Facts Collection and Verification Efforts

Rizwan Mehmood and Hermann Maurer
Institute for Information Systems and Computer Media, Graz University of Technology, Graz, Austria

Keywords: Data Quality, Information Reliability, Data Integration Systems.

Abstract: Geographic web portals and geospatial databases are emerging on the web recently, 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 encyclopedias and they are empowering information seekers by providing information at the door step. However, there is no guarantee of validity and authenticity of the information provided by them. The reason behind 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 teachers, students and tourists rely on search engines such as Google which often points to Wikipedia. We will identify some inconsistencies in online facts such as area, cities and mountains rankings using multiple data sources. Our investigations reveal that there is a need for a reliable geographic web portal 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 portal 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.

1 INTRODUCTION

The web has significant impact on the lives of the people and its importance has been mounting miraculously. The culture of opening books has changed with the invention of web. The problem is not, the lack of information, but the reliability of information as claimed and highlighted in (Wurzinger, 2010). 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 it has to be of good quality and reliability before it can be used for any purpose. 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.

If we look at online information, we come across many problems related to quality and reliability. Sometimes a time-line is missing which is very critical in some facts such as population figures; references are missing which can be used to factor out more details related to an article and to judge the authenticity of facts; or there are presentation of facts without units attached to them.

Let us look at an example. We are presenting the ranking of the five highest mountains of India taken from GeoNames and WolframAlpha as shown in Table 1. We have chosen them since GeoNames is one of the largest and most frequently used geospatial database and it is generally assumed to be of sufficient quality. We can see the discrepancies in the rankings as shown in Table 1. Kangchenjunga is ranked 1st in WolframAlpha, where as Nanda Devi comes on top in GeoNames ranking. Although Kangchenjunga is present in GeoNames[1], but it is associated with Nepal and is the 2nd highest mountain in Nepal after Mount Ever-

Table 1: Different mountain rankings of India in WolframAlpha and GeoNames.

Wolfram Alpha	Elevation (m)	GeoNames	Elevation (m)
Kangchenjunga	8586	Nanda Devi	7816
Kangchenjunga West	8505	Kamet	7756
Kangchenjunga South	8494	Saser Kangri	7672
Kangchenjunga Central	8482	Kabru	7412
Distaghil Sar	7885	Badrinath	7138

[1] geonames.org/maps/google_27.703_88.147.html

Mehmood R. and Maurer H..
Facts Collection and Verification Efforts.
DOI: 10.5220/0005505501420151
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (DATA-2015), pages 142-151
ISBN: 978-989-758-103-8

est. One obvious reason for discrepancies in mountain rankings is that mountain ranges often define the border between two countries. Also, Distaghil Sar is counted as the 7th highest peak of Pakistan, see (Kulathuramaiyer et al., 2014). Another example is the mountain Golemi Korab which is highest mountain in both Albania and Macedonia. However this short example forces us to dig deeper into verification of facts for our geographic portal, using different reliable sources.

There are some inherent complications which are associated with some facts. When we list countries, a main problem arises, it is not clear whether a particular geographic place is recognized as a country or just a territory. For instance Taiwan is not recognized as a country by UN. Likewise different portals present the size (area) of countries differently; for either political reasons, or because they exclude/include lakes, glaciers and/or ocean channels. Similar problems happen with mountains (when their peak is at a border) or cities (in some cases only the core is counted, in others the whole much larger metropolitan areas). Another interesting case concerns Nobel Laureates: It is also not clear how to associate them with a particular country due to the following reasons: i) Is it decisive where a person was born. ii) Where the person obtained the award. iii) Where the person did the work for which the award was given, etc. iv) Also, the country where e.g. a person was born may not exist any more! 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. We list her in both India and Macedonia. 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.

Several attempts have been made by researchers in the past to utilize web data that is present either in textual or in structured form, see (Zhao and Betz, 2007), (Carlson et al., 2010). We start with the facts which we import from different data sources. Then we try to find similar facts on other online data sources. We verify information before using it. Verification is done using comparative analysis of multiple data sources, encyclopedias such as Britannica, querying maps and and also with the help of domain experts in some cases. We will use the terms data and fact interchangeably in the rest of paper. Further, we call our geographic web portal, the AF-Geo portal[2].

The rest of the paper is organized as follows. Section 2 describes the kind of geographic data which we are considering for our portal. Section 3 describes the geographic data sources that have large sets of information. We will highlight new ways of verifying and collecting facts from online sources in Section 4. We present our results regarding facts verification in Section 5 followed by evaluation in Section 6.

2 GEOGRAPHIC FACTS IN CONSIDERATION

There are different types of data. Considering the temporal dimension, we can classify types of data as stable, long term changing and frequently changing data as mentioned in the book (Carlo Batini, 2006). A sample data[3] is shown in Figure 1; that can be associated with a country.

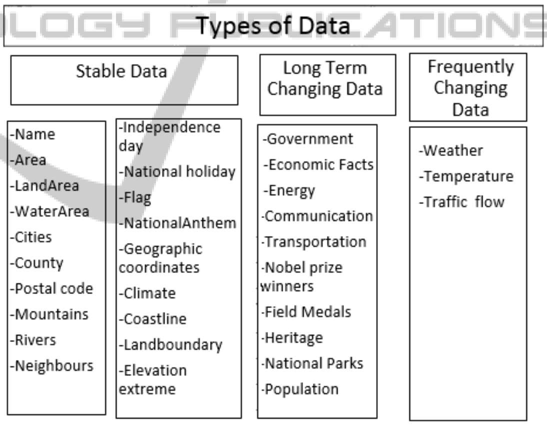


Figure 1: Types of facts related to any particular country based on temporal dimension.

2.1 Stable Data

This type of data rarely changes. It remains constant over a long period of time. The area of a country comprising of land area and water area etc. is the example of stable data. There are some exceptional cases when country areas are reported differently, considering issues such as the recently occurring war between Ukraine and Russia. Another example are Pakistan and India, both claiming parts of Kashmir long ago since countries became independent. Similarly the postal code of an area is hardly changed. A list of facts including in this category is shown in Figure 1.

[2] http://austria-forum.org/af/Geography

[3] The list of such facts is quite long but to demonstrate types of data we are presenting a sample

2.2 Long Term Changing Data

This type of data changes but it has very low change frequency. If we look at population, the factual data representing population changes with time. Similar examples are birth-rate, employment-rate etc. The reason behind this variability of data is a process called census which is the procedure of systematically acquiring and recording information about a country, usually population and other facts, that belong to a country. Once the process is over, the responsible authority publishes the numbers obtained. The population of a country is normally updated annually whereas economic facts such as GDP are calculated biannually or on annual basis. Similarly, Nobel prize winners list is updated every year.

2.3 Frequently Changing Data

This type of data changes frequently. If we look at the temperature and weather of a particular region, it constantly changes every day. There are some web services that are used to report this type of data. However we are not considering this in our geographic portal and we pay attention to correctness of stable and long term changing data and apply different measures to verify facts as described in Section 4.

3 DATA SOURCES

This section looks at sources of information about countries. We group data sources according to their type, as shown in Figure 2. 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.

3.1 Geospatial Databases

GeoNames: It contains over 8 million place names that are available for download free of charge. The GeoNames[4] 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 1 shows a query to extract the 5 highest mountains from GeoNames.

```
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.
```

Listing 1: SQL query to extract 5 highest mountains of Austria from GeoNames.

3.2 Semantic Web

```
1 PREFIX dbpprop: <http://dbpedia.org/property
      />
2 PREFIX db:<http://dbpedia.org/resource/>
3 PREFIX dbpedia-owl:<http://dbpedia.org/
      resource/>
4 SELECT  ?pop ?countries ?area
5 WHERE {
6 db:Asia  dbpprop:population ?pop.
7 db:Asia dbpprop:countries ?countries.
8 db:Asia dbpedia-owl:areaTotal ?area.}
```

Listing 2: SPARQL query to extract facts of Asia from DBpedia.

DBpedia: It is one of the online structured information resources. 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. RDF is a language that is used to express data in the form of triples as shown in Figure 3. The RDF triple consists of a subject, predicate and an object. The sparql query is used to query RDF triples. In Line 6 shown in Listing 2, db:Asia is a subject whereas dbpprop:population represents the property and variable ?pop stores population of the Asia, which is returned when query is executed.

3.3 Geographic Web Portals

CIA World Factbook: It is the most widely used online information source for geography. The data can be downloaded from CIA world factbook freely. It provides an archive of geographic information for the last 10 decades which can be used for statistical inferencing. It is updated biannually. The Factbook[5] was selected as our starting point as there is no copyright issues.

3.4 Computational Knowledge Engine

WolframAlpha: 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[6] provides an API for extracting data. The API returns data in the form of XML. It needs subscription to allow open access to data. We have taken

[4] http://www.geonames.org/

[5] https://www.cia.gov/library/

[6] http://www.wolframalpha.com/

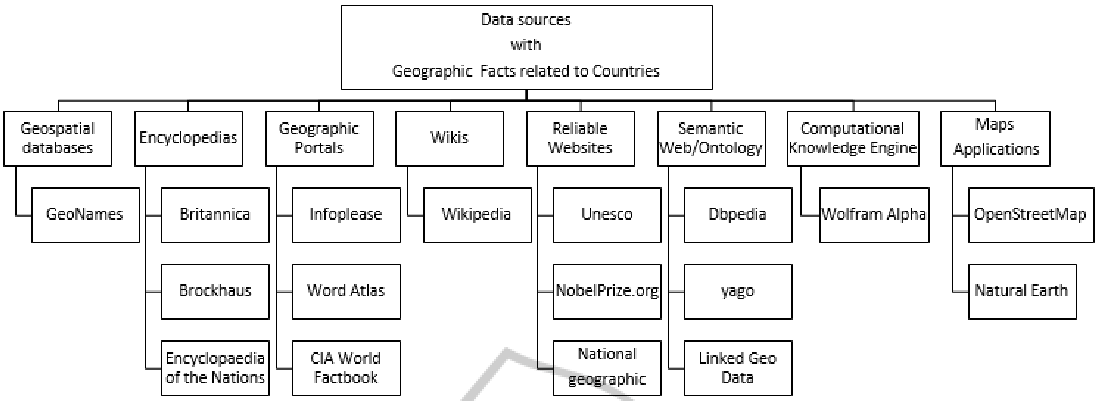


Figure 2: Geographic data sources.

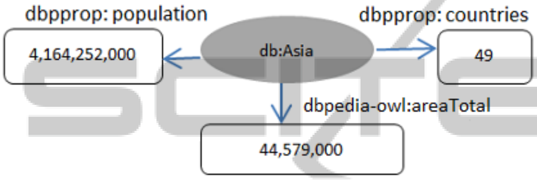


Figure 3: Semantic Network showing RDF Triples representing Asia.

```
C:\ >osmosis ——rbf austria.osm.pbf —— nkv
keyValueList=”amenity.bank” ——wx bank.osm
```

Listing 3: Osmosis command to extract banks of Austria from OSM file.

city and mountain rankings from Wolfram Alpha to verify GeoNames rankings.

3.5 Encyclopedias

Britannica: It provides a general article about a country which includes historical and cultural information. We have used QuickFacts[7] page of each country to verify the area of countries taken from Factbook.

3.6 Maps

OpenStreetMap (OSM): It is 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 banks in Austria from OpenStreetMap is shown in Listing 3.

Natural Earth: It is an online resource for free geographic dataset, maps and shape files. Natural Earth[8] provides both vector and raster graphics that can be used to draw map visualizations. It allows to download vector map files representing each country.

3.7 Reliable Web Sites

This category contains those web sites which contain valuable data such as UNESCO heritage sites, national parks and Nobel prize winners related to countries. The culture section of AF-Geo portal contains this valuable information. We are using both links and facts from those web sites. We redirect users to these web sites to learn more about facts.

4 METHODOLOGY

There were two motives behind collecting facts. i) We want to build a geographic web portal with a large number of facts related to countries and territories. ii) We want to automate the process of verification of collected facts. In this section we will deal with verifying geographic facts using maps along with techniques ranging from information extraction on Web, to facts comparison stored in multiple data sources, see (Weikum and Theobald., 2010), (Suchanek et al., 2009).

4.1 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 (Mehmood, 2014). Spatial operators allow automatic calculation of geographic features. Using spatial methods such as STArea(), STLength() we can verify facts such as area of country, boundary length

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

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

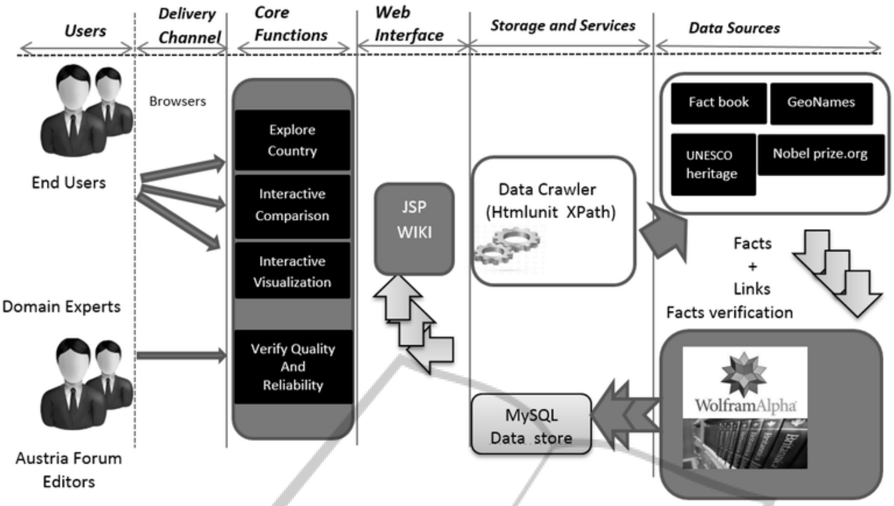


Figure 4: System Architecture.

of countries etc. The STArea() method returns the total surface area of a geographic entity such a polygon representing a country.

4.2 Verification using Encyclopedias

Traditional encyclopedias rely on domain experts to generate content. In contrast, Wikipedia uses "the wisdom of crowds" i.e. relies on a large community of users. It allows users to edit existing entries and therefore depends on its contributors which are not necessarily experts. Several studies have shown that Encyclopedia Britannica is an example of an accurate reference, see (Calzada and Dekhtyar, 2010), (Dalip et al., 2009). We thus use Britannica, Brockhaus[9] for verification of facts. Let's walk through the code which is used to extract area of countries from Britannica as shown in Listing 4. We are using Htmlunit[10](a window-less browser) API. It allows to query a web page using WebClient object, followed by information extraction using XPath[11] expressions. Once we have the access to table rows, using the Xpath expression (shown in line 4, 5 of Listing 4), we iterate through the rows and extract a specific table row that contains total area of a particular country. Afterwards, we populate our data store Britannica which is latter used for verification.

4.3 Verification using Reliable Sources

We also decided to go one step further and look at other data sources discussed in Section 3. We want to ensure the reliability by comparing facts with other online geographic portals. Therefore we choose infoplease.com which is a reliable web portal. In Infoplease, the area facts are stored in an HTML page; embedded in paragraph with class area, therefore we are capturing those <p> elements which contain the area of country using XPath expression below.

```
//p[@class='area']/text();
```

Figure 4 summarizes the overall process of data collection and verification. We extract facts from number of data sources (a sample is shown). After extracting facts we verify them using Britannica, WolframAlpha and other web portals such as Infoplease[12], World Atlas etc. for an additional check. The code snippet to extract large five cities from WolframAlpha is shown in the Listing 5. The wolfram API returns data in the XML format. We are using python library xml.etree.ElementTree[13] for parsing xml response, thus making it easy to extract cities from WolframAlpha. Our geographic portal has static collection of facts in the form of wiki pages which is based on JSP wiki syntax[14]. Therefore we generate them automatically using our verified and reliable collection of facts. A sample page is shown in Figure 7. Users and 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 up-

[9] http://www.brockhaus-wissensservice.com/

[10] http://htmlunit.sourceforge.net/

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

[12] http://www.infoplease.com/

[13] docs.python.org/2/library/xml.etree.elementtree.html

[14] http://www.ecyrd.com/JSPWiki/wiki/TextFormattingRules

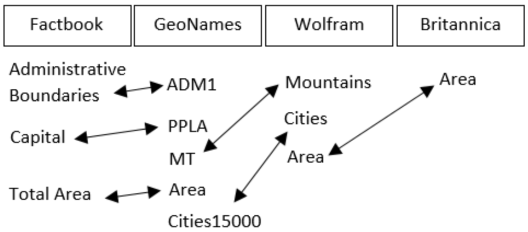


Figure 5: Schema Matching. Double sided arrows show similarity between concepts.

```
String URI=”britannica.com/topic/”+country+”−quick−facts”;
WebClient client = new WebClient();
HtmlPage mainPage = client.getPage(URI);
links1=(List<?>)mainPage.getByXPath(”//tr[@class=’eb−profile−table−even’]”);
HtmlTableRow tr=(HtmlTableRow)(links1.get(j));
links2=tr.getChildNodes();
```

Listing 4: Webpage parsing using Htmlunit and Xpath to extract area of countries from Britannica for verification.

date the existing wiki pages after log in. Interactive data visualisations are also provided for meaningful understanding of geography.

We now discuss the comparison of facts. This was a difficult procedure posing several challenges. Different data sources use different underlying names and schemas, see (Rahm and Bernstein, 2001), (Bellahsene et al., 2011) and (Bernstein et al., 2011). Figure 5 is showing concepts in different data sources and double sided arrows indicate similarity between two concepts in different data sources. For instance, ADM1 is a feature code in GeoNames and it represents administrative boundaries of a country such as provinces and districts; it is similar to the concept Administrative Boundaries in Factbook.

```
_q=Large+five+cities+of+”+country
url=”api.wolframalpha.com/query?input=_q”
response = urllib2.urlopen(url)
tree=ET.parse(response)
root=tree.getroot()
cities=root[0][0][0].text
```

Listing 5: Python code snippet to extract large five cities from WolframAlpha.

Area comparison was easy, as numbers representing area of country are easily compared using the similarity function in excel sheet. But prior to that, integrating facts and filling them into excel sheet was a tedious procedure. We also have to state here that our semi-automatic approach of verification of facts using data sources was only successful because of the powerful tools such as Htmlunit and Xpath. By harnessing the power of both, we succeeded in extracting the facts from Britannica, WolframAlpha and Infoplease which stores data in textual form, for verification.

5 RESULTS AND DISCUSSIONS

This section presents the summarized results of the verification process. The continent-wise city ranking results based on population, using two data sources (WolframAlpha, GeoNames) are shown in Table 2. For this particular study we are presenting top 5 city rankings. City rankings are found correct in case of 73 out of 193 countries (UN member states); whereas in case of 110 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. 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

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

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

From the above two lists the 4th entry is showing two different city names; the wikipedia[15] 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

[15] http://en.wikipedia.org/wiki/Bender,_Moldova

more carefully in our portal. The Not Verified represents those rankings which are different in two data sources. Let us look at the concrete case. Listing 6 is used to generate Table 3 which presents Not Verified case. Brasilia is ranked 4th by WolframAlpha whereas Fortaleza is ranked 4th in GeoNames. According to Wikipedia[16], Belo Horizonte is the 6th largest city in Brazil, but it is ranked 5th in GeoNames (Surprising: One city, three different rankings in three different data sources). Similarly if we look at city rankings of Bosnia in the list below, we find that Banja Luka is ranked 2nd in GeoNames whereas Zenika is ranked 2nd in Wolfram. According to Wikipedia[17] Zenica is 4th largest city in Bosnia and Herzegovina. There are few cases in the Not Verified category. They are due to lack of completeness.

Wolfram: Sarajevo; Zenica; Banja Luka; Tuzla; Mostar
GeoNames: Sarajevo; Banja Luka; Zenica; Tuzla; Mostar

For instance, Solomon Islands (a country associated with Australia) has no city list in WolframAlpha. All this 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. Now we turn to mountain rankings. Following is the list showing highest mountains of Austria.

Wolfram: Grossglockner; Wildspitze; Weisskugel; Grossvenediger; Similaun
GeoNames: Grossglockner; Wildspitze; Palla Bianca; Grossvenediger; Ramolkogel

The mountains ranked 1 and 2 are correct. The mountains ranked 3rd and 4th are also correct as Italian name of Weisskugel is Palla Bianca (different name case again!). Rank 5 seems incorrect, as Hinterer Brochkogel (3628 m) is higher than Similaun, see (Kulathuramaiyer et al., 2014). According to Wikipedia[18] the Similaun is the 6th highest summit, thus representing a Not Verified case. Sometimes data sources display different peaks of the same mountain which affects the ranking. Table 4 is showing the highest mountains of some European countries. Monte Rosa is the highest mountain in Switzerland according to Wolfram whereas Dom comes on top in Geonames, therefore it represents a Not Verified case. In the last row of Table 4, Hoverla and Goverla are different names for the same mountain in Ukraine, thus comes under Verified category.

Against our expectations we found discrepancies in area (size of country) figures also. Norway, France, Sweden, etc. are some of the examples as shown in Table 5; whereas countries like Philippines, Sri Lanka etc. have accurate area figures. A map (coded using shades of grey) of European countries displaying area difference is shown in Figure 6. The transition from white to dark grey shade represents the increase of the area difference. Countries like Austria, Poland have almost no area difference in multiple data sources where as France, Norway have much larger area difference, see Table 5. The continent-wise area verification results are shown in Table 6. The methodology of identifying Verified and Not Verified cases regarding area using multiple data sources is discussed in (Kulathuramaiyer et al., 2014). We found 79 out of 193 UN countries where area facts mismatch (Not verified). In case of 114 out of 193 UN countries area facts match (Verified).

While verifying information, we have learnt that even simple quantitative questions like “what is the area of a country in square kilometers” cannot be answered easily. We hope that with the help of community and domain experts, we will be able to clarify the reasons for discrepancies in as many cases as is possible. Another important consideration is that some data sources simply copy some facts from others, It is hard to determine the exact source for facts, due to multiple entries in reference section. But after exploring the reference section of mentioned data sources, we have noticed that WolframAlpha takes data from CIA World Factbook. The reference section of Wikipedia page of several countries points to CIA Factbook, besides other sources e.g. UN data[19]. The geographic web portals like Infoplease and World Atlas also take data from CIA World Factbook and U.S. Census Bureau[20].

A Reliable Geographic Web Portal

We therefore seem to have at least partially succeeded in building a large geographic portal which is reliable and contains a large number of facts. The portal is online and it is open to the general public, see Austria-Forum[21]. The main interface showing each country is shown in Figure 7. This geographic portal provides factual information about the 193 countries which are recognized by United Nations, along with a number of territories, oceans and islands.

Smart Display of Facts for further Verification

In order to get help from domain experts, when we find discrepancies in area figures we list the various numbers with their sources. In some cases it is pos-

[16] http://en.wikipedia.org/wiki/Belo_Horizonte

[17] http://en.wikipedia.org/wiki/Zenica

[18] http://en.wikipedia.org/wiki/Similaun

[19] https://data.un.org/

[20] http://www.census.gov/

[21] http://austria-forum.org/af/Geography/

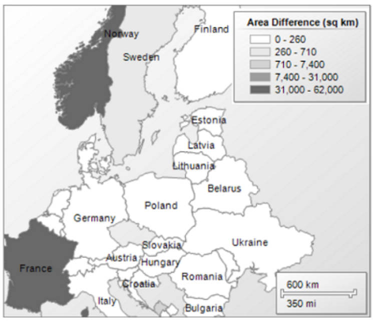
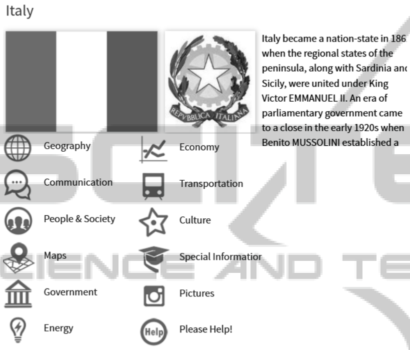
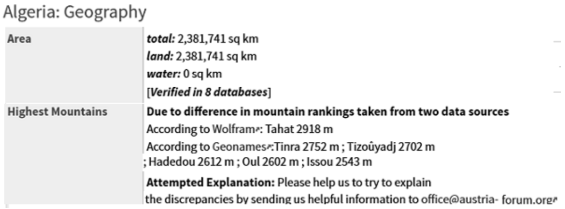


Figure 6: Map showing area difference of countries of Europe using data sources Cia World Factbook and DBpedia.


Italy

Italy became a nation-state in 186 when the regional states of the peninsula, along with Sardinia anc Sicily, were united under King Victor EMMANUEL II. An era of parliamentary government came to a close in the early 1920s when Benito MUSSOLINI established a

Figure 7: Main page of Italy in AF-Geo portal. It displays flag, emblem, short description and icons. The icons lead to sub-pages that contain details on related topics.


Algeria: Geography

Area	***total:*** 2,381,741 sq km ***land:*** 2,381,741 sq km ***water:*** 0 sq km [***Verified in 8 databases***]
Highest Mountains	**Due to difference in mountain rankings taken from two data sources** According to Wolfram: Tahat 2918 m According to Geonames:Tinra 2752 m ; Tizoûyadj 2702 m ; Hadedou 2612 m ; Oul 2602 m ; Issou 2543 m **Attempted Explanation:** Please help us to try to explain the discrepancies by sending us helpful information to office@austria- forum.org

Figure 8: Algeria Geography section. When facts in multiple data sources match we display [verified]. When facts mismatch (different mountain rankings) we ask for help from domain experts and community.

sible to find out why the numbers differ but in others, this might be an impossible job. The variation of numbers depends also on the point in time and the definitions involved makes all this a more than challenging task that will never be fully completed. Further, when we find identical figures we do indicate so, raising the probability that the figures are correct. Similarly we indicate where city and mountain rankings mismatch and ask the domain experts for verification in case of Not Verified rankings as shown in Figure 8.

We realize that this verification can only be successful if we get help from many persons (domain experts, geographers and organizations).

Table 2: Continent-wise city ranking verification results of 193 UN member states.

Continent	Verified	Partially Verified	Verified + Explanation	Not Verified
Europe	22	0	1	20
Africa	16	1	3	34
Australia	5	0	0	8
America	11	0	0	24
Asia	19	0	5	24
Total	73	1	9	110

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

Listing 6: Using joins to get city ranks from both tables Wolfram and Geonames based on ISO codes.

Table 3: Different city rankings of Brazil found in WolframAlpha and GeoNames.

Country	Wolfram_city	GeoNames_city	Rank
Brazil	Sao Paulo	Sao Paulo	1
Brazil	Rio de Janeiro	Rio de Janeiro	2
Brazil	Salvador	Salvador	3
Brazil	Brasilia	Fortaleza	4
Brazil	Fortaleza	Belo Horizonte	5

Table 4: Highest mountains of some countries, ✓shows Verified case, X shows Not Verified case.

Country	GeoName	Wolfram	
Albania	Maja e Jezercs (2694)	Golemi Korab (2764)	X
Switzerland	Dom (4545)	Monte Rosa (4633)	X
Germany	Zugspitze (2962)	Zugspitze (2962)	✓
France	Mont Blanc (4810)	Mont Blanc (4810)	✓
Montenegro	Durmitor (2522)	Deravica (2656)	X
Macedonia	Titov Vrv (2784)	Golemi Korab (2764)	X
Portugal	Ponta do Pico (2351)	Do Pico (2351)	✓
Serbia	Dolni Kara (2100)	Deravica (2656)	X
Ukraine	Hora Hoverla (2061)	Goverla (2061)	✓

6 EVALUATION

The AF-Geo portal provides about 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 approximately 7000 static pages displaying information about different aspects of countries. We present some quantitative measures as shown in Figure 9. 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).

Table 7: Comparison of Facts and Features of AF-Geo portal with other geographic data sources.

Facts and Features	Infoplease	Factbook	DBpedia[22]	AF-Geo portal	WorldAtlas	Geonames
• Geography	✓	✓	✓	✓	✓	✓
• Economy	✓	✓	✓	✓	✓	
• Transportation		✓		✓		
• Government	✓	✓	✓	✓		
• People and Society	✓	✓	✓	✓	✓	
• Pictures		✓		✓	✓	
• Energy		✓		✓		
• Maps		✓		✓	✓	✓
• Heritage				✓		
• Nobel prize winners				✓		
• National parks				✓		
• Facts validation				✓		
• Virtual lab				✓		
• Interactive visualizations				✓		

Table 5: Total Area (sq km) of countries in different data sources. For full list see, (Kulathuramaiyer et al., 2014).

Country	Factbook	DBpedia	Infoplease	Britannica
Philippines	300000	300001	300000	300000
Sri Lanka	65610	65610	65610	65610
Finland	338145	338242	338145	390903
Norway	323802	385183	324220	385186
France	643801	674843	547030	543965
Sweden	450295	449964	449964	447420
Austria	83871	83855	83870	83879
Belgium	30528	30528	30528	30528
Poland	312685	312602	312685	312679

Table 6: Continent-wise area verification results of 193 UN member states.

Continent	Verified	Not Verified
Europe	30	13
Africa	35	19
Australia	7	6
America	20	15
Asia	22	26
Total	114	79

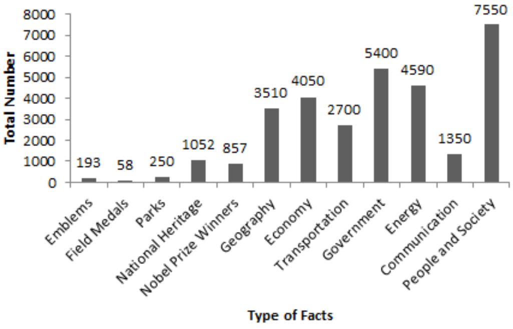
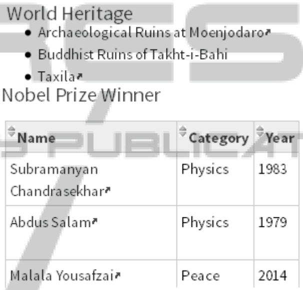


Figure 9: Facts Collection Summary.

Further, we present a comparison with other geographic data sources as mentioned in the Table 7. The on demand exploration of data (countries comparison, top and bottom rankings of UN member states) using virtual lab, interactive visualizations, and sophisticated reliability measures for validation of facts, makes AF-Geo portal fairly unique.



World Heritage

- Archaeological Ruins at Moenjodaro
- Buddhist Ruins of Takht-i-Bahi
- Taxila

Nobel Prize Winner

Name	Category	Year
Subramanyan Chandrasekhar	Physics	1983
Abdus Salam	Physics	1979
Malala Yousafzai	Peace	2014

Figure 10: Culture section of Pakistan in AF-Geo Portal.

Besides factual data, we have consolidated pictures from different data sources. The current collection contains approximately 20,000 pictures. The culture section of each country is very significant. It shows different aspects (heritage, national parks, different prize winner holders lists such as Wolf prize, Abel prize, Turing award) all related to the culture of a country. Figure 10 shows a sub-page that represents aspects of the culture of Pakistan.

7 CONCLUSION AND FUTURE WORK

The AF-Geo portal will hopefully succeed by providing reliable facts or at least it indicates those facts that differ in different sources. Additionally, if users want to explore a country, they have sufficient information. We believe that this geographic web portal can be valuable for large number of communities, particularly teachers, students or some tourists, who can use this for learning and understanding geography. Overall, it is clear that verification of information is an extremely important task. So far, very little has been done towards measuring quality and reliability of online information. However, reliability seems to us is at most as important as just raw quality. We hope to extend this verification to other quantitative facts in the future.

REFERENCES

Bellahsene, Z., Bonifati, A., and Rahm, E., editors (2011). *Schema Matching and Mapping*. Data-Centric Systems and Applications. Springer.

Bernstein, P. A., Madhavan, J., and Rahm, E. (2011). Generic schema matching, ten years later. *PVLDB*, 4(11):695–701.

Calzada, G. and Dekhtyar, A. (2010). On measuring the quality of wikipedia articles. In *Proceeding WICOW 10 Proceedings of the 4th workshop on Information credibility*, pages 11–18.

Carlo Batini, M. S. (2006). *Data Quality Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer, USA.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Dalip, D. H., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304.

Kulathuramaiyer, N., Maurer, H., and Mehmood, R. (2014). Some aspects of the reliability of information on the web. *JUCS*, 20(9):1284–1305.

Mehmood, R. (2014). Geographic data verificaiton. *IPSI*, 10(2):20–25.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB*, 10:334–350.

Suchanek, F. M., Sozio, M., and Weikum, G. (2009). Sofie: A self-organizing framework for information extraction. In *In Proceedings of WWW*, pages 631–640.

Weikum, G. and Theobald., M. (2010). From information to knowledge: harvesting entities and relationships from web sources. In *In Proceedings of PODS*, pages 65–76.

Wurzinger, G. (2010). Information consolidation in large bodies of information. *JUCS*, 16(21):3314–3323.

Zhao, S. and Betz, J. (2007). Corroborate and learn facts from the web. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 995–1003.