Journal of Universal Computer Science, vol. 20, no. 9 (2014), 1284-1303
submitted: 27/6/14, accepted: 29/8/14, appeared: 1/9/14 © J.UCS

Some Aspects of the Reliability of Information on the Web

Narayanan Kulathuramaiyer
(Faculty of Computer Science and Information Technology, University of Malaysia Sarawak
Kota Samarahan, Malaysia
nara@fit.unimas.my)

Hermann Maurer
(Institute for Information Systems and Computer Media. Graz University of Technology
Graz, Austria
hmaurer@iicm.edu)

Rizwan Mehmood
(Institute for Information Systems and Computer Media. Graz University of Technology
Graz, Austria
rmehmood@iicm.tu-graz.ac.at)

Abstract: 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 paper 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.

Keywords: Reliability of information, verification of information, checking facts, statistical techniques, heuristic approaches
Categories: H.1, H.3. H.4, L.1, L.4, L.6

1 Introduction

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

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 [Maurer et al. 2014] 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 has already tried to explain many years ago in [Maurer 2004] 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 2014] and [Utexas 2014] 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, should we not 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 statement has been made against 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 [Taleb 2007]; 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 (2013) 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* [Baker & Potts, 2013] and other terms that result from mere rumors that becomes easily spread on the web [Niggemeier, 2012] 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 [Baker & Potts, 2013]. 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, 2013] 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 Fig. 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 reinforcing their importance. At the same time it is also highly likely that users become distracted and drift away from the original search intent. Search engine 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.

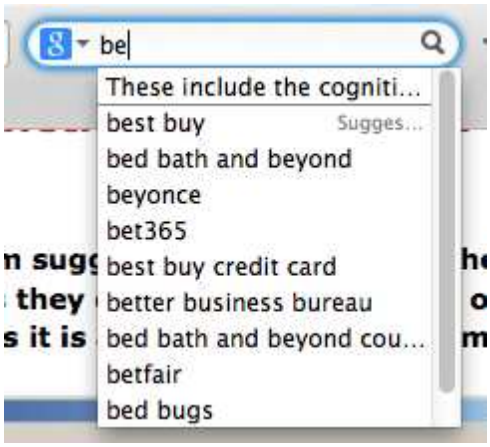

Figure 1: Results of Auto-Complete suggestions

[Baker & Potts, 2013] 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 [Couldry, 2003] and in a way indirectly altering the recording of history [Witten et al., 2007] is irresponsible and can be extremely dangerous [Weber, 2006]. 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 [Keen, 2007, pg. 16].

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 [Kulathuramaiyer and Maurer, 2014]

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 related area which relates 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 [Molavi Kakhki, et al., 2013]. 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 needs 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 form 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, this paper checks 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 starting with sources 1: [Factbook 2014], 2: [DBpedia 2014], 3: [GeoNames 2014] and a number of special sites like [UNESCO 2014]. We then compare the information obtained and check it against other sources like 4: [Infoplease 2014], 5: [Britannica 2014], 6: [WolframAlpha 2014], and others.

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 2 we explain our main approach a bit more carefully, then present in Sections 3-5 three different topics that we checked. Section 6 is a short conclusion.

2 Trying to test reliability of information in a special case

We started a project to set up a server "Geography of the world".[Geography 2014]. We initially imported the information on all countries listed in [Factbook 2014] 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 "Please provide help." Under "Culture" we imported for each country data (or links) to all UNESCO heritage sites from the server

[UNESCO 2014], and similarly information on Nobel Prize laureates from [Nobelprize 2014]. Awardees of the Wolf prize from the [Wolf Foundation 2014], of Fields fellows from [Fields 2014] and Field medallists from [Areppim 2014] and others from reliable list will follow. 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 [Factbook 2014] when available, with many others coming from elsewhere.

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 "Please provide help" is a plea to the community to improve and add information. Mind you, all information obtained will be screened by an editorial team whose composition will be publicized.

However, most emphasis 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 square kilometres for each country (Section 3) and were in for some surprises. We added major cities to each country (Section 4) and major mountains (Section 5), provided that mountains above 1000 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 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 2014] 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 Italien 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 [UN 2014]? 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 2014] "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.

3 Area of countries

Before looking at details let us point out that even the definition of area of a country is done in a rather arbitrary way: 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 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 claims are included. 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 the 82 of the 193 UN countries, a number whose size that may come as surprise. If we consider Difference=10 as threshold we still find 41 countries whose area in square kilometres differs according to the sources by more than 1%. For a list of all UN countries see [UN 2014], we list below for brevity only those in Europe ones with a Difference>1.

Country	Factbook	Dbpedia	Geoname	Infoplease	Britannica	Wolfram	Difference
Croatia	56594	56594	56542	56542	56954	56594	6
Cyprus	9251	9251	9250	3571	5896	9251	194
Finland	338145	338242	337030	338145	390903	338145	127
France	643801	674843	547030	547030	543965	551500	154
Ireland	70273	84421	70280	70280	70273	70273	162
Liechtenstein	160	160	160	161	160	160	5
Macedonia	25713	-	25333	25333	25713	25713	6
Malta	316	316	316	321	315	316	14
Montenegro	13812	12999	14026	14026	13812	13812	20
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 Cyprus (194), Finland (127), France (154), Ireland (162), Malta (14), Montenegro (20), 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 paper.

Cyprus	9251	9251	9250	3571	5896	9251	194

Cyprus really consists of four parts: The Southern part, the Northern (Turkish) part, the demilitarized zone in-between, and two small areas Akrotiri and Dhekelia that are part of the “British Overseas Territory on Cyprus”. The total area of the island

is around 9250 km² as listed in sources 1, 2 ,3 and 6 and others like in [Countrycode 2014] and [Worldbank 2014]. Source 4 lists the area of around 3570 km² which is reasonably close to the area of Northern (Turkish) Cyprus which other sources list with a bit less: 3.350 km². Source 5 gives the area of Southern (Greek) Cyprus at around 5900. The last two figures add up to exactly 9250. Two minor problems remain: where do the extra 220 km² come from in Source 4, and where is the area of Akrotiri and Dhekelia (with some 250 km²) plus the area of the demilitarized zone taken into account? It seems clear that those comparatively small areas are bundled into the other figures in different ways in the ten sources that we consulted is not obvious. One of the many cases where the community may provide help!

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 2014] and [Countrycode 2014] are quoting. The figures include the area of lakes (roughly 34.500 km²) which seems reasonable. [Worldbank 2014] gives only around 304.000, 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 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 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 regions (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 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 2014] agrees with Source 1, in [Worldbank 2014] 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 2014] 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 2014] 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. It is for this reason that we have developed a tool that allows to take a map and determine the area inside a curve you have drawn, allowing to check figures yourself. While we have used a first prototype of the tool for clarifying some instances we will return to the value of such a tool when made available to the public in a future paper.

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	39

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 2014. 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 2014] agrees with Source 1, [Worldbank 2014] with 5.

Chad	1000000	1283994	1284000	1284000	1284000	1284000	38
Niger	1000000	1267000	1267000	1267000	1267000	1267000	36

In both cases, [Factbook 2014] seems to consider the official boundaries of those countries between themselves and Lybia somewhat artificially drawn in literally lifeless Sahara, too ill-defined so it just lists a rough estimate. The figures of all other

Sources including [Brockhaus 2014], [Countrycode 2014] and [Worldbook 2014] agree!

Gambia, The	11295	10689	11300	11300	11632	11295	34

Except for Source 2 there is agreement that Gambia has about 11.300 km², a figure also supported by [Brockhaus 2014], [BBC 2014] and [Countrycode 2014]. 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 2014] also reports a figure significantly below 11.000 explains the lower figures: They ignore the area of inland water ([Worldbank 2014] 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!

For the purpose of this paper 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	117
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 2014], [Daftlogic 2014] and [Freemaptools 2014] (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 and Kashmir" belongs to Pakistan, how much to India. Additionally, the boundary in the North to China is not clear at all, but there is almost continuous fighting at altitudes over 4.000 m, actually more between India and China than Pakistan and China!

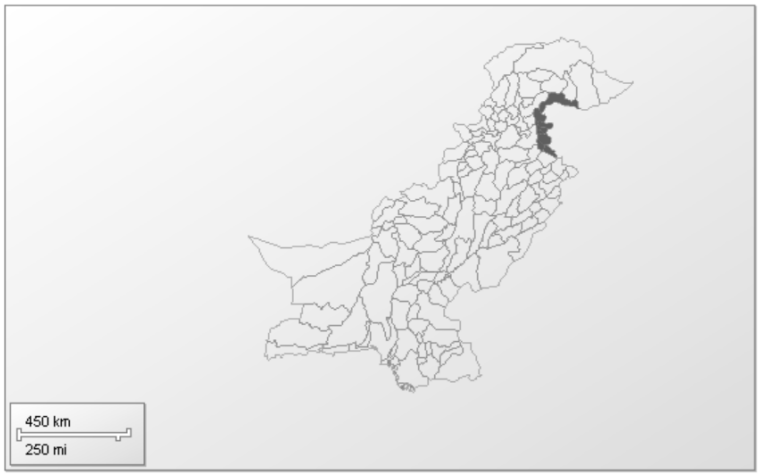

Figure 2: Pakistan (Map taken from [Natural Earth 2014], the dark part is the contentious area of Kashmir)

Let us just briefly mention some points concerning the territories which are not in [UN 2014]:

Jan Mayen	377	373	-	-	-	62045	1964
Svalbard	62045	61022	-	-	-	62045	6

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	163

The area of Greenland is indeed what all Sources above except one indicate, and 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, that is indeed not far West of Greenland. Note that Europa considers Greenland definitely part of Europe, but (according to [Factbook 2014]) it is part of North America. It is also worth noting that [Worldbank 2014] gives only an area of 410.450 km², i.e. discounts all areas covered by deep ice.

Let us finish with a curiosity:

British Indian Ocean Territory	54400	54400	60	220	-	-	995

This territory today only comprises some small islands and reefs. The area indicated includes all the ocean surrounding a total of less than 100 km² of land including Diego Garcia, now a US naval base.

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

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 2014] 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 and 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 shows a few samples of countries of cities arranged by size and problems encountered.

Pakistan

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

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 2014]. Gujaranwala wins with a small margin, so both cities can be seen as more or less the same size (1.4 million).

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

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 Georname 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.

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	Missisauga	Edmonton	Quebec	Mississauga
7	Winnipeg	Mississauga	Hamilton	Winnipeg
8	Vancouver	NorthYork	Winnipeg	Vancouver

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 Missisauga 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.

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

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 [DOSM 2014] and is reflected correctly in Source 4. Source 1 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.

5 Some mountains of countries ranked by height

In this section we are considering only countries with mountains higher than 1.000 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, 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?

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

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.

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 above is because of borders on or near the peak 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 7 th 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.

6 Conclusion

In this paper we have tried to show that even when using multiple Web sources and simple quantitative questions they are often not easy to resolve without the help of specialists. Hence information obtained on servers and search engines on the WWW is much less reliable and trustworthy than is usually assumed, confirming earlier arguments in e.g. [Rieh 2002], [Liu et al 2005] and [Parker at el 2006].

To provide a really trustworthy service in our project “Geography of the world” we will indeed have to involve specialists in Geography and persons with good local knowledge beyond the research we can do ourselves..

Let us conclude by mentioning that some of the measurements facilities described will be available through or project, but we will also point to powerful facilities like Google Earth Pro and similar efforts. Hence we hope that our project will provide a valuable tool for all interested, including teachers and students. It will also contain a number of interactive facilities that will allow to experiment as will be explained in detail in a forth coming paper.

References

[Areppim 2014] The complete list of Fields Medal winners, http://stats.areppim.com/listes/list_fieldsxmedal.html, visited: 10 September 2014

[Baker and Potts 2013] Baker, P., Potts, A.: “Why do white people have thin lips? Google and the perpetuation of stereotypes via auto-complete search forms”; Critical Discourse Studies, 10, 2 (2013), 187-204, DOI:10.1080/17405904.2012.744320

[BBC 2014] The Gambia Profile, http://www.bbc.com/news/world-africa-13378351, visited 12 September 2014

[Britannica 2014] ENCYCLOPEDIA BRITANNICA, http://www.britannica.com/, visited: 10 September 2014

[Brockhaus 2014] BROCKHAUS, http://www.brockhaus-wissensservice.com/, visited: 10 September, 2014

[Cornell 2014] Five criteria for evaluating Web pages, olinuris.library.cornell.edu/ref/research/webcrit.html, visited 12 September 2014

[Couldry 2003] Couldry, N.: Media Meta-Capital: Extending the Range of Bourdieu's Field Theory, Theory and Society, 32, 5 (2003), 5-6

[Countrycode 2014] Country Codes, Phone Codes, Dialing Codes, Telephone Codes, ISO Country Codes, http://Countrycode 2014.org/, visited: 12 September 2014

[DaftLogic 2014] Google Maps Area Calculator Tool, http://www.daftlogic.com/projects-google-maps-area-calculator-tool.htm, visited: 12 September 2014

[DBpedia 2014] DBpedia, http://dbpedia.org/About, visited: 10 September 2014

[DOSM 2014] The Source of Malaysia's Official Statistics, http://www.statistics.gov.my/portal/download_Population/files/population/03ringkasan_kawasan_PBT_Jadual1.pdf, visited: 14 September 2014

[Factbook 2014] The WORLD FACTBOOK. https://www.cia.gov/library/publications/the-world-factbook/, visited: 10 September, 2014

[Fields 2014] Fields Institute Fellows, http://www.fields.utoronto.ca/honours/fieldsinstfellows.html, visited: 10 September 2014

[Freemaptools 2014] Free Map Tools, http://www.freemaptools.com/area-calculator.htm, Visited Sept 12, 2014

[Geography 2014] http://austria-forum.org/af/Geography, open as of November 5, 2014

[GeoNames 2014] GeoNames, http://www.geonames.org/, visited: 10 September, 2014

[Infoplease 2014] Countries of the World, http://www.infoplease.com, visited: 10 September 2014

[Keen 2008] Keen, A.: "The cult of the amateur"; Double Day (2008).

[Kulathuramaiyer and Maurer, 2014] Kulathuramaiyer, N., Maurer, H., A Survey of Communications and Collaborative Web Technologies, accepted for publication in CIT, 2014

[Liu et al 2005] Liu, Z., Huang, X.: "Evalutating the credibility of scholarly information on the web: A cross cultural study"; Science Direct, 37, 2 (2005), 99-106

[Maurer et al. 2014] Maurer, H., Mehmood, R., Korica-Pehserl, P.: "How Dangerous is the Web for Creative Work"; CIT 21, 2 (2013), 59-69.

[Maurer 2004] Maurer, H.: Der Berg von hinten; In: XPERTEN- Der Anfang, Freya Publishing, Austria (2004), 220-222.

[Molavi Kakhki et al. 2013] Molavi Kakhki, A., Kliman-Silver, C., Mislove, A. Iolaus: "Securing online content rating systems"; In: Proceedings of the 22nd international conference on World Wide Web(2013), 919-930

[Niggemeier 2012] Niggemeier, S.: "Autocompleting Bettina Wulff: Can a Google Function Be Libelous?", http://www.spiegel.de/international/zeitgeist/google-autocomplete-former-german-first-lady-defamation-case-a-856820.html, visited : 12 September 2014

[Nobelprize 2014] The official web site of Nobel Prize, http//www.nobelprize.org, visited: 10 September 2014

[Natural Earth 2014]Natural Earth, http://www.naturalearthdata.com/, visited: 12 September 2014

[Parker at el. 2006] Parker, M.B., Moleseh, V., De la Hapre, R., Will, G. B..: "An evaluation of Information quality frameworks for the World Wide Web"; Proc. 8th Annual conference on WWW Applications (2006),1-11

[Rieh 2002] Rieh, S.J.: "Judgement of Information Quality and Cognitive Authority in the web"; Journal of American Society of Information Science and Technology, 53,2 (2002), 145-161

[Taleb 2011] Taleb, N. N.: The Black Swan: Penguin Books (2011)

[UNESCO 2014] UNESCO World Heritage List, http://whc.unesco.org/en/list/, visited: 10 September, 2014

[UN 2014] Member States of the United Nations, http://www.un.org/en/members/ index.shtml, visited September 10 2014

[Utexas 2014] How Can I Tell if a Website is Reliable, http://www.edb.utexas.edu/ petrosino/Legacy_Cycle/mf_jm/Challenge%201/website%20reliable.pdf,visited:12 September 2014

[Weber 2006] Weber, S.: Das Google-Copy-Paste-Syndrom, Wie Netzplagiate Ausbildung und Wissen gefährden, Heise, Hannover (2006)

[Witten et al. 2007] Witten, I. H., Gori, M., Numerico, T.: Web Dragons, Inside the Myths of Search Engine Technology, Morgan Kaufmann, San Francisco (2007)

[WolframAlpha 2014] Wolfram Alpha Computational Knowledge Engine, http://www.wolframalpha.com/, visited: 10 September 2014

[Wolf Foundation 2014] Wolf Foundation, http://www.wolffund.org.il, visited: 10 September 2014

[Worldbank 2014] Land area (sq.km), http://data.Worldbank.org/indicator/AG.LND.TOTL.K2, visited September 12, 2014