Figure 10.4.

Did you know that Italy is very low in ranking among 193 countries in the following 5 respects.

1. Age Structure: 0-14 years - %: 13.8 (Rank 184).
2. Age Structure: 15-24 years - %: 9.8 (Rank 189).
3. Maternal mortality rate - deaths/100,000 live births: 4.0 (Rank 177).
4. Infant mortality rate - deaths/1,000 live births: 3.31 (Rank 184).
5. Industrial production growth rate - %: -2.7 (Rank 182).

Figure 10.4: Bottom rankings of Italy.

10.2.4 Bottom rankings of Syria

The bottom rankings of Syria is shown in the Figure 10.5.

Did you know that Syria is very low in ranking among 193 countries in the following 3 respects.

1. Population growth rate - %: -9.73 (Rank 193).
2. Net migration rate - migrant(s)/1,000 population: -113.51 (Rank 192).
3. Industrial production growth rate - %: -20.6 (Rank 191).

Figure 10.5: Bottom rankings of Syria.

10.2.5 Bottom rankings of Estonia

The low rankings of Estonia is shown below. Estonia is ranked very low in irrigated land.

Did you know that Estonia is very low in ranking among 193 countries in the following 4 respects?

1. Irrigated Land - sq km: 4.58 (Rank 165).
2. Population growth rate - %: -0.68 (Rank 190).
3. Maternal mortality rate - deaths/100,000 live births: 2.0 (Rank 181).
4. % Irrigated land - : 0.01 (Rank 168).

10.2.6 Bottom rankings of Brunei

Did you know that Brunei is very low in ranking among 193 countries in the following 5 respects?

1. Arable Land - %: 0.52 (Rank 183).
2. Airports - : 1.0 (Rank 189).
3. Airports - with paved runways - : 1.0 (Rank 184).
4. Health expenditures - % of GDP: 2.5 (Rank 184).
5. Death rate - deaths/1,000 population: 3.47 (Rank 186).

10.3 Similarity between two countries

In the “Country Comparison” Table [4] see Figure 10.6, we are showing those pairs of countries that have (of 100 important properties) between 7 and 10 properties that are similar. We count a property of two countries similar if the values differ by at most 1 %.

10.3.1 Countries having similar properties

It is interesting that no pair of countries has more than 10 similar properties and that pairs of countries with 7 or more similar properties are countries that one would intuitively think are similar, like former Yugoslav countries. Austria is a surprising exception: It has many similarities with highly developed European countries (Italy, France, UK and Germany) but also with Bulgaria!

Figure 10.6 shows countries and their similarity in terms of properties. Croatia and Slovakia (top on the list) are similar in 10 properties.

Countries Comparisons

These results are generated by comparing 100 properties. We count a property of two countries similar if the values differ by at most 1 %.
To see comparison click on a country.

Country1	Country2	Number of Similar Properties
Croatia	Slovakia	10
Czech Republic	Hungary	10
Estonia	Latvia	10
Austria	Bulgaria	9
Bulgaria	Croatia	9
Canada	United States	9
Czech Republic	Poland	9
France	United Kingdom	9
Guinea-Bissau	Togo	9
Hungary	Serbia	9
Mauritius	Maldives	9
Mexico	Colombia	9
Slovakia	Azerbaijan	9
Turkey	Colombia	9
Australia	United Kingdom	8
Austria	France	8

Figure 10.6: Similar countries

To see what those 10 properties are, users can click on the countries which redirects them to a page showing similar properties as shown in the Figure 10.7.

Did you know that Croatia is very similar to Slovakia in the following 10 respects.

1. Heliports - : (1.0, 1.0).
2. Age Structure: 15-24 years - %: (12.1, 12.2).
3. Life expectancy at birth - years: (76.41, 76.69).
4. Literacy - %: (98.9, 99.6).
5. Nobel Prize Winners - : (1.0, 1.0).
6. Refined petroleum products - production per capita - bbl/day: (0.02, 0.02).
7. Refined petroleum products - consumption per capita - bbl/day: (0.02, 0.02).
8. Refined petroleum products - exports per capita - bbl/day: (0.01, 0.01).
9. Refined petroleum products - imports per capita - bbl/day: (0.01, 0.01).
10. Telephones - mobile cellular per capita - : (1.11, 1.12).

Did you know that Croatia is very similar to Czech Republic in the following 8 respects.

1. Land boundaries - km: (1,982, 1,989).
2. Heliports - : (1.0, 1.0).
3. Mothers mean age at first birth - years: (27.7, 27.6).
4. Literacy - %: (98.9, 99.0).
5. Crude oil - imports per capita - bbl/day: (0.01, 0.01).
6. Refined petroleum products - production per capita - bbl/day: (0.02, 0.02).
7. Refined petroleum products - consumption per capita - bbl/day: (0.02, 0.02).
8. Refined petroleum products - imports per capita - bbl/day: (0.01, 0.01).

Figure 10.7: Page displaying similarity of Croatia with other countries.

Perhaps it is not surprising that the countries Argentina and Uruguay, although extremely different in size, as neighbours have 7 common characteristics, the same number of similarities between the UK and Germany. But that very different countries like Italy and Kazakhstan are similar in 7 aspects is perhaps more surprising?

10.3.1.1Similar countries to Italy

Did you know that Italy is very similar to Kazakhstan in the following 7 respects?

1. Physicians density - physicians/1,000 population: (3.8, 3.84).
2. Age Structure: 25-54 years - %: (43.0, 42.6).
3. Mothers mean age at first birth - years: (27.7, 27.6).
4. Literacy - %: (99.0, 99.7).
5. Refined petroleum products - exports per capita - bbl/day: (0.01, 0.01).
6. Refined petroleum products - imports per capita - bbl/day: (0.01, 0.01).
7. Telephones - mobile cellular per capita - : (1.58, 1.6).

10.3.2 Procedure to find similar countries

Following is a code snippet demonstrating the use of multiple MySQL cursors in stored procedure for Iteration of countries and multiple properties.

```
1: DECLARE cur1 CURSOR FOR SELECT fc.id from factbook_countries fc
2: where id in
3: (select id from un_countries)
4: DECLARE cur2 CURSOR  FOR SELECT distinct property from reportdata
```

Following is SQL query that populates similar countries table when values of two countries are 1% similar. It compares the value of selected countries with the remaining countries and selects only those countries that fulfils the criteria.

```
1: select cid,val,id,value,property,type from reportdata
2: where property=_prop
3: and id!=cid
4: and (value>=min and value<=max) and cid<id
```

SQL query which is used to find similar countries is given below. The variables “max” and “min” are set at line 3 and 4, which are used to identify the values that are 1% similar.

```
1: select value into val from reportdata where
2: id=cid and property=_prop;
3: set max=val+(val*1)/100;
4: set min=val-(val*1)/100;
5: end;
```

10.3.3 Lab showing similar properties

In the lab when two countries are compared, rows with similar values are highlighted red as shown in the Figure 10.8.

Virtual Laboratory for Exploring Countries

back

S#	Property	Austria	Bulgaria
28	Age Structure: 55-64 years - %	19.2 (9)	19.3 (8)
29	Age Structure: 65 and above - %	18.9 (8)	18.9 (9)
30	Languages -	8.0 (27)	5.0 (69)
31	Ethnic groups -	8.0 (38)	7.0 (58)
32	Population -	8,223,062 (93)	6,924,716 (100)
33	Mothers mean age at first birth - years	28.5 (17)	26.2 (40)
34	Population growth rate - %	0.01 (165)	-0.83 (191)
35	Birth rate - births/1,000 population	8.76 (185)	8.92 (181)
36	Death rate - deaths/1,000 population	10.38 (42)	14.3 (6)
37	Net migration rate - migrant(s)/1,000 population	1.76 (38)	-2.89 (156)
38	Life expectancy at birth - years	80.17 (23)	74.33 (83)
39	Urbanization - %	68.0 (67)	73.1 (54)
40	Unemployment- youth ages 15-24 - %	8.3 (109)	28.1 (27)
41	Literacy - %	98.0 (57)	98.4 (54)
15	Heliports -	1.0 (86)	1.0 (97)

Figure 10.8: Comparison of Austria and Bulgaria. Similar properties are shown in the red colour.

Besides 1 % similarity we also compute similarity based on independence year, for example the report shown in the Figure 10.9 shows the properties of countries Pakistan and India. As they both became independent in year 1947 therefore this property is marked as red in our lab.

65	Imports - $	515,100,000,000 (8)	39,270,000,000 (57)
66	GDP (purchasing power parity) - $	4,960,000,000,000 (3)	574,100,000,000 (25)
67	GDP (official exchange rate) - $	1,758,000,000,000 (11)	236,500,000,000 (43)
68	Industrial production growth rate - %	0.9 (143)	3.5 (84)
69	GDP per capita -	4,000 (137)	3,100 (144)
70	Independence -	1947-08-15	1947-08-14
71	Capital -	New Delhi	Islamabad
72	Internet hosts -	6,746,000 (17)	365,813 (55)
73	Telephones - main lines in use -	31,080,000 (10)	5,803,000 (28)
74	Telephones - mobile cellular -	893,862,000 (2)	125,000,000 (9)
75	Internet country code -	.in	.pk

Figure 10.9: Report showing properties of India and Pakistan

References

[1] "An Introduction to Stored Procedures in MySQL 5." [Online]. Available: http://code.tutsplus.com/articles/an-introduction-to-stored-procedures-in-mysql-5--net-17843. [Accessed: 22-Oct-2015].

[2] "MySQL Cursor with Example." [Online]. Available: http://www.mysqltutorial.org/mysql-cursor/. [Accessed: 22-Oct-2015].

[3] "Geography im Austria-Forum." [Online]. Available: http://austria-forum.org/af/Geography. [Accessed: 02-Nov-2015].

[4] "Countries comparison." [Online]. Available: http://austria-forum.org/af/Geography/countries_comparison. [Accessed: 19-Oct-2015].

Chapter 11 : Interactive Data Visualisations

We start this chapter using a quote by Drew Skau[1].

"Data visualisations are only good if people see them, and there is no better place to see them than on the internet, in your browser''.

Geographic databases tend to contain large sets of data, usually in the form of tables. In a few instances animation is added like in Factfish, see [2], a website with pleasant design and, as far numerical information goes, fairly sophisticated. It does lack cultural information, stories and large sets of pictures. The figures on this site often do allow animations, like showing the population growth of a country in a graph. However, the most significant part of visualising information is to compare information, and this is not offered in Factfish [2]. Based on the data taken from three main data source s(UN, Worldbank, International Energy Statistics (EIA)), we have developed interactive visualisations for countries of the world which allow comparison between countries, see [3].

Following are the research questions that are addressed in this chapter.

RQ1. How can we visualise time oriented data such as population, energy production etc.

RQ2. How can we visualise exports of countries and where the export goes to meaningfully.

RQ3. How can we visualise and bind data with globe (a fancy visualisation).

11.1 Introduction

There are three types of visualisations that we will cover in this chapter.

- Visualisation of time oriented data
- Visualising trade among countries
- Interactive world globe

Before diving into the details of these visualisations we will explain first the visualisation components, libraries and frameworks that we have used for making these visualisations.

11.2 Visualisation of time oriented data

There are some facts related to countries that are time oriented. Some typical examples are population, population growth rate and energy production etc. These properties increase or decrease with time. We will explain how we have managed to visualise them. As mentioned earlier these visualisations are fully interactive and allow users to compare trends in different countries. These visualisations allow us to look into past of countries.

We have created these visualisations considering the following tasks that are common.

Visualisation Tasks

- Direct lookup: What was the population of Greece in 1990?
- Inverse lookup: In which year was population of Greece highest?

- What is the trend of population in a particular country?
- What is the trend of population growth in comparison?

11.2.1 Visualisation Components

In this section we will explain how we have created these visualisations. There are many text elements to consider when making a chart: axis titles, data labels, gridline labels and legends. We are using D3.Js (a JavaScript library) for creating visualisations in our server. It is a de-facto standard for web-based data visualisation. D3 references to data driven documents [4]. It manipulates documents based on data. It creates charts using SVG. Recently there are few examples on the web demonstrating visualisation with D3 and canvas. It is good for complex visualisations. It is compatible with every browser that supports SVG. It was introduced by Mike Bostock.

11.2.1.1 Scales

D3 offers different types of scales. Scales are functions that map from an input domain to an output range. Linear scale is the most common scale, and a good default choice to map a continuous input domain to a continuous output range.

d3.scale.**linear** () create a linear scale. Minimum and maximum population figures are passed to the scale's input domain. Whereas height of graph container is passed as range of yScale as shown below.

```
1: yScale.domain
2: ([d3.min(population,function(d){return d;}),
3: d3.max(population,function(d){return d;})]).
4: range([height, 0])
```

11.2.1.2 Axis

D3's axis d3.svg.**axis** () component displays reference lines for scales automatically. This lets developers to focus on displaying the data, while the axis component takes care of the tedious task of drawing axes and labelled ticks. Years are displayed on X axis and population is displayed on Y axis as shown in the Figure 11.1. Graph axis titles are set according to selected visualisation using code below.

```
1: gaxis.text (function (d)
2: {if (choice=='population') return "Total Population (1000)";
3: else return "Population growth rate (5 years)";});
```

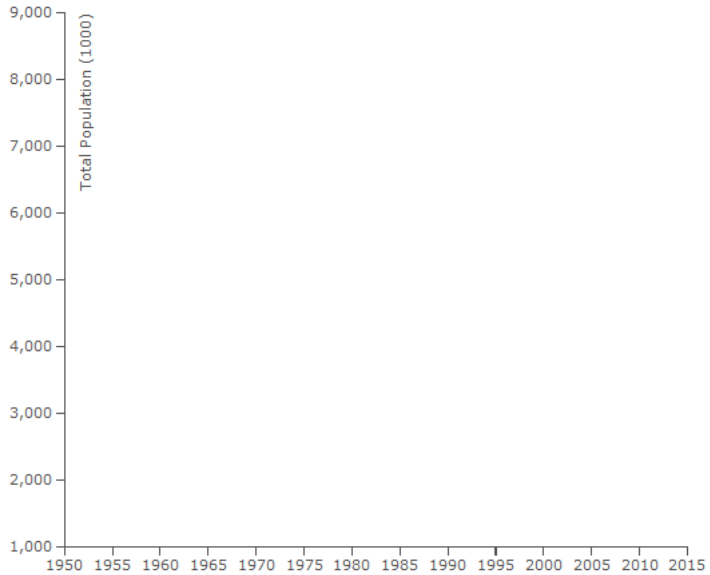


Figure 11.1: Axis setting for Visualisation

11.2.1.3Line Chart

The line chart is represented by a series of data points connected with a straight line. Line charts are most often used to visualise data that changes over time. The line chart showing population of Egypt for the last few years is shown in the Figure 11.2.

```
1: var line = d3.svg.line()
2:     .x (function(d) { return xScale(d.year); })
3:     .y (function(d) { return yScale(d.population); });
```

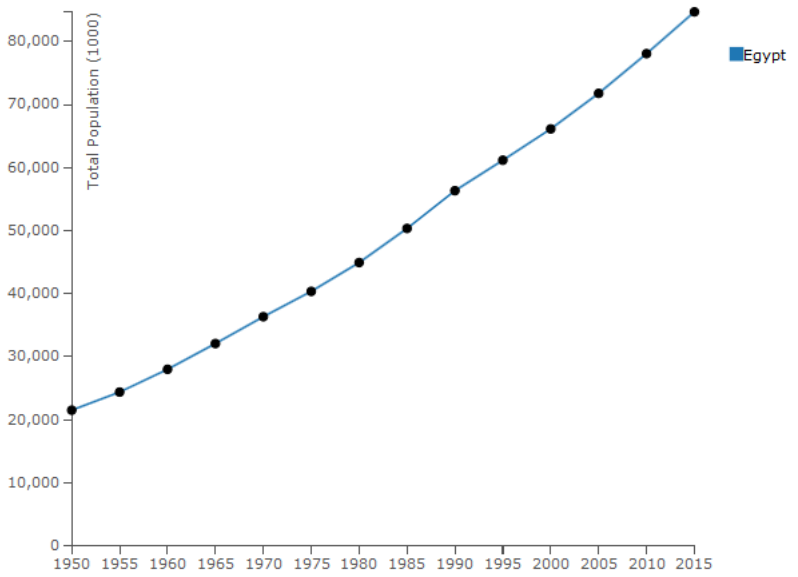


Figure 11.2: Line chart showing population of Egypt

11.2.1.4 Legends and Colours

Legends and colours are often used in visualisation to make it meaningful. As our visualisation allows to draw multiple line charts of countries simultaneously therefore in order to differentiate between them we are using legends and different colours to represent different countries. Figure 11.3 shows the usefulness of legends and colours in our visualisation. Users can easily differentiate between Egypt and Fiji using colours.

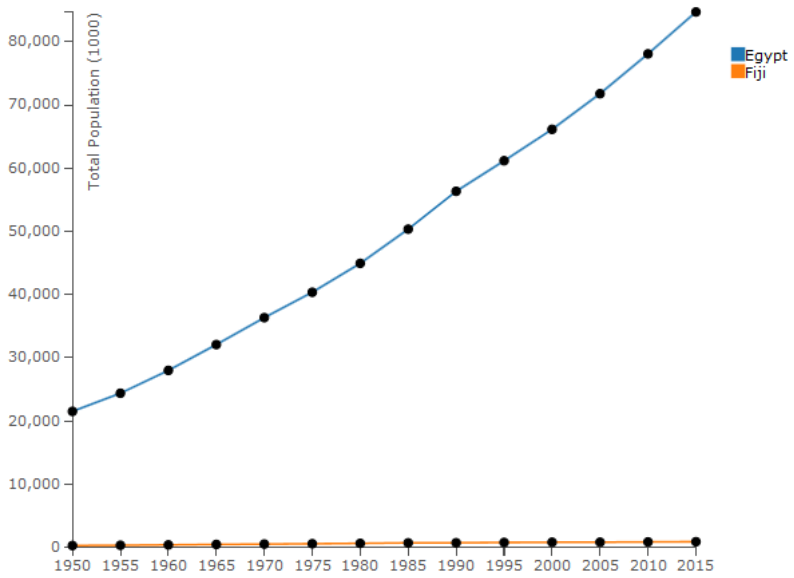


Figure 11.3: Line chart showing population of Egypt and Fiji: Users can view the difference easily using specific colour assigned to all countries shown.

11.2.2 Data Sources

Every visualisation needs data. The UN [5] and World bank [6] are two major data repositories that provide yearly data about population and other facts of countries. We have downloaded population and population growth rates figures of countries from UN. Likewise, electricity production data for past few years is downloaded from World Bank. We have downloaded CSV files of population, population growth rate and electricity production from these data sources. The data source "International Energy Statistics" [7] has all kinds of data related to "Energy" such as Natural gas production, Natural gas consumption, CO2 emissions etc.

11.2.3 Population

Our interactive visualisations allow to compare the size of the population of countries. To choose countries we employ a different interface. We do not provide a hierarchical structure for two reasons: one, the obvious structure by continents is not convenient, since it is often not clear which continent a country belongs to: Does Russia belong to Europe or Asia? (If we decide by population it is Europe, if we decide by area it is Asia). To which continents do islands in the Pacific or Caribbean belong to? On the other hand, since the names of all countries can be shown simultaneously on the screen we have decided not even to provide a search function. Actually, this decision is doubtful, since a country has often different names, like UK vs. Great Britain or Holland vs. Netherlands or even vs. The Netherlands. Note that this difficulty could be overcome in searches or by listing a country under all its common names.

Figure 11.4 shows a comparison of population between China and Switzerland. The chart shows that population of Switzerland is very low as compared to China. One can switch to population growth rate view very easily by selecting the appropriate radio button.

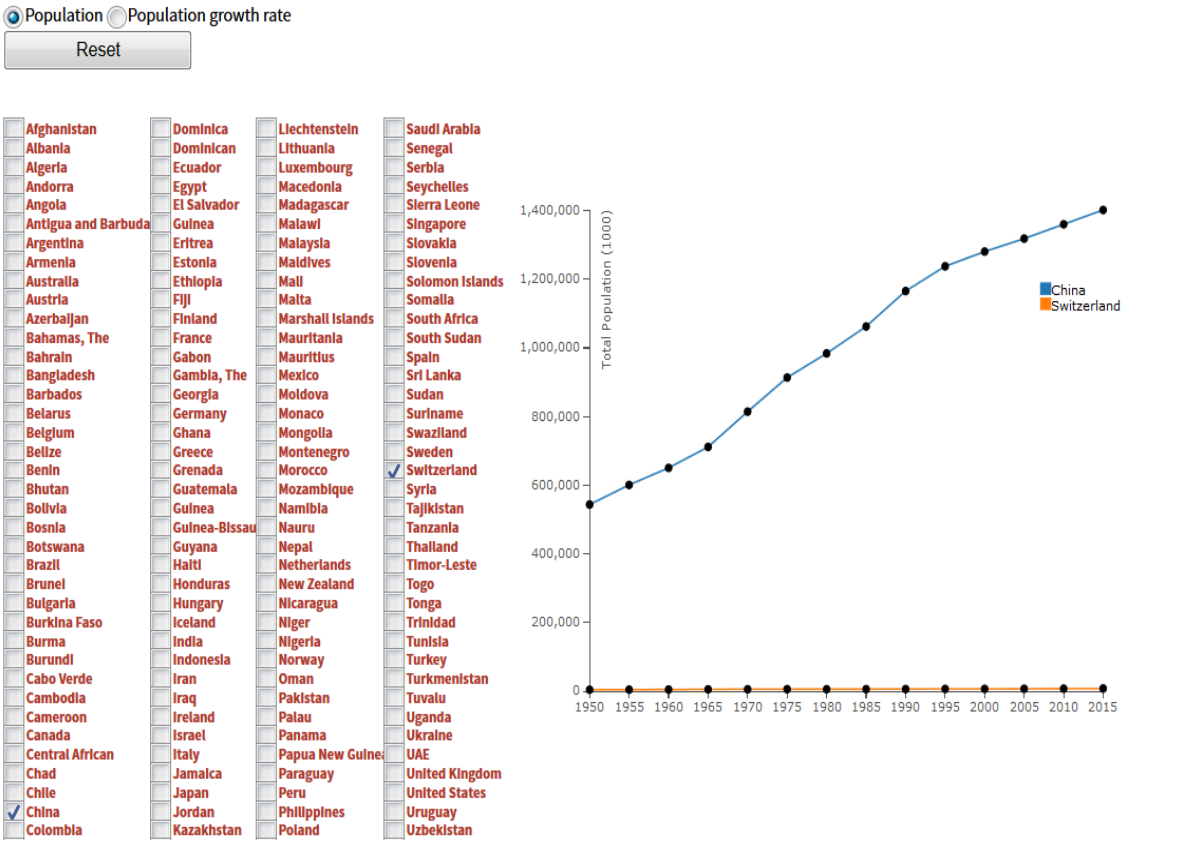


Figure 11.4

Users can compare multiple countries as shown in the Figure 11.5. It is interesting to see Germany and France are dominating among the selected European countries concerning population.

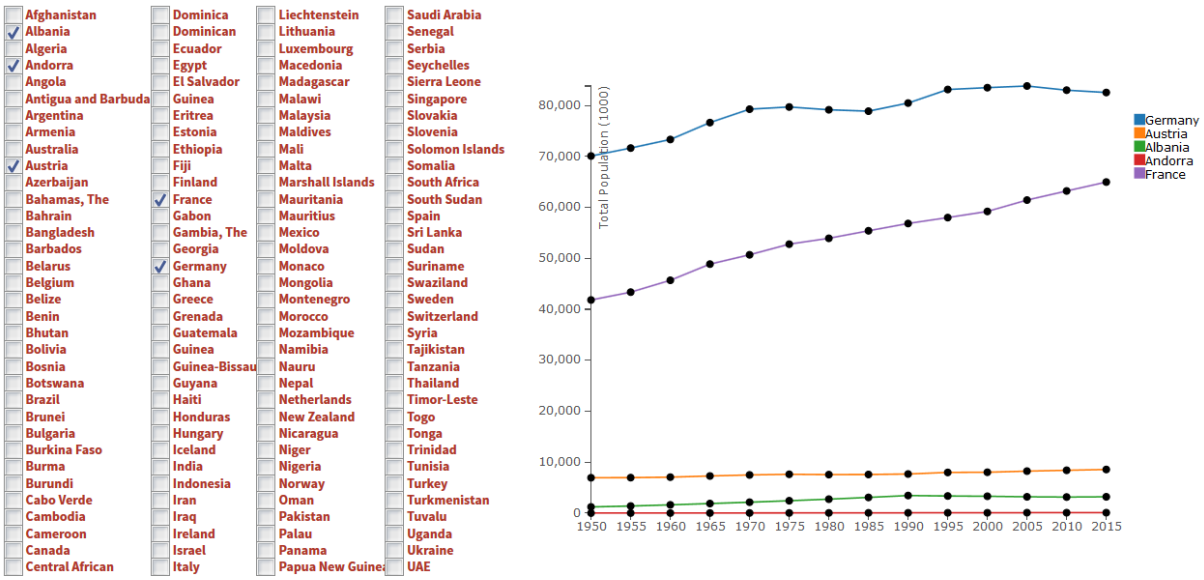


Figure 11.5

As shown in the Figure 11.6 population has dramatically increased in Pakistan after 1980. It overtakes the population of Germany which was larger than Pakistan before 1980. Users can switch to population

growth rate to figure out the cause of population increase. Pakistan has highest population growth among the countries selected as shown in the Figure 11.7.

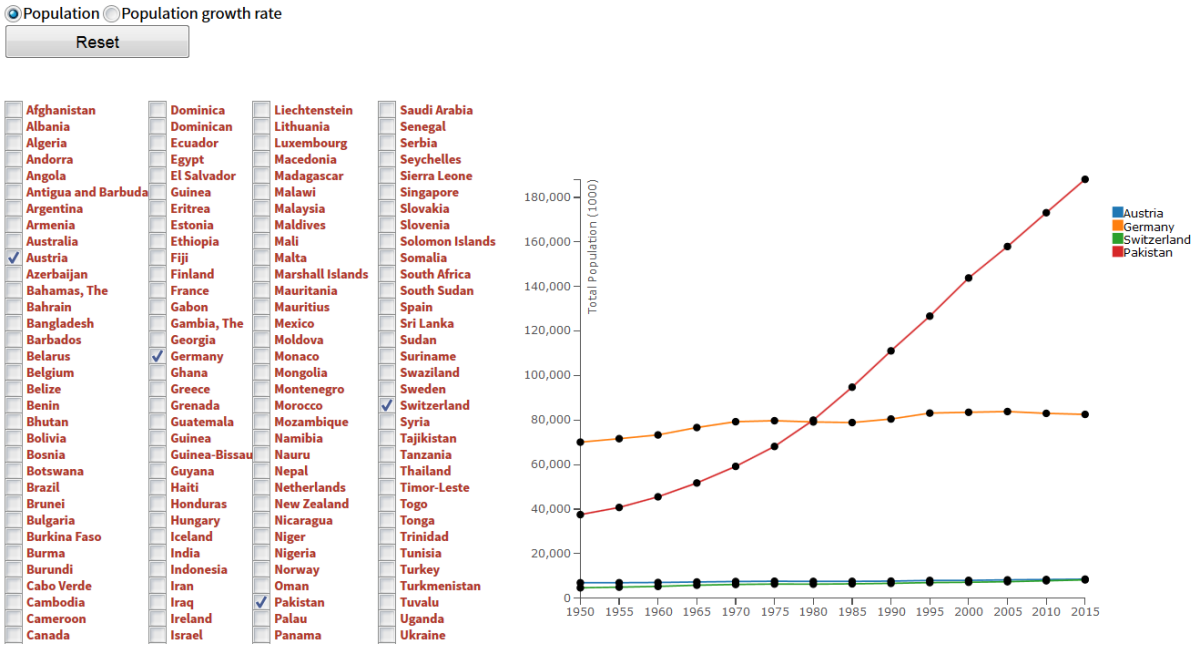


Figure 11.6

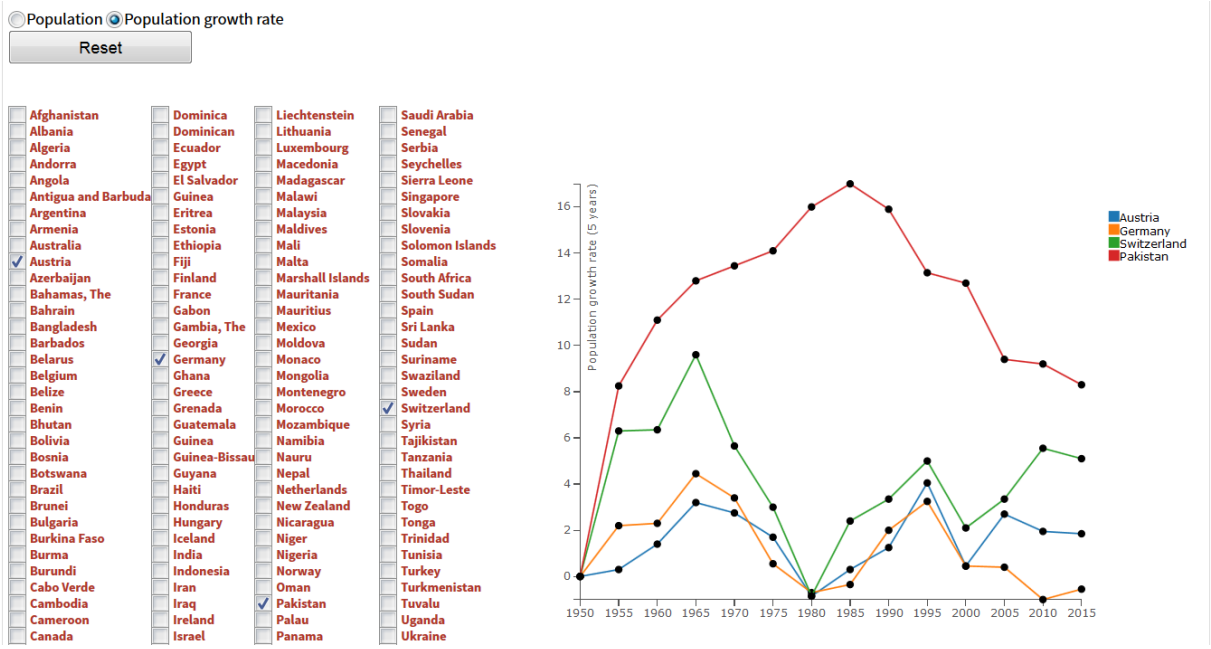


Figure 11.7

11.2.4 Population Growth Rate

Our visualization allows to visualise both the real population figures and the population growth rate. After all, in absolute figures a country like Switzerland will be almost invisible when compared to China (shown earlier in Figure 11.4). Yet the population growth can be visually compared as is shown in Figure

11.8. Switzerland has been growing at a faster relative rate than China for the last 11 years (from year 2005)! Population growth rate is calculated using the formula below.

$$Population\ growth\ rate = \frac{P(t_2) - P(t_1)}{P(t_1)(t_2 - t_1)}$$

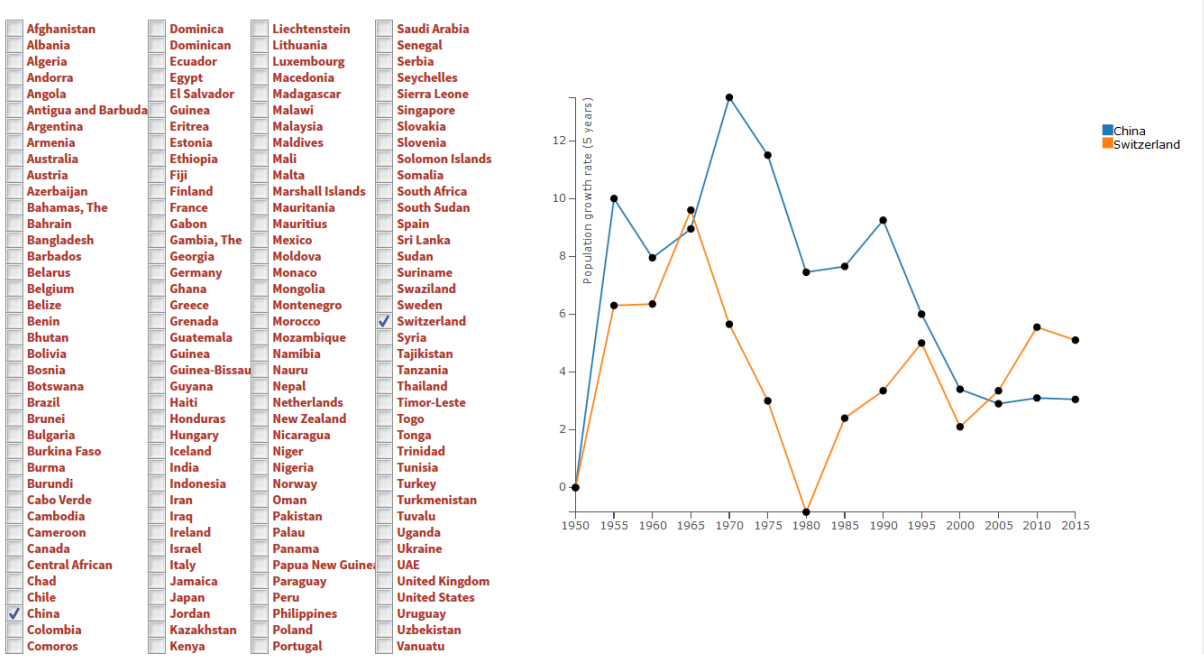


Figure 11.8

11.2.5 Electricity Production

On our geographic server we provide numerous options to visualise data. For instance users might be interested in visualising electricity production of countries. The visualisation shown in the Figure 11.9 shows the electricity production of Albania over the last few years, starting from year 2001.

One can see that in year 2010 Albania was good in electricity production followed by a decline in 2011. The electricity production of Albania was around 7.5 billion KWH in 2010, whereas in 2011 it reduces up to 4 billion KWH. Line graphs are good for showing trends and time oriented data, therefore we have used them for visualising data related to countries.

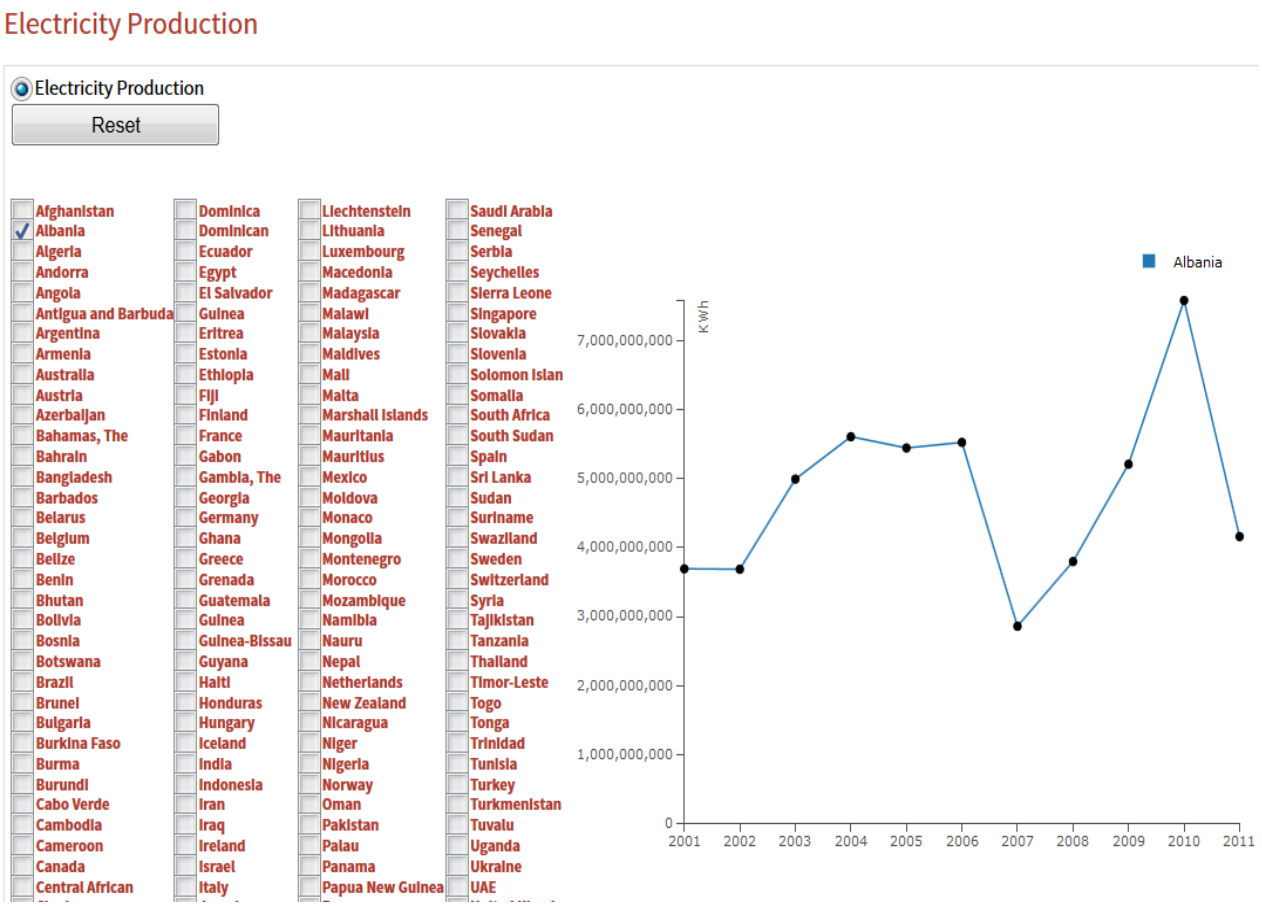


Figure 11.9

Again this visualisation allows comparison between countries.

Let us look at the chart in Figure 11.10. The electricity production (kWh) in Pakistan was last measured at 95,258,000,000 in 2011, according to the World Bank. Pakistan's electricity production is almost twice as large when compared to e.g. Algeria, see Figure 11.10. The electricity production (kWh) in Algeria was last measured at 51,224,000,000 in 2011, according to the World Bank. However, Pakistan has a five times larger population than Algeria, hence per capita Pakistan is producing only about one third of what Algeria is producing. Hence we have to note once more: When looking at comparative country statistics one has to take into account the size of the countries, both in area and population, or else all conclusions drawn will be wrong. This is why on our server we have tried to normalize many figures by showing them per capita or/and square kilometre.

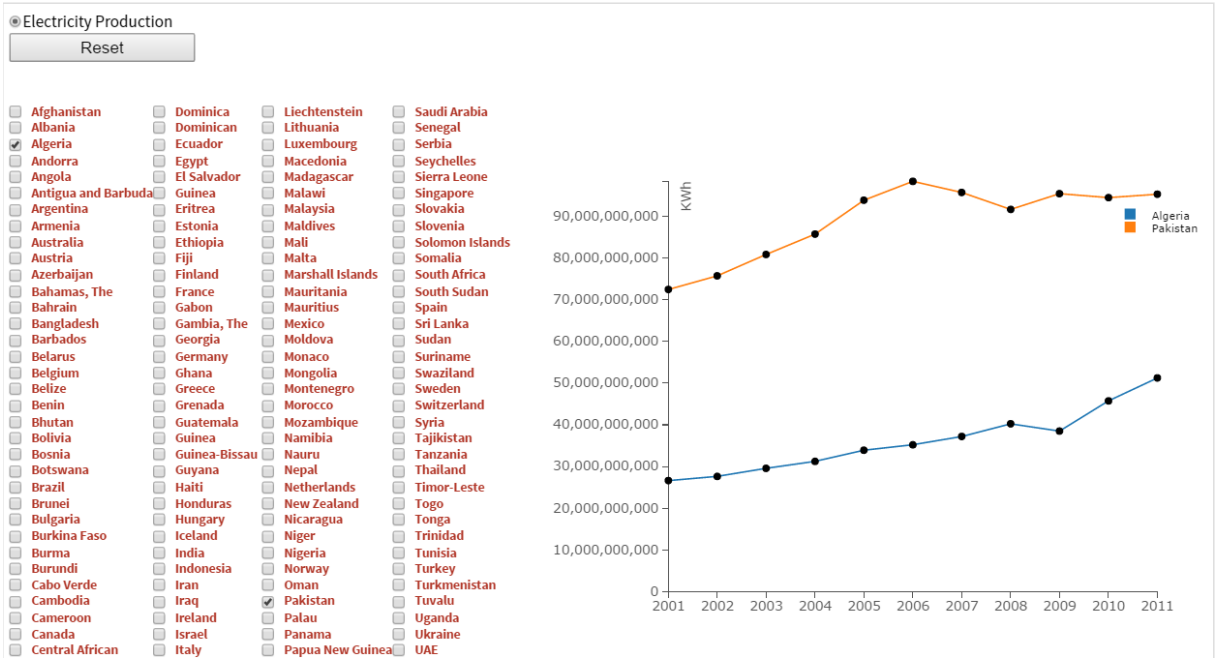


Figure 11.10

11.2.6 Electricity Consumption

The graph of electricity consumption is shown in the Figure 11.11. Lithuania is ahead to Albania in electricity consumption, although both countries have approximately the same population. The reasons for this are probably: a) Climate; b) Higher standard of living: measured by GDP per capita Lithuania's standard of living is 2.5 times higher than that of Albania.

Figure 11.11

11.3 Exports among Countries

In this section we will emphasize on an interactive visualisation that unveils trade among countries in the world. Note that these visualisation are based on parameter “Exports Partners” in Factbook database. These visualisations can be more effective if one combines both parameters “Exports Partners” and “Imports Partners”.

According to Factbook the “Exports Partners” is

“This entry provides a rank ordering of trading partners starting with the most important; it sometimes includes the percent of total dollar value.”

11.3.1 Data Preparation

Every visualisation needs data. The JSON file that is used for this visualisation is prepared using factbook data as shown in the Listing 11.1.

```
[{"exports": ["Italy", "Spain", "Greece"], "name": "Albania"},
{"exports": ["Germany", "Italy", "Switzerland", "France"], "name": "Austria"},
{"exports": ["Netherlands", "Ukraine", "Latvia"], "name": "Belarus"},
{"exports": ["Greece", "United Kingdom"], "name": "Cyprus"},
{"exports": ["Sweden", "Finland", "Latvia", "Lithuania", "Germany"], "name":
"Estonia"},
{"exports": ["Sweden", "Germany", "Netherlands", "United Kingdom"], "name":
"Finland"},
{"exports": ["Italy", "Germany", "Bulgaria"], "name": "Greece"}]
```

Listing 11.1: Few records of JSON file that is used for visualisation

11.3.2 Displaying Countries in Circular Layout

The circular hierarchical layout is shown in the Figure 11.12. It represents countries of Africa that are arranged in circular fashion. SVG transformations are applied to rotate and position country names as discussed in [8]. The countries which are not in Africa have abbreviation of their continent in square brackets.

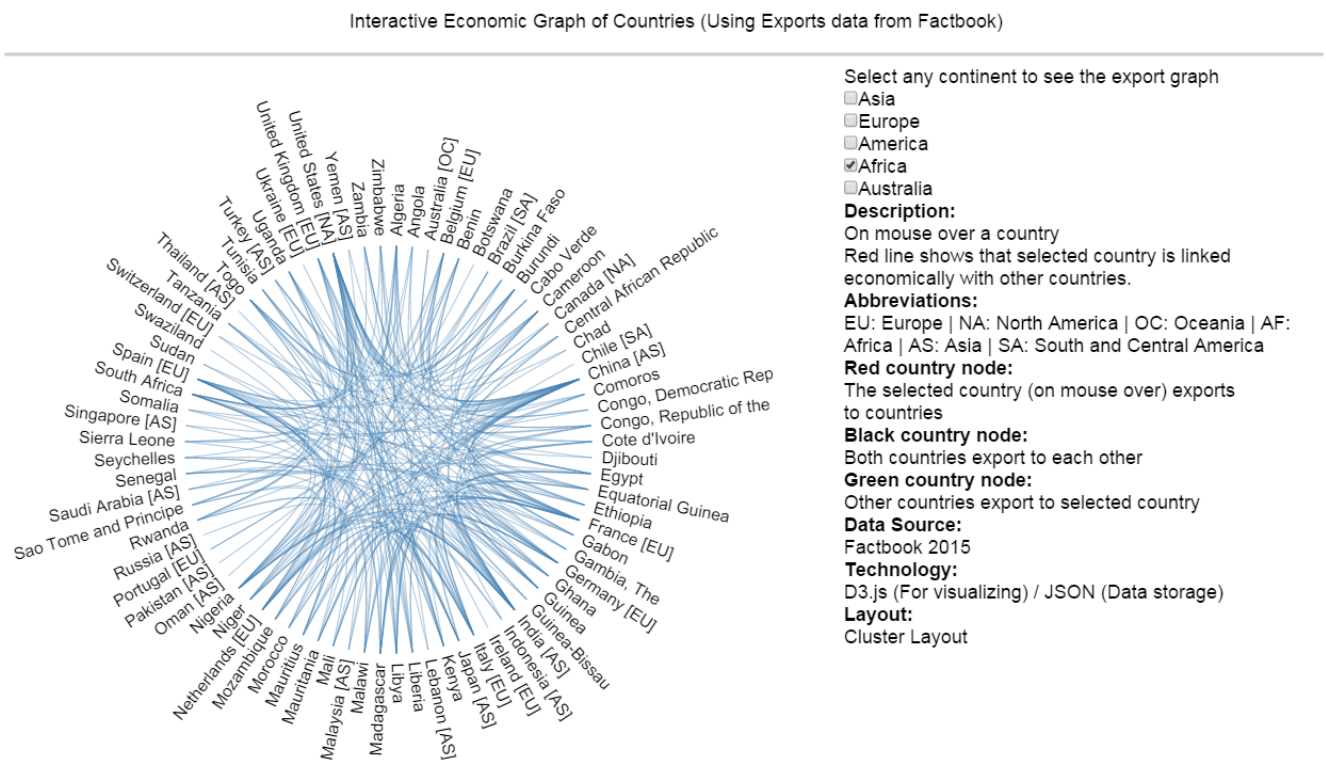


Figure 11.12

11.3.3 Interactive Visualisation

As shown in the Figure 11.13, on mouse over a country, the user can see how a particular country is linked economically with other countries using export parameter. For example when user moves the mouse over Austria, the countries are highlighted in different colours. Austria exports goods to Switzerland, France and Italy. Germany and Austria export to each other. Whereas countries like Bosnia, Czech Republic, Slovenia and Slovakia export their goods to Austria.

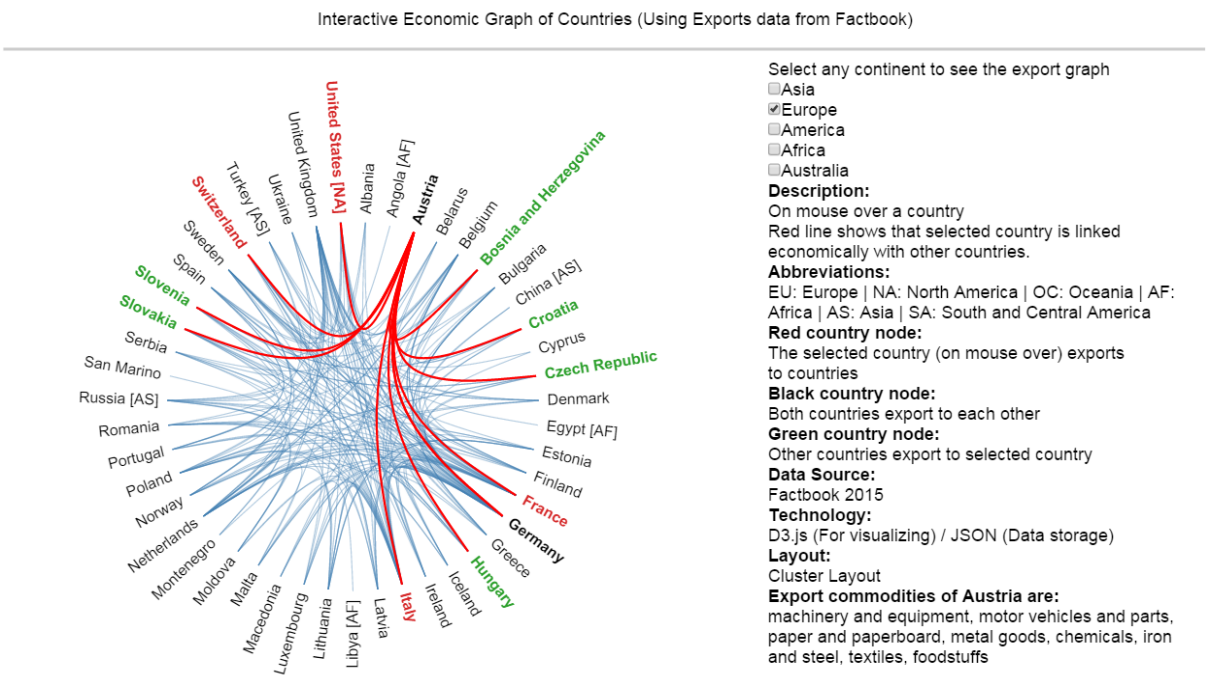


Figure 11.13: Interactive circular layout that displays trade of Austria.

Let us now look at Bahrain. Bahrain exports to United Arab Emirates, Saudi Arabia, India and Korea, South as shown in the Figure 11.14.

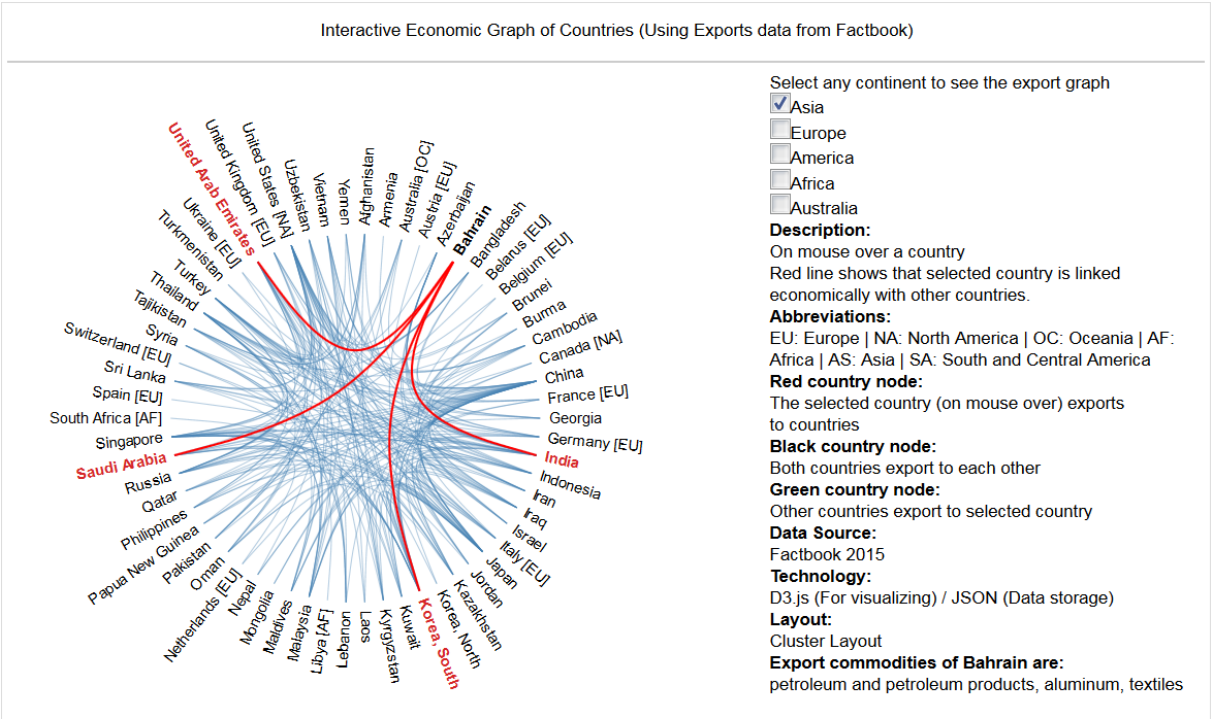


Figure 11.14

11.3.4 Displaying Continent-wise Exports

We have extended this export visualisation to all continents. Users can select one continent at a time to see the graph of countries in that continent using checkboxes as shown in the Figure 11.15. All the countries are sorted in alphabetical order which makes it easy to find a particular country. Some

European countries are shown in the graph of “Asia” because some Asian countries export to European countries. For example Kazakhstan exports to Austria as shown in the Figure 11.16.

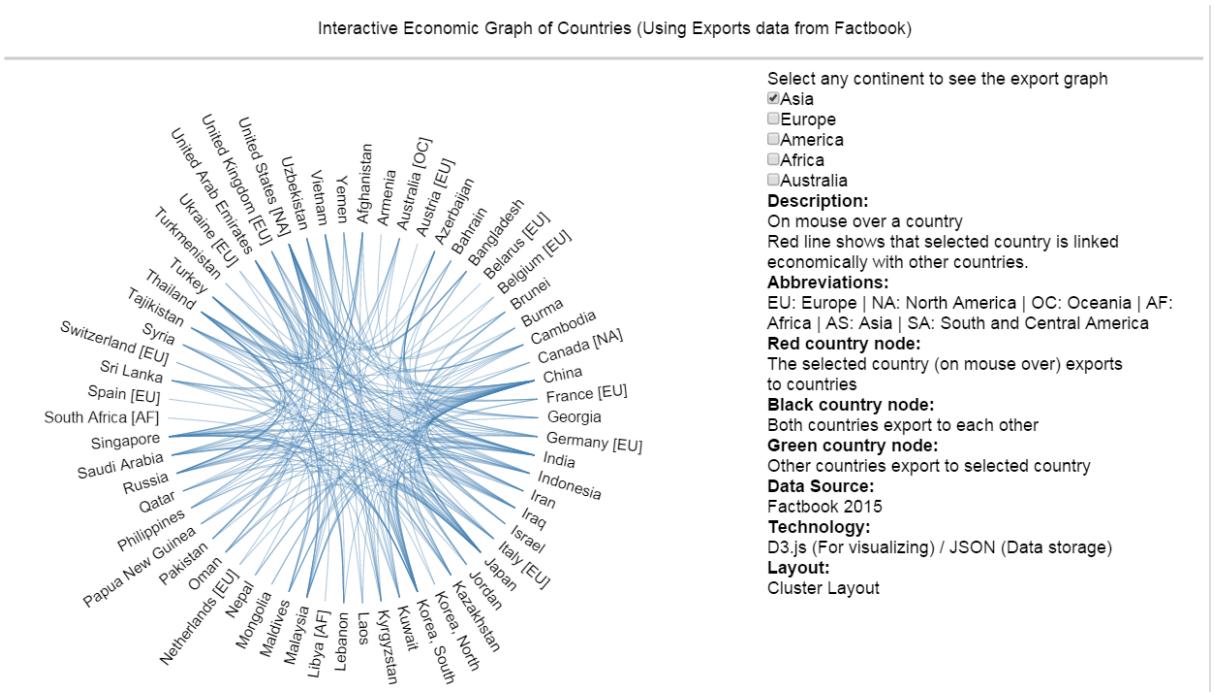


Figure 11.15

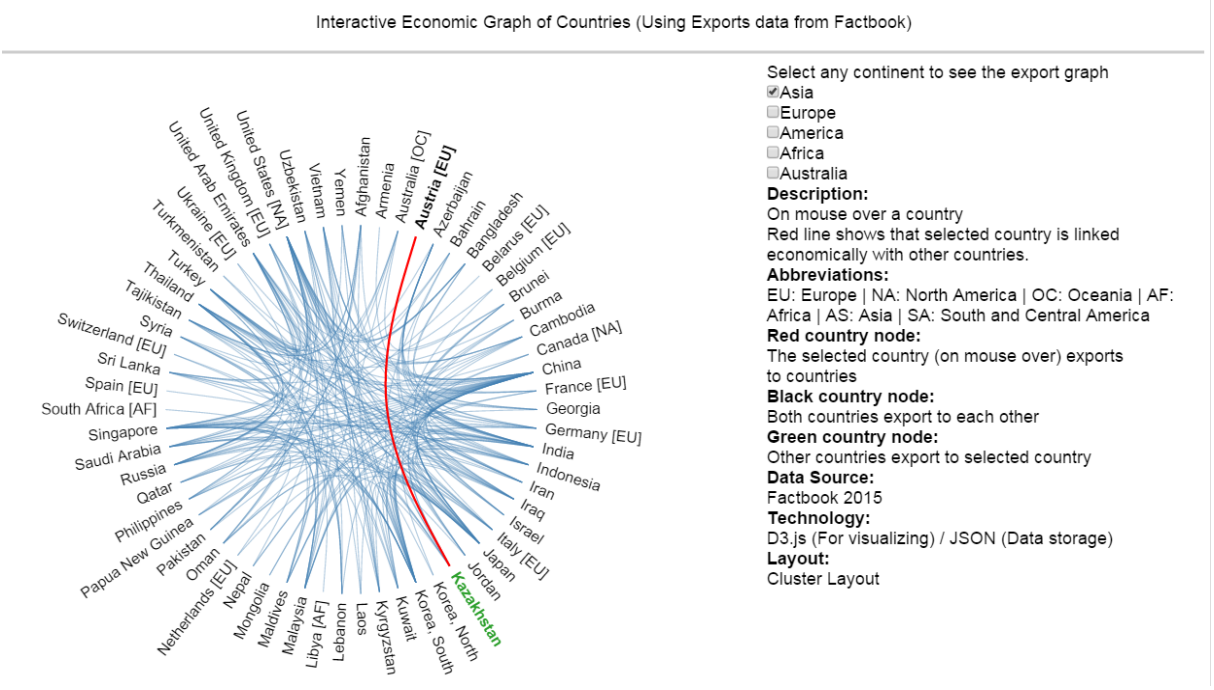


Figure 11.16

11.4 Interactive Globe

D3.js is not only used for data visualisations but it also has remarkable capabilities when it comes to mapping [9]. We have used orthographic projection to design globe. We in fact want to give users a spherical view of earth as it appears from outer space, see Figure 11.17.

We have used orthographic projection using code below.

```
var projection = d3.geo.orthographic ()
```

Property: Highest Rank Country	Property: Highest Rank Country
Total Area: Russia	Natural gas - consumption: United States
Land Area: Russia	Natural gas - exports: Russia
Land boundaries: China	Natural gas - imports: Japan
Highest Point: Nepal	Natural gas - proved reserves: Russia
Coastline: Canada	CO2 emissions: China
Irrigated Land: India	Exports: China
Arable Land: Denmark	Imports: United States
Freshwater withdrawal: India	GDP (purchasing power parity): United States
Freshwater withdrawal (pc): Turkmenistan	GDP (official exchange rate): United States
Airports: United States	Industrial production growth rate: Sierra Leone
Heliports: United States	GDP (pc): Qatar
Railways: United States	Internet hosts: United States
Roadways: United States	Telephones - main lines in use: China
Waterways: China	Telephones - mobile cellular: China
Airports - with paved runways: United States	Nobel Prize Winners: United States
Airports - with unpaved runways: United States	Electricity - production (pc): Iceland
Education expenditures: Lesotho	Electricity - consumption (pc): Iceland
Health expenditures: Liberia	Electricity - exports (pc): Bhutan
Physicians density: Monaco	Electricity - imports (pc): Luxembourg
Hospital bed density: Monaco	Crude oil - production (pc): Kuwait
Age Structure: 0-14 years: Niger	Crude oil - exports (pc): Qatar
Age Structure: 15-24 years: Maldives	Crude oil - imports (pc): Singapore

Figure 11.17: Globe Visualisation in D3.js

11.4.1 Data

11.4.1.1 Map Data

For smooth loading, processing and rotation of map in browser we are using TopoJSON file instead of GeoJSON. TopoJSON is best explained by “Mike Bostock” in [10].

“TopoJSON is an extension of GeoJSON that encodes topology. Rather than representing geometries discretely, geometries in TopoJSON files are stitched together from shared line segments called *arcs”.*

The primary advantage of TopoJSON is size. TopoJSON files are often an order of magnitude smaller than GeoJSON files, therefore provides best user experience.

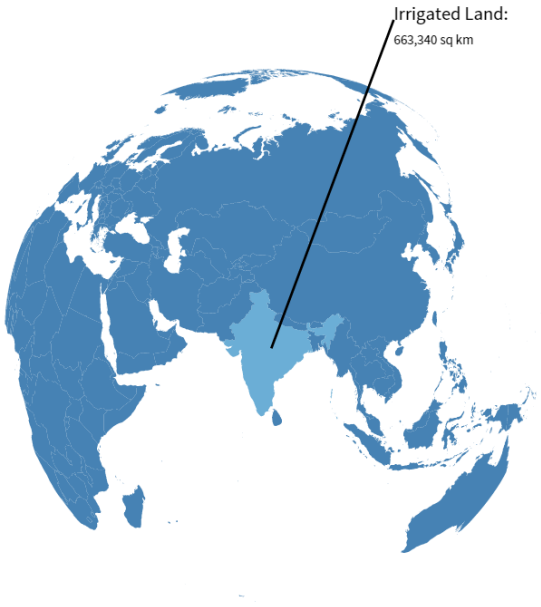
11.4.1.2 Statistical Data

On the right side of globe, some properties (around 100) are listed along with the countries which have highest rank in a particular property. For example China has highest number of water ways. United States has highest number of roadways. Russia has the largest area among countries of the world. We have already generated rankings of countries according to different properties, see chapter 9, 10.

11.4.2 Interactivity

On mouse over a country or property we rotate the globe and show the country which has highest rank in a certain property, see Figure 11.18. The globe is rotated and continent Asia is shown on the front in this view.

The globe rotation is done via parameters (yaw- λ, pitch- ϕ and roll-γ). The line <svg line> is drawn from centre of the map. Further, the map of India is highlighted. India has largest irrigated land area 663,340 sq. km.



Property: Highest Rank Country	Property: Highest Rank Country
Total Area: Russia	Natural gas - consumption: United States
Land Area: Russia	Natural gas - exports: Russia
Land boundaries: China	Natural gas - imports: Japan
Highest Point: Nepal	Natural gas - proved reserves: Russia
Coastline: Canada	CO2 emissions: China
Irrigated Land: India	Exports: China
Arable Land: Denmark	Imports: United States
Freshwater withdrawal: India	GDP (purchasing power parity): United
Freshwater withdrawal (pc): Turkmenistan	States
Airports: United States	GDP (official exchange rate): United States
Heliports: United States	Industrial production growth rate: Sierra
Railways: United States	Leone
Roadways: United States	GDP (pc): Qatar
Waterways: China	Internet hosts: United States
Airports - with paved runways: United	Telephones - main lines in use: China
States	Telephones - mobile cellular: China
Airports - with unpaved runways: United	Nobel Prize Winners: United States
States	Electricity - production (pc): Iceland
Education expenditures: Lesotho	Electricity - consumption (pc): Iceland
Health expenditures: Liberia	Electricity - exports (pc): Bhutan
Physicians density: Monaco	Electricity - imports (pc): Luxembourg
Hospital bed density: Monaco	Crude oil - production (pc): Kuwait
Age Structure: 0-14 years: Niger	Crude oil - exports (pc): Qatar
Age Structure: 15-24 years: Maldives	Crude oil - imports (pc): Singapore

Figure 11.18: Property and values are shown on mouse over a country

Figure 11.19 shows another view of globe. United States is shown highlighted. The globe is rotated and continent America is shown on the front this time. United States has largest number of airports in the world.

Property: Highest Rank Country	Property: Highest Rank Country
Total Area: Russia	Natural gas - consumption: United States
Land Area: Russia	Natural gas - exports: Russia
Land boundaries: China	Natural gas - imports: Japan
Highest Point: Nepal	Natural gas - proved reserves: Russia
Coastline: Canada	CO2 emissions: China
Irrigated Land: India	Exports: China
Arable Land: Denmark	Imports: United States
Freshwater withdrawal: India	GDP (purchasing power parity): United
Freshwater withdrawal (pc): Turkmenistan	States
Airports: United States	GDP (official exchange rate): United States
Heliports: United States	Industrial production growth rate: Sierra
Railways: United States	Leone
Roadways: United States	GDP (pc): Qatar
Waterways: China	Internet hosts: United States
Airports - with paved runways: United	Telephones - main lines in use: China
States	Telephones - mobile cellular: China
Airports - with unpaved runways: United	Nobel Prize Winners: United States
States	Electricity - production (pc): Iceland
Education expenditures: Lesotho	Electricity - consumption (pc): Iceland
Health expenditures: Liberia	Electricity - exports (pc): Bhutan
Physicians density: Monaco	Electricity - imports (pc): Luxembourg
Hospital bed density: Monaco	Crude oil - production (pc): Kuwait
Age Structure: 0-14 years: Niger	Crude oil - exports (pc): Qatar
Age Structure: 15-24 years: Maldives	Crude oil - imports (pc): Singapore

Figure 11.19: Globe rotation: North America is shown on the front in this view.

11.5 Choropleth/Thematic map

A choropleth map is a thematic map in which areas are shaded in proportion to the measurement of the statistical variable being displayed on the map. A choropleth map showing 'Physician density' of Europe is shown in the Figure 11.20. Greece has high physician density around 6 physicians/1,000 population. Austria and Norway have physician density of around 4.9 and 4.1 physicians/1,000 population respectively. Leaflet provides several built in features for making choropleth map, see [11].

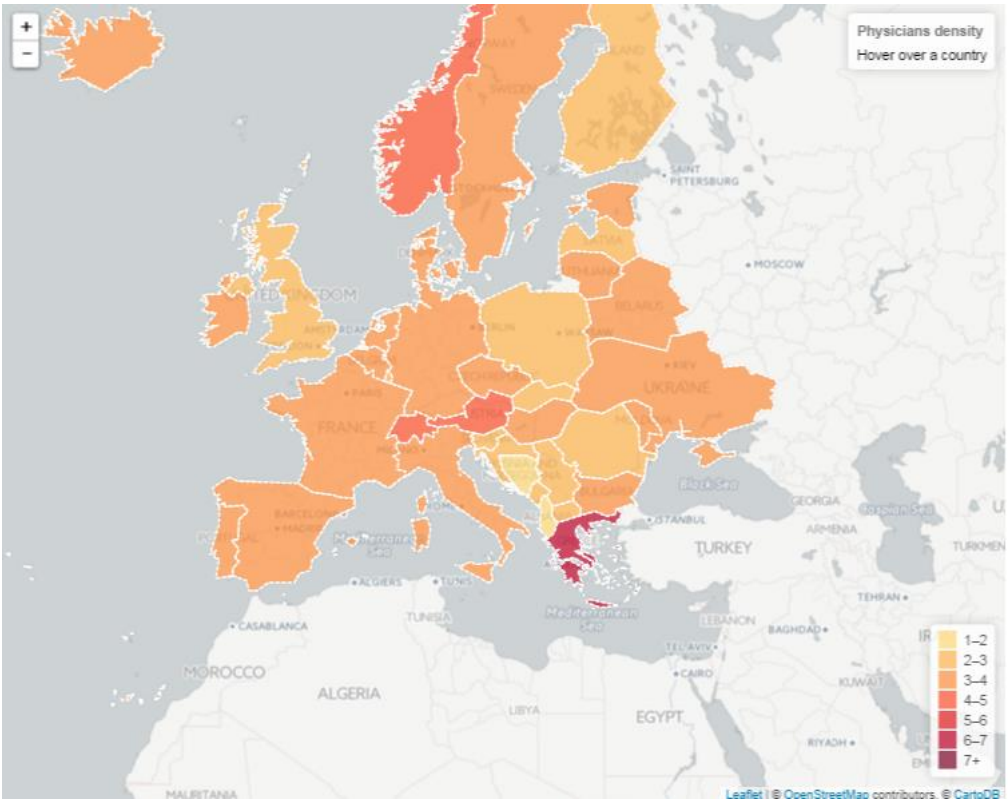


Figure 11.20: Choropleth map

The choropleth map of Europe based on hospital bed density is shown in the Figure 11.21. Ukraine, Belarus, Germany have higher hospital bed density compared to other European countries.

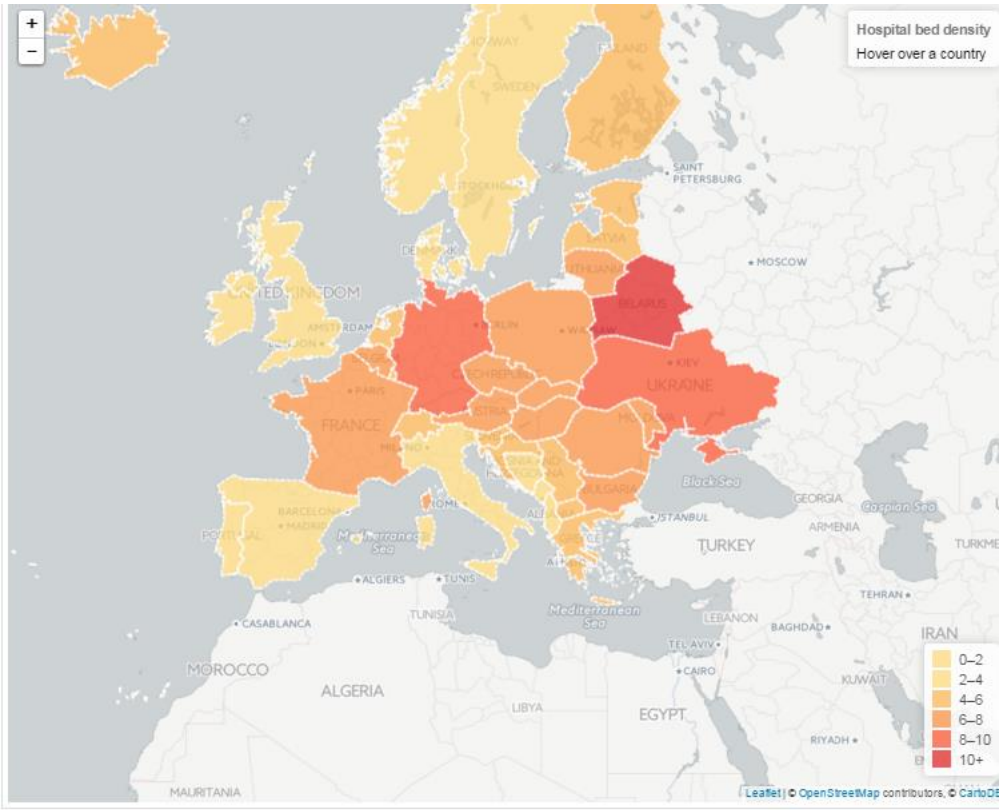


Figure 11.21

11.6 Accessing Visualisations

These visualisation can be accessed via “Visualisations” Web page shown in the Figure 11.22.

Figure 11.22

References

[1] "Drew Skau - Visually Blog." [Online]. Available: http://blog.visual.ly/author/drew/. [Accessed: 17-Feb-2016].

[2] "Factfish World Statistics and Data Research." [Online]. Available: http://factfish.com/. [Accessed: 22-Oct-2015].

[3] "Population." [Online]. Available: http://austria-forum.org/af/Geography/Visualizations/Population. [Accessed: 22-Oct-2015].

[4] "D3 — Scott Murray — alignedleft." [Online]. Available: http://alignedleft.com/tutorials/d3. [Accessed: 07-Feb-2016].

[5] "UNdata." [Online]. Available: http://data.un.org/. [Accessed: 24-Feb-2016].

[6] "Data- The World Bank." [Online]. Available: http://data.worldbank.org/. [Accessed: 24-Feb-2016].

[7] "U.S. Energy Information Administration (EIA)." [Online]. Available: http://www.eia.gov/. [Accessed: 24-Feb-2016].

[8] "Understanding SVG Coordinate Systems and Transformations (Part 2) — The <code>transform</code> Attribute." [Online]. Available: http://sarasoueidan.com/blog/svg-transformations/. [Accessed: 22-Oct-2015].

[9] "Building Great Web Maps: A D3.js Tutorial." [Online]. Available: https://www.toptal.com/javascript/a-map-to-perfection-using-d3-js-to-make-beautiful-web-maps. [Accessed: 05-Mar-2016].

[10] "TopoJSON." [Online]. Available: https://github.com/mbostock/topojson/wiki. [Accessed: 05-Mar-2016].

[11] "Interactive Choropleth Map." [Online]. Available: http://leafletjs.com/examples/choropleth.html. [Accessed: 05-Mar-2016].

Chapter 12 : Mapping Services of Geographic Server

This chapter tries to provide a description of geo visualisations (maps) that global-geography.org is offering. There were several challenges regarding data collection and visualisation of maps that we will highlight in this chapter.

For decades, most geographic information was confined to use on desktop-based PCs and could not be easily shared with other organizations[1]. With the arrival of Web the display of maps on the Web has grown dramatically. The first step was the display of static images of maps on HTML pages. With the emergence of mapping libraries such as leaflet, open layers etc. it became easy to create mapping applications.

This chapter will address the following research questions.

RQ1. How can we extend the map section of global-geography.org using both static maps and vector maps?

RQ2. How can we provide easy navigation of different types of maps for a particular country? (Using a suitable plugin)

RQ3. How can we create a service that allows creation of markers on the map?

RQ4. How can we generate images of vector maps that can be used in a slide show plugin for better user experience?

RQ5. How can we store information in markers generated by users? (Permanent storage mechanism)

12.1 Introduction

We are storing two types of maps for each country as shown in the Figure 12.1.

1. Static map images
2. Vector maps

The users can interact with vector maps and create markers of their own choice.

Figure 12.1: Slide show of different maps of Austria

We have used Leaflet for creation of maps. Leaflet is a leading open source JavaScript library[2]. It allows to create maps on browser with much ease. Leaflet has numerous map features e.g. tile layers, popup, markers, etc. It allows to draw vector layers like polylines, circles or rectangles on the maps. There are hundreds of nice plugins, see [3], for specific functions.

Leaflet is only a framework for showing and interacting with map data, but we have to provide map data, see Section 12.3 about map data collection. The most basic setting for the creation of maps in Leaflet is shown below. The L.map () is used to initialize the map object. It requires the id of the div ('map' in this case) where we want to show the map. The setView() method is used to center the initial map view on area that is specified by latitude(42.35) and longitude(-71.08), see [2] [4] for more details about map creation in Leaflet.

```
var map = L.map('map').setView([42.35, -71.08], 13);
```

12.2 Maps in Global Geography

As mentioned earlier, we are providing at the moment two types of maps, but plan to add more as the project grows and becomes more international.

12.2.1 Static Maps

The two maps shown in the Figure 12.2 and 12.3 are taken from Fact book [5]. These maps are just static images. One represents a view of the country along with location of bigger cities, in Austria like Vienna, Graz, Linz, Innsbruck, Salzburg and Klagenfurt. The other map in Figure 12.3 describes the position of country on its continent.

The java code to extract maps from Fact book is given below:

```
URI="https://www.cia.gov/library/publications/the-world-
factbook/geos/"+rs.getString(1)+".html";
final WebClient client = new WebClient();
final HtmlPage mainPage = client.getPage(URI);
List<?> links1=null;
links1 = (List<?>) mainPage.getByXPath("//td[@class='area']/a/img");
```

Figure 12.2: Map of Austria

Austria in Europe

Figure 12.3

12.2.2 Vector Maps

The vector map shows the country, its districts and capital.

It has three special features:

- It allows to access the capital city of the country (blue marker). By clicking; it leads to the OpenStreetMap[6] with the capital city in the centre.
- The square icons give the name of the districts/provinces.
- The red markers can be placed by any person authorized to do so. A click provides whatsoever information was decided to attach to the marker, i.e. can be textual, a pointer to some picture or pictures, or to some Web page(s).

  We are showing all three types of marker on the map for Germany in Figure 12.4. Few more examples will be presented in section 12.4.
  Note: Markers can be turned on and off using a layer control, see section 12.5.

Figure 12.4: Map of Germany

12.3 Visualisation Pipeline

The visualisation pipeline is shown in the Figure 12.5. It shows the process of generating visual representations in the form of maps in browser.

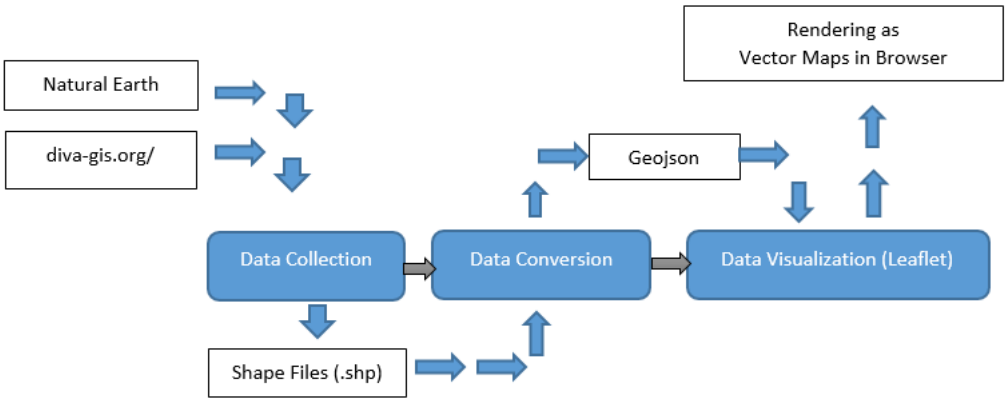


Figure 12.5

Before you create geographic maps, you have to collect map data. We have first downloaded the shape files with (.shp) extension from Natural Earth and Diva (open source map data providers). The main interface of Diva GIS is shown in the Figure 12.6

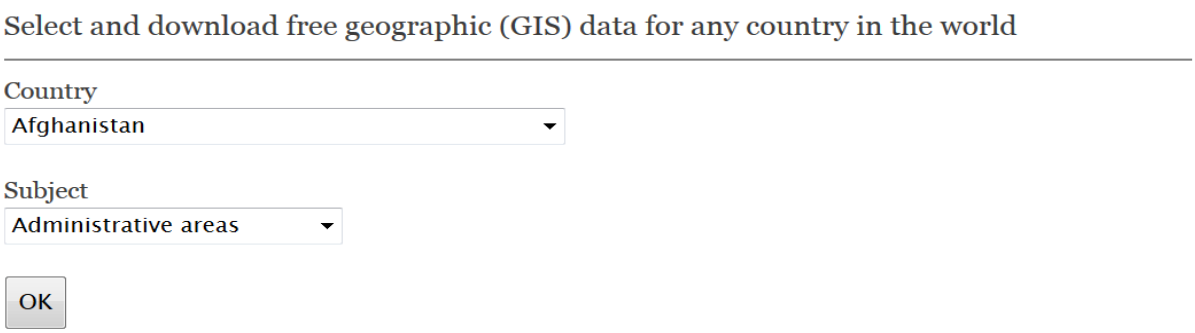


Figure 12.6

More specifically, we have downloaded the required shape files of all countries automatically from the above interface of DIVA GIS[18] which offers free shape files of each country, using the script shown below.

```
String saveTo = "C:\\Users\\IICM\\Desktop\\shapefiles\\"
try {
            String name=rs.getString(1)+"_adm"+".zip";
            URL url1 = new URL("http://biogeo.ucdavis.edu/data/diva/adm/"+name);
            URLConnection conn = url1.openConnection();
            InputStream in = conn.getInputStream();
            FileOutputStream out = new FileOutputStream(saveTo + name);
            byte[] b = new byte[1024];
            int count;
            while ((count = in.read(b)) >= 0) {
            out.write(b, 0, count);
    }}
```

The next step was the creation of GeoJSON files. It is essential to convert shape files into GeoJSON files for loading them into leaflet. We have used the following command to convert shape files into GeoJSON.

```
$ ogr2ogr -f GeoJSON  [AUS_admn].GeoJSON [AUS_admin].shp
```

The leaflet loads these GeoJSON files and displays them as vector maps in the respective html container as shown in the Figure 12.7. The user can further zoom and pan these maps around.

What is a GeoJSON file? Is best described by Scott Murray in his book, see [7].

"Just like JSON is formalization of existing JavaScript object syntax, GeoJSON is a formalized syntax of JSON objects, optimized for storing geodata. GeoJSON can store points in geographical space (typically as longitude/latitude coordinates) but also shapes (lines, polygons)"

A typical GeoJSON file is shown in the Listing 12.1.

```
{"type": "FeatureCollection",
"features": [
{ "type": "Feature",
 "properties": { "ID_0": 17, "ISO": "AUT", "NAME_0": "Austria", "ID_1": 162,
"NAME_1": "Burgenland", },
 "geometry":
 {
 "type": "Polygon",
 "coordinates": [ [ [ 17.074972, 48.076299 ], [ 17.070900, 48.074092 ], [
17.074699, 48.072745 ], [ 17.089402, 48.067535 ].........
 ]}
```

[18] http://www.diva-gis.org/datadown

Listing 12.1: A typical GeoJSON file storing polygons about province Burgenland of Austria.

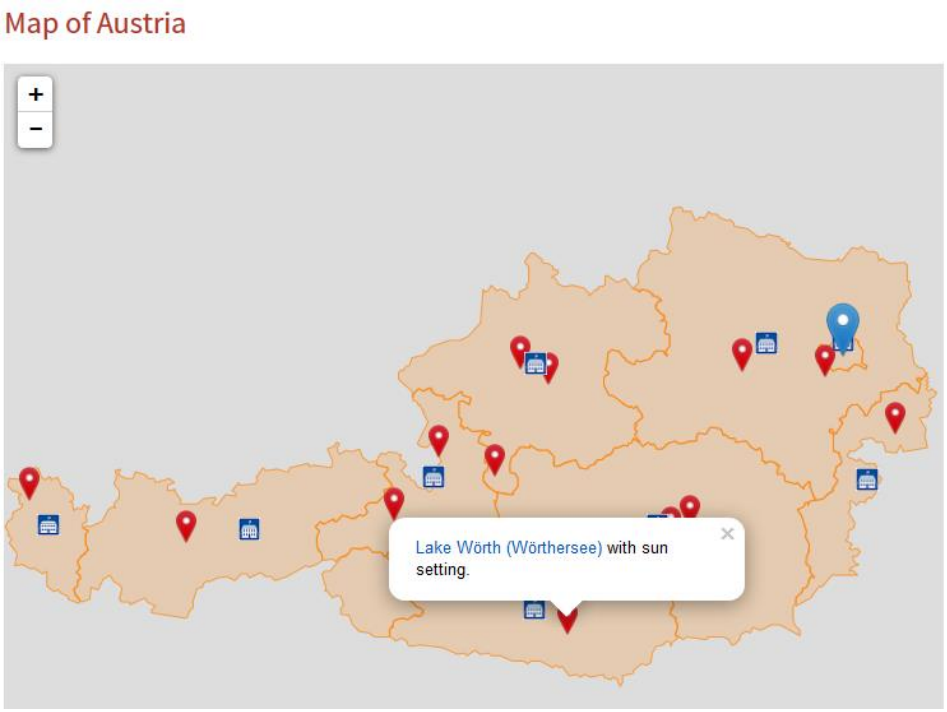


Figure 12.7: Vector map of Austria with popups and different markers. We explain the reason for the Zoom Function in Section 12.10

12.4 Dynamic Markers Generation on Maps

In order to draw capital markers of each country on the map we have used decimal degrees using the formula below.

Degree + Minutes/60 + Seconds/3600 →Decimal Degree

For example CIA Fact book gives geographic coordinates of Islamabad as

33 41 N AND 73 03 E

When it is expressed in degree notation it is written as

33.68 And 73.05

The big blue marker represents capital marker that is displayed on each country map. These markers provide links that take users to open street map. The code to create and link popup with OpenStreetMap is shown below:

```
L.marker([ lat, lon ], {
icon : mIcon
}).addTo(capitalmarker).bindPopup(
"Capital: <a href=http://www.openstreetmap.org/search?query='"
+ capital + "'#map=7/" + lat + "/" + lon
+ "&layers=H target=_blank />" + caplabel);
```

Registered users of Austria-Forum or global-geography can create markers on map by simply clicking anywhere on the map as shown in the Figure 12.8. Afterwards, they fill the textual information and links in the textbox shown in the Figure 12.8.

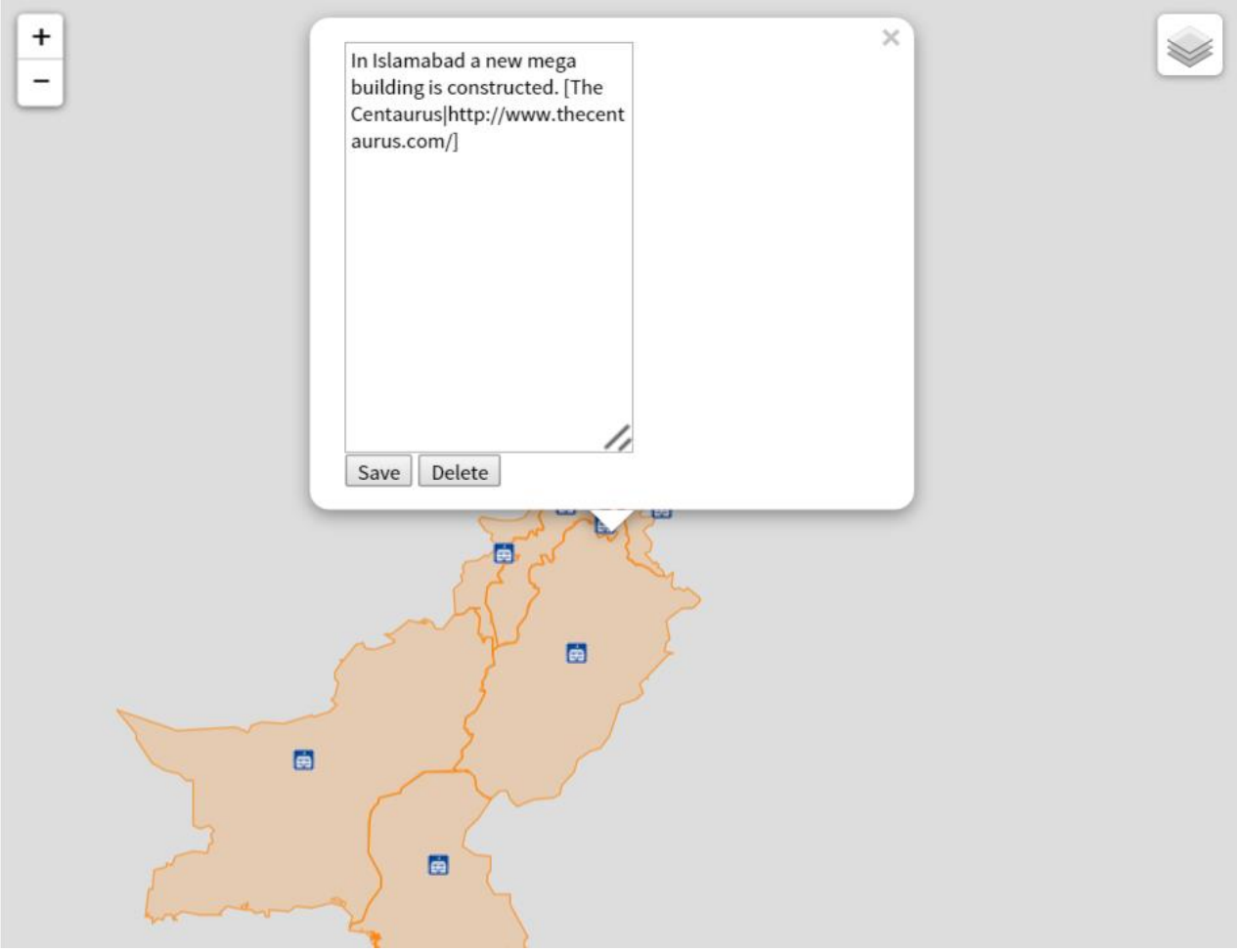


Figure 12.8: Vector map of Pakistan. Registered users can create marker and popups on the map.

12.4.1 Markers Generation using Latitudes and Longitudes

We have provided flexible options to users for markers creation. They can create markers either by clicking on the map or by entering latitude longitude values in the required textbox shown in the Figure 12.9. One can easily find latitude and longitude of a location using Google maps. For example in the URL of Google maps below, showing location of “Hekla volcano”, the text marked as red represents latitude and longitude.

```
https://www.google.at/maps/place/Hekla,+Island/@63.9921823,-
19.9459926,10z/data=!3m1!4b1!4m2!3m1!1s0x48d6cfb1e8291f6d:0x46cf34771f2838b3?hl=de
```

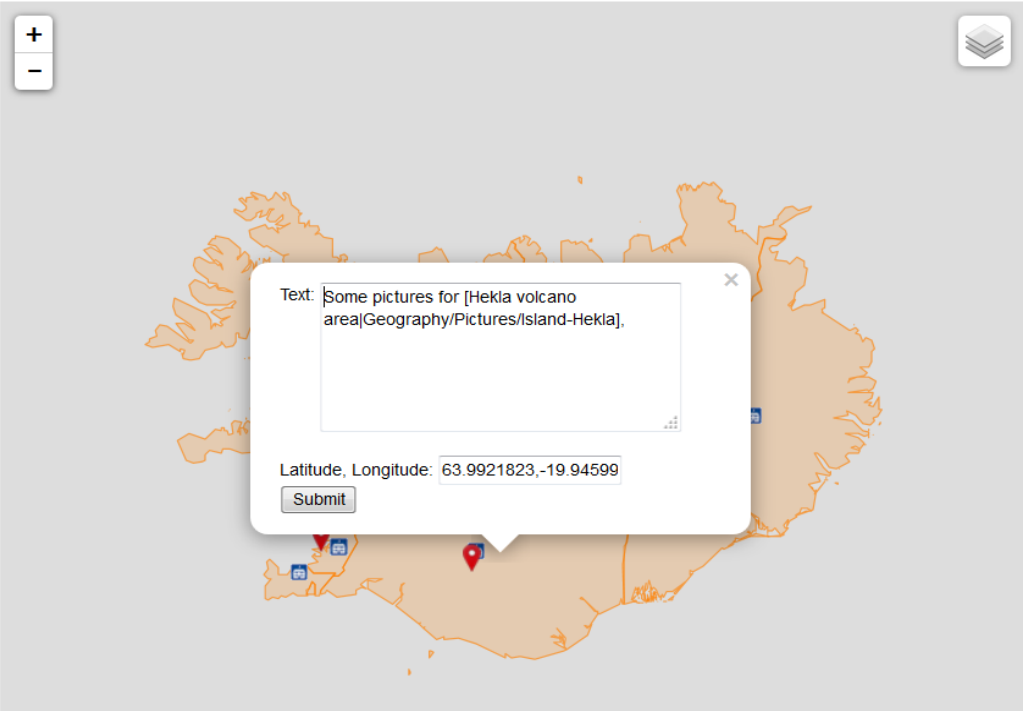


Figure 12.9: Map of Iceland, Latitude and longitude values are provided in textbox by the user.

The regular expression that ensures proper entry of latitude and longitude values is shown below.

```
/-?\d+\.?\d*,-?\d+\.?\d*/
```

12.4.2 Icons

Marker icons in leaflet is defined by *L.icon* command. We are using different icons for different markers. As an example, the code to create a province icon is shown below. We can set the icon size, image and its position on the map.

```
var provinceIcon = L.icon({
       iconUrl : '/images/map/province-icon.png',
       iconSize : [ 25, 41 ],
       iconAnchor : [ 10, 42 ],
       popupAnchor : [ 0, -30 ],
})
```

12.5 Markers Layering

The leaflet allows to group several marker layers into one. It also provides a layer control that allows users to easily switch on/off different layers on map [8].

Sometimes there are cases when a map is overcrowded due to numerous province and user markers. The solution is switching (on/off) of markers according to users need or to use zoom function (+ or -).

An example of “switching layer on and off” is demonstrated using map of United States, see Figure 12.10 and 12.11. The default setting of layer control is set to “on” i.e. all markers (capital, provinces) are displayed initially on page.

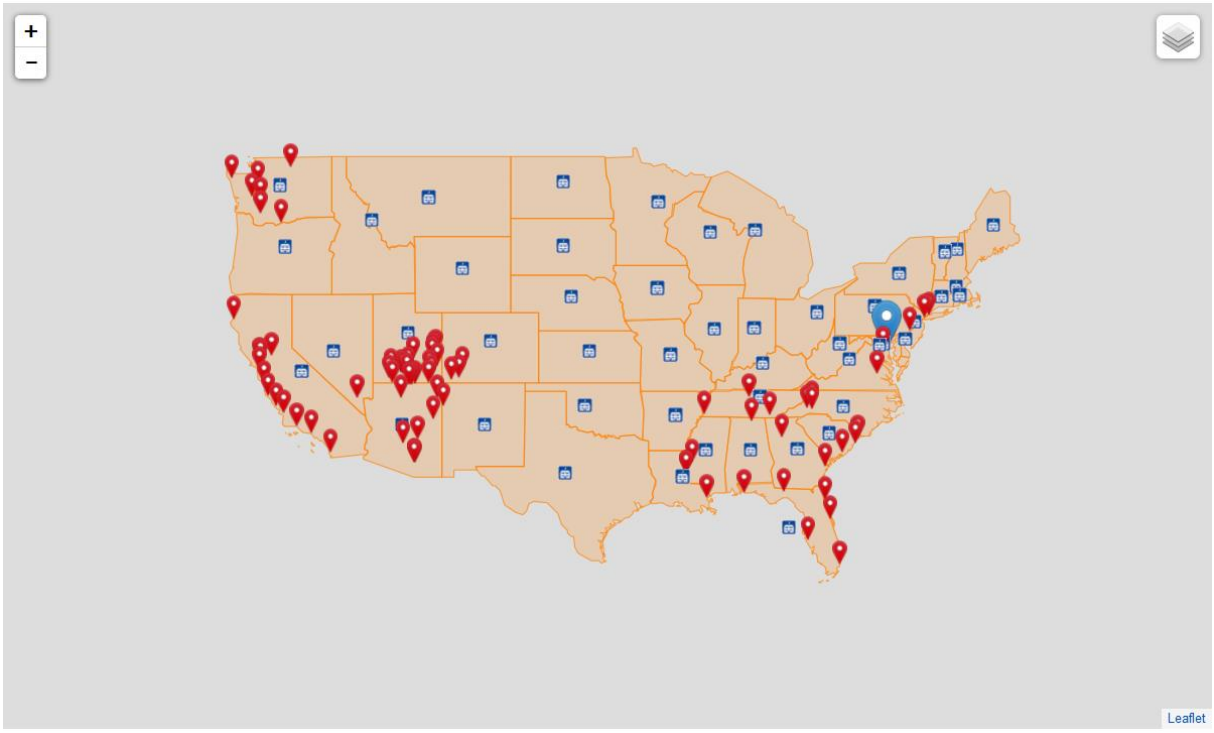


Figure 12.10: Vector map of United States with default setting (All layers enabled).

When “user markers” setting is set “off” then map looks like as shown in the Figure 12.11. The red markers disappear. Only province and capital markers are left on the map. Users can also switch off province and capital markers easily by unselecting the appropriate checkboxes.

The code to add layers in map is shown below:

```
var overlayMaps = {
    "Provinces": provinces,
    "Capital marker": capitalmarker,
    "User markers": usermarker
};
L.control.layers(temp,overlayMaps).addTo(map);
```

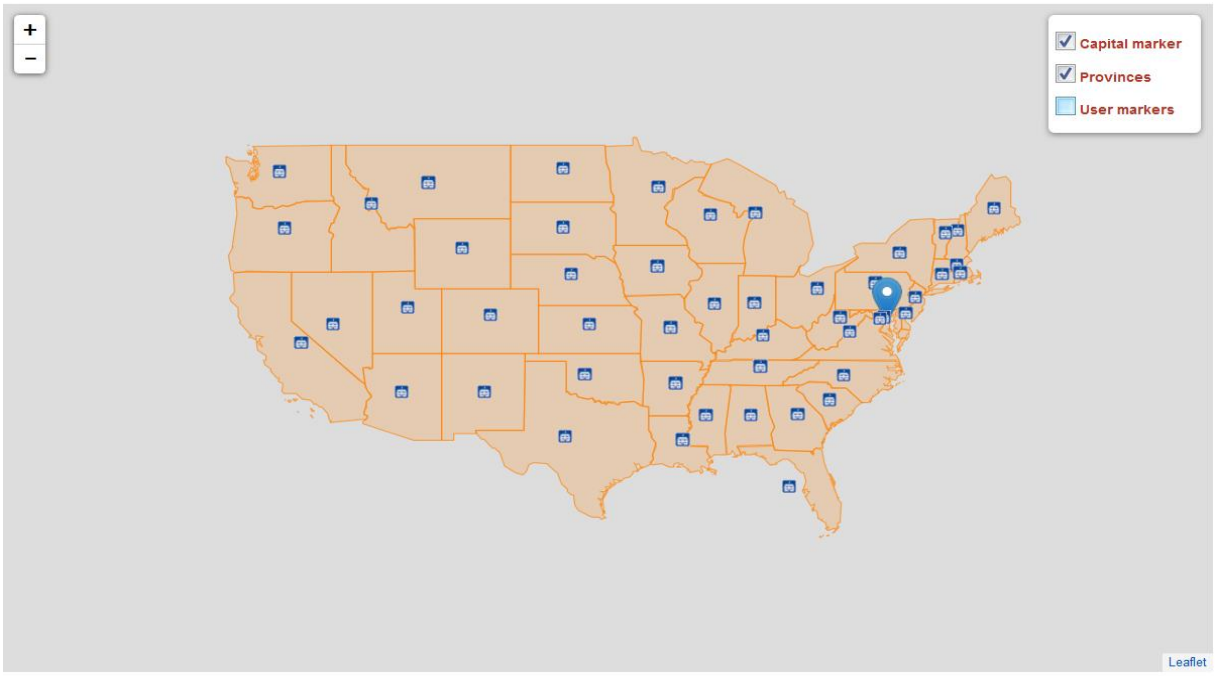


Figure 12.11: Vector map of United States showing no user defined markers, when user markers checkbox is unselected.

12.6 Map cutting

Some countries like Russia, United States cover a large geographic area on earth. Therefore we decided to cut their maps into proper regions. An example demonstrating map cutting of Russia is shown in Figure 12.12. QGIS[9] open source GIS tool is used for map cutting.

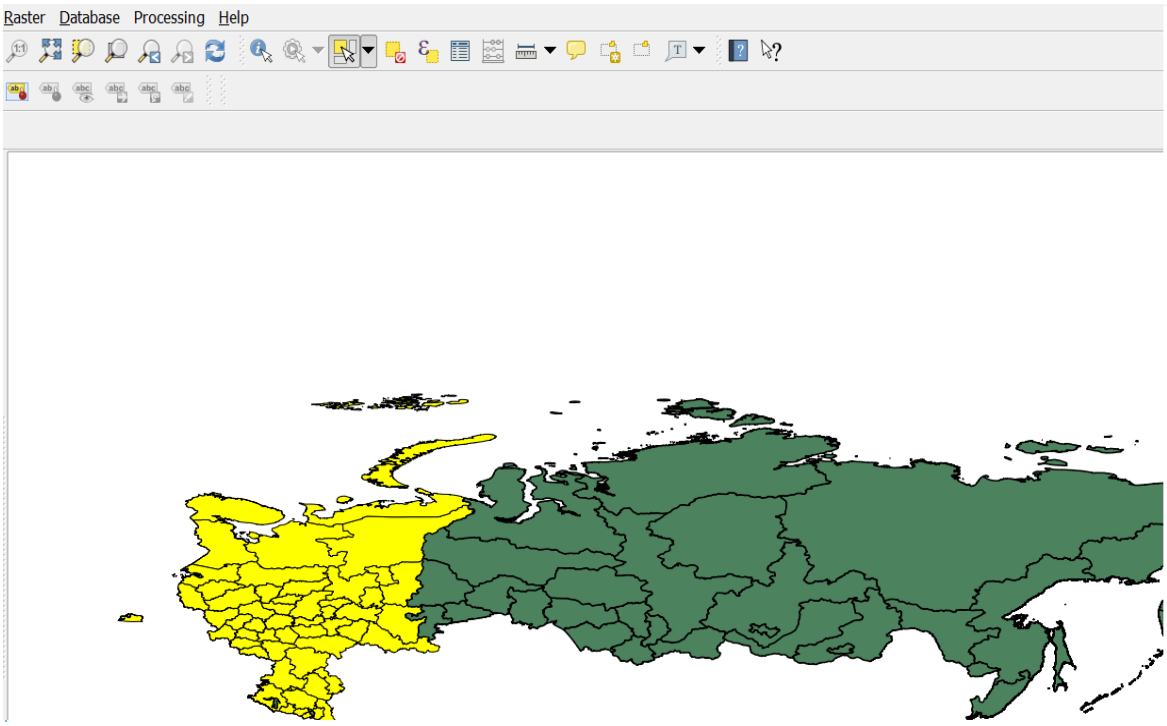


Figure 12.12: Vector map of Russia opened in QGIS.

The selected yellow portion indicates those regions that we have filtered out from whole map and they are displayed as “European Russia” in our server. After selecting current geometries using selection tool in QGIS [9], we have used “save selection” command to store selected regions as GeoJSON file.

12.7 Image Creation of Vector Maps for Slide Show Plugin

The slide show plugin displays map images as thumbnails for the corresponding wiki pages as shown in the Figure 12.13. Without proper image, the plugin shows the logo of Austria-Forum, hence it affects the preview. However when image is present as an attachment to the page, the preview changes as shown in the Figure 12.14. The slideshow plugin now shows the image of map of Algeria. In the second case we have captured the image of vector maps and saved it as an attachment in the page of Algeria.

Figure 12.13: Map section of Algeria: Without image, Slideshow plugin displays nothing.

Figure 12.14: Map section of Algeria: Better appearance using Image of Vector map.

As mentioned, it was necessary to use images of vector maps as an attachment (jpegs, png, jpg) to page for better appearance in slide show displaying map previews. To fulfil this requirement we have

used leaflet-image plugin as discussed in [10]. The leaflet-image plugin exports images out of Leaflet maps without a server component, by using Canvas and CORS.

By setting parameter L_PREFER_CANVAS = true; vector layers are drawn in canvas rather than SVG or VML. Once drawn in canvas it is easier to save them as image. We first export vector maps and then save their images as attachments in corresponding wiki pages that contain vector map.

12.8 How to reduce vector maps loading time

We have used mapshaper (a tool used to reduce the size of shape geometries)[11]. Mapshaper by Mathew Bloch is used for fast data reduction and easy editing of polygon data[12]. The maps of big countries usually take more time to load thus affects the usability and use experience; therefore we have simplified their shapes using "mapshaper" thus reducing the size of GeoJSON files considerably well.

12.8.1 Steps of reducing size of shapefiles

1. Select the GeoJSON file e.g. "MEX_adm.json" (Mexico shape file) from the directory and load it in the mapshaper.
2. Simplify it for around 10%, so that the shapes remain stable and not distorted. The map of Mexico is shown in the Figure 12.15 and it is reduced to 2.5%.
3. The size of GeoJSON file of Mexico is reduced to only 517 KB after simplification.

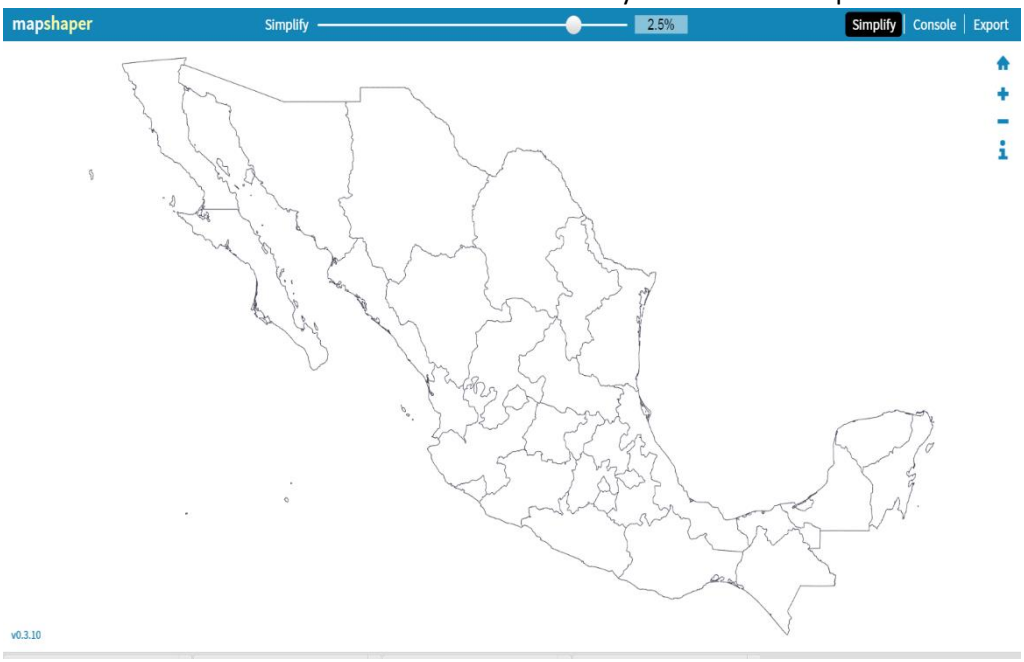


Figure 12.15: Map shaper reducing size of Mexico shape file

Table 12.1

S#	Country	Original Size	Reduced Size
1	China	23.1 MB	4MB
2	India	12MB	2MB
3	Greece	10 MB	1.2MB
4	Australia	36 MB	1.5MB
5	Indonesia	42 MB	680 KB

12.9 Performance Comparison

We have conducted a small experiment to see the difference of loading time of files in browser using Network Monitor [13].

The Network Monitor shows:

- All the network requests which Firefox makes (for example, when it loads a page, or due to XMLHttpRequests)
- How long each request takes.
- Details of each request.

To open the monitor select "Network" from the "Web Developer" menu, or open the Developer Toolbox and switch to the "Network" tab shown in blue colour in Figure 12.16.

Figure 12.16 shows the map visualisation using original GeoJSON file of size 42MB. It took 3827 milliseconds to load as compared to reduced GeoJSON file which took only 172 milliseconds shown in the Figure 12.17. The reduction of file size does not affect the presentation of maps as can be seen in the Figure 12.16 and 12.17.

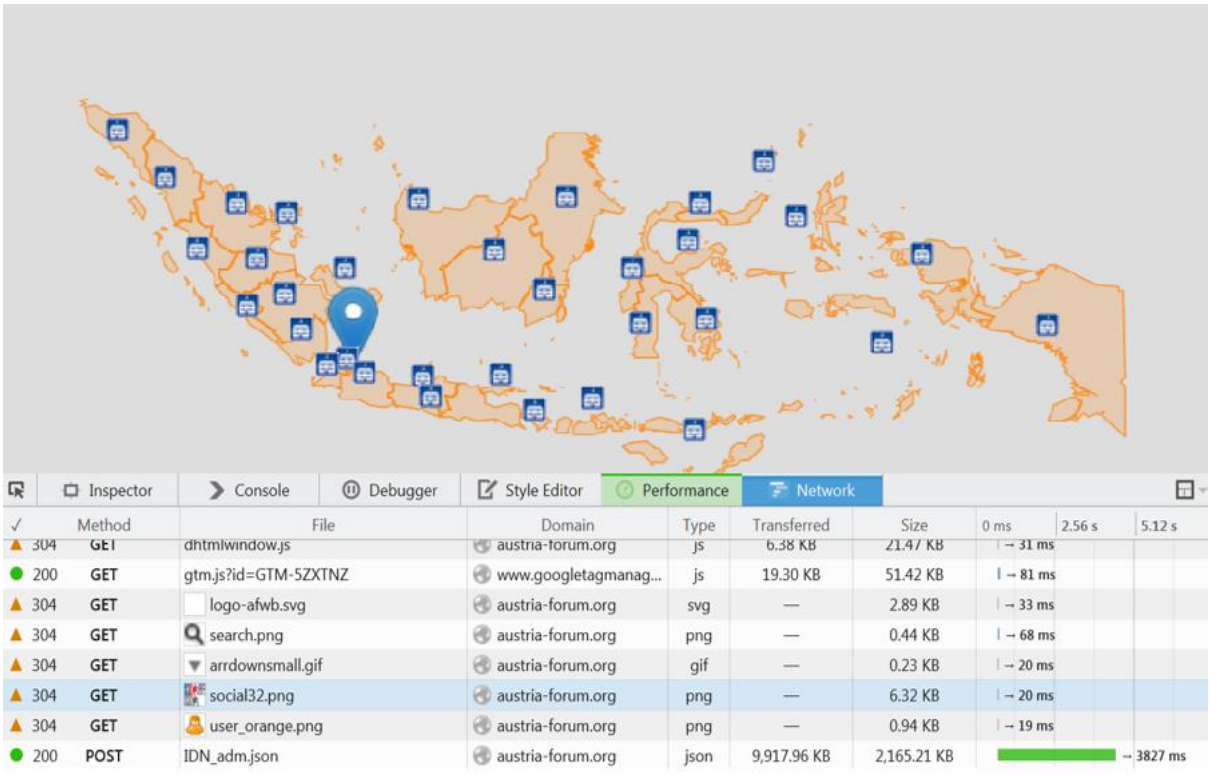


Figure 12.16: The map of Indonesia is shown along with Firefox Inspector Network Tab. File IDN_adm.json (Indonesia shape file) takes 3827ms to load in browser.

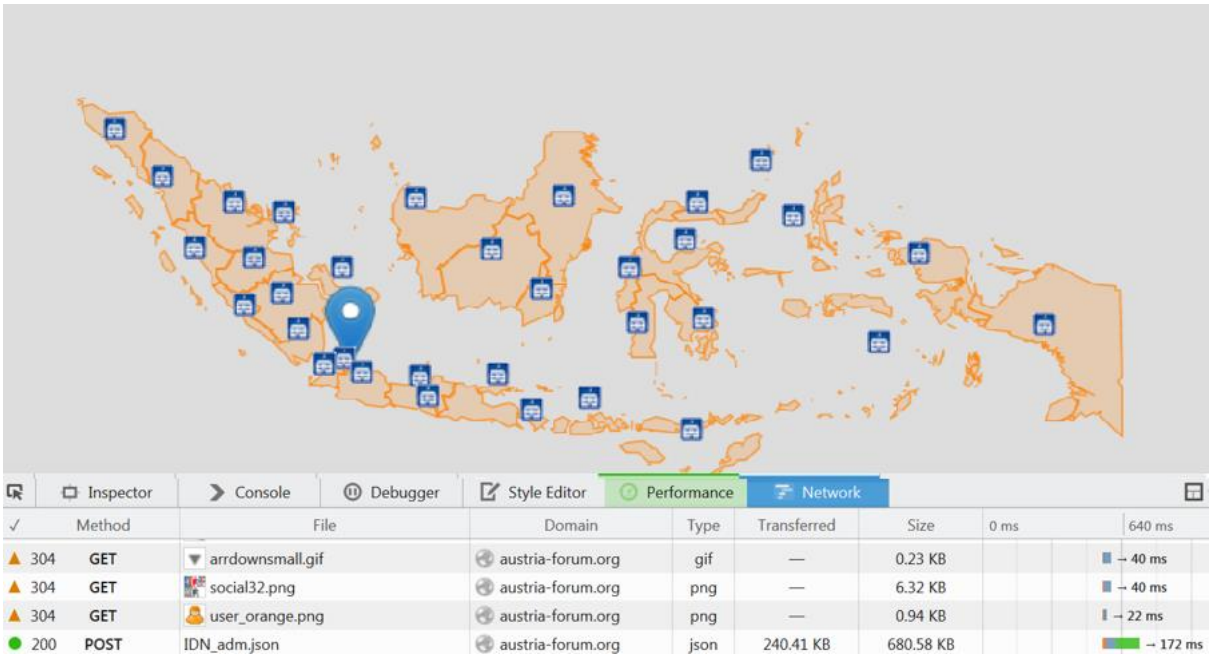


Figure 12.17: The map of Indonesia is shown along with Firefox Inspector Network Tab. File IDN_adm.json takes only 172ms to load in browser.

12.10 Zooming

Zooming functionality is provided for ease of navigation. For example in the map shown in the Figure 12.18, there are three red markers and they are hidden behind each other. After zooming users can easily see three markers as shown in the Figure 12.19.

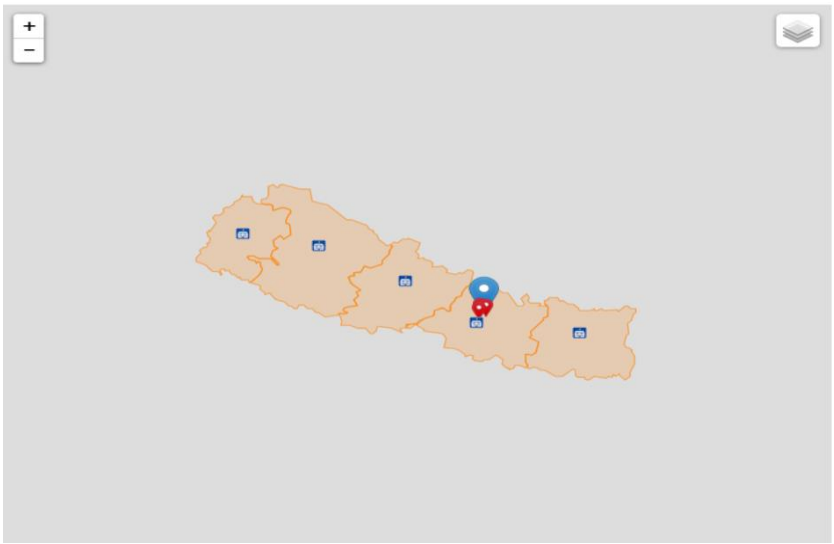


Figure 12.18: Map of Nepal with markers overlap

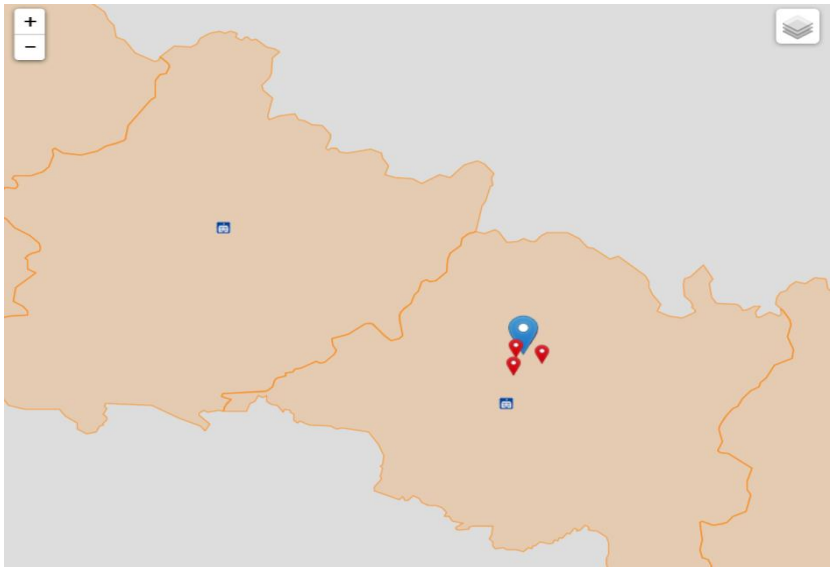


Figure 12.19: Map of Nepal after zoom in and makers are now quite visible.

12.11 Storing Mechanism

For proper functioning, user generated markers need a storage mechanism as shown in Figure 12.20. We have used JSONPRC which is a lightweight remote procedure call protocol, see [14]. User enters the "textual description and links" in the textbox using wiki syntax. These values are stored in MySQL[15] data store. We allow permanent storage of markers data. Associated data with markers is sent to browser when page is loaded. The markers are then displayed along with the maps in the browser.

latitude	longitude	page
48.22851013590295	13.760375641286373	Geography/Europe/Austria/Maps/Vector_map
47.29058918517856	14.990844391286373	Geography/Europe/Austria/Maps/Vector_map
47.268229028244654	11.017455998808146	Geography/Europe/Austria/Maps/Vector_map
47.0141533284404	13.071899358183146	Geography/Europe/Austria/Maps/Vector_map
47.63966502667759	13.544311467558146	Geography/Europe/Austria/Maps/Vector_map
48.140610752405735	13.983764592558146	Geography/Europe/Austria/Maps/Vector_map

Figure 12.20: Few records of user generated markers created by user.

12.11.1 Markers Protection

We have a proper mechanism to protect the markers created by user. More specifically, if a user X creates a marker, only that user X is allowed to delete or edit the marker. Other users only see those markers but they are not allowed to delete or edit the information in marker.

References

[1] "The history and importance of web mapping." [Online]. Available: https://www.e-education.psu.edu/geog585/node/643. [Accessed: 22-Oct-2015].

[2] "Leaflet - a JavaScript library for interactive maps." [Online]. Available: http://leafletjs.com/. [Accessed: 21-Oct-2015].

[3] "Plugins - Leaflet - a JavaScript library for interactive maps." [Online]. Available: http://leafletjs.com/plugins.html. [Accessed: 08-Mar-2016].

[4] "Leaflet: Make a web map!" [Online]. Available: http://maptimeboston.github.io/leaflet-intro/. [Accessed: 08-Mar-2016].

[5] "The World Factbook." [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/. [Accessed: 19-Oct-2015].

[6] "OpenStreetMap." [Online]. Available: http://www.openstreetmap.org/#map=7/55.700/37.600&layers=H. [Accessed: 21-Oct-2015].

[7] S. Murray, Interactive Data Visualization for the Web. OREILLY Media Inc., 2013.

[8] "Layer Groups and Layers Control - Leaflet - a JavaScript library for interactive maps." [Online]. Available: http://leafletjs.com/examples/layers-control.html. [Accessed: 21-Oct-2015].

[9] "Welcome to the QGIS project!" [Online]. Available: http://www.qgis.org/en/site/. [Accessed: 21-Oct-2015].

[10] "Leaflet Image." [Online]. Available: https://www.mapbox.com/mapbox.js/example/v1.0.0/leaflet-image/. [Accessed: 21-Oct-2015].

[11] "mapshaper." [Online]. Available: http://www.mapshaper.org/. [Accessed: 29-Oct-2015].

[12] "Digital Geography." [Online]. Available: http://www.digital-geography.com/mapshaper-org-easy-editing-polygon-polyline-datasets-mapping-puposes/#.VjNWcCuPxsk. [Accessed: 30-Oct-2015].

[13] “Network Monitor - Firefox Developer Tools” [Online]. Available: https://developer.mozilla.org/en-US/docs/Tools/Network_Monitor. [Accessed: 30-Oct-2015].

[14] “simple is better - JSON-RPC.” [Online]. Available: http://www.simple-is-better.org/rpc/. [Accessed: 04-Jan-2016].

[15] “MySQL.” [Online]. Available: https://www.mysql.de/. [Accessed: 04-Jan-2016].

Chapter 13 : Semi-Automatic Story Generation

When we look up information in existing geographic servers we usually find numerical data on aspects like size of country, structure and size of population, parameters describing infrastructural or economical facts, etc. Where databases offer more (like Wikipedia does) they most often use a kind of standard format that is not well adaptable to offer "unusual" and "unexpected" facts. We also believe that important historical events, cultural or scientific achievements, typical pictures with good descriptions and interesting maps are essential. Above credos have been the driving force for this research.

We want to show that if we extract facts from existing data bases by applying a set of rules (heuristics) and enrich those using suitable media (maps, pictures, videos) then we are able to generate interesting facts in the form of stories that appeal to the curiosity of users. To achieve this, we have applied some filters to extract "unusual" facts from existing Web pages, never exceeding a few lines as allowed for quotation purposes. We will show that the stories can be completely different: For some countries historical facts may dominate, for others the beauty of landscapes, for others cultural or economic achievements, for still others unusual facts concerning Nobel Prize winners, food, entertainment, sports, other activities, national symbols, special laws, or the kind of written language, and so on. We want to capture such "not-expected" facts into our stories.

We hope that our stories will offer astonishing or amusing facts about countries. Clearly, the border-line between interesting and non-interesting facts is murky and depends on the persons reading the stories. Hence we have employed, and will continue to do so an evaluation system. The evaluations are done using students of Graz University of Technology to show us how successful our heuristics are, and where improvements are necessary.

This chapter will address the following research questions.

RQ1. How can we generate stories about countries?

RQ2. How can we add suitable maps in stories?

RQ3. How can we effectively utilize Meta data about Nobel Prize winners to generate interesting facts that can be used in stories?

RQ4. How can we create tag clouds of countries?

13.1 Introduction

Geography influences world history, society and human development. Geography offers students, teachers, researchers, and the general public, a tool for better understanding our world.

The Web has enormous amounts of data where one can easily extract information of interest. Typically, online encyclopaedias such as Wikipedia are providing information about all the countries of world. Their objective is usually to provide as much information about a particular country as can be collected.

This is what sometimes is wanted. In other cases a short overview, or an emphasis on a particular topic might be more desirable, or some highlights typical of the country. Note that any lengthy information in Wikipedia is preceded by a few line summary, but this is not enough to cater for the very different

interests of users. As long as we cannot state something like "I want an n character long exposition on country x with heavy emphasis on topic y" other avenues have to be used to satisfy various tastes or interests, depending on what a person wants to obtain at a particular moment in time. We will explain our approach after mentioning some important facts and other attempts.

Wikipedia is a remarkable resource as a corpus for finding knowledge [1]. There were some efforts in the past to extract information from Wikipedia, as it is a huge open source data repository with free access [2]. Recent years have seen the use of computers in data mining for extraction of hidden knowledge, see [3]. The authors in [3] have tried to highlight important years for different nations.

They performed their experiments on Google news archive. They have identified frequently mentioned years for different countries. For Argentina, years 1974, 1976 and 1978 occur more frequently in news[19] and top words are "team, first, cup, Maradona (a footballer)": the reason for this are clearly the success stories of Argentina's soccer team in those years. Authors of BioRaT [4] have tried to analyse textual data. Their system works in the biological domain. It is a special Information Extraction (IE) tool, designed for analysis of abstracts and full length papers.

The goal of Natural Language Processing-NLP is to make computers better "understand" natural language [5]. Web pages contain great amounts of knowledge in textual form, hence without the computer "understanding" the context to some extent it is difficult to automatically extract facts from them [6]. However, POS (Parts Of Speech) tagging[20] can be used to extract interesting facts. The process of assigning or marking a word corresponding to a particular parts of speech is called POS tagging. In linguistics, this process is also called grammatical tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. We ourselves have used adjectives and superlatives to locate interesting facts in sentences.

Let us now turn to approach that we are taking. The web pages for a particular country in our geographic server cover information about different aspects of geography as shown in the Figure 13.1. Note that by adding a special section on maps, culture, pictures and stories we allow the users to just pick those parts that are of interest to them. We have discussed this to some extent in the paper [7].

In this chapter we focus on stories that are typical for a country. They can be found in the category "Special Information" as shown in the Figure 13.1. Quite a few have been written by persons based on their own experiences, such as those listed in [8]. However to get a substantial quantity we try to compile many more by using the archives of the Factbook, DBpedia and other data sources mentioned in references. We are also utilizing textual information of countries from Wikipedia. We explicitly highlight (marked as italic) and properly reference (references are given at end of each story indicating wiki page of country with the license) those pieces of text that we are taking from Wikipedia or other sources that require the mentioning of some license. Readers can have a look at stories under special information for each country in [9].

[19] Argentina won the FIFA world cup in 1978

[20] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

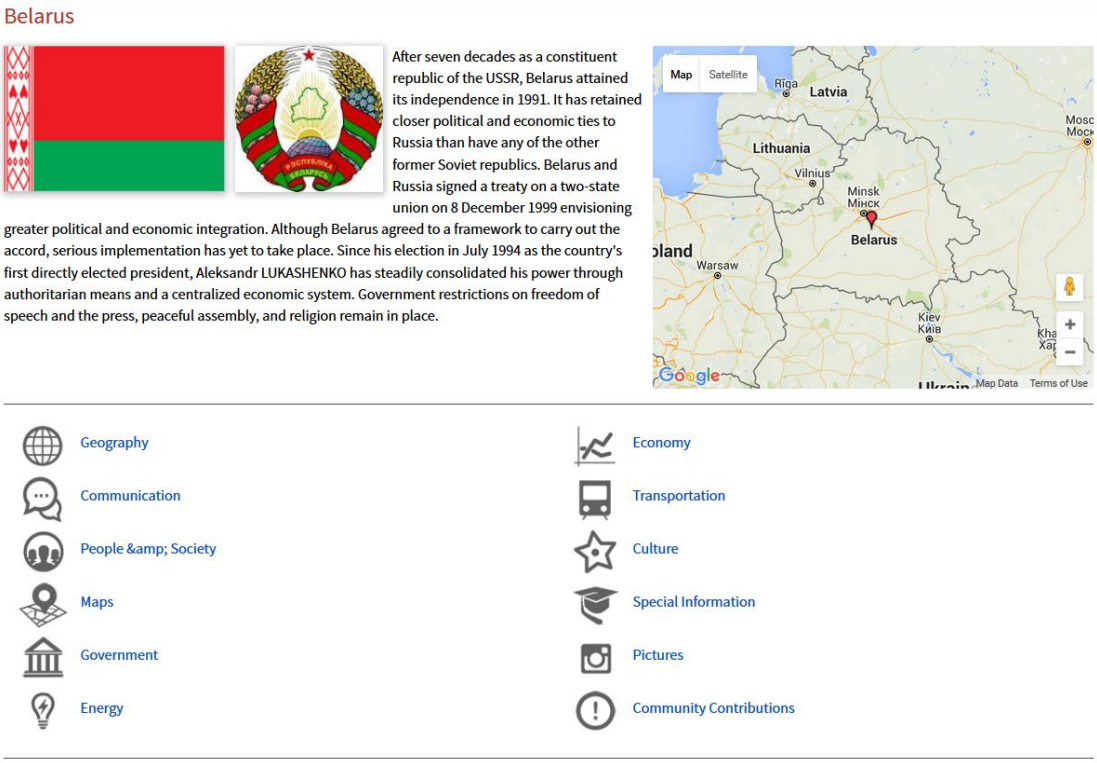


Figure 13.1: Main entry page of Belarus (a country in Europe)

To give some impression of what we consider “interesting facts”, let us look at some examples.

Geographic servers usually display border countries of a particular country. It is quite obvious that many countries share border with other countries like Afghanistan, Austria, Switzerland; exceptional cases are Australia, New Zealand and of course islands etc. Users might be surprised to find bordering countries with surprisingly long or short borders, like Greece sharing the longest border with Bulgaria as shown in Figure 13.2, or having longer coastlines than expected. The sentence taken from Wikipedia “Greece has one of longest histories” is doubtful, simply because it is not clear what definition of history is. Wikipedia should say “oldest civilisation in Europe with writing” but not more, or some such statement.

Maps are important for grasping geographic facts. The map of neighbouring countries is provided for easy understanding of facts related to boundaries and seashores. The maps in stories are fully interactive and users can hover over a country to see the population of selected and neighbouring countries.

The idea of using maps is based on recently emerging technologies. Freely available mapping libraries such as Openlayers[21] , Leaflet[22] are empowering developers to use Web as a medium for displaying maps. There were early efforts proposing and applying visual perspective in geographic domain for better geographic understanding as in [10]. A recent study tries to reveal the benefits of digital map to ease learning, see [11]. Our story generation idea also follows “Geobrowsing”, the term first coined by Donna in 2002, it is termed as a process of integrating information with maps for better understanding

[21] http://openlayers.org/
[22] http://leafletjs.com/

and knowledge discovery. Projects such as GeoSTART[23], Mission Geography[24] etc. provide materials about geography in textual and image format that are somehow helping middle/high school students in learning state-of-the-art approaches to geography, earth science, and enhancing their spatial thinking skills. We use these ideas and systematically use maps, videos and pictures in stories. Binding population statistics with a map helps users to comprehend a country within a geographic context. We have used Leaflet to display maps. The tiles are taken from an open source tile provider, CartoDB[25].

Another example of a typical surprising fact might be that Austria not only imports and exports to and from many European countries but also imports products form probably unexpected candidates like Kazakhstan. Similarly some Asian countries like Sri Lanka and Israel export to Belgium.

It is interesting and possibly entertaining to find out that a country got a Nobel Prize in some area unexpected earlier or unexpected late: Germany has two Nobel Prize winners (Emil Adolf von Behring, Physiology or Medicine and Wilhelm Conrad Röntgen, Physics) in the first year Nobel prizes were awarded. Some other facts about Nobel Prize winners that might be interesting or amusing are:

- United States has the largest number of Nobel Prize winners (counting who was born in United States) in all areas except literature (in literature it is France)
- France has the largest number of Nobel Prize winners (counting who was born in France) in literature (11)
- Marie Curie was the first woman ever, born in Poland, to get a Nobel Prize in physics. She obtained it in 1903.
- Although India has only 8 Nobel Prize winners (counting who was born in India) it does have a winner in each category. In all other countries that have a winner in each category there are more than 11 Nobel Prize winners.

We will present the rules for identifying interesting facts about Nobel Prize winners in section 13.3.11.

We generate stories by consolidating information (e.g. facts on various aspects of a country) by extracting information from different databases and taking only those if they match specific constraint. Our overall objective is to entertain or surprise people by summarizing some of the most important and unique facts.

[23] http://www.aag.org/cs/education/teaching_and_learning_materials/geostart_teaching_earth_science
[24] http://missiongeography.org/
[25] https://cartodb.com/

Greece

Athens is the nation's capital and largest city, followed by Thessaloniki, which is commonly referred to as the co-capital.

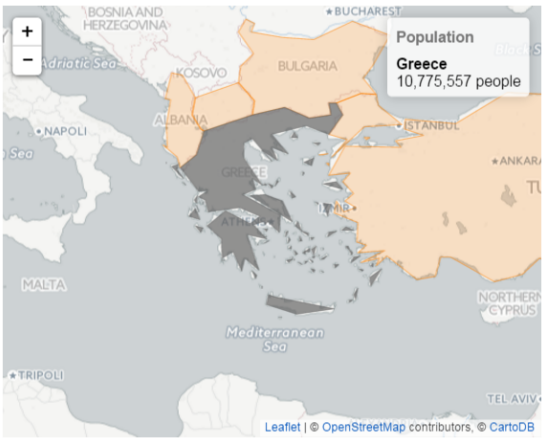
Greece has the longest coastline on the Mediterranean Basin and the 11th longest coastline in the world at 13,676 km (8,498 mi) in length, featuring a vast number of islands, of which 227 are inhabited.

Eighty percent of Greece is mountainous, with Mount Olympus being the highest peak at 2,918 metres (9,573 ft).

Greece has one of the longest histories of any country, tracing its roots to the civilization of Ancient Greece, which is considered the cradle of all Western civilization; its legacy includes democracy, Western philosophy, the Olympic Games, Western literature, historiography, political science, major scientific and mathematical principles, and Western drama, including both tragedy and comedy.

Greece, which is one of the world's largest shipping powers, middle powers and top tourist destinations, has the largest economy in the Balkans, where it is an important regional investor.

- Greece has largest boundary with Bulgaria which is approximately 494 km .
- Greece has shortest boundary with Turkey which is approximately 206 km .
- The largest river in Greece is Struma which is 410 km.
- The largest export partner of Greece is Turkey.
- The largest import partner of Greece is Russia.
- There are 66 seaports in Greece.



Mr Odysseus Elytis got Nobel Prize in the field of literature in year 1979 "for his poetry, which, against the background of Greek tradition, depicts with sensuous strength and intellectual clear-sightedness modern man's struggle for freedom and creativeness"

Figure 13.2: Parts of story about Greece, The map of neighbouring countries helps in verifying the facts related to longest and shortest boundaries. Population is shown on mouse over a country.

This chapter presents an approach (mostly depending on heuristics) to acquire exciting knowledge from various data sources about the countries which we will explain in section 13.3. The rest of chapter is organised as follows. We will explain the architecture of system in Section 13.2. We will narrow our focus in Section 13.3 and describe our heuristics in detail, in corresponding sub-sections of section 13.3. We end with a conclusion and short list of representative references.

13.2 Architecture of System

Figure 13.3 shows the overall architecture of story generation. We start the story of each country using sentences that are extracted from Wikipedia articles (see Section 13.3.5 for more details). For extraction of geographic and economic facts we are using CIA World Factbook, which is in the public domain, but written in a crude HTML. The “FactbookXML”, a service made available by Michael Schierl, makes it easy to parse Factbook data. We are using SQL to query facts and further use functions like “substring_index” for parsing and extraction of facts. We have a rule service that works on parsed facts and capture only those facts that match a criteria (max, min and user defined rules for Nobel Prize winners, see section 13.3.11 for Nobel Prize winners). Further we are using facts from DBpedia, National Geographic[26] (travel-guide page), and Nobel Prize website[27].

[26] http://travel.nationalgeographic.com/travel/countries/pakistan-guide/
[27] http://www.nobelprize.org/

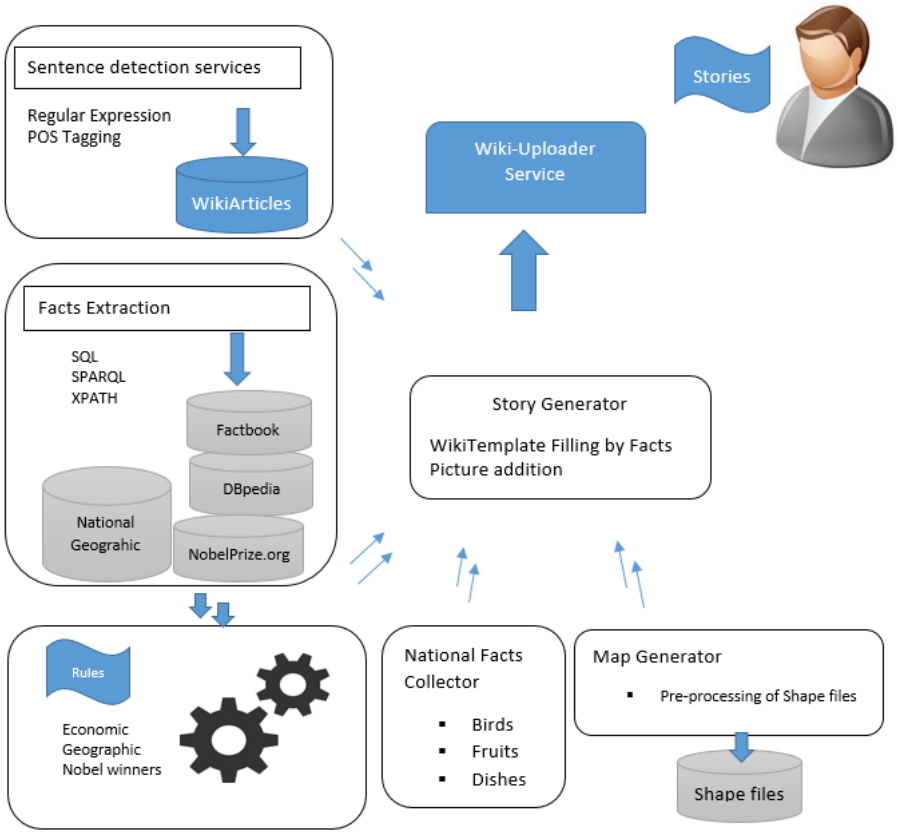


Figure 13.3: Overall architecture of system

The RDF triples describing facts about "http://dbpedia.org/page/Khewra_Salt_Mine" are shown in Figure 13.4. Each RDF triple represents a subject, a predicate and an object. Khewra_Salt_Mine is a subject; predicates are "dbo:abstract" and "dbo:country". The objects are dbr:Pakistan and "textual description" about saltmine.

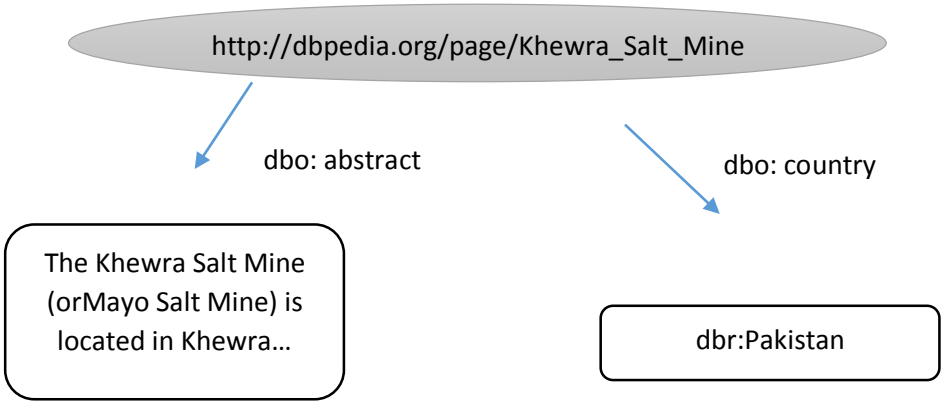


Figure 13.4: RDF triples showing facts about Khewra Saltmine.

The map generator module shown in the Figure 13.3 is responsible for the dynamic maps that are displayed along with the interesting and amusing descriptions. We have downloaded the shape file (.shp) of the countries (see Figure 13.5) from Natural Earth[28] which is a great resource for geographic shape files. Some pre-processing is required to extract the selected polygons. To select only the neighbouring countries of a particular country we have used "Select Features by Free hand" (a special QGIS feature that allows the selection of particular polygons). The yellow portion selected in the Figure 13.5 shows the neighbouring countries of Greece. We have extracted these specific polygons and used them in the corresponding maps of countries. The map of France is shown in the Figure 13.6. It represents the finished product of "map generator service" shown in the Figure 13.3.

Figure 13.5: Pre-processing phase extracting only selected polygons using "Feature selection Tool in QGIS"

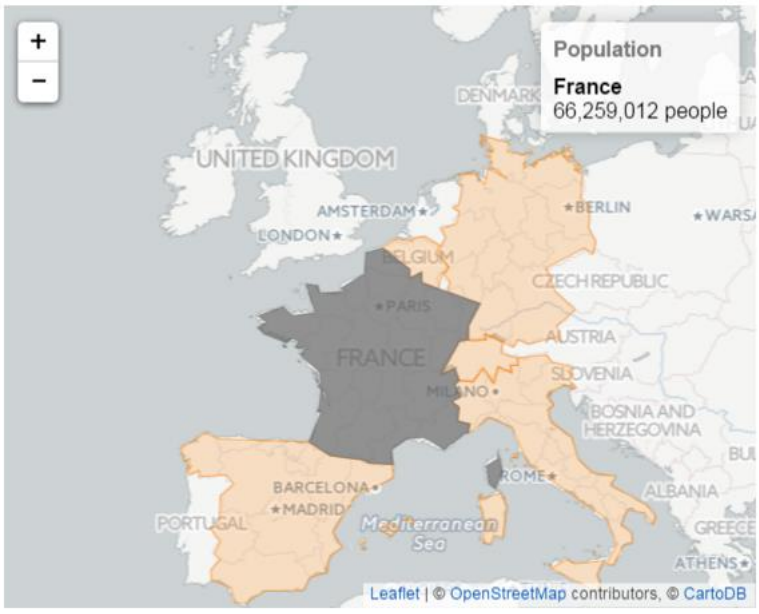


Figure 13.6: Map showing neighbour countries of France. Population is shown on mouse over.

[28]http://www.naturalearthdata.com/

The last step was the aggregation of different facts in a template designed specifically for stories. Once we have stories in final form we augment them with suitable pictures and videos.

This process is currently being done manually, therefore requiring some effort. We hope to cut this down by using a semi-automatic approach as we did when merging picture databases, see [12].

13.3 Heuristics for Story Generation

In the quest of interesting and surprising facts, we have looked at following:

1. National belongings/facts/symbols of each country.
2. Tourist attractions in a country.
3. Filtering of facts using rules to pick extreme values like “highest”, “lowest”, “most popular”, “amazing” etc.
4. Geographic facts manipulation to create new interesting facts.
5. Identification of important sentences from Wikipedia articles using NLP (Natural Language Processing).
6. Food items/beverages/dishes
7. Country specific inventions
8. History of vehicles e.g. trains, cars, planes, satellites, submarines, etc.
9. Ways of transportation e.g. seaports
10. Labour force- by occupation
11. Nobel Prize winners
12. Tag clouds

13.3.1 National Belongings/Facts/Symbols.

Typically, countries are represented by former great leaders, inventors and artists, or sport “heroes”, or very special features including landscapes, animals or flowers.

China is clearly associated with words such as great wall, silk worm. Ayers rock [29], Kangaroo, koala are typical for Australia. New Zealand comes to mind when talking about the bird Kiwi, and USA when mentioning the Statue of Liberty etc.

In this section as a first step towards story generation (considering uniqueness and level of interest), we have identified the national symbols that represent countries as shown in the Figure 13.7. It is important to mention here that some countries specify certain items as national symbol and some do not.

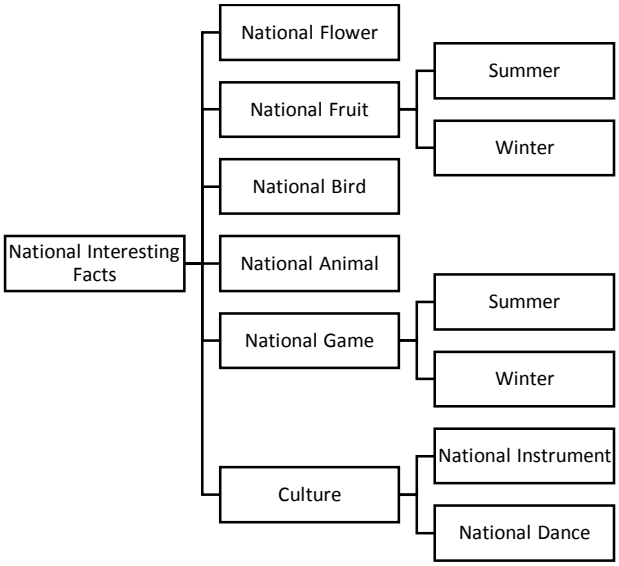


Figure 13.7: Taxonomy of National Facts

As an example we are listing the national symbols of Pakistan.

- Pakistan has two national fruits, Mango (Summer Fruit) and Guava (Winter Fruit).
- The national bird of Pakistan is Chukar Partridge.
- Markhor and Snow Leapord are considered as national animals.
- Jasmine is the national flower of Pakistan.
- The national sport of Pakistan is Field Hockey.
- The dafli, also popularly known as daf is a considered as national instrument.
- The national dance of Pakistan is Khattak.

Let's proceed and go through the taxonomy shown in the Figure 13.7. On the top spot we encounter flowers. Flowers have long been admired by people as they affect the beauty of environment. The "Edelweiss" is the national flower of Austria. The "Rose" is considered the national flower in United States, United Kingdom, Maldives and other countries. There are few websites that are managing the list of national flowers for every country. One of them is "the flower expert" [13].

The national symbols of Austria are shown in the Figure 13.8.

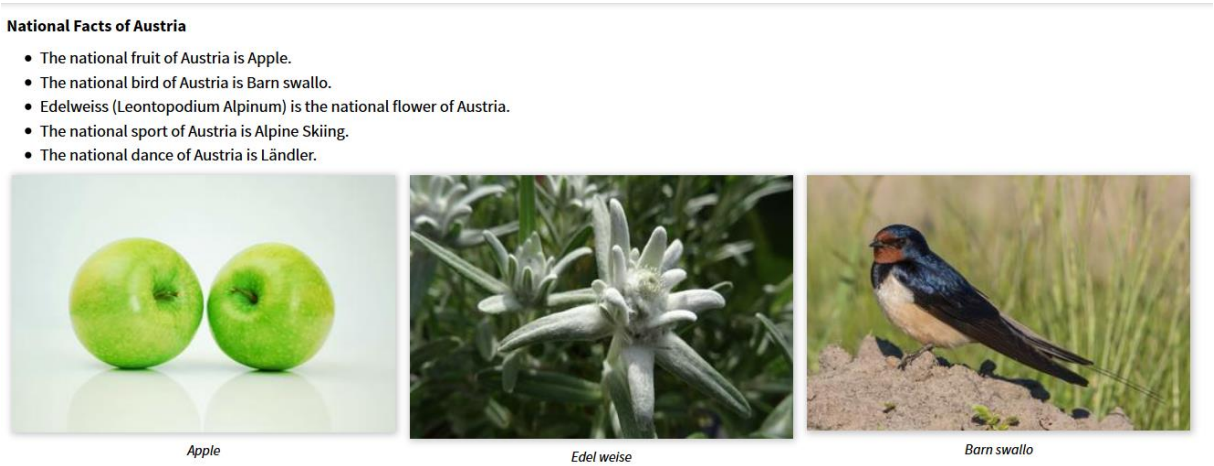


Figure 13.8: Representation of national facts of Austria in Austria-Forum Server.

National sports are vital for countries as they affect the social values and represent a nation world-wide. Cricket is very popular in Australia, India, and New Zealand. American countries such as Antigua and Barbuda, Barbados, and Bermuda are also fond of cricket which is considered as national sport in these countries. Soccer is considered as national sport in many European countries such as Hungary or Poland. Dance is a type of art which represents the culture of a nation. For instance, Kolo dance is considered as national dance in Croatia. Landler is a folk dance which was very popular in Austria, Viennese Waltz is the genre of a ballroom dance that Austria is best known for. Likewise, some countries officially recognize fruits as their national fruits. Armenia's national fruit is apricot. Pomegranate is national fruit of Azerbaijan and Iran. Birds and animals also represent countries. India is known for different animals including aquatic, bird's reptiles etc. King Cobra is considered as the national reptile in India. Indian elephant is the heritage animal. Ganges river dolphin is considered as national dolphin in India. Belarus is famous for its white storks.

We have gathered national facts of countries from different websites including Wikipedia lists and consolidate them into story section "National Facts" of each country.

13.3.2 Tourist Attractions

In the quest of interesting tourist attractions we have explored different online websites specifically those that are dealing with interesting places such as [14] and [15] (particularly travel guide page of National Geographic). Further we have utilized the power of semantic web specifically (DBpedia page of countries[29] where information is in structured form in RDF triples and easy to extract) and find the concepts such as "dbo:saltmines" and "dbo:Museum" that belong to Tourist-Attractions. Similarly lakes, mountains, castles, fortifications, canyons, volcanos etc. fall in this category. The examples of "Tourist attraction" concerning Pakistan and France are shown in the Figure 13.9.

[29] http://dbpedia.org/page/Finland

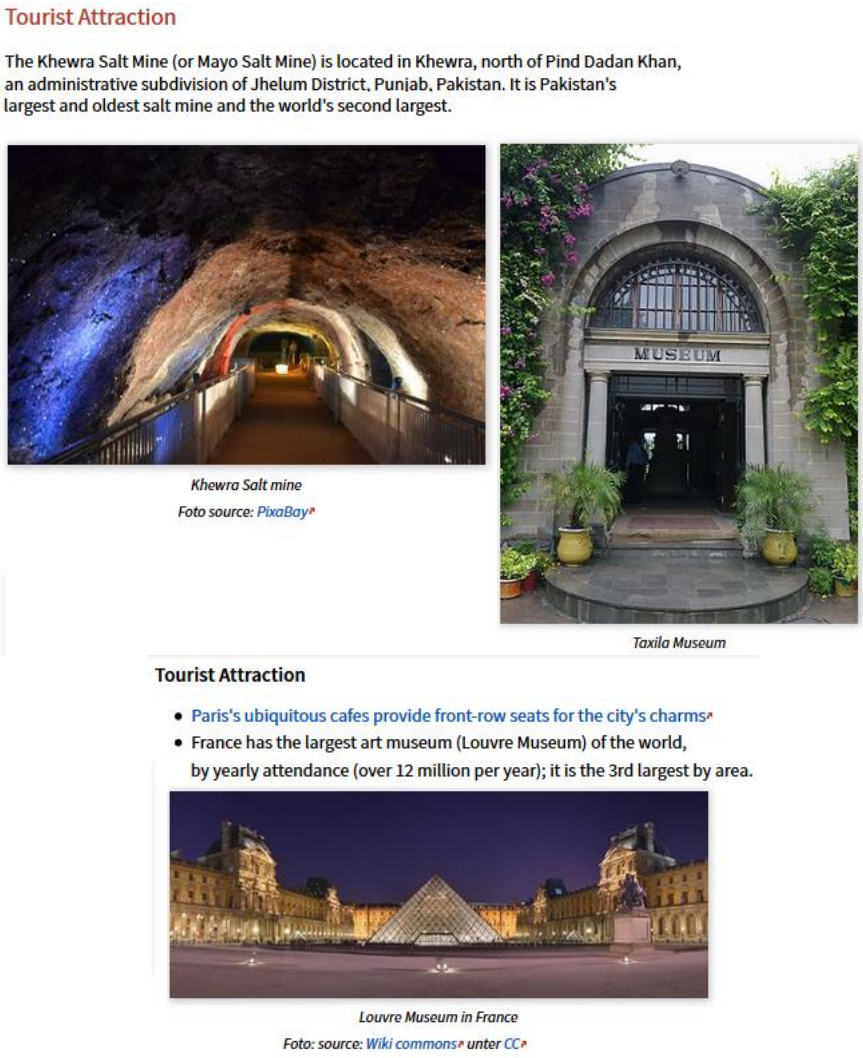


Figure 13.9: Tourist attractions in Pakistan and France.

There are some famous trees that are found in different countries either due to historic reasons or other special reasons [14]. The "Boab Prison Tree" is a large hollow tree just south of Derby in Western Australia as shown in the Figure 13.10. It was used to keep criminals overnight when moving them to the nearest court. Also see story about Mt.LEISLER [28]. The dragon tree of Teneriffe is famous for e.g. releasing "blood" (red sap) when cutting into it. The Welwitschia from Namibia is often called the ugliest but most interesting plant in the world. Concerning tourist destinations, caves are often of interest. The Waikato glow worm caves became famous because of a curious misunderstanding, see [26]. Austria is known for two of the largest ice caves in the world considering non-polar climate areas.

In this category we are considering those bridges that are special for some reasons either they are tall or high. Danyang–Kunshan Grand Bridge is the world longest bridge in China [16][27]. The Golden Gate bridge of San Francisco is probably the best know bridge of the world. One of the most impressive bridges is certainly the Bridge of Shahara. Museums are always of interest to people. Taxila Museum is one of the famous tourist attraction in Pakistan. France has the largest art museum (Louvre Museum) of the world, by yearly attendance (over 12 million per year); it is the 3rd largest by area.

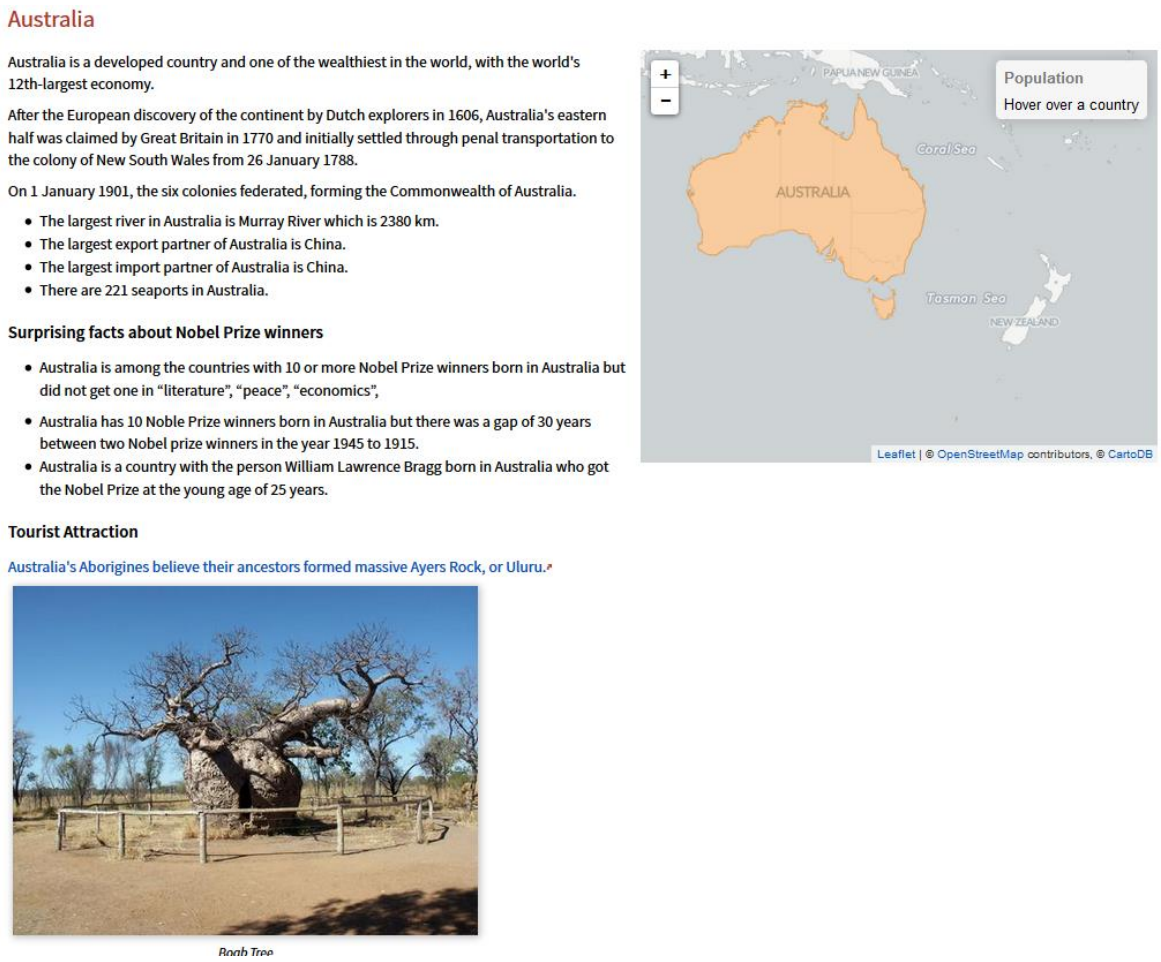


Figure 13.10 : Boab Prison tree as tourist attraction in story of Australia.

13.3.3 Facts filtering using Rules

We have applied rules to pick the extreme values. Pakistan exports to different countries but the uniqueness comes from applying rule max(Export) . We first extract and capture the percentage exports from Factbook dataset [17] and then identify the highest export country. Pakistan largest export partner is US. These facts are present in textual form in Fact book shown below.

Exports – partners

US 13.6%, China 11.1%, UAE 8.5%, Afghanistan 7.8% (2012)
Imports – partners are

China 19.7%, Saudi Arabia 12.3%, UAE 12.1%, Kuwait 6.3% (2012)
Pakistan imports from different countries but the uniqueness comes from applying rule max(Import). Pakistan imports more from China.

There are different types of people living in the country that constitute (Ethnic groups).

Ethnic Groups of Pakistan
Punjabi 44.68%, Pashtun (Pathan) 15.42%, Sindhi 14.1%, Sariaki 8.38%, Muhajirs 7.57%, Balochi 3.57%, other 6.28%

- Punjabi are most
- Balochi are least

Figure 13.11 shows different parts of story about Pakistan.

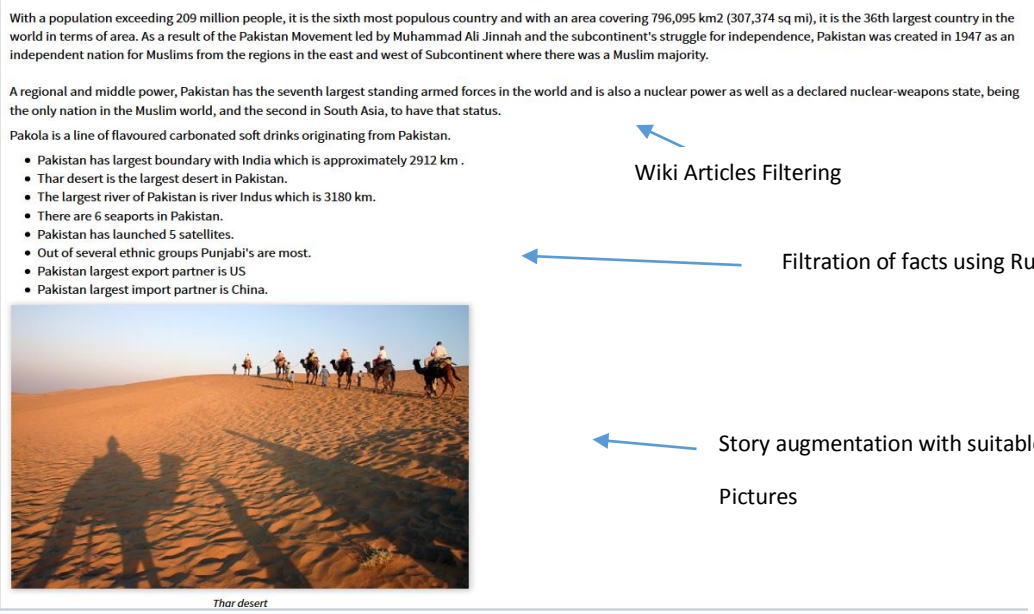


Figure 13.11: Representation of different parts of story about Pakistan in Austria-Forum.

13.3.4 Geography Stories

Following are some of interesting geographic facts that represent a country:

- Landlocked
- Largest boundary with a particular neighbour
- Largest river

An interesting element is whether country is landlocked or not. We recognize this easily by using by using the rule If coastline=0 then landlocked, for example Switzerland is a landlocked country. Similarly Austria is a landlocked country. This fact can be verified by the map provided in stories. Afghanistan is a land locked country shown in the Figure 13.12.

Figure 13.12: Map showing Afghanistan is a land locked country. The country with dark colour is Afghanistan. Neighbours have light colour.

Although most countries share boundaries with other countries but it might be interesting to note which country has longest or even the shortest boundary in common with the selected country. We can figure out this using the rule *Land boundaries + Largest in length*. Pakistan has as longest land boundary, the one with India. Austria has as shortest land boundary the one with Liechtenstein approximately 34 km. Afghanistan has as the longest boundary the one with Pakistan which is approximately 2670 km. Also Afghanistan has as the shortest boundary the one with China (91 km), as can be shown in the Figure 13.12.

13.3.5 Unique facts identification from Wikipedia articles using NLP

In this section we will mention some ideas about how to find relevant and interesting facts using text processing which we are using as introductory part of our story.

According to Olena Medelyan

"Wikipedia is a goldmine of information; not just for its many readers, but also for the growing community of researchers"

For example Wikipedia article of Italy contains lots of facts about Italy. We do not want to include all the facts into our story about Italy that exists in the article. Therefore we have applied filters using parts of speech tagging. Specifically we have extracted two types of sentences:

1. Sentences with superlatives/adjectives.
2. Sentences with years (for capturing any historic events)

13.3.5.1 Interesting sentences identification

Following python code is used for tokenizing and tagging of wiki article of Italy.

```
1 def displayImportantSentence(wikitext):
2 sentences=nltk.sent_tokenize(wikitext)
3     i=0
4     for i in range (0,len(sentences)):
5         word=nltk.word_tokenize(sentences[i])
6         tags = nltk.pos_tag(word)
7
8         for key,value in  tags:
9             if value=='JJS':
10                print sentences[i]
11                print '\n'
12                break;
13        i=i+1
```

The algorithm has identified following 4 sentences out of several paragraphs of Wiki article due to presence of superlatives highlighted as italics in the following.

- With 61 million inhabitants, it is the 4th *most* populous EU member state.
- After various unsuccessful attempts, the second and the third wars for Italian independence resulted in the unification of most of present-day Italy between 1859–66.
- Italy has the third *largest* economy in the Eurozone (the eighth-largest in the world) and the highest life expectancy in the EU, and a very high human development index.
- Italy is home to the *greatest* number of World Heritage Sites.

The sentences extracted are part of our story generation. Let us look at the important sentences of Greece extracted from Wiki article of Greece. Again they are creating elements of interest and fit well into the story.

- Athens is the nation's capital and largest city, followed by Thessaloniki, which is commonly referred to as the co-capital.
- Greece has the longest coastline on the Mediterranean Basin and the 11th longest coastline in the world at 13,676 km (8,498 mi) in length, featuring a vast number of islands, of which 227 are inhabited.
- Eighty percent of Greece is mountainous, with Mount Olympus being the highest peak at 2,917 m (9,570 ft).
- Greece has one of the longest histories of any country, tracing its roots to the civilization of Ancient Greece, which is considered the cradle of all Western civilization; its legacy includes democracy, Western philosophy, the Olympic Games, Western literature, historiography, political science, major scientific and mathematical principles, and Western drama, including both tragedy and comedy.
- Greece, which is one of the world's largest shipping powers and top tourist destinations, has the largest economy in the Balkans, where it is an important regional investor

The significant sentences of Germany are:

- Its capital and largest city is Berlin.
- After the United States, it is the second most popular migration destination in the world.
- The rise of Pan-Germanism inside the German Confederation resulted in the unification of most of the German states in 1871 into the Prussian-dominated German Empire.

The interesting sentence of some countries are listed below:

- Honduras: Honduras has the highest murder rate in the world
- San Marino: The country is considered to have the earliest written governing documents (constitution) still in effect.
- Haiti: It has the lowest Human Development Index in the Americas.
- Cuba: In 2015, it became the first country to eradicate mother-to-child transmission of HIV and syphilis, a milestone hailed by the World Health Organization as "one of the greatest public health achievements possible".
- Grenada is also known as the "Island of Spice" because of the production of nutmeg and mace crops, of which it is one of the world's largest exporters.
- Costa Rica: Costa Rica is the world's happiest and greenest country in the world".
- Ghana: is one of the world's largest gold and diamond producers, and is projected to be the largest producer of cocoa in the world as of 2015.
- Papua New Guinea: The country is one of the world's least explored, culturally and geographically, and many undiscovered species of plants and animals are thought to exist in the interior.

We do expect some false positives; that is those sentences that are not interesting. Those sentences can be eliminated later using human intervention at the stage of polishing the stories.

13.3.5.2 Historic year identification

Let us look at the regex [12]\d{3}; it captures the sentences with years.

If we apply this regex on Wikipedia articles we can extract those sentences that contain years (telling us about historical events). For instance, after applying this regex on the Wikipedia article of Belarus we get the following two sentences.

1) Much of the borders of Belarus took their modern shape in 1939 when some lands of the Second Polish Republic were reintegrated into it after the Soviet invasion of Poland and were finalized after World War II.
2) In 1945, Belarus became a founding member of the United Nations, along with the Soviet Union and the Ukrainian SSR.

The story of Belarus is shown in the Figure 13.13 that shows the inclusion of Belarus in UN member states in 1945 besides other interesting facts.

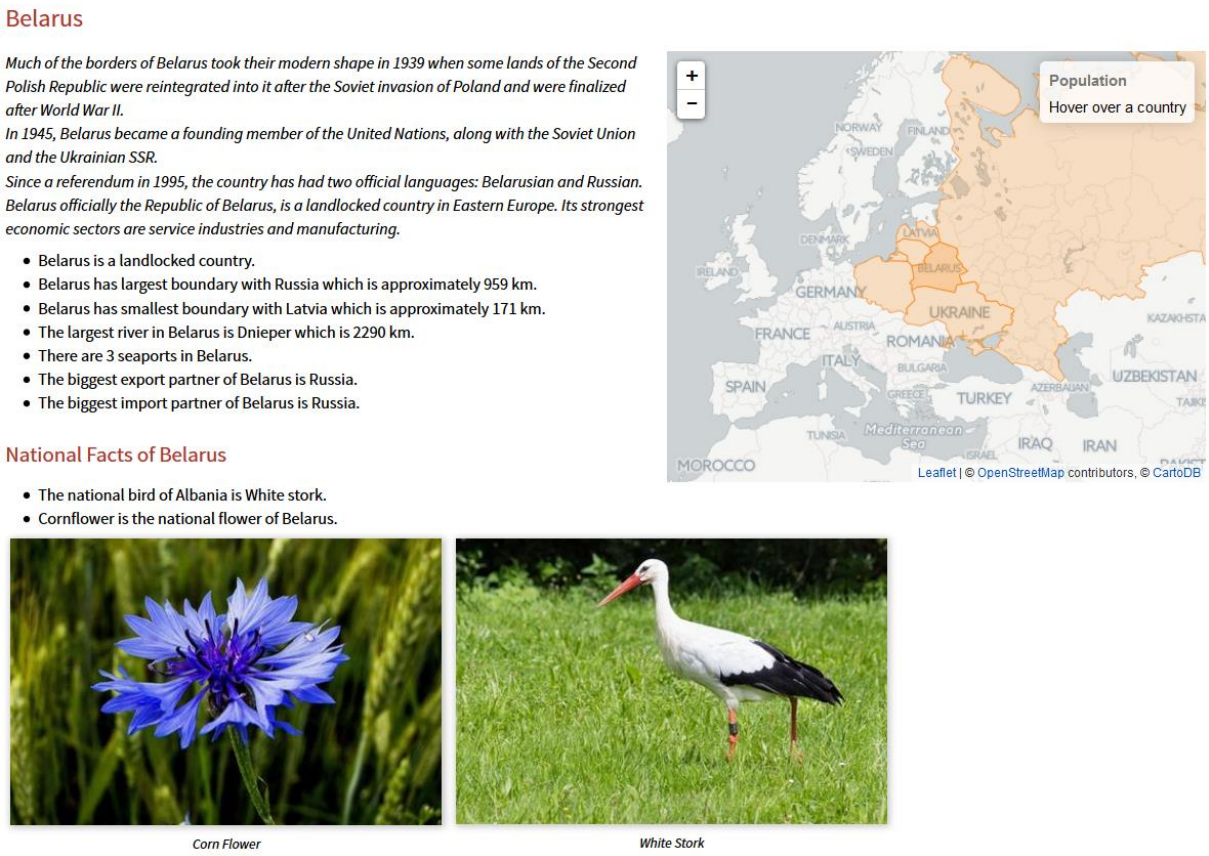


Figure 13.13: Story about Belarus

13.3.6 Food Items

In this section we will explain how we are using food items in our stories. First, we have looked at beverages, using the SPARQL query below.

```
1: select?s  ?c ?abs  where {
2: ?s a dbo:Beverage.
3: ?s dbo:abstract ?abs.
4: ?s dbo:origin ?c.
5: Filter(
6: ?c=<http://dbpedia.org/resource/Sri_Lanka>
7: )}
```

The sparql query shown above returns the beverage named as “Three Coins Beer” which is a Sri Lankan beer brewed by McCallum Breweries (Ceylon) Ltd. The same query when used for Pakistan returns “Pakola” which is one of flavoured carbonated soft drinks originated from Pakistan.

We have also identified special food websites that we plan to link to our stories besides presenting food items. The site “Food in Every Country” presents food of some countries with recipes [18].

We are using list of food items as a starting point and filling our stories with dishes mentioned in the list [19]. Nasi lemak is a national dish of Malaysia, Tumpeng is considered as national dish of Indonesia. Kabuli Palaw is favourite dish of Afghan nation. Austria is famous for Tafelspitz and Wiener Schnitzel.

13.3.7 Historical Inventions

The association of thousands of inventions in human history with countries is not trivial. We are attempting to integrate list of inventions country-specific in our story generation section. For example, Quartz watch was invented in Canada. Remote control is an invention that belongs to Austria-Hungary. Microscope and Telescope are attributed to Netherlands [20]. Wrist watch is considered as invention coming from Switzerland, etc.

13.3.8 History of vehicles e.g. trains, cars, planes, satellites, submarines

Exploring universe is always human’s concern. Sometimes people want to see launch history of vehicles. Therefore we are listing the launch history of vehicles of different countries. For example Soviet Union was the first country who launched “Sputnik 1” in 1957.

According to Wikipedia

“As of December 2013, sixty-one countries have operated artificial satellites”.

Some interesting example of vehicles such as submarines and cars are

- USS Nautilus (SSN-571) was the world's first operational nuclear-powered submarine.
- German Flocken Elektrowagen of 1888, is regarded as the first electric car of the world.

China ranks no 1 in car producing countries. The production of cars in China was about 23,722,890 in 2015, see OICA [21]. The OICA is managing a list of countries and statistics about cars production in these countries.

13.3.9 Seaports

Airports and other transportation means are quite common but an important mean of communication is seaport. There are alone 3024 sea ports (a seaport is a facility which can accommodate ships which go out to sea) in Europe [22]. Considering Europe, the largest number of seaports are in United Kingdom (731). There are 4 seaports in Austria namely: i) Port of Vienna ii) Port of Krems iii) Port of Linz iv) Port of Enns. This may be surprising, since Austria is a land locked country, yet the Danube river that empties into the Black Sea is large enough for Ocean going vessels. Table 13.1 shows continent-wise distribution of ports. Europe is leading in number of seaports with North America occupying 2nd position. Note that seaports can be in landlocked countries because a large river allows sea going vessels to navigate in the country. The Danube river in Austria is one such example.

Table 13.1: Continent wise Sea Ports

S#	Continents	Total
1	Europe	3024
2	Africa	421
3	Asia	1720
4	Oceania	373
5	North America	2293
6	South America	453

The code to extract seaports of countries of Oceania/Australia is given below.

```
1: mainPage = (HtmlPage) client.getPage("http://ports.com/browse/oceania/"
2: + rs.getString(4).toString().toLowerCase().replace(" ","-")+ "/");
3: if (mainPage != null) {
4: String ports = mainPage.getByXPath("//p[@class='large
quiet']/text()").get(0).toString();
5: addPorts(rs.getString (4).toString(),Integer.parseInt(ports));//data filling
```

13.3.10 Labour Force- by Occupation

The Labour force is used to describe the distribution of people working in different professions. For example the labour force entry for Albania in Factbook is

```
<i>agriculture:</i>47.8%<br /><i>industry: </i>23%<br /><i>services: </i>29.2%<br />
```

About 50% people in Albania belong to agriculture contrary to industry and services.

According to Factbook

“Agriculture includes farming, fishing, and forestry. Industry includes mining, manufacturing, energy production, and construction. Services cover government activities, communications, transportation, finance, and all other economic activities that do not produce material goods”.

We have further parsed these entries which is in textual form in Factbook into structured data and after applying max rule, following are the sentence picked for our story:

- 93.6% of people belong to occupation “Agriculture” in Brundi.
- 90% of people belong to occupation “Agriculture” in Burkina Faso
- 90% of people belong to occupation “Agriculture” in Malawi

13.3.11 Nobel Prize winners

Since 1901, the Nobel Prizes have been awarded to different people and organizations [23]. The Nobel Prize is an international award that is awarded in different categories (medicine, physics, chemistry, economics, literature, and peace). The category-wise distribution of Nobel Prizes is shown in the Table 13.2. There are very few Nobel prizes in economics. Noble Prizes in all categories were awarded in 1901 except economics. First Nobel Prize was awarded in category Economics in 1969.

Table 13.2: Category-wise distribution of Nobel Prizes

Category	Total
medicine	210
physics	198
chemistry	169
literature	111
peace	100
economics	76
Total	864

We have extracted Nobel Prize winners list from [23] along with the Meta data (born country, affiliation country, year, field etc.). Figure 13.14 shows the age of different Nobel Prize winners of Austria (born in Austria). Karl von Frisch got the Nobel Prize at the old age of 87 years.

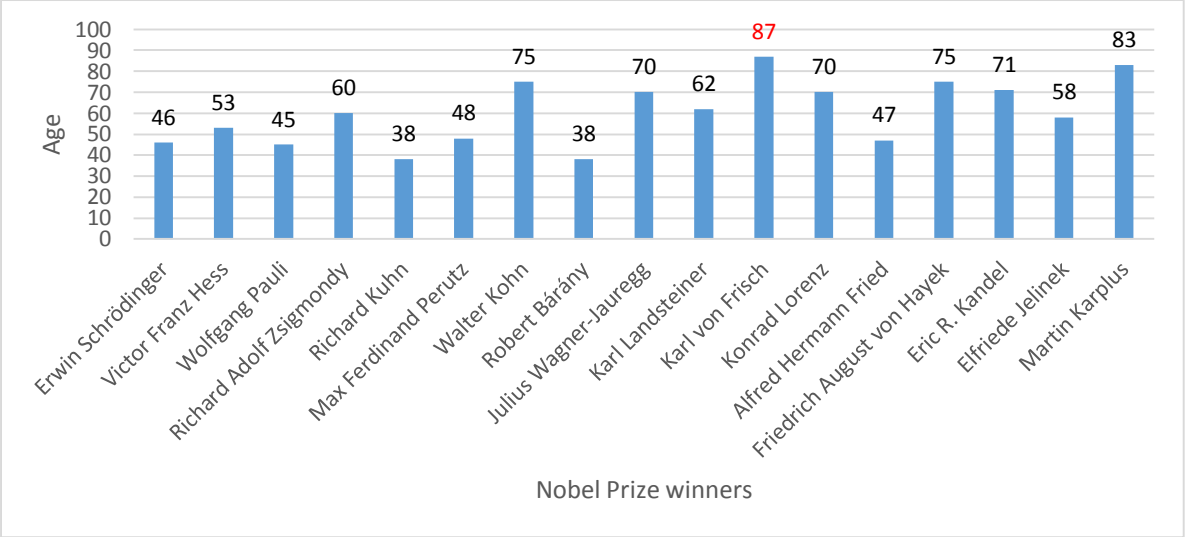


Figure 13.14: Nobel Prize winners of Austria.

The analysis given above is solely based on one country. We went further and involved other countries statistics in comparison. The unusual facts according to carefully designed rules are given below. We are listing here few of them. For unusual facts about Nobel Prize winners matching rules, see stories in [9].

1. **Country x has the largest number of Nobel Prize winners (counting who was born in x) in area z**

- United States has largest number of Nobel Prize winners (counting who was born in United States) in all areas except literature (in literature it is France)
- France has largest number of Nobel Prize winners (counting who was born in France) in literature (11)

2. **Country x has the largest number of women Nobel Prize winners born in x**

- United States has largest number of women Nobel Prize winners (11) (counting women born in United States)

3. <Country x>- born y was the first Nobel Prize winner for <area> in the first year of Nobel Prizes, 1901

- France-born Frederic Passy was the first Nobel Prize winner for the category peace in 1901 (the first year Nobel Prizes for peace were awarded).
- France-born Sully Prudhomme was the first Nobel Prize winner for the category literature in 1901 (the first year Nobel Prizes for literature were awarded).
- Germany-born Wilhelm Conard was the first Nobel Prize winner for category physics in 1901 (the first year Nobel Prizes for physics were awarded).
- Netherlands-born Jacobus Henricus was the first Nobel Prize winner for category chemistry in 1901 (the first year Nobel Prizes for chemistry were awarded).
- Poland-born Emil Adolf was the first Nobel Prize winner for category medicine in 1901 (the first year Nobel Prizes for medicine were awarded).
- Switzerland-born Jean Henry was the first Nobel Prize winner for category peace in 1901 (the first year Nobel Prizes for peace were awarded).

4. There are <n> <Country x> -born Nobel prize winners but none in that country from the last 10 years

- There are 16 Switzerland-born Nobel Prize winners but none from that country in the last ten years.
- There are 12 Denmark-born Nobel Prize winners but none from that country in the last ten years.
- There are 7 Spain-born Nobel Prize winners but none from that country in the last ten years.
- There are 9 Hungary-born Nobel Prize winners but none from that country in the last ten years.
- There are 17 Netherlands-born Nobel Prize winners but none from that country in the last ten years.
- There are 25 Poland-born Nobel Prize winners but none from that country in the last ten years.

5. y was the first woman ever, born in country x, to get a Nobel Prize.

- Marie Curie was the first woman ever, born in Poland, to get a Nobel Prize. She obtained it in 1903.

6. There are <n> <Country x>-born Nobel prize winners, but there is no >Country> born winner in c,d.

- There are 10 Australia-born Nobel Prize winners, but there is no Australia-born Nobel Prize winner in literature, peace, or economics.
- There are 12 China-born Nobel Prize winners, but there is no China-born Nobel Prize winner in economics.
- There are 12 Denmark-born Nobel Prize winners, but there is no Denmark-born Nobel Prize winner in economics.
- There are 23 Japan-born Nobel Prize winners, but there is no Japan-born Nobel Prize winner in economics.

- There are 17 Netherlands-born Nobel Prize winners, but there is no Netherlands-born Nobel Prize winner in literature.
- There are 16 Switzerland-born Nobel Prize winners, but there is no Switzerland-born Nobel Prize winner in economics.

7. Country born Nobel Prize Winner got Nobel Prize m times, m>1

- Poland born Maria curie got Nobel Prize twice in two different categories (physics, chemistry) in years 1903, 1911.
- United States born John Bardeen got Nobel Prize twice in physics in years 1956 and 1972.
- United States born Linus Carl Pauling got Nobel Prize twice in two different categories (chemistry, peace) in years 1954 and 1962.
- United Kingdom born Frederick Sanger got the Nobel Prize in chemistry twice, in years 1958 and 1980.

8. Country x has n Nobel Prize winners in year(s) y1, y2,…! (Counting winners if born in x)

- France has 3 Nobel Prize winners in year 1912, 1965, and 2008. (Counting winners if born in France)
- Sweden has 3 Nobel Prize Winners in 1974, 1982. (Counting winners if born in Sweden)
- United Kingdom has 3 Nobel Prize Winners in years 1932, 1937, 1973, 1974, 1998, 2003. (Counting winners if born in United Kingdom)
- Japan has 4 Nobel Prize Winners in 2008. (Counting winners if born in Japan)
- Germany has 4 Nobel Prize Winners in 1988. (Counting winners if born in Germany)
- Russia has 4 Nobel Prize Winners in 1958. (Counting winners if born in Russia)
- United States has 8 Nobel Prize Winners in years 1972, 1990, 2001. (Counting winners if born in United States)

9. Country x has only n Nobel Prize winner. (Counting winners if born in x) n=1

- Azerbaijan has only 1 Nobel Prize winner Lev Davidovich Landau. (Counting winners if born in Azerbaijan)
- Bangladesh has only 1 Nobel Prize winner Muhammad Yunus. (Counting winners if born in Bangladesh)
- Brazil has only 1 Nobel Prize winner Peter Brian Medawar. (Counting winners if born in Brazil)
- Bulgaria has only 1 Nobel Prize winner Elias Canetti. (Counting winners if born in Bulgaria)
- Colombia has only 1 Nobel Prize winner Gabriel García Márquez. (Counting winners if born in Colombia)
- Burma has only 1 Nobel Prize winner Aung San Suu Kyi. (Counting winners if born in Burma)
- Costa Rica has only 1 Nobel Prize winner Oscar Arias Sánchez. (Counting winners if born in Costa Rica)
- Croatia has only 1 Nobel Prize winner Leopold Ruzicka. (Counting winners if born in Croatia)
- Cyprus has only 1 Nobel Prize winner Christopher A. Pissarides. (Counting winners if born in Cyprus)
- Ghana has only 1 Nobel Prize winner Kofi Annan. (Counting winners if born in Ghana)
- Greece has only 1 Nobel Prize winner Odysseus Elytis. (Counting winners if born in Greece)
- Iceland has only 1 Nobel Prize winner Halldór Kiljan Laxness. (Counting winners if born in Iceland)

- Indonesia has only 1 Nobel Prize winner Willem Einthoven. (Counting winners if born in Indonesia)
- Kenya has only 1 Nobel Prize winner Wangari Muta Maathai. (Counting winners if born in Kenya)
- Latvia has only 1 Nobel Prize winner Wilhelm Ostwald. (Counting winners if born in Latvia)
- Macedonia has only 1 Nobel Prize winner Mother Teresa. (Counting winners if born in Macedonia)
- Madagascar has only 1 Nobel Prize winner Claude Simon. (Counting winners if born in Madagascar)
- Nigeria has only 1 Nobel Prize winner Wole Soyinka. (Counting winners if born in Nigeria)
- Peru has only 1 Nobel Prize winner Mario Vargas Llosa. (Counting winners if born in Peru)
- Slovakia has only 1 Nobel Prize winner Philipp Eduard Anton von Lenard. (Counting winners if born in Slovakia)
- Slovenia has only 1 Nobel Prize winner Fritz Pregl. (Counting winners if born in Slovenia)
- Taiwan has only 1 Nobel Prize winner Yuan T. Lee. (Counting winners if born in Taiwan)
- Trinidad and Tobago has only 1 Nobel Prize winner Sir Vidiadhar Surajprasad Naipaul. (Counting winners if born in Trinidad and Tobago)
- Venezuela has only 1 Nobel Prize winner Baruj Benacerraf. (Counting winners if born in Venezuela)
- Vietnam has only 1 Nobel Prize winner Le Duc Tho. (Counting winners if born in Vietnam)
- Yemen has only 1 Nobel Prize winner Tawakkol Karman. (Counting winners if born in Yemen)
- Zimbabwe has only 1 Nobel Prize winner Albert John Lutuli. (Counting winners if born in Zimbabwe)

10. <person z> is the first woman Nobel prize winner in w

- Bertha von Suttner is the first woman Nobel Prize winner in peace in year 1905. She was born in Czech Republic.
- Selma Ottilia Lovisa is the first woman Nobel Prize winner in literature in year 1909. She was born in Sweden.
- Gerty Theresa is the first woman Nobel Prize winner in medicine in year 1947. She was born in Czech Republic.
- Maria Curie is the first woman Nobel Prize winner in physics in year 1903. She was born in Poland.
- Maria Curie is the first woman Nobel Prize winner in chemistry in year 1911. She was born in Poland.
- Elinor Ostrom is the first woman Nobel Prize winner in economics in year 2009. She was born in United States.

11. <Person z> is a <Country X>-born Nobel Prize winner who got the Nobel prize quite early at age <p> where p<35

- Malala Yousafzai is a Pakistan born Nobel Prize winner who got the Nobel Prize in peace quite early, at the age of 17 years.
- William Lawrence Bragg is an Australia born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 25 years.
- Paul Adrien Maurice Dirac is a United Kingdom born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 31 years.

- Tsung-Dao (T.D.) Lee is a China born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 31 years.
- Carl David Anderson is a United States born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 31 years.
- Werner Karl Heisenberg is a Germany born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 31 years.
- Frederick Grant Banting is a Canada born Nobel Prize winner who got the Nobel Prize in medicine quite early, at the age of 32 years.
- Mairead Corrigan is a United Kingdom born Nobel Prize winner who got the Nobel Prize in peace quite early, at the age of 32 years.
- Tawakkol Karman is a Yemen born Nobel Prize winner who got the Nobel Prize in peace quite early, at the age of 32 years.
- Rudolf Ludwig Mössbauer is a Germany born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 32 years.
- Brian David Josephson is a United Kingdom born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 33 years.
- Betty Williams is a United Kingdom born Nobel Prize winner who got the Nobel Prize in peace quite early, at the age of 33 years.
- Joshua Lederberg is a United States born Nobel Prize winner who got the Nobel Prize in medicine quite early, at the age of 33 years.
- Rigoberta Menchú Tum is a Guatemala born Nobel Prize winner who got the Nobel Prize in peace quite early, at the age of 33 years.
- Donald Arthur Glaser is a United States born Nobel Prize winner who got the Nobel Prize in physics quite early, at the age of 34 years.
- James Dewey Watson is a United States born Nobel Prize winner who got the Nobel Prize in medicine quite early, at the age of 34 years.

12. <Person z> is a <Country X>-born Nobel Prize winner who got the Nobel prize quite late at age <p> p>85

- Ferdinand Buisson is a France born Nobel Prize winner who got the Nobel Prize quite late, at the age of 86 years.
- Karl von Frisch is an Austria born Nobel Prize winner who got the Nobel Prize quite late, at the age of 87 years.
- Yoichiro Nambu is a Japan born Nobel Prize winner who got the Nobel Prize quite late, at the age of 87 years.
- Joseph Rotblat is a Poland born Nobel Prize winner who got the Nobel Prize quite late, at the age of 87 years.
- Vitaly L. Ginzburg is a Russia born Nobel Prize winner who got the Nobel Prize quite late, at the age of 87 years.
- Peyton Rous is a United States born Nobel Prize winner who got the Nobel Prize quite late, at the age of 87 years.
- Doris Lessing is an Iran born Nobel Prize winner who got the Nobel Prize quite late, at the age of 88 years.
- Raymond Davis Jr. is a United States born Nobel Prize winner who got the Nobel Prize quite late, at the age of 88 years.
- Lloyd S. Shapley is a United States born Nobel Prize winner who got the Nobel Prize quite late, at the age of 89 years.

- Leonid Hurwicz is a Russia born Nobel Prize winner who got the Nobel Prize quite late, at the age of 90 years.

13. There are <n> <country-x>-born Nobel Prize winners but there was a gap of <m> years (xxxx-yyyy) between two awards.

- There are 18 Canada born Nobel Prize winners but there was a gap of 26 years (1923-1949) between two awards.
- There are 12 Denmark born Nobel Prize winners but there was a gap of 31 years (1944-1975) between two awards.
- There are 12 Norway born Nobel Prize winners but there was a gap of 46 years (1922-1968) between two awards.
- There are 12 China born Nobel Prize winners but there was a gap of 32 years (1957-1989) between two awards.
- There are 10 Australia born Nobel Prize winners but there was a gap of 30 years (1915-1945) between two awards.
- There are 9 South Africa born Nobel Prize winners but there was a gap of 28 years (1951-1979) between two awards.
- There are 8 India born Nobel Prize winners but there was a gap of 38 years (1930-1968) between two awards.
- There are 7 Spain born Nobel Prize winners but there was a gap of 34 years (1922-1956) between two awards.
- There are 6 Czech Republic born Nobel Prize winners but there was a gap of 42 years (1905-1947) between two awards.

14. Of the <n> <country> born Nobel Prize winners <m> are still alive today (m > n/2).

- Of the 23 Japan born Nobel Prize winners 17 are still alive today.
- Of the 12 China born Nobel Prize winners 11 are still alive today.
- Of the 9 South Africa born Nobel Prize winners 5 are still alive today.
- Of the 6 Israel born Nobel Prize winners 5 are still alive today.

15. The first <country-x> born Nobel Prize winner got this award quite late, <n> years after the first Nobel Prize was awarded in 1901.

- The first Mexico born Nobel Prize winner got this award quite late, 81 years after first Nobel Prize was awarded in 1901.
- The first Colombia born Nobel Prize winner got this award quite late, 81 years after first Nobel Prize was awarded in 1901.
- The first Madagascar born Nobel Prize winner got this award quite late, 84 years after first Nobel Prize was awarded in 1901.
- The first Nigeria born Nobel Prize winner got this award quite late, 85 years after first Nobel Prize was awarded in 1901.
- The first Taiwan born Nobel Prize winner got this award quite late, 85 years after first Nobel Prize was awarded in 1901.
- The first Costa Rica born Nobel Prize winner got this award quite late, 86 years after first Nobel Prize was awarded in 1901.
- The first South Korea born Nobel Prize winner got this award quite late, 86 years after first Nobel Prize was awarded in 1901.

- The first Israel born Nobel Prize winner got this award quite late, 93 years after first Nobel Prize was awarded in 1901.
- The first Trinidad and Tobago born Nobel Prize winner got this award quite late, 100 years after first Nobel Prize was awarded in 1901.
- The first Ghana born Nobel Prize winner got this award quite late, 100 years after first Nobel Prize was awarded in 1901.
- The first Iran born Nobel Prize winner got this award quite late, 102 years after first Nobel Prize was awarded in 1901.
- The first Kenya born Nobel Prize winner got this award quite late, 103 years after first Nobel Prize was awarded in 1901.
- The first Bangladesh born Nobel Prize winner got this award quite late, 105 years after first Nobel Prize was awarded in 1901.
- The first Peru born Nobel Prize winner got this award quite late, 109 years after first Nobel Prize was awarded in 1901.
- The first Cyprus born Nobel Prize winner got this award quite late, 109 years after first Nobel Prize was awarded in 1901.
- The first Liberia born Nobel Prize winner got this award quite late, 110 years after first Nobel Prize was awarded in 1901.
- The first Yemen born Nobel Prize winner got this award quite late, 110 years after first Nobel Prize was awarded in 1901.

16. The couple <x> and <y> both got Nobel Prize in category <w> in year z.

- The couple "Frederic Joliot" and "Irene Joliot" both got Nobel Prize in chemistry in 1935. They were born in France.
- The couple "Gerty Theresa" and "Carl Cori" both got Nobel Prize in medicine in 1947. They were born in Czech Republic.
- The couple "May-Britt" and "Edvard I Moser" both got Nobel Prize in medicine in 2014. They were born in Norway.
- The couple "Pierre Curie" and "Marie Curie" both got Nobel Prize in physics in 1903. Pierre Curie was born in France where as Marie Curie was born in Poland.
- The couple "Gunnar Myrdal" and "Alva Mydral" both got Nobel Prize in economics and peace in years 1974 and 1982 respectively. They were born in Sweden.

17. <country x> has only female Nobel Prize winners.

- Liberia has only female Nobel Prize winners. (Counting winners if born in Liberia)
- Burma has only female Noble Prize winner. (Counting winner if born in Burma)
- Yemen has only female Nobel Prize winner. (Counting winner if born in Yemen)
- Macedonia has only female Noble Prize winner. (Counting winner if born in Macedonia)
- Iran has only female Nobel Prize winners. (Counting winners if born in Iran)
- Kenya has only female Nobel Prize winner. (Counting winner if born in Kenya)

13.4 Pictures and videos addition to make stories more interesting

Besides factual information filtering we are extending our stories using meaningful pictures from PixaBay [24]. It offers different versions of pictures (Small, Medium, Large, and Extra Large). Pictures are in public domain. Large size pictures often require an account but small version of pictures is free to download. Pictures of national facts such as birds or fruits are freely available on PixaBay server. Other sources include all the pictures in Austria-Forum, particularly global-geography.org, Wikipedia, Flickr, Factbook etc. Further, we are using links to videos from YouTube in the appropriate sections. YouTube[30] is a free video sharing website (originally started in 2005) that makes it easy to watch online videos. Look at the story of Hungary with map (highlighting Hungary in dark orange colour) and a video (describing facts about Budapest) in Figure 13.15. We are also using the pictures from a dataset YFCC100M (The New Data in Multimedia Research), see [25]. The authors in [25] have consolidated images along with their metadata from Flickr, Instagram, Yahoo and created shared dataset namely (YFCC100M) in 2014. The YFCC100M is the largest public multimedia collection ever released.

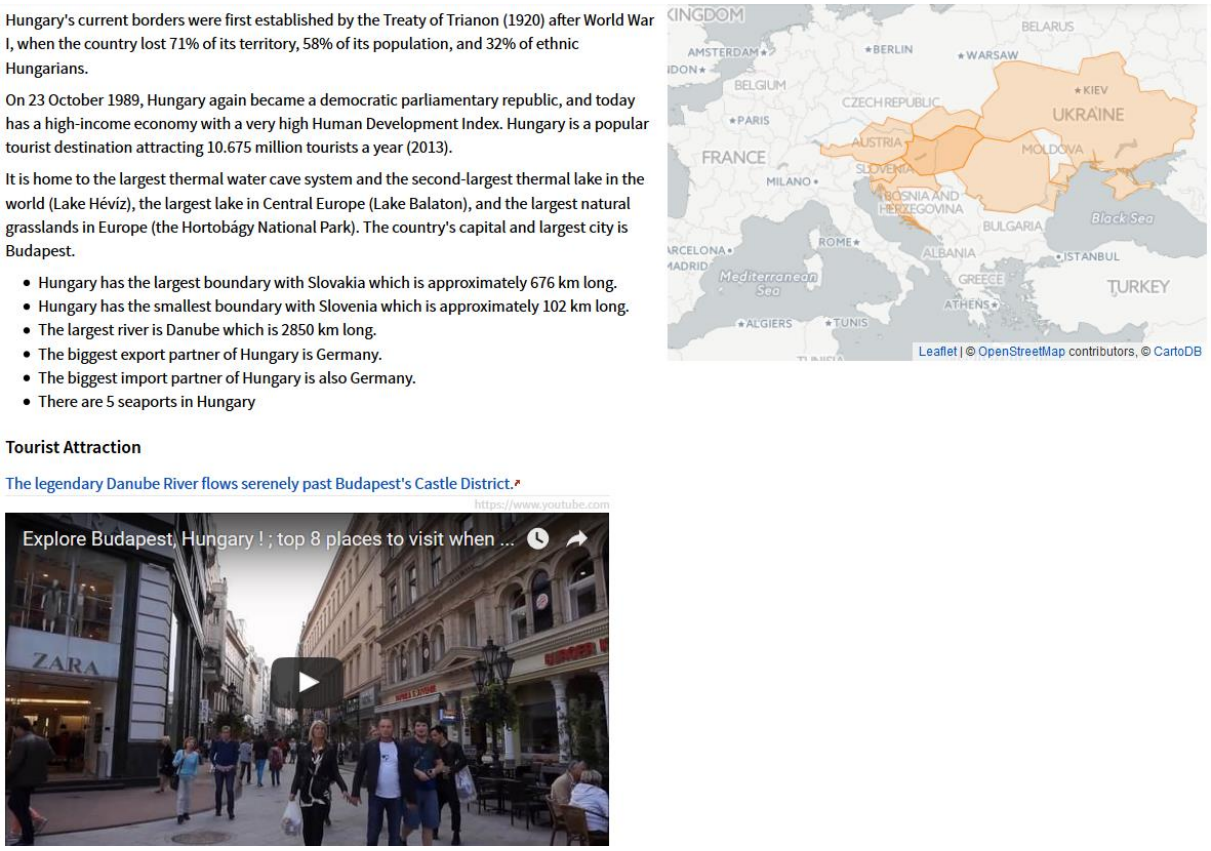


Figure 13.15: Story of Hungary. You Tube video is added in Tourist Attraction section using Iframe plugin

13.5 Tag Clouds

A tag cloud/word cloud is a visual representation of textual data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the

[30] https://www.youtube.com/

importance of each tag is shown with font size or colour. We have created tag clouds representing countries using the properties in which they rank high as shown in the Figure 13.16. The tag cloud of Nauru is shown in the Figure 13.16. Nauru has very high percentage (71%) of obesity among people. It also has high percentage of permanent crop and urbanization. Contrary to Nauru, Iran has high rank in many properties such as Water area, Crude oil production, Natural gas production etc. But also CO2 emission is high in Iran. Tag clouds are useful to provide a quick overview of a country. We have chosen around 100 properties and create these tag clouds which represent countries meaningfully, see Figure 13.17.

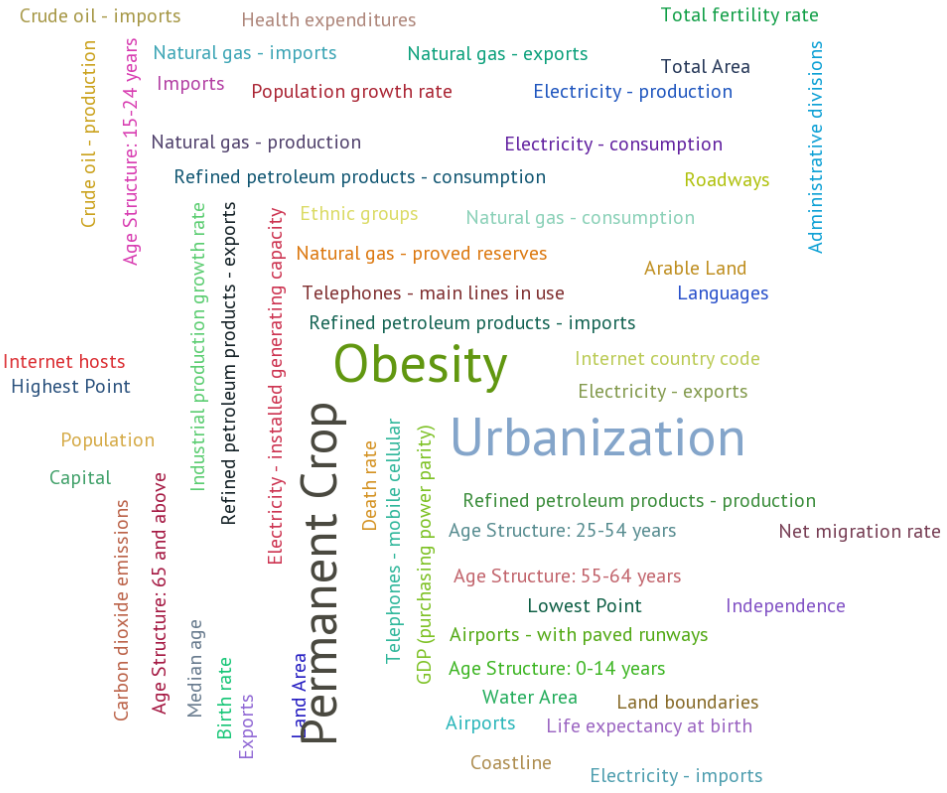


Figure 13.16: Tag Cloud representing Nauru in various properties

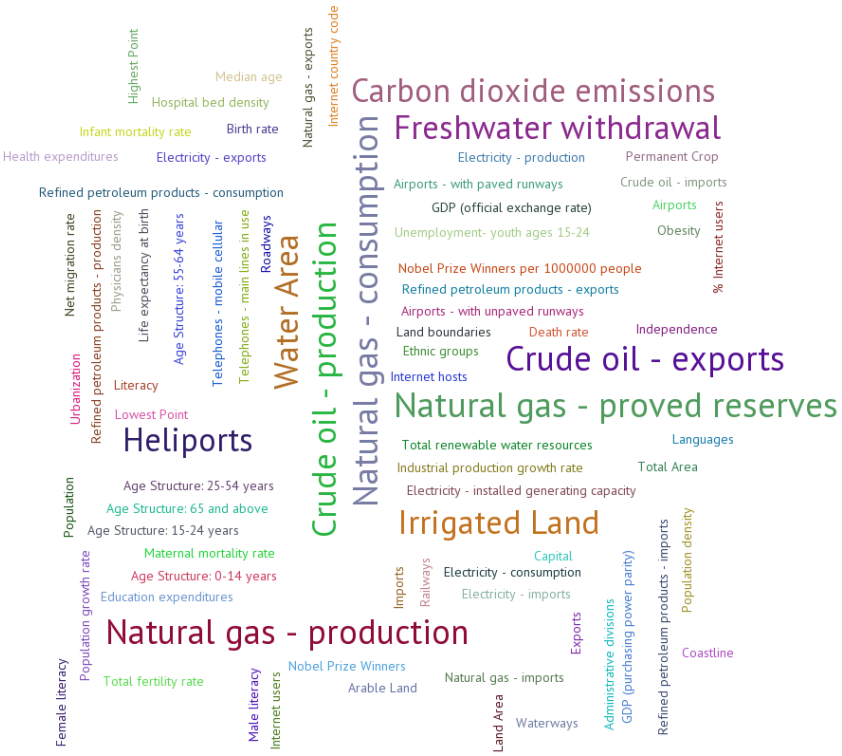


Figure 13.17: Tag Cloud representing Iran in various properties

13.6 Conclusion

Heuristics helped us in generating stories about countries. In this chapter intelligent text processing was proposed for the generation of stories along with simple but effective facts manipulation. We have identified unusual facts about Nobel Prize winners. We have explored and analysed textual material in Wikipedia. The analysis and text processing performed on Wiki articles revealed a wealth of interesting information about countries that cannot be extracted manually. The discovery of related and interesting facts was only possible using intelligent heuristics. The emerging mapping tools and libraries have reduced the amount of effort needed to create interactive maps. Student's feedback confirmed the appeal of the generated stories, see Chapter 15 for student's feedback.

References

[1] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology,” CEUR Workshop Proc., vol. 334, pp. 59–73, 2008.

[2] O. MEDELYAN, D. MILNE, C. LEGG, and I. WITTEN, “Mining meaning from Wikipedia,” International Journal of Human Computer Interactions, vol. 67, no. 9, pp.716-754, 2004.

[3] C. A. Yeung and A. Jatowt, “Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining,” Proc. 20th ACM Int. Conf. Inf. Knowl. Manag., pp. 1231–1240, 2011.

[4] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, “BioRAT: Extracting biological information from full-length papers,” Bioinformatics, vol. 20, no. 17, pp. 3206–3213, 2004.

[5] R. Feldman and I. Dagan, “Knowledge Discovery in Textual Databases (KDT).,” Int. Conf. Knowl. Discov. Data Min., pp. 112–117, 1995.

[6] T. Nasukawa and T. Nagano, “Text analysis and knowledge mining system,” IBM Syst. J., vol. 40, no. 4, pp. 967–984, 2001.

[7] R. Mehmood, N. Kulathuramaiyer, and H. Maurer, “A New Look at Geography of the World,” IPSI, vol. 12, no. 1, pp. 21–29, 2016.

[8] “Current List of Stories.” [Online]. Available: http://austria-forum.org/af/Geography/Main_Ideas/Current_List_of_Stories. [Accessed: 15-Feb-2016].

[9] “Geography im Austria-Forum.” [Online]. Available: http://austria-forum.org/af/Geography. [Accessed: 02-Nov-2015].

[10] M. Gahegan, M. Wachowicz, M. Harrower, and T.-M. Rhyne, “The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation,” Cartogr. Geogr. Inf. Sci., vol. 28, no. 1, pp. 29–44, 2001.

[11] M. P. Verdi, C. S. M, and D. R. White, “Leraning Effects of Print and Digital Geographic Maps,” J. Res. Technol. Educ., vol. 35, no. 2, pp. 290–302, 2002.

[12] H. Maurer and R. Mehmood, “Merging image databases as an example for information integration,” CEJOR, vol. 23, no. 2, pp. 441–458, 2015.

[13] “National Flowers - List Of National Flowers by Country.” [Online]. Available: http://www.theflowerexpert.com/content/aboutflowers/national-flowers. [Accessed: 02-Nov-2015].

[14] “10 Most Famous Trees in the World – Touropia Travel Experts.” [Online]. Available: http://www.touropia.com/famous-trees-in-the-world/. [Accessed: 20-Oct-2015].

[15] “United Arab Emirates Guide -- National Geographic.” [Online]. Available: http://travel.nationalgeographic.com/travel/countries/united-arab-emirates-guide/. [Accessed: 15-Feb-2016].

[16] “25 Longest Bridges In The World.” [Online]. Available: http://list25.com/25-longest-bridges-in-the-world/5/. [Accessed: 15-Feb-2016].

[17] "FactbookXML - All textual information of the CIA World Factbook in one XML file." [Online]. Available: http://jmatchparser.sourceforge.net/factbook/. [Accessed: 21-Oct-2015].

[18] "Food in Every Country." [Online]. Available: http://www.foodbycountry.com/. [Accessed: 20-Oct-2015].

[19] "National Dishes." [Online]. Available: https://en.wikipedia.org/wiki/National_dish.

[20] "Great Human Inventions by country."[Online]. Available: http://www.eupedia.com/europe/list_of_inventions_by_country.shtml

[21] "Production Statistics - OICA." [Online]. Available: http://www.oica.net/category/production-statistics/. [Accessed: 17-Mar-2016].

[22] "Ports." [Online]. Available: http://ports.com/. [Accessed: 19-Oct-2015].

[23] "Nobelprize.org." [Online]. Available: http://www.nobelprize.org/. [Accessed: 19-Oct-2015].

[24] "Free photo: Edelweiss - Free Image on Pixabay - 818414." [Online]. Available: https://pixabay.com/en/edelweiss-leontopodium-microdochium-818414/. [Accessed: 02-Nov-2015].

[25] T. Bard, S. David, G. Friedland, B. Elizalde, N. Karl, D. Poland, D. Borth, and L. Jia Li, "vfcc100M: The New Data in Multimedia Research," Commun. ACM, vol. 59, no. 2, pp. 64–73, 2016.

[26] "The_ Waitomao_Caves." [Online]. Available: http://austria-forum.org/af/Geography/Australia/New_Zealand/Special_Information/The_Waitomao_Caves. [Accessed: 19-Oct-2015].

[27] "Shahara - The incredible city." [Online]. Available: http://austria-forum.org/af/Geography/Asia/Yemen/Special_Information/Shahara-_the_incredible_city. [Accessed: 19-Oct-2015].

[28] "From Bungabiddy Rockhole to Mount Leisler via the Sandy Blight Junction Road." [Online]. Available: http://austria-forum.org/af/Geography/Australia/Australia/Special_Information/. From_Bungabiddy_Rockhole_to_Mount_Leisler. [Accessed: 19-Oct-2015].

[29] "Ayers Rock." [Online]. Available: http://austria-forum.org/af/Geography/Australia/Australia/Pictures/Factbook_Pictures/350as. [Accessed: 19-Oct-2015].

Chapter 14 : Responsive Visualisations

This chapter presents responsive visualisations and charts. The idea of responsive visualisation is similar to responsive design. The aim is to create visualisations that work on all devices. It is needed at a time like now, when devices with very different screen sizes are proliferating.

This chapter first presents the state of art programmable graphics technologies such as HTML5 Canvas, SVG, and WebGL. Recent years have seen great advancement in JavaScript libraries that support visualisation. We will explain the use of these libraries. We will present examples of different charts that are built using D3, Chart, Highcharts and Chartist. We will uncover the techniques that make charts responsive. There are a few challenges which comes across while making responsive visualisations. Those are also discussed.

We will briefly describe the state of the art graphics technologies that are used inside web browser in Section 14.1. Section 14.2 briefly explains graphics libraries. Different techniques for making responsive visualisations are discussed in Section 14.3.

Following are the research questions that are addressed in the chapter.

RQ1. How can we make a line chart responsive?

RQ2. How can we make a bar chart responsive?

RQ3. How can we make a pie chart responsive?

RQ4. How can we make parallel coordinates responsive?

RQ5. What are the existing responsive JavaScript libraries used for drawing graphics and visualisations?

RQ6. What are the various ways and tricks to make different visualisations responsive?

14.1 Programmable Graphics Technologies inside Web Browser

This section focuses on state of the art graphics technologies that are used inside web browser. We will discuss those technologies that are used in combination with JavaScript.

14.1.1 HTML5 Canvas (2D via JS)

HTML5 brings revolution in graphics handling inside browser by introducing canvas element [1]. It is a rectangular area on the HTML page. It offers numerous possibilities for rendering graphs, game graphics, visual images editing etc. It is not more than a two-dimensional grid. The coordinate (0,0) is the upper-left corner of the canvas (2D) as shown in the Figure 14.1. It requires JavaScript to draw graphics. It offers drawing context (surface) to draw graphics. The canvas element is the actual DOM node that is embedded in the HTML page. The canvas context is an object with properties and methods that one can use to render graphics inside the canvas element.

Let us walk through the example shown below:

```
<canvas id= "myCanvas" width="200" height="100" style="border:1px solid #000000;">
</canvas>
<script>
 var c = document.getElementById("myCanvas");
 var ctx = c.getContext("2d");
 ctx.fillRect(0,0,150,75);
</script>
```

The markup for canvas element is shown above. The canvas element is first drawn on the html page. It is essential to specify an id attribute in order to get the canvas reference in the script. The drawing context is get using getContext method. We are using 2D context for this example. The canvas element has no border by default. We can add border using CSS style. Afterwards, a rectangle is drawn using fillRect() method. The value along X-axis increases towards right whereas, it increases towards bottom along Y-axis [2].

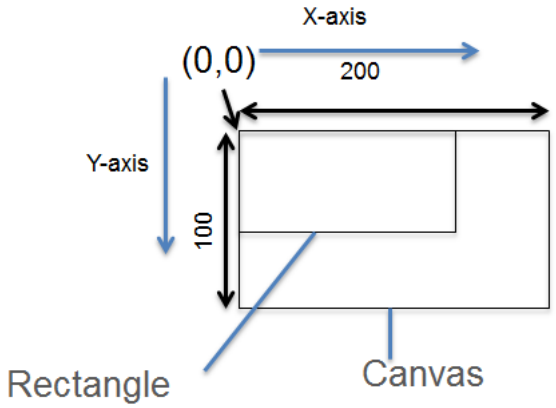


Figure 14.1

14.1.2 SVG

SVG stands for Scalable Vector Graphics. SVG defines vector-based graphics in XML format. Scalable Vector Graphics (SVG) is a language for describing two-dimensional graphics in XML. There is no quality loss if it is zoomed or resized. It can be animated. It is integrated with DOM. It has great browser support. It is recommended by W3C. The code below shows how to draw rectangles using SVG. One can immediately think of making bar charts using SVG <rect> element. Three rectangles are shown in the Figure 14.2 using the code below.

```
<html>
<body>
<svg width=" 200" height=" 400 " style=border:solid 1px black>
<rect width="150" height="75" style="fill:green";  />
<rect x="0" y="80" width="180" height="75" style="fill:green" />
<rect x="0" y="160" width="200" height="75" style="fill:green" />
</svg>
</body>
</html>
```

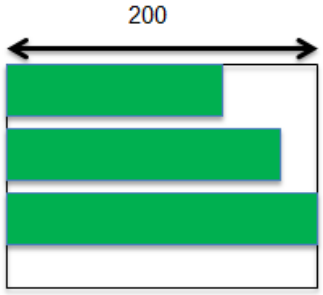


Figure 14.2

14.1.3 WebGL

It is used for 3D rendering. It is based on OpenGL. In contrast to Canvas (2D) the top left corner is at (-1,-1) the bottom is at (1,1). These are basically clip space coordinates. It provides support for hardware accelerated graphics. Browser support is not as good as that of canvas and SVG. Google Chrome and Internet Explorer 11 fully support it. Firefox, Safari, Opera have partial support. It eliminates the need to install a 3rd party plugin e.g. Flash. It is fast and powerful in performance. It makes use of HTML5 canvas element. It requires shaders (special programs) to draw graphics. Different ways to get a WebGL context is shown below.

```
var canvas = document.getElementById("canvas");
var gl = canvas.getContext("webgl");
var gl = canvas.getContext("experimental-webgl");
```

14.1.4 A Comparison

There are few notable differences between SVG, canvas and WebGL.

- SVG is object model based. It is integrated with DOM. Event handlers are attached to individual elements. SVG offers retained mode graphics. One important aspect about performance should be taken into consideration while using SVG. There is a great performance degradation if many SVG objects are used. It is best suited for charts and visualization.

- As compared to SVG, canvas provides pixel based graphics. The whole canvas element has to be redrawn. On the fly graphics are made using JavaScript. It provides immediate mode graphics. It is fast and easy to use and best suited for game development.

- WebGL offers hardware accelerated graphics. It supports 3D graphics. The big drawback is that it is difficult to code. However to overcome the difficulties of using WebGL there are many high level libraries such as Three.js which provides easy to use functions to operate over WebGL.

14.2 Visualisation Libraries

This section explores JavaScript libraries that are used for visualisations. These libraries use the technologies discussed in Section 14.1. It is important to understand the working of these libraries before applying responsive techniques. A basic comparison is shown in Table 14.1.

Table 14.1: Visualisation Libraries: A comparison

Libraries	Underlying Technology	Mode	Responsive
D3	SVG, Canvas	2D	No
Charts.js	Canvas	2D	Yes
Chartist	SVG	2D	Yes
Highcharts	SVG	2D, 3D	Yes

14.2.1 D3

It is a de-facto standard for web-based data visualisations. The D3 comes from **D**ata **D**riven **D**ocuments [3]. It manipulates documents based on data. It creates charts using SVG. Recently there are few examples on web demonstrating visualisations with D3 and canvas. It is good for complex visualisations. It is compatible with every browser that supports SVG. It was introduced by Mike Bostock. It is not responsive by default. There are two main solutions that make visualisations responsive in D3 as discussed by Peter Le Bek, see [4] .

- Using ViewBox attribute of SVG
- Using Windows Resize Event

We will explain these solutions with examples in Section 14.3.

14.2.2 Chart.js

In contrast to D3, it uses HTML5 canvas. It provides a global variable of Chart. It uses 2D context of canvas to draw the chart. It provides global chart configuration. It was introduced by Nick Downie. Figure 14.3 shows a responsive bar chart in Chart.js[31]

```
<canvas id="myChart" width="400" height="400"></canvas>
var ctx = document.getElementById("myChart").getContext("2d");
var myNewChart = new Chart(ctx).Bar(data);
```

Bar chart in Chart.js

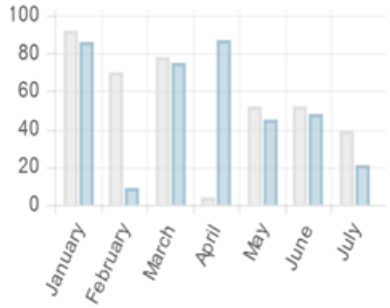


Figure 14.3: Responsive bar chart in Chart.js. Example taken from [10]

[31] http://www.chartjs.org/

14.2.3 Chartist

It has logo “simple responsive charts”. It uses SVG. It has good support for animation. Chartist uses its own classes to set containers with fixed aspect ratios (intrinsic ratios). Let us look at two different versions of chart shown in the Figure 14.4.

Figure 14.4: Charts in chartist.js [12].

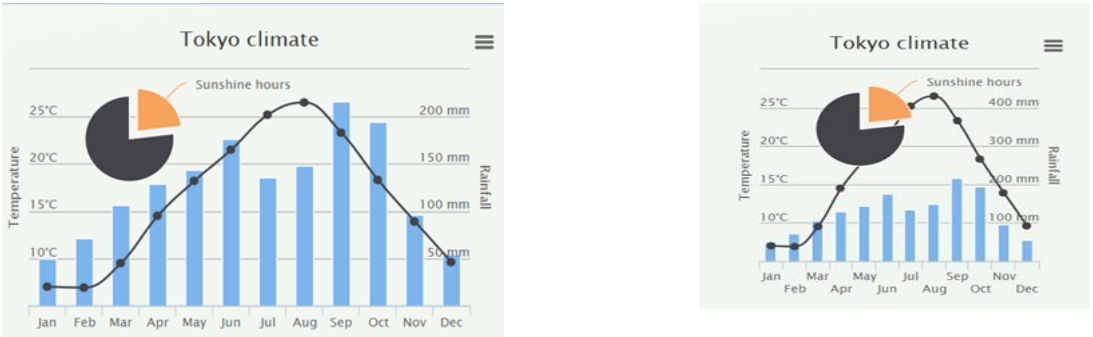
Chartist[32] provides an easy to use responsive configuration mechanism. One can easily override the existing configuration. Listing 14.1 shows two configurations. Points are not shown on big screens. Labels are modelled as Week 1, 2 and so on. This configuration is set using variable *responsiveOptions*. On narrow screens only first letter of week is shown and chart is showing only points instead of line as shown in the Figure 14.4.

```
var responsiveOptions = [ ['screen and (min-width: 40em) and (max-width: 64em)',
 { showPoint: false, axisX: {labelInterpolationFnc: function(value)
 { return 'Week ' + value; } } }], ['screen and (max-width: 40em)',
 { showLine: false,  axisX: { labelInterpolationFnc: function(value) { return 'W' +
value; } }}]
```

Listing 14.1: Responsive configuration for charts in Chartist.js

14.2.4 Highcharts

Highcharts[33] is a JavaScript charting library. It is developed by Highsoft. It uses SVG. It is free for non-commercial use. It provides facility to export charts in multiple formats as png, jpeg, svg, and pdf. It is responsive by default. It can be used with jQuery, MooTools. Figure 14.5 shows different versions of line chart using Highcharts.



Axis labels displayed on single line on wide screens

Axis labels displayed on multiple lines on mobile screens

Figure 14.5: Different versions of chart using Highcharts. Example taken from [11]

[32] https://gionkunz.github.io/chartist-js/
[33] http://www.highcharts.com/

14.3 Responsive Visualisations

This section is about responsive visualisations that we have created using d3.js. We will highlight different ways to create responsive visualisations that work on multiple devices having different screen sizes. We have applied various techniques on multiple charts and visualisations to make them responsive that will be discussed in this section.
Due to increasing popularity in "Responsive Web Design", Firefox has initiated a Responsive Bookmarklet to the Browser [5]. This tool allows to test responsive website in different viewports without resizing the browser window. This view is available from this menu: Tools > Web Developer > Web Design View. We will present a few examples of testing a website using this tool. Further we have used mattkersely.com for testing responsive visualisations[6].

14.3.1 Bar Chart

It is a graphical display of data using bars of different height or width. It is used to compare multiple classes or groups of data. The bar height or width is proportional to value.

There are two main types of bar charts

- Vertical bar chart
- Horizontal bar chart.

Vertical bar charts are also known as column charts. When resizing bars, relationship/scale must be preserved. If X is 5 times larger than Y, the bar chart must retain these proportions on a narrow screen.

Figure 14.6 is showing different versions of vertical bar chart. On desktop with screens having width 64em, we have axis labels positioned horizontally. On medium size screens having minimum width 30em, labels are rotated 45 degree; whereas on mobiles, where space is much less, labels are rotated vertically.

Figures 14.7 and 14.8 show bar charts with two series. Weekdays are displayed in full on large screens whereas on small screens following techniques are applied:

- Bar width is reduced.
- Weekdays/months are abbreviated
- Overlapping of series bar.
- Font size of both graph labels and axis labels is reduced

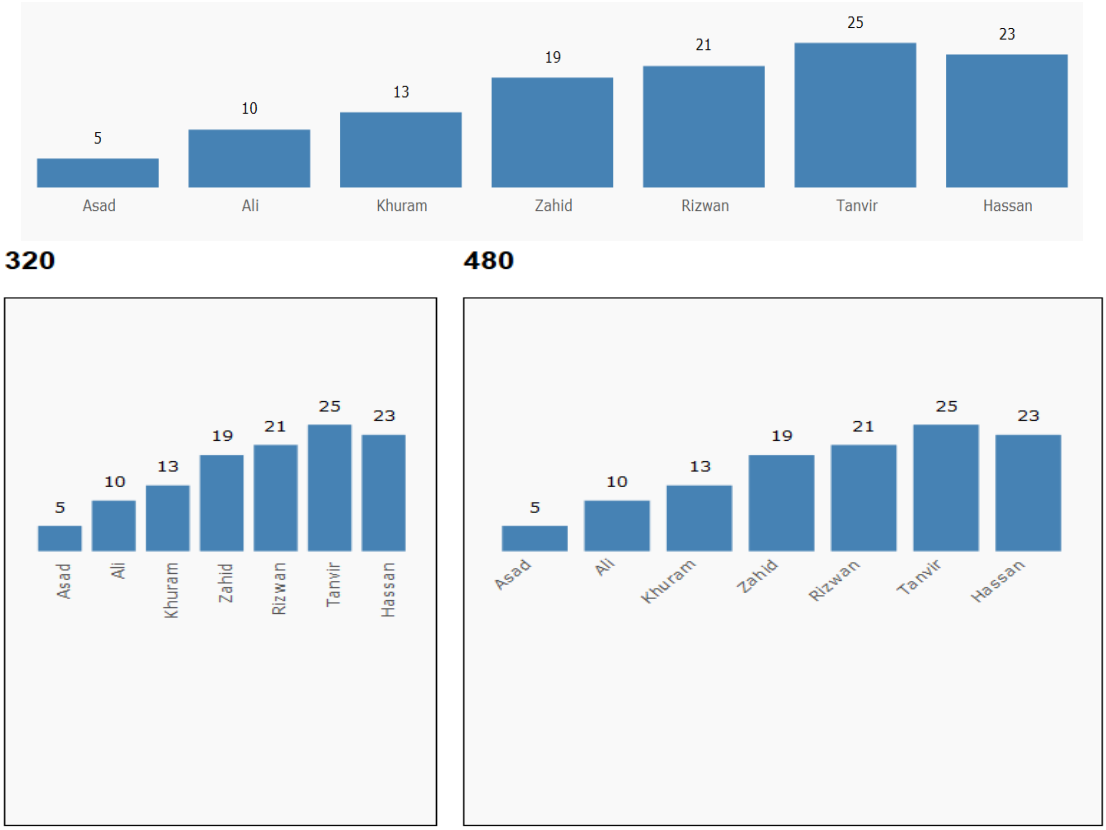


Figure 14.6: Axis labels rotation for smaller screens to avoid mixing of labels

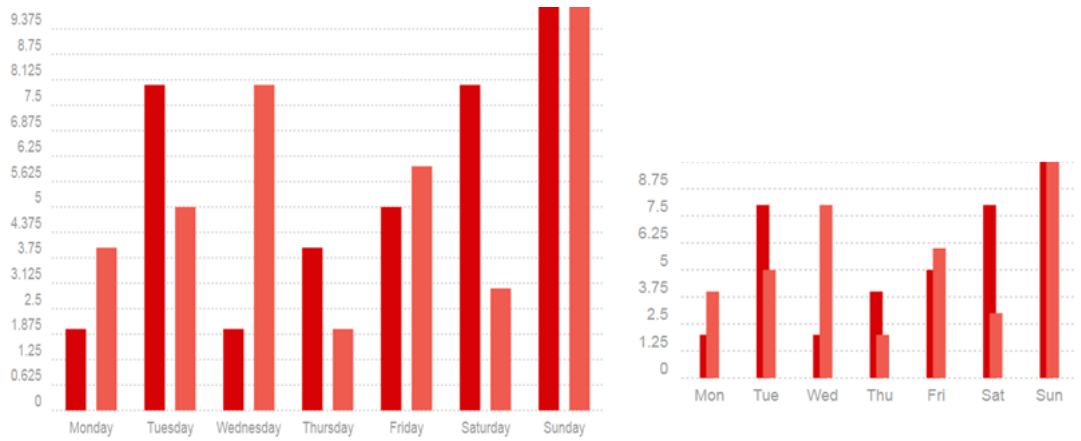


Figure 14.7: Axis labels shortened; bars are overlapped. Example taken from [9]

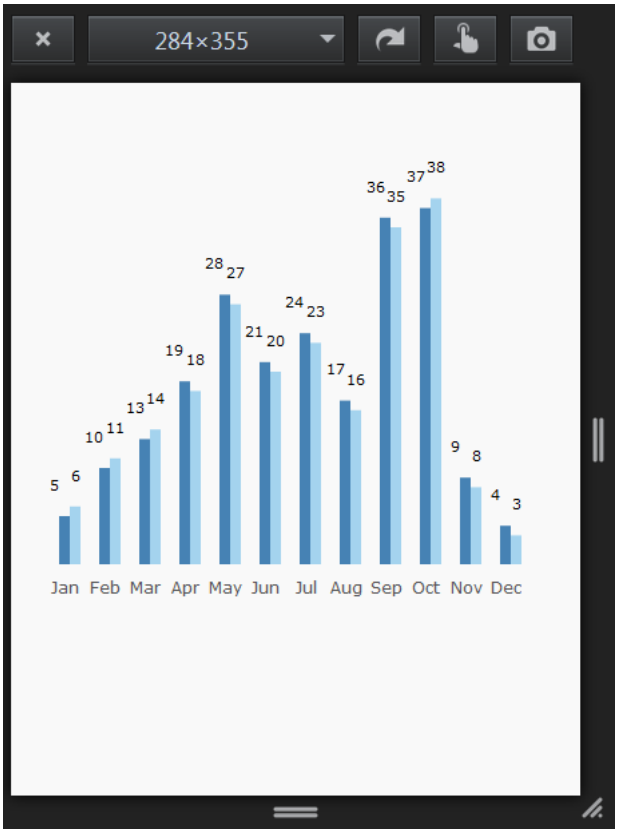


Figure 14.8: Responsive vertical bar chart with multiple series in D3.js

An example of zigzag pattern of axis labels is shown in the Figure 14.9. Month labels are arranged in multiple lines using the code below in D3.js.

```
.attr("dy", function(d,i)
{if(i%2==0)return convertRem(1.25);else return convertRem(0.625);})
```

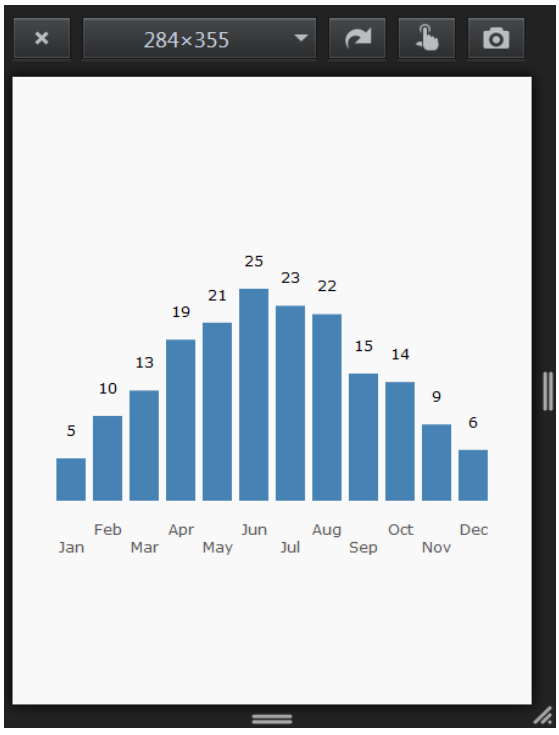


Figure 14.9: Axis labels are arranged in two lines to utilize space efficiently on mobiles

Figure 14.10 shows different versions of horizontal bar charts displaying election results about different parties of Austria. The bars are colour coded according to different parties.

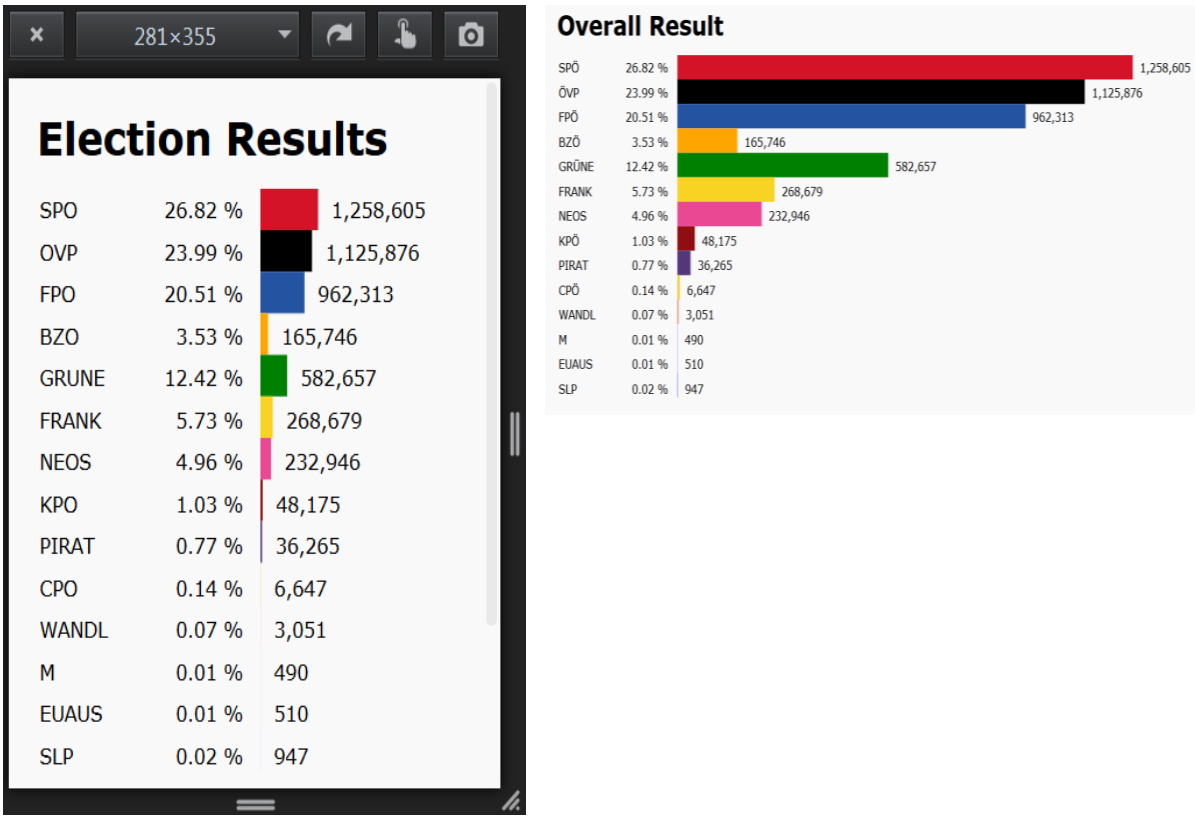


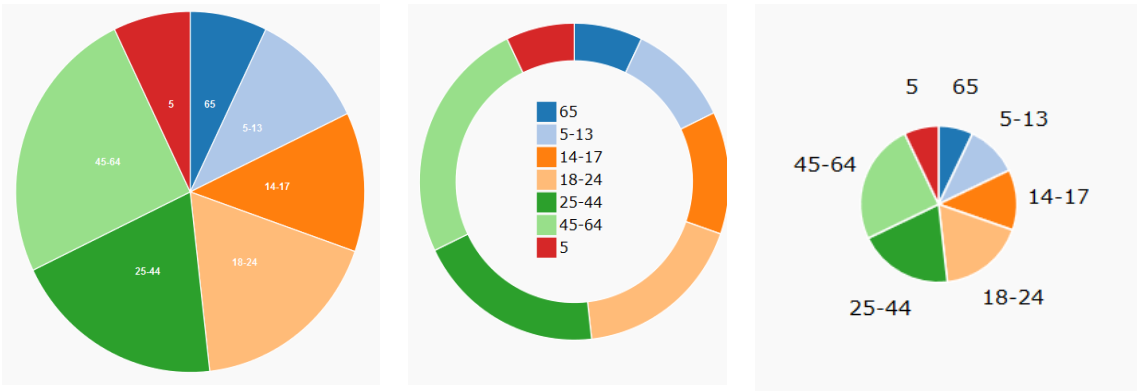
Figure 14.10: Horizontal bar charts on small and large screens: Election dataset is used.

14.3.2 Pie Chart

Pie chart is a circular chart divided up into sectors. It is best for showing Part-to-whole relationship. Figure 14.11 shows the pie chart for desktop screens and mobiles. There are two techniques that make this visualisation responsive.

14.3.2.1 Technique 1

- Move the labels towards centre of chart.
- Increases the inner radius from 0 up to approximate value so that labels adjust inside.
- Pie chart conversion to Donut chart



(a) (b) (c)

Figure 14.11: Pie Chart on desktop screens on left in (a) taken from [8]. Inner radius is increased to create space for legends to adjust inside as shown in (b). Labels move outside from the sectors as shown in (c).

14.3.2.2 Technique 2

- Move the labels outside from the sectors of chart.
- It would be more appropriate to use lines that draw from labels to center of arcs to remove any ambiguity, see Figure 14.11 (c).

14.3.3 Chord Diagram

It is combination of arcs and a set of ribbons that show mapping between them. Figure 14.12 shows a chord diagram. In terms of responsiveness, the following techniques are applied:

- Arcs radius is resized according to the screen width.
- Arc labels are positioned along the arcs.
- At some point, the font-size of arc labels is reduced.

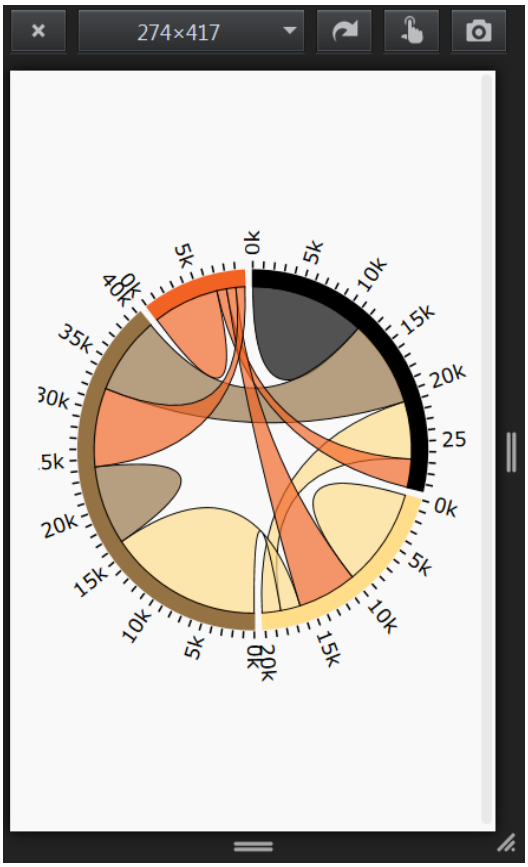


Figure 14.12: Chord Diagram, taken from [8].

14.3.4 Parallel coordinates

It is used for analysing high dimensional data. The columns are mapped onto vertical axis. It is ideal for identifying patterns among different variables. Each dimension is a vertical axis. Each record is a polyline.

Figure 14.13 shows parallel coordinates. There are two ways of making this visualisation responsive:

- Scaling with axis adjustment on different screens.
- Selective columns display.

14.3.4.1 Scaling with axis adjustment on different screens

One way is to scale the graph. In order to avoid axis labels overlapping on narrow screens, take following actions:

- Rotate the axis labels (30-60) degree to accommodate them on tablets and medium size screens
- Rotate the axis labels 90 degree to fit them on narrow screens as shown in the Figure 14.13.

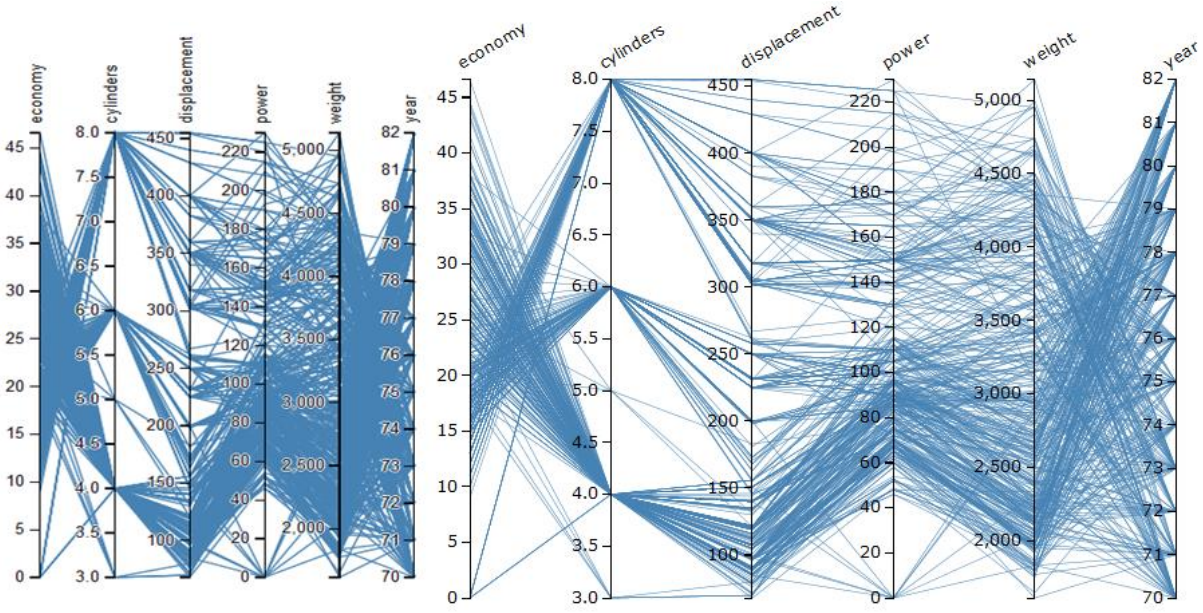


Figure 14.13: Axis rotation in parallel coordinates, created using d3.

14.3.4.2 Selective Columns Display

There is another approach (similar to responsive tables) of making “parallel coordinates” visualisation responsive. All axis (dimensions) are displayed on large screens. On narrow screens remove few axis (dimensions). In addition, provide a mechanism (a dialog box with checkboxes) for selecting the columns that needs to be drawn on narrow screens as shown in the Figure 14.14. Afterwards, only selective columns (economy, cylinder, displacement) are displayed as shown in the Figure 14.14.

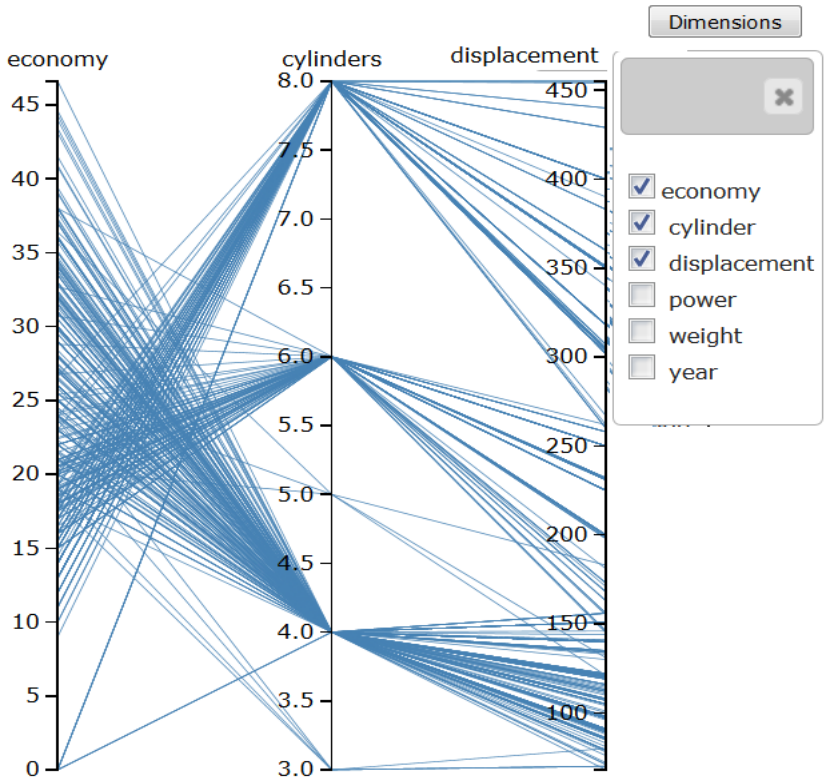


Figure 14.14: Selective Columns Display.

14.3.5 Line Chart

We start by highlighting the problem in the line chart of Olympic medals in Figure 14.15. Axis labels start to overlap on narrow screens, although line chart scales well according to the screen width.

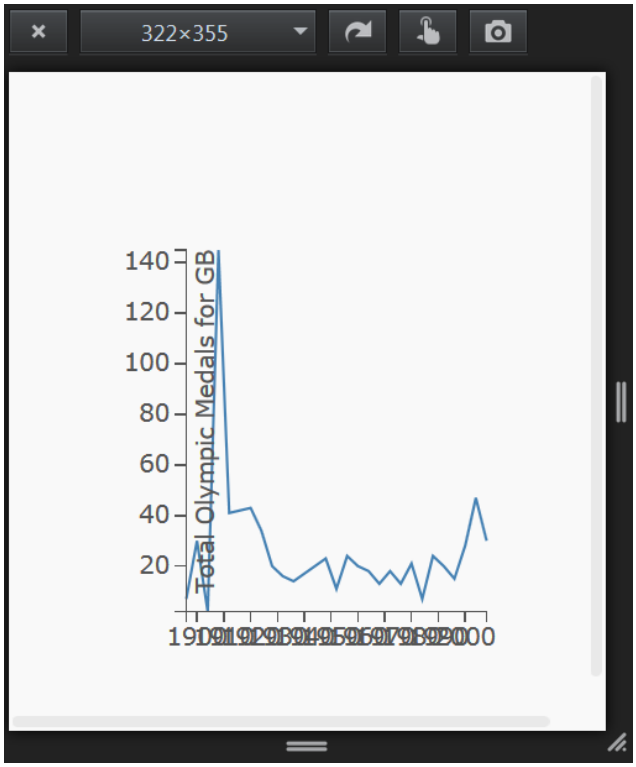


Figure 14.15: Axis labels mixing on narrow screens.

Figure 14.16 shows few axis labels that remain according to screen width. On further narrow screens spark lines may suffice.

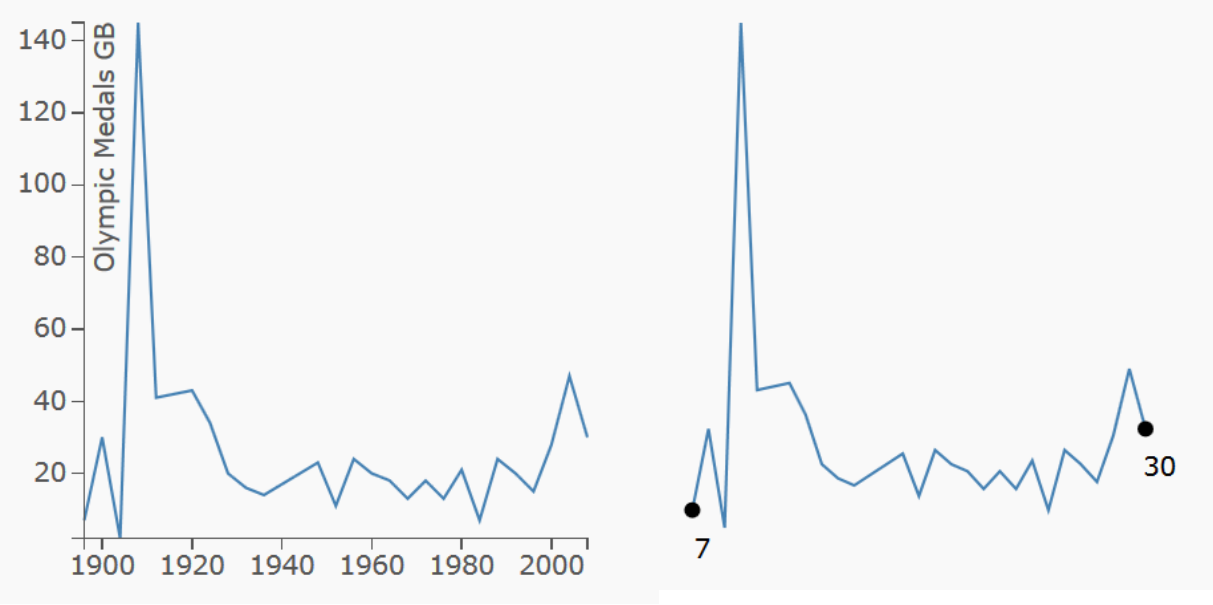


Figure 14.16: Few axis labels remain on narrow screens. Sparkline for mobiles.

Another example of line chart using HighCharts is shown in the Figure 14.17. All month labels are displayed on the large screens. On narrow screens few labels remain that fit on the narrow screens.

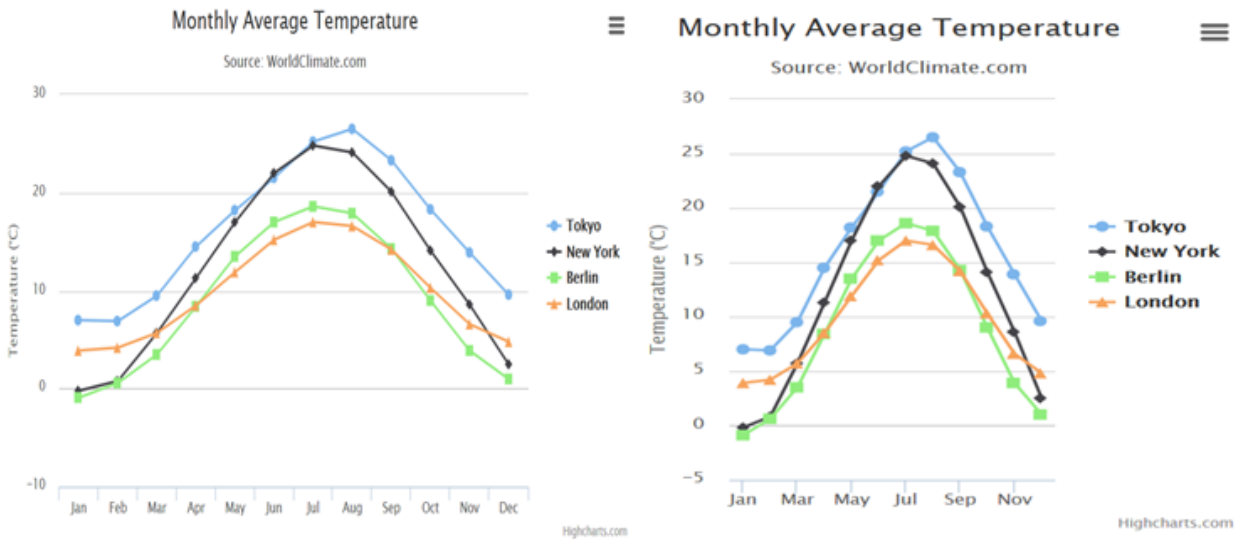


Figure 14.17: Few axis labels (months) remain on narrow screens. Example taken from [7]

14.3.6 Tree layout

The tree layout produces node-link diagrams. For example, a tree layout can be used to represent software classes in a hierarchy as shown in the Figure 14.18, Example taken from d3js.org.

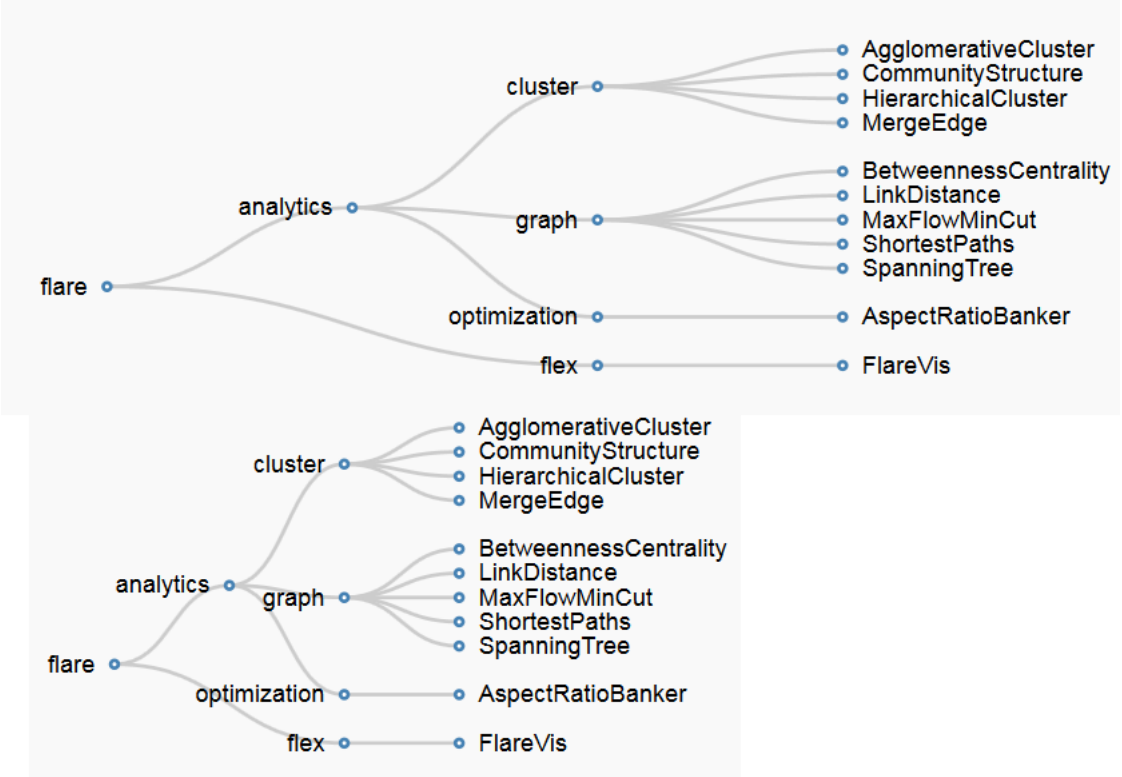


Figure 14.18*: Tree layout showing software hierarchy Example taken from [8]. The layout size is reset for small screens.*

This visualisation can be made responsive using:

- On windows resize event, capture new width and height.
- Reset the size of tree layout according to new width and height.

- Decrease the size of circles representing nodes on small screens.

14.3.7 Indented Tree

The indented tree layout is commonly used for hierarchical lists, e.g. file directories, as shown in the Figure 14.19.

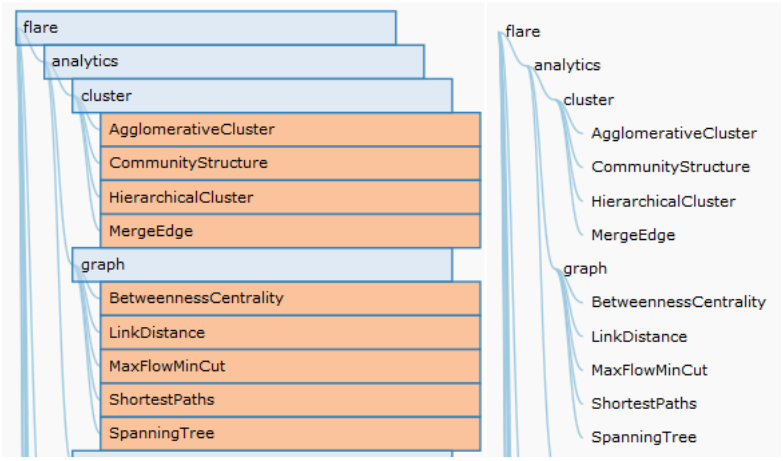


Figure 14.19: Indented tree layout showing software hierarchy, Example taken from [8].

This visualisation can be made responsive using:

- On windows resize, reduce the width of horizontal bars.
- Remove the horizontal bars on small screens.

14.3.8 Sunburst

Unlike resize event we have used viewbox attribute of SVG to make sunburst scalable as shown in the Figure 14.20.

```
<svg id="graph" viewbox="0 0 900 600"></svg>
```

Figure 14.20: Different views of scalable Sunburst, Example taken from [8].

References

[1] "HTML5 Canvas." [Online]. Available: http://www.w3schools.com/html/html5_canvas.asp. [Accessed: 07-Feb-2016].

[2] "HTML5 Canvas Tutorials." [Online]. Available: http://www.html5canvastutorials.com/. [Accessed: 07-Feb-2016].

[3] "D3 — Scott Murray — alignedleft." [Online]. Available: http://alignedleft.com/tutorials/d3. [Accessed: 07-Feb-2016].

[4] "Building Responsive Visualizations with D3.js - Safari Blog." [Online]. Available: https://www.safaribooksonline.com/blog/2014/02/17/building-responsible-visualizations-d3-js/. [Accessed: 07-Feb-2016].

[5] "Essential Firefox Tools For Web Developers - Hongkiat." [Online]. Available: http://www.hongkiat.com/blog/firefox-web-development-tools/. [Accessed: 16-Mar-2016].

[6] "Responsive Design Testing." [Online]. Available: http://mattkersley.com/responsive/. [Accessed: 16-Mar-2016].

[7] "Interactive JavaScript charts for your webpage - Highcharts." [Online]. Available: http://www.highcharts.com/. [Accessed: 07-Feb-2016].

[8] "Data – Driven Documents" [Online]. Available: https://d3js.org/. [Accessed: 07-Feb-2016].

[9] "Chartist.js, An Open-Source Library For Responsive Charts", [Online]. Available: https://www.smashingmagazine.com/2014/12/chartist-js-open-source-library-responsive-charts/. [Accessed: 07-April-2016]

[10] "Chart.js", [Online]. Available: http://www.chartjs.org/. [Accessed: 07-April-2016]

[11] "Highcharts Demos: Basic line", [Online]. Available: http://www.highcharts.com/demo/line-basic/. [Accessed: 07-April-2016]

[12] "Chartist – Examples", [Online]. Available https://gionkunz.github.io/chartist-js/examples.html. [Accessed: 07-April-2016].

Chapter 15 : Summary and Evaluation

This chapter discusses the evaluation of our geographic server (global-geography.org) using different ways and approaches. Further, we summarize our research efforts and take a glimpse into the future. The chapter summarizes research presented in earlier chapters and outlines strengths and opportunities provided by the system. Further, we mention some potential threats that were critical for the reputation of system (copyright issues) and discuss how we have solved them.

15.1 System Evaluation

The first version of "global-geography.org" was made available to users and general public in 2015. Since then the system is up and running. The system has been extended using maps, stories, virtual lab, and interactive visualisations. The system is continuously expanding. A good example is the new "Community Contribution" page which was recently included in the "global-geography.org" in February 2016, see e.g. Community Contribution Page of Nepal [1].

We have evaluated different modules of system using three ways.

- Users Feedback
- Facts Summarization
- Comparison of global-geography.org with other Geographic Servers

Let us first quote: "Evaluation is never an easy task and consequently there are a lot of suggestions for how to evaluate IT-system" [2]

15.1.1 Users Feedback

An evaluation concerning quality of stories found under [3] was conducted using 20 students from Graz university of Technology. Initially, they are invited to read the stories (we evaluated stories by inviting one student at a time); afterwards they were requested to fill the evaluation form. Some pictures of evaluation sessions are shown in Figure 15.1. The video of full activity can be seen at [4]. There are three categories of students according to their expertise level in geography and knowledge about countries. Some have basic knowledge about countries, some are good and just a few of them are expert in geography having lots of information about the countries. The distribution according to level of expertise is shown in the Table 15.1. The evaluation form filled by the student (expert in geography) "Paso Sivro" is shown in Figure 15.2.

The students have different views regarding the introductory section (sentences taken from Wikipedia). Some have asked to start the story with Independence Day of a country. They all found facts about Nobel Prize winners surprising. Pictures, maps and short videos are highly appreciated by all students who participated. Overall students appreciated the stories a lot.

Besides feedback using forms, the system was evaluated with the help of user interviews and their general opinion. Additionally, Professor Maurer presented the system in different seminars. He took the feedback from the researchers in different conferences. This activity helped us a lot in improving the system. An example is trade visualisation where initially we were not showing export commodities. Based on the users' feedback we have added the export commodities on mouse over a country, see

Chapter 11 (interactive visualisations) and arrange the countries in alphabetical order in a circular fashion.

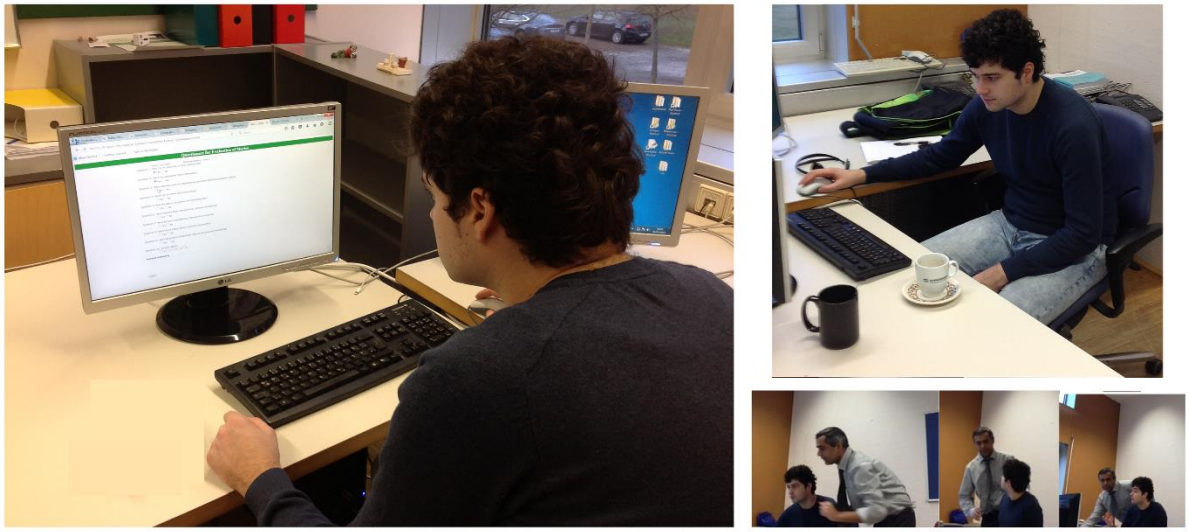


Figure 15.1: Evaluation of stories. The student Bruno Bakula (TU Graz ID No 1531612) is giving his feedback using the evaluation form in the picture on left. On the right side the author of this thesis is shown conducting an evaluation session and giving guidance.

Questionare for Evaluation of Stories

Name: Pašo Sivro Matrikelnummer: 0330969

Question 1: Was the introduction of story appropriate?
(•) Yes ( ) No

Question 2: Were the geographic facts interesting?
(•) Yes ( ) No

Question 3: Were dynamic maps of neighbouring countries showing population useful?
(•) Yes ( ) No

Question 4: Were the economic facts interesting?
(•) Yes ( ) No

Question 5: Was the labour occupation an interesting fact?
(•) Yes ( ) No

Question 6: Were national facts representing countries interesting?
(•) Yes ( ) No

Question 7: Were pictures representing national facts essential?
(•) Yes ( ) No

Question 8: Were short videos about countries appropriate?
(•) Yes ( ) No

Question 9: Were facts about Nobel Prize winners amusing and interesting?
(•) Yes ( ) No

Question 10: Overall rating?
[ ] 1 [ ] 2 [ ] 3 [ ] 4 [x] 5

General comments

I find Greece facts interesting. Population display on mouse over a country was fascinating. Nobel Prize winners facts are surprising for me.

Submit

Figure 15.2: Evaluation form filled by Pašo Sivro.

Table 15.1: Expertise-wise Distribution

S#	Type	Total
1	Beginner	10
2	Intermediate	8
3	Expert	2

Let us turn to ratings given by students. The satisfactory part was that, out of 20 students 17 have given rating 5 to stories. Only 3 students have given rating 4 to stories, none has rated stories below average. The results of ratings given by students are shown in the Figure 15.3.

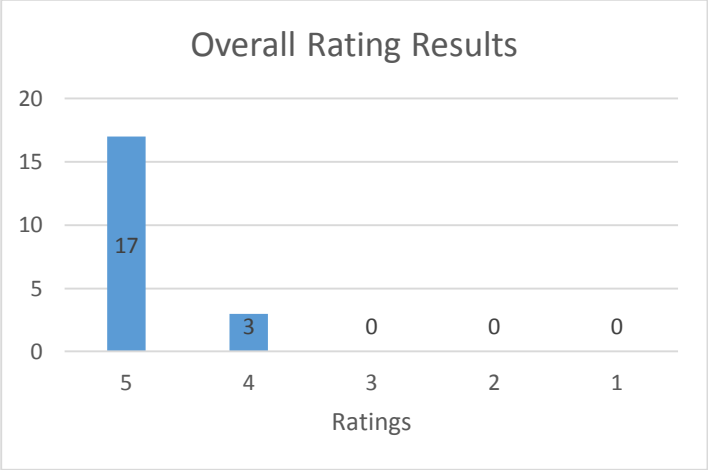


Figure 15.3: Ratings given by students about stories.

15.1.2 Facts Summarization

The global-geography.org [5] provides approximately 30,000 facts about countries and territories. A sample is checked manually for further reliability by Austria-Forum editors and Quality Assurance analysts. It has around 18,000 static pages displaying information about different aspects of countries. We present some quantitative statistics about facts as shown in Figure 15.4.

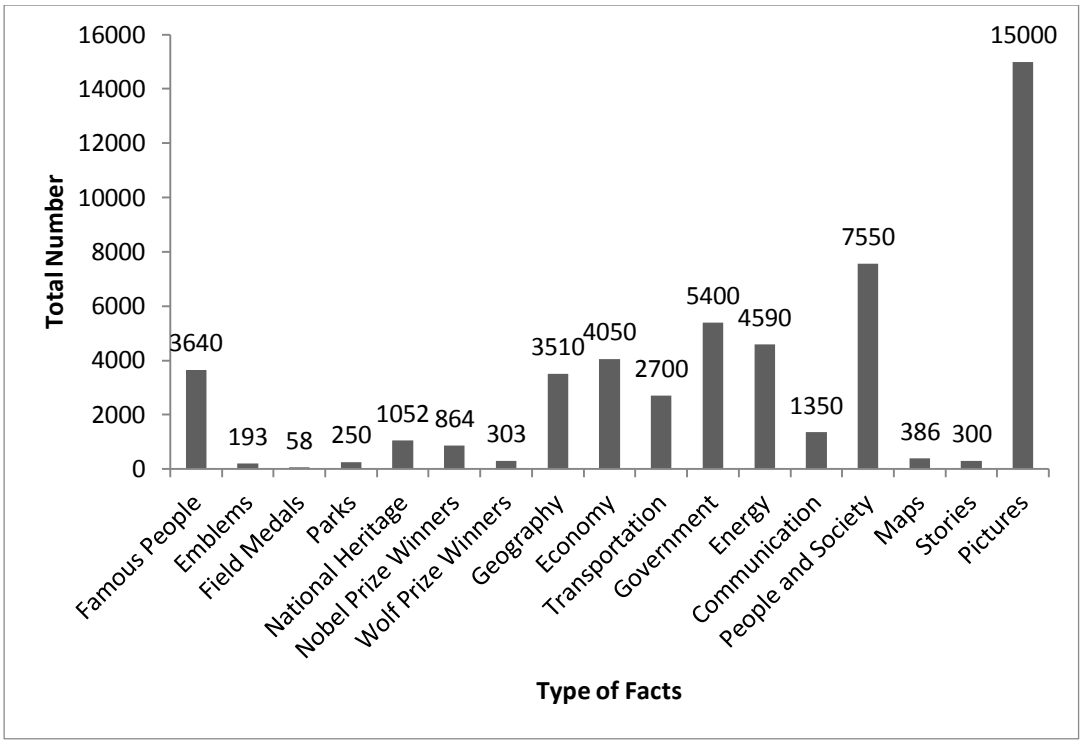


Figure 15.4: Summary of Facts (Taken on 15 March 2016)

The number 3510 against Geography comes from the following procedure. We collected approximately (13) facts such as land area, water area etc. for around 270 countries and territories. Therefore we are showing (3510=270*13).

The statistics of Geographic Server is shown in the Table 15.2 (sample data taken on 15 March 2016). The "Categorycountplugin" of JSPwiki is used to generate these results. There are 17,921 wiki pages in the server. The type "Attachment" can be a picture or any multimedia object. There are 15,580 attachments in the system. The "Objects" type represents the sum of Wikipages and Attachments (17921+15580=33501).

Table 15.2: Statistics of Geographic Server (Data taken on March 15, 2016)

S#	Type	Number
1	Wikipages	17,921
2	Attachments	15,580
3	Objects	33,501

We have partial lists of famous people for around 82 countries of the world which we are extracting from the site "The Famous People" [6]. The chart shown in the Figure 15.5 represents the count of famous people against 10 countries in "global-geography". The list of famous people is certainly incomplete. We expect that after officially opening the server in April 2016 we will get many suggestions for adding persons.

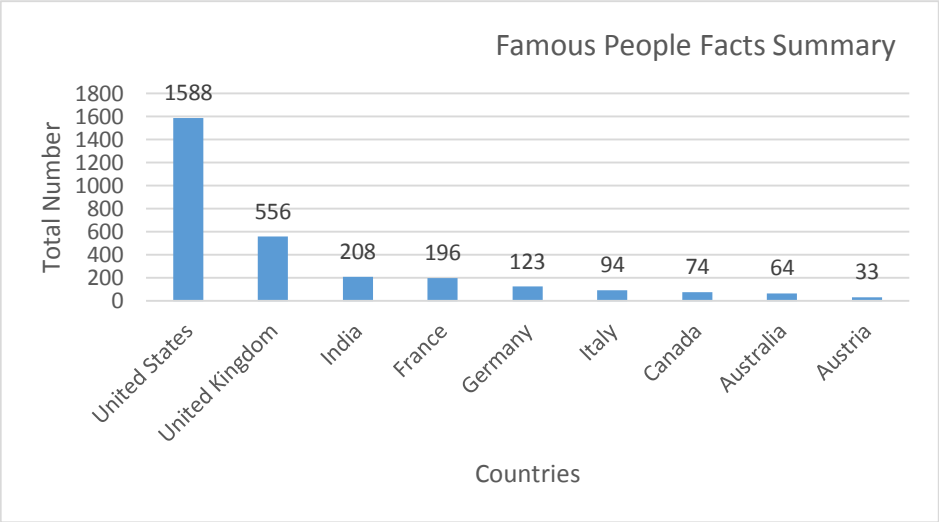


Figure 15.5: Number of Famous People in "global-geography.org" for 10 countries, data taken on March 15, 2016.

We have heritage lists (facts+links) for around 161 countries of the world which we are extracting from "UNESCO World Heritage"[7]. The chart shown in the Figure 15.6 represents the count of heritage sites against 10 countries in "global-geography". Italy is dominating in the list.

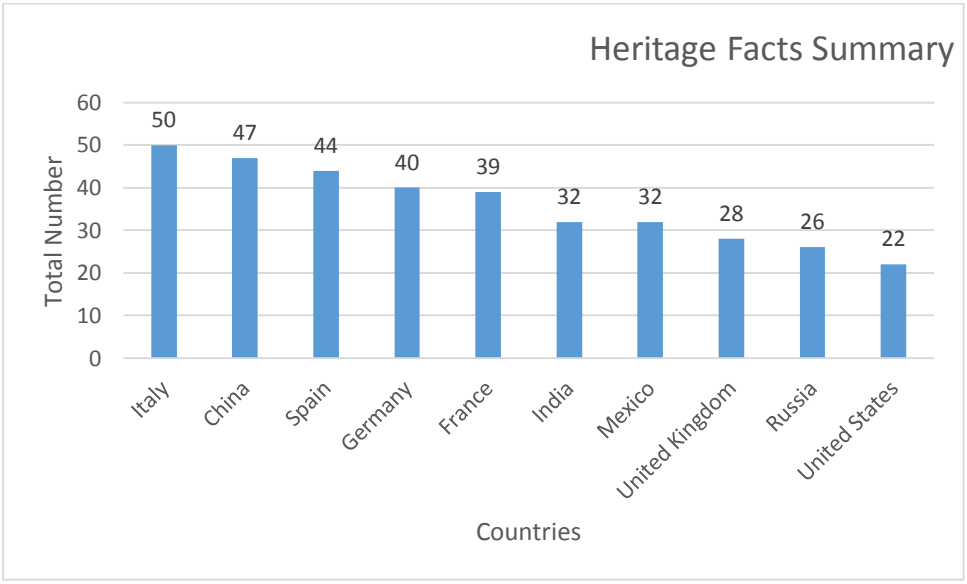


Figure 15.6: Number of Heritage sites in "global-geography.org" for 10 countries, data taken on March 15, 2016

15.1.3 Comparison with Existing Geographic Servers

In this section we will present a comparison with other geographic servers (Infoplease, Factbook, DBpedia, WordAtlas, Factfish, and Geonames) as mentioned in the Table 15.3. The selected servers are well known geographic servers or data sources. They contain vast amount of geographic data. The comparison is based on services and types of facts offered by global-geography.org.

Table 15.3: Comparison with Existing Servers

	Infoplease	Factbook	DBpedia[34]	global-geography	WorldAtlas	FactFish	GeoNames
Facts							
Geography	√	√	√	√	√	√	√
Economy	√	√		√		√	
Transportation		√		√		√	
Government	√	√		√		√	
People and Society	√	√	√	√	√	√	
Energy		√		√		√	
Communications	√	√		√		√	
Flags	√	√		√		√	√
Emblems				√	√		
Pictures							
Pictures with titles		√		√	√	√	
Pictures with description				√			
Historic Pictures				√			
Facts Validation-Service							
Area				√			
City Rankings				√			
Mountain Rankings				√			
Facts Exploration							
Virtual Lab				√			
Rankings		√		√		√	
Countries Similarity				√			
Visualisations							
Static Visualisations				√		√	√
Interactive Visualisations (that allow comparison)							
Population				√			
Population growth rate				√			
Electricity Production				√			
Economic Trade				√			
Choropleth/Thematic maps				√			
Fancy visualisations				√			
Culture							
Nobel Prize Winners				√			
Field Medals				√			
World Heritage List				√			
Wolf Prize winners				√			
National Parks				√			
Famous People	√		√	√	√		
Mapping Services							
Maps (static images)		√		√	√	√	
Maps (Zooming, Navigation)				√	√	√	√
Markers Creation				√			√
Neighbour countries				√			
Provinces/Districts				√	√		
Stories							
Amusing descriptions				√			

[34] DBpedia page of a particular country e.g. http://dbpedia.org/page/Finland

15.1.3.1 Facts

As far as general facts are concerned, all the geographic servers have more or less same amount of facts. Infoplease[8] presents country profile. It gives general facts and historical information about countries of the world. It also provides articles related to government/social activities. There are some dynamic services such as distance calculator, place finder which are also present. CIA Factbook is a very rich geographic resource. It offers all kinds of facts including military, transnational issues. It does not have cities and mountains rankings. The Word Atlas offers more about geography and people. It also covers city information. Large cities of countries are found in Word Atlas. Flags of countries are available in all servers, however emblems are missing. Our server, however, also shows emblems of each country.

15.1.3.2 Pictures

In addition to factual data, we have consolidated pictures from different data sources. The current collection contains approximately 25,000 pictures including both historic and current. Note that some pictures even if taken recently are in a sense historic, since they cannot be taken any more: Many photos of Syria (taken in 2009) show buildings that have been virtually erased because of civil war; some pictures of Nepal were taken just days before the major earthquake in fall 2015 destroyed important building in e.g. Kathmandu. We initially imported 2,883 Factbook pictures in our server, see Figure 15.7. The pictures taken from different sources other than Factbook are approximately 10,000. This number keeps growing as the project is engaging institutions and people all over the world (Pakistan, India, USA, Japan, Germany, Netherlands etc.). The full list of members of the global-geography.org consortium as of March 18, 2016 can be seen at [9].

To get an overview of picture collection in "global-geography.org" against each country, see Figure 15.8, but do look at the changes by clicking at the URL below the table. United States with 2087 has currently the largest number of pictures. Greece is at 2nd position in terms of pictures in our server. Figure 15.9 shows picture section of Botswana.

The entry "Pictures" is supposed to give a pictorial overview of each country by providing a selection of an average of 100 pictures per country: some of them are taken from Factbook [10]. The main point is not to just offer pictures (zillions of them can be found on the web already) but to offer pictures with substantial descriptions, and even pictures of the past, of buildings, monuments or nature, as destroyed by time, industrial developments or wars. The pictures of Syria can be seen at [11].

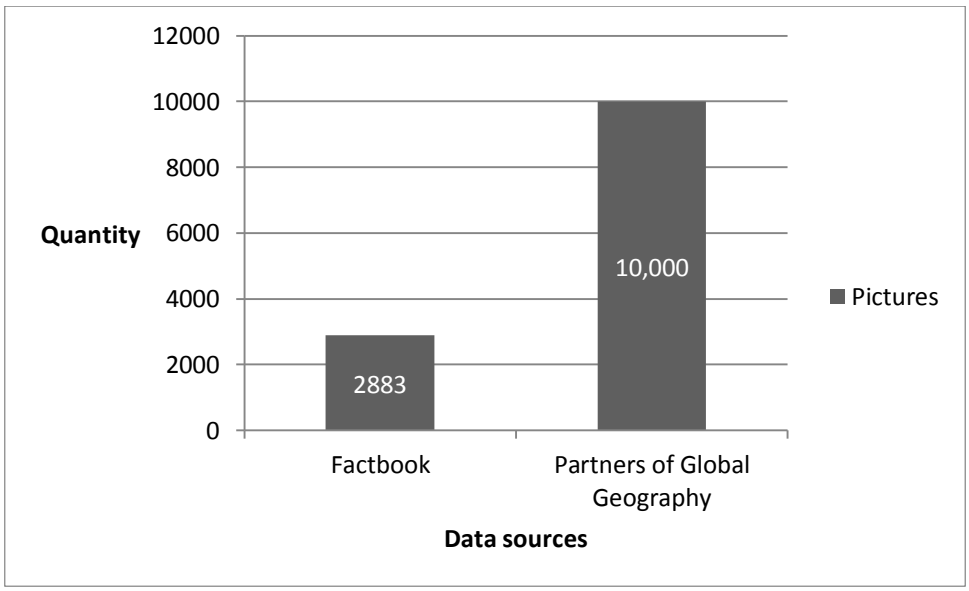


Figure 15.7: Picture collection in Austria-Forum

Pictures

country	number
United States	2087
Greece	1070
Philippines	720
Syria	694
Taiwan	597
Egypt	484
Nepal	473
India	451
Spain	412
Australia	382
Zimbabwe	381
Korea, South	365
Georgia	301
Yemen	300
Botswana	298
France	296
Netherlands	294

Figure 15.8: Country-wise number of pictures, see [12]

Botswana: Pictures

- 148 Pictures ot the Okavango Delta by Gerhard Huber 2012
- 52 Pictures ot the Makgadikgadi-Pans by Gerhard Huber 2012
- 91 Pictures of Chobe by Gerhard Huber 2012
- 50 Botswana Travel Pictures by Karl Trummer 1997
- 8 Factbook Pictures 2014

Figure 15.9: Picture section of Botswana

15.1.3.3 Facts Exploration

Let us now discuss the virtual lab. The on demand exploration of data (comparison of properties of countries, top and bottom rankings of UN member states) using virtual lab, makes "global-geography.org" fairly unique. The virtual lab is not offered by any of earlier mentioned geographic servers. Factbook and Factfish provide general ranking of countries regarding different properties but they give absolute rankings only. We are providing both absolute and relative rankings, see Chapter 9 for more examples or visit virtual lab in "global-geography.org"[5].

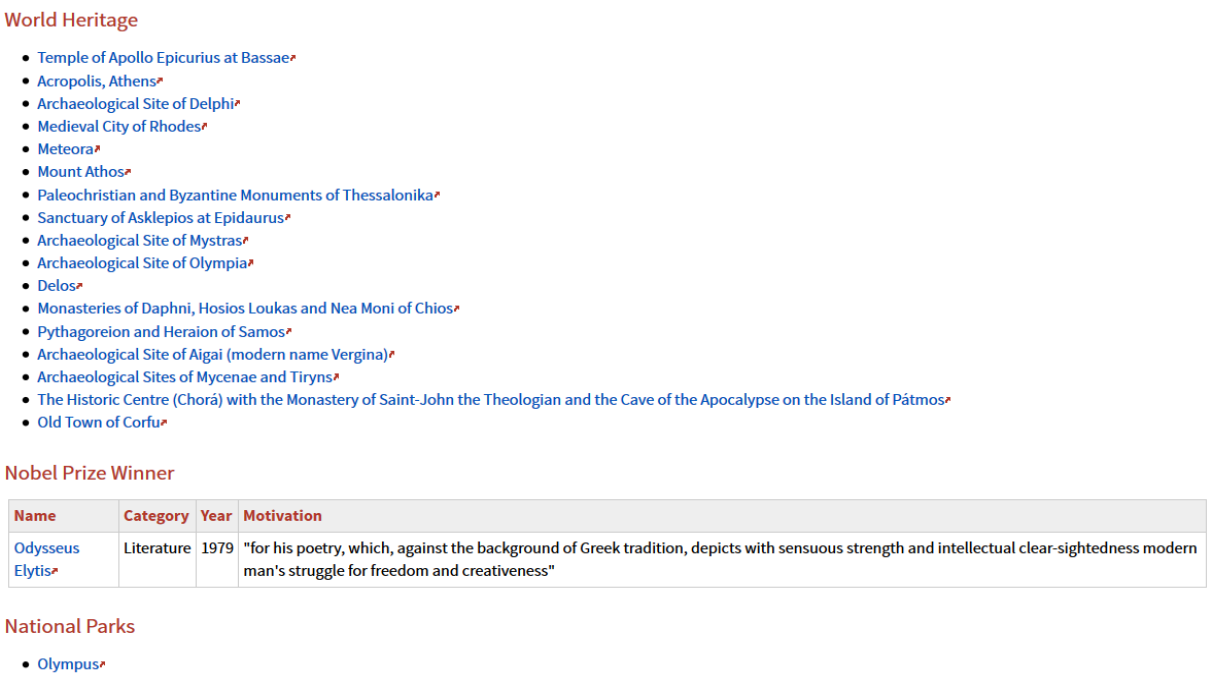
15.1.3.4 Interactive Visualisations

Interactive visualisations are a central part of global-geography.org. Among all the servers mentioned in Table 15.3, Factfish has visualisations that allow animations, like showing the population growth of a country in a graph. However, the most important part of visualizing information is to compare information of different countries, and this is not offered in Factfish. Trade visualisations unveiling trade among countries are not found in any of the servers mentioned, for detailed explanation about

visualisations, see Chapter 11. Interactive globe showing highest rank countries in different properties makes it quite easily to explore facts.

15.1.3.5 Culture

Another noteworthy section of global-geography is culture. It shows different aspects (UNESCO heritage, national parks, famous people and different prize holder's lists such as Wolf prize, Field medals) all related to the culture of a country. Figure 15.10 shows a sub-page that represents aspects of the culture of Greece. Infoplease has no separate section of Nobel Prize winners but it indicates prominent Nobel Prize winners like Mother Tressa in the category of "Noteworthy People". DBpedia page of countries also lists important persons. The list is a mixture of (leaders, players, artist etc.). Word Atlas lists famous people under the category "Famous Natives". World Heritage sites are considered only in Global-geography.org. We have a category "Partial List of Famous People" where we are listing famous people of a particular country. For example famous people of Austria in "global-geogrphy.org" is shown in the Figure 15.11.



World Heritage

- Temple of Apollo Epicurius at Bassae
- Acropolis, Athens
- Archaeological Site of Delphi
- Medieval City of Rhodes
- Meteora
- Mount Athos
- Paleochristian and Byzantine Monuments of Thessalonika
- Sanctuary of Asklepios at Epidaurus
- Archaeological Site of Mystras
- Archaeological Site of Olympia
- Delos
- Monasteries of Daphni, Hosios Loukas and Nea Moni of Chios
- Pythagoreion and Heraion of Samos
- Archaeological Site of Aigai (modern name Vergina)
- Archaeological Sites of Mycenae and Tiryns
- The Historic Centre (Chorá) with the Monastery of Saint-John the Theologian and the Cave of the Apocalypse on the Island of Pátmos
- Old Town of Corfu

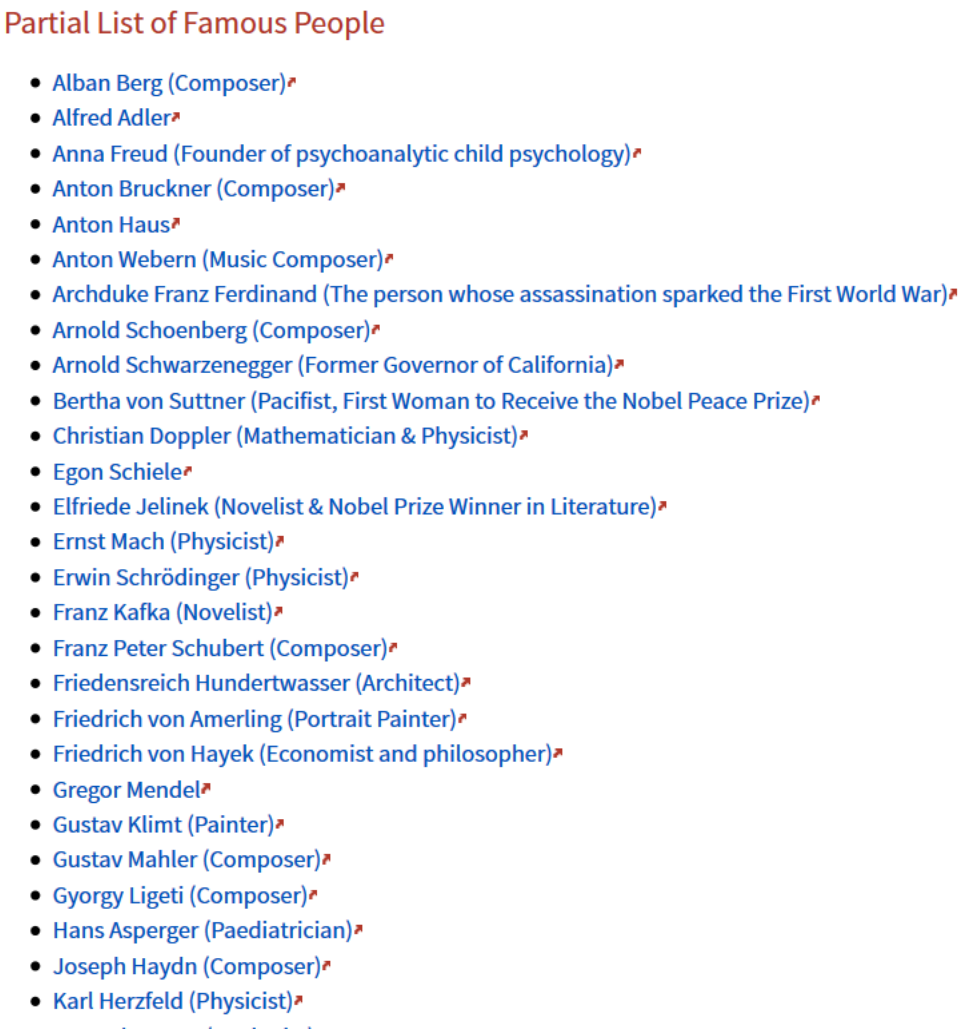
Nobel Prize Winner

Name	Category	Year	Motivation
Odysseus Elytis	Literature	1979	"for his poetry, which, against the background of Greek tradition, depicts with sensuous strength and intellectual clear-sightedness modern man's struggle for freedom and creativeness"

National Parks

- Olympus

Figure 15.10: Culture Section of Greece



Partial List of Famous People

- Alban Berg (Composer)
- Alfred Adler
- Anna Freud (Founder of psychoanalytic child psychology)
- Anton Bruckner (Composer)
- Anton Haus
- Anton Webern (Music Composer)
- Archduke Franz Ferdinand (The person whose assassination sparked the First World War)
- Arnold Schoenberg (Composer)
- Arnold Schwarzenegger (Former Governor of California)
- Bertha von Suttner (Pacifist, First Woman to Receive the Nobel Peace Prize)
- Christian Doppler (Mathematician & Physicist)
- Egon Schiele
- Elfriede Jelinek (Novelist & Nobel Prize Winner in Literature)
- Ernst Mach (Physicist)
- Erwin Schrödinger (Physicist)
- Franz Kafka (Novelist)
- Franz Peter Schubert (Composer)
- Friedensreich Hundertwasser (Architect)
- Friedrich von Amerling (Portrait Painter)
- Friedrich von Hayek (Economist and philosopher)
- Gregor Mendel
- Gustav Klimt (Painter)
- Gustav Mahler (Composer)
- Gyorgy Ligeti (Composer)
- Hans Asperger (Paediatrician)
- Joseph Haydn (Composer)
- Karl Herzfeld (Physicist)

Figure 15.11: Very partial list of famous people of Austria.

15.1.3.6 Stories

Our stories are unique and contain fairly unique description of countries. Maps, videos and pictures in stories make them more interesting. Stories are not found in any of geographic servers. Our stories are mixture of different types of facts related to countries, see Chapter 13 for more details.

15.2 Community Contributions

In this very large project we depend on the community. We have provided a platform for the community (this is just one aspect). Anonymous and registered users can participate and send us interesting and amusing facts about a country using the form as shown in the Figure 15.12. Specifically they can enter the following in the form fields.

- URLs for interesting pictures
- URLs for interesting videos
- Items typical for a country
- One liner that represents a country.

When the name of person is posted along with any picture or video URL, it is properly displayed along with wiki link. In future, partners of global geography will contribute through this form and add their contributions easily.

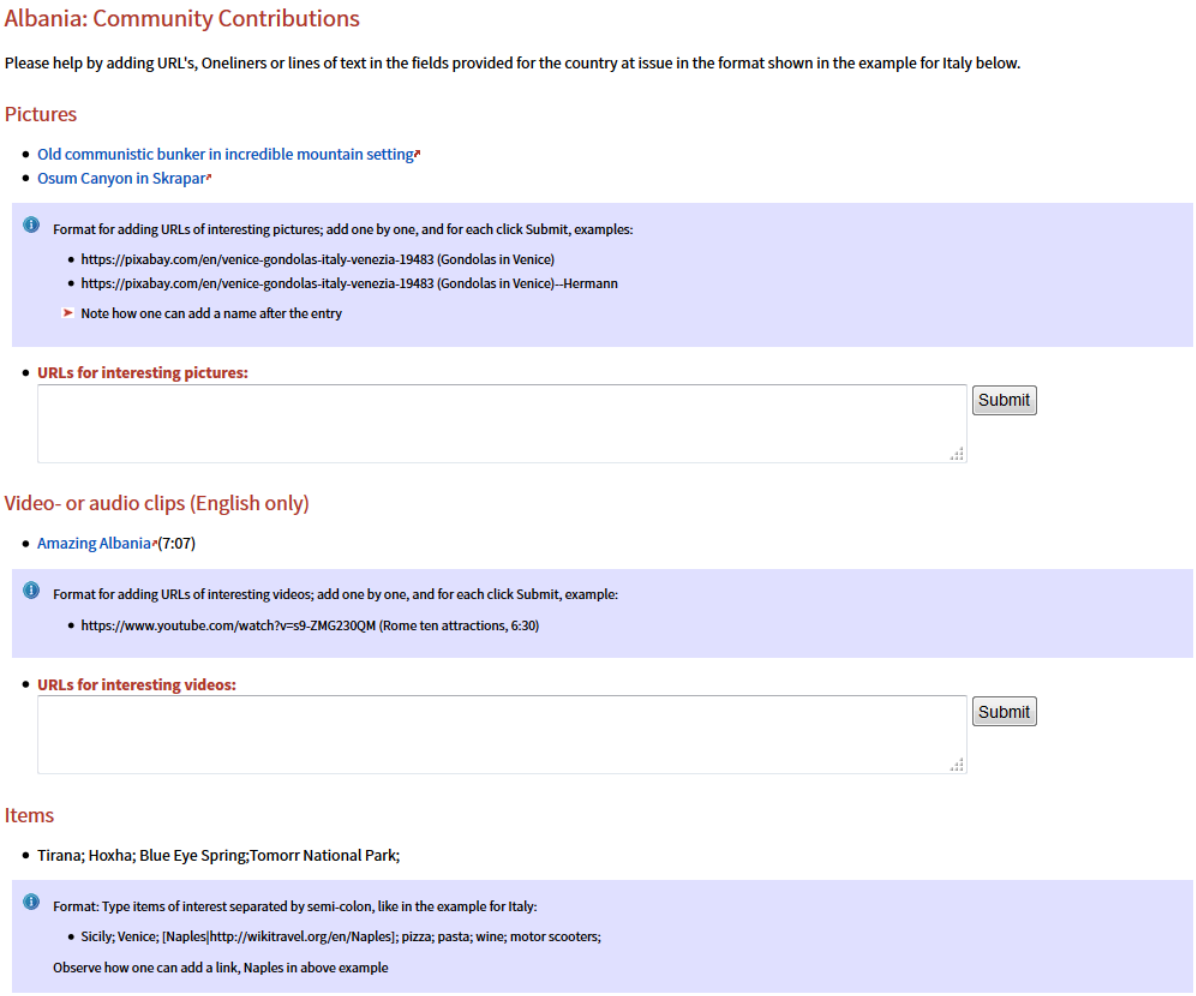


Figure 15.12: Community Contributions page of Albania displaying specially designed form for entering pictures, videos, items and one liners.

15.3 Opportunities

The system has provided several opportunities, some of them are:

- Global Geography has provided a global platform for information sharing. Using this platform people from all over the world can contribute and provide valuable information about their countries. They have the possibility to verify information as well about their countries.

- In the future, the system can be used by schools all over the world. Children in schools will use global-geography as geographic learning tool.

- The virtual lab can be extended using more properties. It can be used as a learning tool for quick learning of facts about different countries of the world. It contain valuable facts and rankings which surely are of interest and important for students, teachers and people working in different domains.

- We have developed visualisations that can be easily extended. D3.js [13] is used to make these visualisations. The library is very flexible and provides reusability. Visualisations can be easily extended using more properties, therefore providing still more opportunities to explore and compare countries visually.

15.4 Threats and their Solutions

Although there are numerous opportunities and benefits of system, possible problems are outlined in this section. The major issue is copyright Issue. As our system is built using different data sources (fortunately mostly are in public domain) still there is a need to focus on copyright issue. Proper indication of sources from which material is taken is very important.

> "Copyright has never been an easy, black-and-white kind of issue"[14].
>
> ***Kaitlyn Ellison***

We have taken special measures to avoid copyright violation which we are going to discuss in what follows next.

15.4.1 Picture handling in Global Geography

If we are using pictures in our stories from "PixaBay" and "Wikicommons" as shown in the Figure 15.13. We always indicate the source. For example "Sushi" (a food item) picture is taken from wikicommons [15] under license [16], see Figure 15.13. Similarly "Pixabay" is mentioned as image source in the caption of flower image "Lily of the Valley". Note, however that the largest number of pictures came from public domain (Factbook) or from global-geography consortium.

Figure 15.13: National Facts of Japan and Finland in global-geography

15.4.2. Text handling in Global Geography

When we take textual material from Wikipedia we mark that text as italic. For example the story of Finland is shown in the Figure 15.14. The text in the introductory section is marked italic which is taken from Wikipedia. Also we mention explicitly exact source and the license in the references.

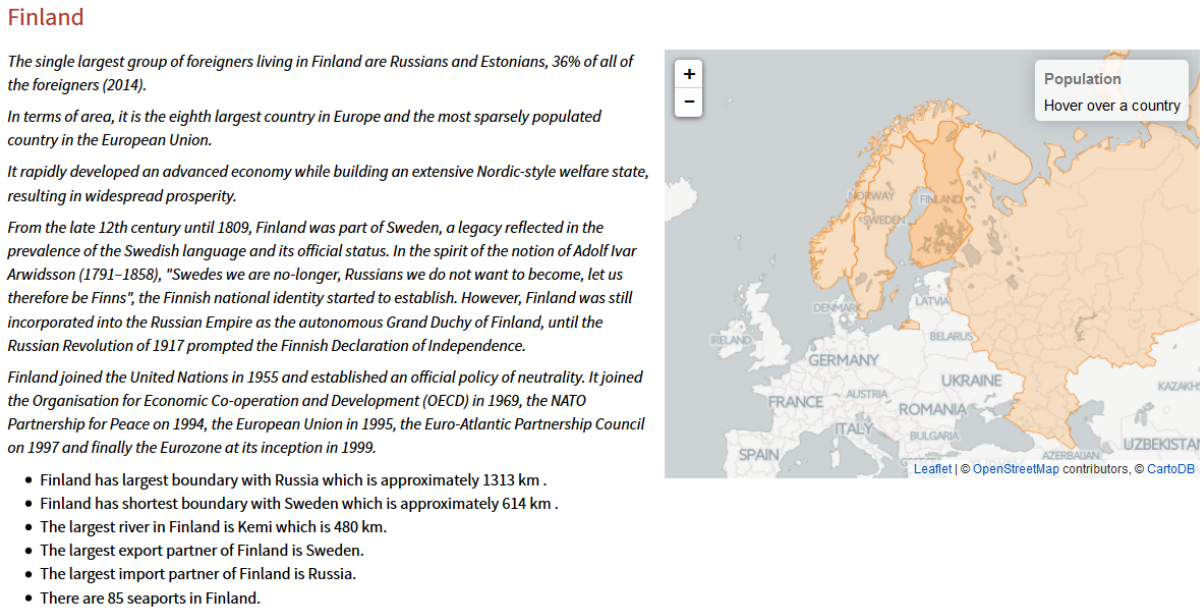


Figure 15.14: Story of Finland, Text marked as italic is taken from Wikipedia

15.4.3 Server Update

The major issue that we have encountered so far is “how we update the data on the server”. Let us first explain the issue. Entries in global-geography.org come from some data-bases. However, they can be changed in a number of way: By our consolidation algorithms, like in the case of size of countries, of city rankings and mountain rankings; other changes are possible since items will be added in the culture section, e.g. by updating the list of Nobel prize winners or the list of famous persons using additional data-bases; also, users in the category “editor” can change entries manually. For instance, “Water Area” of Ethiopia, and a number of entries in Georgia were changed recently, since they turned out to be wrong. If we run the updates using new versions of various databases then previous changes that have been made on purpose should not be erased. For this reason we will handle updates roughly as follows:

(1)We add a special mark in the page at the beginning and end of each item or section, if it has been changed from what the database contained.

(2) The update script will not overwrite such entries but will produce a set of “warning messages” such as:

<Country>, <Section>: “Should xxx (the marked data) be replaced by yyy (the new data in the database).”

Ethiopia, “Geography”, “Should water area (7,444 sq km) of Ethiopia be replaced by water area (104,300 sq km)?”

(3) In the culture section “Famous People” entries should only be added by the update script, but none deleted.

(4) The population CSV file that is used in the story section should be replaced with new version.

15.5 Summary and Future work

The thesis explores different possibilities of information consolidation and exploration in different domains.

Reading the research work on Global Geography, many will have wondered how we can ever expect to reach a level of acceptable completeness. It is clear that this cannot be done by one group alone, but by tying in many groups and the community.

For this reason in addition to continue the "obvious" implementation and information gathering activities we are in the process of pursuing two important avenues. One is curation to easily incorporate material that is located outside the main server by using tools that are similar to Pinterest and others, yet the user can view the results without registering. This is clearly a large topic on its own. Curation on the Internet is closely linked to traditional curation of art and such, see [17] . However, we have taken a first but particularly easy to use step by introducing the category "Community Contributions" where pictures, video clips and links can be added by everyone, anonymously.

Also, we provide the geography part of our server with an English interface so that it can be used world-wide.

In this modern era where multi-size screens are emerging, it is necessary to create visualisations that work on all devices. The idea of responsive visualisation is similar to responsive design. We have just scratched the surface of this new paradigm. We will further look into complex visualisations and identify responsive strategies for them.

Currently, the system "global-geography" utilizes the Fact book dataset as baseline. But in future, there is a need to explore other datasets related to countries. The culture section of Global Geography.org can be enhanced with famous people from categories like:

- Famous leaders
- Famous musicians
- Famous writers
- Famous poets

Currently, in the virtual Lab there is no distinction of positive or negative properties, it would be nice if the system does this and countries are ranked top and bottom using positive and negative properties. For example unemployment rate is a negative property and if a country ranks high in this property surely the high rank in this case is not showing the strength but weakness of the country.

Concerning future development, the author of this thesis is likely to work online as team member of Austria-Forum to carry out further integration of facts.

Besides development and integration of facts, a special focus should be on testing the system. As it is huge project and continuously expanding (statistics shown in Table 15.2), it needs special dedicated "Quality Assurance Analysts" and software testers.

References

[1] "Community Contributions" [Online]. Available: http://austria-forum.org/af/Geography/Asia/Nepal/Community_Contributions. [Accessed: 18-Mar-2016].

[2] S. Cronholm and G. Goldkuhl, "Strategies for Information Systems Evaluation - Six Generic Types," Electron. J. Inf. Syst. Eval., vol. 6, no. 2, pp. 65–74, 2003.

[3] "Surprising Facts" [Online]. Available: http://austria-forum.org/af/Geography/Europe/Austria/Special_Information/Surprising_Facts. [Accessed: 18-Mar-2016].

[4] "mehmood rizwan - User im Austria-Forum." [Online]. Available: http://austria-forum.org/af/User/mehmood rizwan. [Accessed: 08-Feb-2016].

[5] "Geography im Austria-Forum." [Online]. Available: http://austria-forum.org/af/Geography. [Accessed: 02-Nov-2015].

[6] "Famous People" [Online]. Available: http://www.thefamouspeople.com/. [Accessed: 23-Feb-2016].

[7] "UNESCO World Heritage Centre - World Heritage List." [Online]. Available: http://whc.unesco.org/en/list/. [Accessed: 19-Oct-2015].

[8] "Infoplease." [Online]. Available: http://www.infoplease.com/. [Accessed: 17-Feb-2016].

[9] "Consortium Members." [Online]. Available: http://austria-forum.org/af/Geography/Consortium_Members. [Accessed: 06-Jan-2016].

[10] "FactbookXML - All textual information of the CIA World Factbook in one XML file." [Online]. Available: http://jmatchparser.sourceforge.net/factbook/. [Accessed: 21-Oct-2015].

[11] "Pictures." [Online]. Available: http://austria-forum.org/af/Geography/Asia/Syria/Pictures. [Accessed: 19-Oct-2015].

[12] “Pictures.” [Online]. Available: http://austria-forum.org/af/Geography/Visualizations/Pictures. [Accessed: 18-Mar-2016].

[13] “D3 — Scott Murray — alignedleft.” [Online]. Available: http://alignedleft.com/tutorials/d3. [Accessed: 07-Feb-2016].

[14] “5 famous copyright infringement cases (what you can learn).” [Online]. Available: https://99designs.com/designer-blog/2013/04/19/5-famous-copyright-infringement-cases/. [Accessed: 20-Feb-2016].

[15] “Sushi.” [Online]. Available: https://en.wikipedia.org/wiki/Sushi#/media/File:2007feb-sushi-odaiba-manytypes.jpg.

[16] “Creative Commons.” [Online]. Available: http://creativecommons.org/licenses/by/2.5/. [Accessed: 20-Feb-2016].

[17] “Rethinking Curating - The MIT Press.” [Online]. Available: https://mitpress.mit.edu/books/rethinking-curating. [Accessed: 19-Oct-2015].

List of Tables

List of Figures

Appendix A: List of Publications

1. [Mehmood and Maurer 2013] Rizwan Mehmood and Hermann Maurer, Aspects of Information Integration in a wiki, Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS 2013), Pages 575-579, doi:10.1145/2539150.2539230, ACM, Vienna, Austria

2. [Mehmood and Maurer 2013] Rizwan Mehmood and Hermann Maurer, Towards the integration of images on the Web,Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS 2013), Pages 580-584, doi: 10.1145/2539150.2539250, ACM, Vienna, Austria

3. [Mehmood and Maurer 2014] Rizwan Mehmood and Hermann Maurer, Merging image databases as an example for information integration, Central European Journal of Operation Research June 2015, Volume 23, Issue 2, Pages 441-458.

4. [Mehmood 2014] Rizwan Mehmood. Geographic data verification. Automatic Verification of Physical Geographic Data using Maps, IPSI BgD Transactions on Internet Research (TIR), Volume 10, Number 2, July 2014. Pages 20-25.

5. [Nara et al 2014] Nara, Maurer, Mehmood, Some aspects of the Reliability of information on the Web. JUCS (Journal of Computer Science), Volume 20, Issue 9. DOI: 10.3217/jucs-020-09-1284, 2014

6. [Mehmood and Korica 2015] Rizwan and Korica, The Web is big business, CIT, Journal of Computing and Information Technology, Volume 23, Number 1, 2015

7. [Mehmood and Maurer 2015] Rizwan Mehmood and Hermann Maurer, Facts collection and Verification Efforts, Proceedings of 4th International Conference on Data Management Technologies and Applications, Pages: 142-151, 2015

8. [Mehmood et al 2016], A new look into Geography Servers, Rizwan, Maurer, Narayan, IPSI BgD Transactions on Internet Research (TIR), Volume 12, Number 1, January 2016. Pages 21-30.

9. [Maurer et al 2014] How Dangerous is the web for creative work, Maurer, Mehmood and Korica, CIT, Journal of Computing and Information Technology, Volume 21, Number 2, pages 59-69, 2013

Appendix B: Verification Results

Countries are shown in the red colour when %difference > 1

Country	Factbook	Dbpedia	Geoname	Infoplease	Britannica	Wolfram	Average	1/10 Average	%Difference
Afghanistan	652230	647500	647500	647500	652864	652230	649971	650	4
Armenia	29743	29743	29800	29800	29743	29743	29762	30	1
Azerbaijan	86600	86599	86600	86600	86600	86600	86600	87	0
Bahrain	760	765	665	665	767	760	730	1	89
Bangladesh	143998	147570	144000	144000	147570	143998	145189	145	16
Bhutan	38394	38394	47000	47000	38394	38394	41263	41	139
Brunei	5765	5765	5770	5770	5765	5765	5767	6	1
Burma	676578	676575	678500	0	676577	676578	676962	677	2
Cambodia	181035	181035	181040	181040	181035	181035	181037	181	0
China	9596961	9706961	9596960	9596960	9572900	9597000	9611290	9611	10
Georgia	69700	0	69700	69700	69700	153900	86540	87	778
India	3287263	3287262	3287590	3287590	3166414	3287000	3267187	3267	31
Indonesia	1904569	1904568	1919440	1919440	1910931	1905000	1910658	1911	5
Iran	1648195	1648195	1648000	1648000	1648200	1648000	1648098	1648	0
Iraq	438317	438314	437072	0	434128	437072	436918	437	7
Israel	20770	20769	20770	20770	21643	20770	20915	21	35
Japan	377915	377944	377835	377915	377873	377835	377886	378	0
Jordan	89342	89342	92300	89342	88794	89342	89744	90	28
Kazakhstan	2724900	2724888	2717300	2717300	2724900	2725000	2722381	2722	2
Korea, North	120538	120540	120540	120540	122762	120538	120910	121	15
Korea, South	99720	100210	98480	0	99678	98480	99314	99	9
Kuwait	17818	17819	17820	17819	17818	17818	17819	18	0
Kyrgyzstan	199951	199898	198500	191300	199945	199951	198258	198	35
Laos	236800	236800	236800	236800	236800	236800	236800	237	0
Lebanon	10400	10452	10400	10400	10452	10452	10426	10	2
Malaysia	329847	329847	329750	329750	330290	329847	329889	330	1
Maldives	298	298	300	300	298	298	299	0	4
Mongolia	1564116	1564116	1565000	1565000	1564160	1564000	1564399	1564	0
Nepal	147181	147181	140800	140800	147181	147181	145054	145	29
Oman	309500	309498	212460	212460	309500	309500	277153	277	233
Pakistan	796095	796095	803940	803940	881889	796095	813009	813	85
Papua New Guinea	462840	462839	462840	462840	462840	462840	462840	463	0

Philippines	300000	300001	300000	300000	300000	300000	300000	300	0
Qatar	11586	11571	11437	11437	11607	11586	11537	12	9
Russia	17098242	17075400	17100000	17075200	17098200	17080000	17087840	17088	1
Saudi Arabia	2149690	2253290	1960582	2149690	2149690	1961000	2103990	2104	71
Singapore	697	710	692	693	714	697	701	1	19
Sri Lanka	65610	65610	65610	65610	65610	65610	65610	66	0
Syria	185180	185130	185180	185180	185180	185180	185172	185	0
Tajikistan	143100	143099	143100	143100	143100	143100	143100	143	0
Thailand	513120	513115	514000	514000	513120	513120	513413	513	1
Timor-Leste	14874	0	0	0	0	14874	14874	15	0
Turkey	783562	783562	780580	780580	785347	783562	782866	783	3
Turkmenistan	488100	488099	488100	488100	491210	488100	488618	489	5
United Arab Emirates	83600	83600	82880	83600	83600	83600	83480	83	7
Uzbekistan	447400	447400	447400	447400	447400	447400	447400	447	0
Vietnam	331210	331210	329560	329560	331212	329560	330385	330	3
Yemen	527968	527829	527970	527969	528076	527968	527963	528	0
Albania	28748	28748	28748	28748	28703	28748	28741	29	1
Andorra	468	468	468	0	468	468	468	0	0
Austria	83871	83855	83858	83870	83879	83871	83867	84	0
Belarus	207600	207595	207600	207600	207595	207600	207598	208	0
Belgium	30528	30528	30510	0	30528	30528	30524	31	0
Bosnia and Herzegovina	51197	51129	51129	51129	51209	51197	51165	51	1
Bulgaria	110879	110911	110910	110910	111002	110879	110915	111	1
Croatia	56594	56594	56542	56542	56594	56594	56577	57	1
Cyprus	9251	9251	9250	9248	9251	9251	8691	9	0
Czech Republic	78867	78865	78866	0	78865	78867	78866	79	0
Denmark	43094	42895	43094	43094	43098	43094	43062	43	4
Estonia	45228	45099	45226	0	45227	45228	45202	45	2
Finland	338145	338242	337030	338145	390903	338145	346768	347	127
France	643801	674843	547030	547030	543965	551500	584695	585	154
Germany	357022	357021	357021	357021	357104	357022	357035	357	0
Greece	131957	131957	131940	131940	131957	131940	131949	132	0
Hungary	93028	93030	93030	93030	93030	93028	93029	93	0
Iceland	103000	103001	103000	103000	103022	103000	103004	103	0
Ireland	70273	84421	70280	70280	70273	70273	72633	73	162
Italy	301340	301338	301230	301230	301336	301340	301302	301	0

Latvia	64589	64589	64589	64589	64589	0	64589	65	0
Liechtenstein	160	160	160	161	160	160	160	0	5
Lithuania	65300	65300	65200	65200	65300	65300	65267	65	1
Luxembourg	2586	2586	2586	2585	2586	2586	2586	3	0
Macedonia	25713	0	25333	25333	25713	25713	25561	26	9
Malta	316	316	316	321	315	316	317	0	14
Moldova	33851	33843	33843	33843	33843	33851	33846	34	0
Monaco	2	2	NULL	2	2	2	2	0	0
Montenegro	13812	12999	14026	14026	13812	13812	13748	14	54
Netherlands	41543	41541	41526	41526	41850	41543	41588	42	6
Norway	323802	385183	324220	324220	385186	323802	344402	344	118
Poland	312685	312602	312685	312685	312679	312685	312670	313	0
Portugal	92090	92211	92391	92391	92212	92090	92231	92	2
Romania	238391	238390	237500	237500	238391	238391	238094	238	2
San Marino	61	61	NULL	0	61	61	61	0	0
Serbia	77474	88360	88361	77474	77498	77474	81107	81	89
Slovakia	49035	49034	48845	48845	49034	49035	48971	49	3
Slovenia	20273	20272	20273	20273	20273	20273	20273	20	0
Spain	505370	505990	504782	504782	505991	505370	505381	505	1
Sweden	450295	449964	449964	449964	447420	450295	449650	450	5
Switzerland	41277	41284	41290	41290	41285	41277	41284	41	0
Ukraine	603550	603628	603700	603700	603628	603550	603626	604	0
United Kingdom	243610	243610	244820	244820	243073	243610	243924	244	4
Algeria	2381741	2381741	2381740	2381741	2381741	2382000	2381784	2382	0
Angola	1246700	1246700	1246700	1246699	1246700	1247000	1246750	1247	0
Benin	112622	112622	112620	112620	114763	112622	112978	113	16
Botswana	581730	581730	600370	600370	581730	581730	587943	588	21
Burkina Faso	274200	274199	274200	274200	270764	274200	273627	274	10
Burundi	27830	27829	27830	27830	27834	27830	27831	28	0
Cabo Verde	4033	0	0	0	4033	4033	4033	4	0
Cameroon	475440	475442	475440	475440	476350	475440	475592	476	2
Central African Republic	622984	622980	622984	0	622436	622984	622874	623	1
Chad	1284000	1283994	1284000	1284000	1284000	1284000	1236666	1284	0
Comoros	2235	2235	2170	2170	1862	2235	2151	2	134

Congo, Democratic Republic of the	2344858	0	0	0	2345410	2345000	2345089	2345	0
Congo, Republic of the	342000	0	0	0		342000	342000	342	0
Cote d'Ivoire	322463	0	0	0	322463	322463	322463	322	0
Djibouti	23200	23200	23000	23000	23200	23200	23133	23	6
Egypt	1001450	1002450	1001450	1001450	996603	1001000	1000734	1001	4
Equatorial Guinea	28051	28050	28051	0	28051	28051	28051	28	0
Eritrea	117600	117598	121320	121320	121144	117600	119430	119	16
Ethiopia	1104300	1104296	1127127	1127127	1063652	1127000	1108917	1109	41
Gabon	267667	267667	267667	267667	267667	267667	267667	268	0
Gambia, The	11295	10689	11300	11300	11632	11295	11252	11	50
Ghana	238533	238533	239460	239460	238533	238533	238842	239	3
Guinea	245857	245857	245857	245861	245857	245857	245858	246	0
Guinea-Bissau	36125	36125	36120	36120	36125	36125	36123	36	0
Kenya	580367	580365	582650	582650	582646	580367	581508	582	2
Lesotho	30355	30355	30355	30350	30355	30355	30354	30	0
Liberia	111369	111369	111370	111370	96917	111369	108961	109	111
Libya	1759540	1759532	1759540	1759540	1676198	1760000	1745725	1746	40
Madagascar	587041	586884	587040	587040	587295	587041	587057	587	0
Malawi	118484	118484	118480	118480	118484	118484	118483	118	0
Mali	1240192	1240187	1240000	1240000	1248574	1240000	1241492	1241	6
Mauritania	1030700	1030696	1030700	1030700	1030700	1031000	1030749	1031	0
Mauritius	2040	2040	2040	2040	2040	2040	2040	2	0
Morocco	446550	446550	446550	446550	442300	446550	445842	446	8
Mozambique	799380	801590	801590	801590	799380	799380	800485	800	1
Namibia	824292	825418	825418	825418	825615	824292	825076	825	1
Niger	1267000	1267000	1267000	1267000	1267000	1267000	1222500	1267	0
Nigeria	923768	923768	923768	923768	923768	923768	923768	924	0
Rwanda	26338	26338	26338	26338	26379	26338	26345	26	1
Sao Tome and Principe	964	0	1001	0	1001	964	983	1	19
Senegal	196722	196723	196190	196190	196722	196722	196545	197	2
Seychelles	455	451	455	456	452	455	454	0	7
Sierra Leone	71740	71740	71740	0	71740	71740	71740	72	0
Somalia	637657	637655	637657	637657	637657	637657	637657	638	0
South Africa	1219090	1221032	1219912	0	1220813	1219000	1219969	1220	1

South Sudan	644329	619745	644329	644329	644330	644329	640232	640	32
Sudan	1861484	1886068	1861484	1861484	1844797	1886000	1866886	1867	12
Swaziland	17364	17363	17363	17363	17364	17364	17364	17	0
Tanzania	947300	945081	945087	945087	945249	947300	945851	946	2
Togo	56785	56785	56785	56785	56600	56785	56754	57	3
Tunisia	163610	163610	163610	163610	163610	163610	163610	164	0
Uganda	241038	0	236040	236040	241551	236040	238142	238	14
Zambia	752618	752617	752614	0	752612	752618	752616	753	0
Zimbabwe	390757	390754	390580	390580	390757	390757	390698	391	0
Australia	7741220	7692024	7686850	7686850	7692202	7741000	7706691	7707	4
Fiji	18274	18274	18270	18270	18272	18274	18272	18	0
Kiribati	811	811	811	811	811	811	811	1	0
Marshall Islands	181	181	0	181	181	181	181	0	0
Nauru	21	21	21	21	21	21	21	0	0
New Zealand	267710	268020	268680	0	270692	268680	268756	269	7
Palau	459	458	458	458	488	459	463	0	53
Samoa	2831	2831	2944	2944	2785	2831	2861	3	29
Solomon Islands	28896	28399	28450	0	28370	28896	28602	29	10
Tonga	747	748	748	748	748	747	748	1	1
Tuvalu	26	26	26	26	26	26	26	0	0
Vanuatu	12189	12190	12200	12200	12190	12189	12193	12	1
Micronesia	702	0	0	0	0	702	702	1	0
Argentina	2780400	2780399	2766890	2766890	2780400	2767000	2773663	2774	2
Bolivia	1098581	0	1098580	1098580	1098581	1099000	1098664	1099	0
Brazil	8514877	8514837	8511965	8511965	8515767	8515000	8514069	8514	0
Canada	9984670	9982034	9984670	9984670	9984670	9985000	9984286	9984	0
Chile	756102	756096	756950	756950	756096	756102	756383	756	1
Colombia	1138910	1141747	1138910	1138910	1141748	1139000	1139871	1140	2
Ecuador	283561	258237	283560	283560	256370	283561	274808	275	67
Guyana	214969	214969	214970	214970	214999	214969	214974	215	0
Mexico	1964375	1972550	1972550	1972550	1964375	1964000	1968400	1968	2
Paraguay	406752	406752	406750	406750	406752	0	406751	407	0
Peru	1285216	1285216	1285220	1285220	1285216	1285000	1285181	1285	0
Suriname	163820	163819	163270	163270	163820	163820	163637	164	2
United States	9826675	9826675	9629091	9631420	9526468	9631000	9678555	9679	16
Uruguay	176215	176215	176220	176220	177879	176215	176494	176	8

Venezuela	912050	916444	912050	912050	916445	912050	913515	914	3
Antigua and Barbuda	442	440	443	440	442	443	442	0	4
Bahamas, The	13880	13878	13940	13940	13939	13880	13910	14	2
Barbados	430	431	431	431	430	0	431	0	1
Belize	22966	22965	22966	22966	22965	22966	22966	23	0
Costa Rica	51100	50901	51100	51100	51100	51100	51067	51	3
Cuba	110860	109884	110860	110860	109884	110860	110535	111	6
Dominica	751	750	754	751	751	751	751	1	4
Dominican Republic	48670	48442	48730	48730	48671	48670	48652	49	4
El Salvador	21041	21040	21040	21040	21040	21041	21040	21	0
Grenada	344	344	344	339	344	344	343	0	12
Guatemala	108889	108888	108890	108890	108889	108889	108889	109	0
Haiti	27750	27749	27750	27750	27700	27750	27742	28	1
Honduras	112090	112090	112090	112090	112492	112090	112157	112	3
Jamaica	10991	10991	10991	10991	10991	10991	10991	11	0
Nicaragua	130370	129999	129494	129494	130373	130370	130017	130	4
Panama	75420	75516	78200	78200	74177	75420	76156	76	27
Saint Kitts and Nevis	261	261	261	0	269	261	263	0	24
Saint Lucia	616	617	616	0	617	616	616	1	1
Saint Vincent and the Grenadines	389	388	389	0	389	389	389	0	2
Trinidad and Tobago	5128	5131	5128	5131	5127	5128	5129	5	0