

Moon Landing or Safari? A Study of Systematic Errors and their Causes in Geographic Linked Data

Krzysztof Janowicz¹, Yingjie Hu¹, Grant McKenzie², Song Gao¹, Blake Regalia¹, Gengchen Mai¹, Rui Zhu¹, Benjamin Adams³, and Kerry Taylor⁴

¹ STKO Lab, University of California, Santa Barbara, USA

² Department of Geographical Sciences, University of Maryland, USA

³ Centre for eResearch, The University of Auckland, New Zealand

⁴ Australian National University, Australia

Abstract. While the adoption of Linked Data technologies has grown dramatically over the past few years, it has not come without its own set of growing challenges. The triplification of domain data into Linked Data has not only given rise to a leading role of places and positioning information for the dense interlinkage of data about actors, objects, and events, but also led to massive errors in the generation, transformation, and semantic annotation of data. In a global and densely interlinked graph of data, even seemingly minor error can have far reaching consequences as different datasets make statements about the same resources. In this work we present the first comprehensive study of systematic errors and their potential causes. We also discuss lessons learned and means to avoid some of the introduced pitfalls in the future.

1 Introduction and Motivation

Over the last few years, the Linked Data cloud has grown to a size of more than 85 billion statements, called triples, contributed by more than 9,900 data sources. A cleaned and quality controlled version made available via the LOD Laundromat [2] contains nearly 40 billion triples.¹ The Linked Data cloud (and proprietary versions derived from it and other sources) have brought dramatic changes to industry, governments, and research. For instance, they have enabled question answering systems such as IBM's Watson [3] and Google's new knowledge graph. Linked Data has also increased the pressure on governments to publish open data in machine readable and understandable formats, e.g., via data.gov. Finally, it has enabled the research community to more efficiently publish, retrieve, reuse, and integrate, scientific data, e.g., in the domain of pharmacological drug discovery [16]. The value proposition of Linked Data as a new paradigm for data publishing and integration in GIScience has been recently discussed by Kuhn et al. [10].

Places and positioning information more broadly play a prominent role for Linked Data by serving as nexuses that interconnect different statements and

¹<http://lodlaundromat.org/>

contribute to forming a densely connected global knowledge graph. GeoNames, for example, is the second most interlinked hub on the Linked Data Web, while DBpedia contains more than 924,000 entities with direct spatial footprints and millions of entities with references to places. Examples of these include birth and death locations of historic figures and places where notable events occurred. Many other datasets also contain spatial references such as sightings of certain species on Taxonconcept,² references to places in news articles published by the New York Times Linked Data hub,³ and affiliations of authors accessible via the RKB Explorer,⁴ to name but a few. In fact, most Linked Data are either directly or indirectly linked through various spatial and non-spatial relations to some type of geographic identifier.

Nonetheless, current statistics show that about 66% of published Linked Datasets have some kind of problems including limited availability of SPARQL query endpoints and non-dereferenceable IRIs.⁵ A recent study of Linked Datasets published through the Semantic Web journal shows that about 37% of these datasets are no longer Web-available [6]. In other words, even the core Linked Data community struggles to keep their datasets error-free and available over longer periods. This problem, however, is not new. It has been widely acknowledged that proper publishing and maintenance of data are among the most difficult challenges facing data-intensive science. A variety of approaches have been proposed to address this problem, e.g., providing a sustainable data publication process [15]. Simplifying the infrastructure and publishing process, however, is just one of many means to improve and further grow the Web of Linked Data. Another strategy is to focus on controlling and improving the quality of published data, e.g., through unit testing [9], quality assessment methods such as measuring query latency, endpoint availability, and update frequency [17], as well as by identifying common technical mistakes [5].

Given the importance of places and positioning information on the Linked Data cloud, this paper provides the first comprehensive study of systematic errors, tries to identify likely causes, and discusses lessons learned. However, instead of focusing on *technical* issues such as non-dereferenceable IRIs, unavailable SPARQL endpoints, and so forth, we focus on Linked Data that is technically correct, available, and in (heavy) use. We believe that understanding quality issues in the *contents* published by leading data hubs will allow us to better understand the difficulties faced by most other providers. We argue that the lead issue is the lack of best practices for publishing (geo)-data on the Web of Linked Data. For instance, geo-data is often converted to RDF-based Linked Data without a clear understanding of reference systems or geographic feature types. Our view is not unique and has recently led to the first joint collaboration of the Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C) by establishing the *Spatial Data on the Web Working Group*.

²<http://www.taxonconcept.org/>

³<http://data.nytimes.com/>

⁴<http://www.rkbexplorer.com/>

⁵<http://stats.lod2.eu/>

In the next sections, we will categorize systematic errors into several types and discuss their impact and likely causes. We will differentiate between (I) errors caused by the triplification and extraction of data, (II) errors that result from an improper use of existing ontologies or a limited understanding of the underlying domain, (III) errors in the design of new ontologies and oversimplifications in conceptual modeling, and (IV) errors related to data accuracy and the lack of an uncertainty framework for Linked Data. Some errors are caused by a combination of these categories. We realize that studies of data quality are often not met with excitement and thus have selected interesting and *humorous* examples that illustrate, with *serious* implications, the far-reaching consequences of seemingly small errors. Finally, we would like to clarify that our work is motivated by improving the quality of the Linked Data to which we contributed datasets ourselves, not in merely blaming errors made by others. We notified the providers of the discussed datasets and some of the issues presented here have been resolved. We hope that our work will help to prevent similar errors in the future.

2 Triplification and Extraction Errors

There are three common ways in which Linked Data is created today. The most common approach is to generate Linked Data from other structured data such as relational databases, comma separated value (CSV) files, or ESRI shapefiles. This approach is often called triplification, i.e., turning data into (RDF) triples. As a second approach, Linked Data is increasingly extracted using natural language processing and machine learning techniques from semi-structured or unstructured data. The most common example is DBpedia [11] which converts (parts of) Wikipedia into Linked Data. Another example is the ontology design patterns-based machine reader FRED that parses any natural language text into Linked Data [14]. Finally, in a small but growing number of cases, Linked Data is the native format in which data are created. This is typically the case for derived data products, such as events mined from sensor observations, metadata records from publishers and libraries, and so on.

The first two approaches share a common workflow. First, the relevant content has to be extracted, e.g., from a tabular representation in hypertext markup language (HTML). Next, the resulting *raw* data have to be analyzed and processed. In a final step, the processed data must be converted into Linked Data by using an ontology. While errors can be introduced during each of these steps, this section focuses on errors introduced during the extraction of data and the conversion into Linked Data, i.e., triplification errors.

One way of studying whether systematic errors have been introduced during the triplification process is to visually map geographic features present in the Linked Data cloud. Figure 1 shows the result for about 15 million features extracted from multiple popular Linked Data sources such as DBpedia, Geonames, Freebase, TaxonConcept, New York Times, and the CIA World Factbook.

These features have been selected through SPARQL queries for all subjects that have a W3C Basic Geo predicate, i.e., `geo:lat` or `geo:long`. For DBpedia, we included multiple language versions. What is noteworthy about Figure 1 is

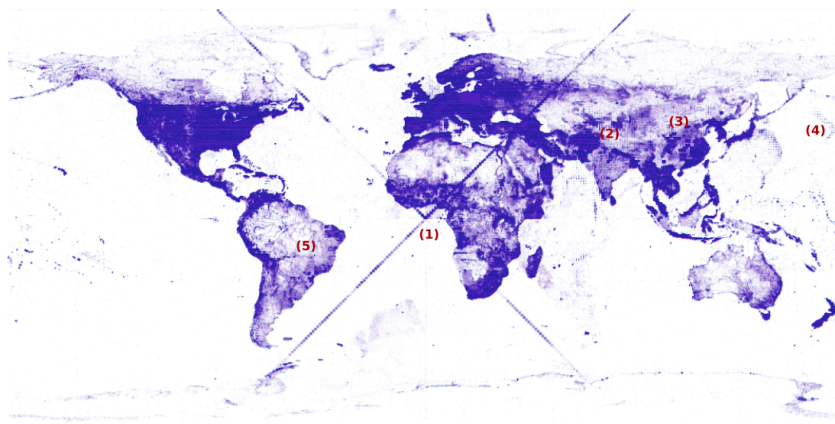


Fig. 1 A representative fraction of places in Linked Data (EPSG:4326, Plate Carree).

the lack of a base map, i.e., the figure is entirely based on point data.⁶ In other words, the Linked Data cloud has high spatial coverage. One can easily identify the outlines of continents and most of the land surface is covered by features with varying density. This is also true for regions in the far North and South of the planet. Nonetheless, one can immediately identify significant errors – the most obvious being perpendicular lines crossing in the middle of the map. In this work, we do not focus on random errors (which are expected in a sample of this size and arise from largely unpredictable and thus not easily correctable reasons), but instead on *systematic* errors inherent to the data. These errors are further examined through a set of cases as follows.

Case 1 shows a massive \times -like structure which represents numerous problems with geographic coordinates such as latitudes and longitudes sharing the same single value. This indicates that latitude values were mistaken for longitude values and vice versa. We also found cases where only latitude values or longitude values were given or where multiple appeared such as entities having two latitude values without any longitudes. The quantity of these errors suggests that they are systematic. Most likely, they stem from problems with scraping or parsing scripts. Cases where features were mapped to (0,0) will be discussed below.

Case 2 depicts one of many examples of grid-like structures. From our observations, these are caused by two separate issues. First, features are often merely represented by coarse location information, e.g., by only using degrees and dropping decimals. Second, the vast majority of geo-data on the (Linked Data) Web today relies on point geometries. This also includes centroids for regions such as counties, countries, mountain ranges, rivers, and even entire oceans. To give a concrete examples, Geonames places the Atlantic Ocean at (10N, 25W), while

⁶A high resolution version that gives a better impression of the coverage as well as various errors is available at http://stko.geog.ucsb.edu/pictures/lstd_map.png.

DBpedia places it about 1200 km away at (0N, 30W). Note, however, that many of the features visible in the oceans are not necessarily errors. They include submarine volcanos, mid-ocean ridges, or reports about events such as oil spills. Whether centroids are errors or are simply an inaccurate and largely meaningless way of representing large regions depends on the context in which the data are to be used. However, it is difficult to imagine use cases for centroids of oceans particularly as the two examples above show the arbitrariness of these locations. The same argument can be made for coarse location data, and in fact, we will discuss one example in greater detail below.

Cases 3 and 4 can be seen through block-like structures in China and a second New Zealand in the Northern Hemisphere. The vast majority of these errors are systematic and appear in the DBpedia dataset. We were able to track down a potential reason for them by exploring the different language versions of DBpedia. It appears as though the scripts used by DBpedia curators to extract content from Wikipedia either expected signs, e.g., (34.413,-119.848), or a hemisphere designator, e.g., (34.413N,119.848W). Some language versions of Wikipedia, e.g., the Spanish version, use other character designators such as (34.413N,119.848O) where O stands for *oeste*. It is likely that the script dropped the O instead of replacing it with a W. Consequently, geographic features in the United States for which a Spanish language version was available in Wikipedia ended up in China. This also explains the lower density of those misplaced features, i.e., the Spanish Wikipedia lists fewer places in the US than the English version. Other, likely non-systematic, errors include the flattening factor for the Earth being reported as 1 which could be caused by a parsing error (or ceiling function) as the data type reported by DBpedia is an `xsd:integer`.⁷

Case 5 in Figure 1 is not an error but rather a reminder that despite the overall coverage, certain regions are underrepresented. Interestingly, the Linked Data map bears a remarkable similarity to maps created for different (social media) datasets such as Flickr, Twitter, Wikipedia, and so forth. This highlights two issues. First, and as outlined previously, most Linked Data are created from existing data sources, and secondly the same underlying biases appear to apply for most of these data sources. In other words, most data used in the Linked Data cloud share the same *blind spots*.

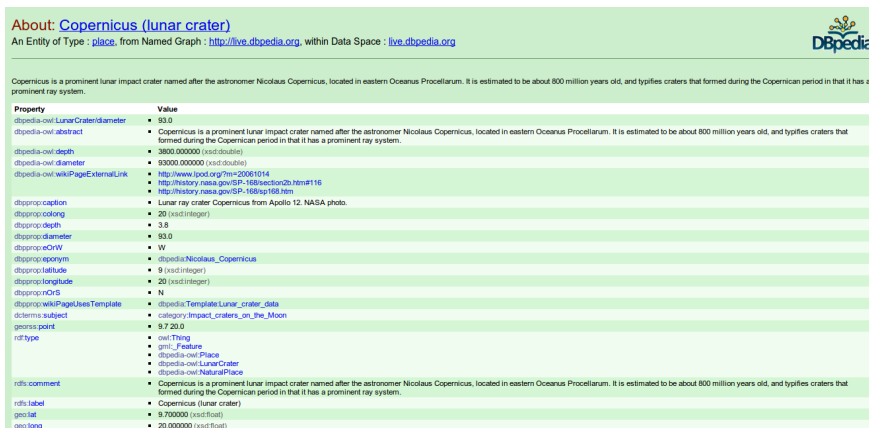
Lessons Learned: Two major sources of errors can be differentiated, those introduced during triplification and knowledge extraction as well as those that were part of the original source data. In the first case, errors are typically introduced by software that does not take the full range of possible syntactic variations into account (e.g., *west* versus *oeste*) or fails to accurately distinguish between point-features and bounding boxes. Furthermore the software may confuse latitudes with longitudes for other reasons (causing the ×-like feature in Figure 1) or parse and cast the data into inappropriate formats (e.g., the flattening factor). For the second type of errors, one could argue that

⁷ SPARQL: `ASK WHERE <http://dbpedia.org/resource/Earth>
<http://dbpedia.org/property/flattening> 1. [using DBpedia 2015-04.]`

they are not specific to Linked Data but simply a result of errors in the source data. Such argument, however, misses the substantial difference between information embedded in the context of a Web page published for human use with the decontextualized raw data statements that form an interlinked and the machine-available knowledge graph. While the Atlantic Ocean was represented by a point-like feature at (0N, 30W) in Wikipedia, it is the DBpedia version that allows for inferences such as plotting the place of death of people who are known to have died somewhere in the Atlantic Ocean (e.g., Benjamin Guggenheim) at 0N, 30W. Summing up, triplification and Linked Data extraction require substantial domain expertise. Approaches such as unit testing and simple integrity constraints could be used to detect many of the errors described above. For instance, most of the places in the US that were duplicated in China also contain topological information such as being part of a county or a state. Thus, checking whether the space-based and place-based information match could be a powerful method to avoid such errors in the future.

3 Ontology Usage and Domain Errors

To improve retrieval and reuse, Linked Data is typically created by using shared ontologies and vocabularies. Most of these, however, are underspecified to a degree where the intended interpretation is largely conveyed by the labels and simple hierarchies rather than a deeper axiomatization. The need for and value of a more expressive formalization is still controversially debated with recent work highlighting the need for stronger ontologies. The following example illustrates the problems that can arise from a lack of deeper axiomatization or the improper use of ontologies outside of their intended interpretation.



About: Copernicus (lunar crater)
 An Entity of Type: *place*, from Named Graph: <http://live.dbpedia.org>, within Data Space: live.dbpedia.org

Copernicus is a prominent lunar impact crater named after the astronomer Nicolaus Copernicus, located in eastern Oceanus Procellarum. It is estimated to be about 800 million years old, and typifies craters that formed during the Copernican period in that it has a prominent ray system.

Property	Value
<code>dbpedia-owl:LunarCraterDiameter</code>	• 93.0
<code>dbpedia-owl:abstract</code>	• Copernicus is a prominent lunar impact crater named after the astronomer Nicolaus Copernicus, located in eastern Oceanus Procellarum. It is estimated to be about 800 million years old, and typifies craters that formed during the Copernican period in that it has a prominent ray system.
<code>dbpedia-owl:depth</code>	• 3000.000000 (xsd:double)
<code>dbpedia-owl:diameter</code>	• 93000.000000 (xsd:double)
<code>dbpedia-owl:wikiPageExternalLink</code>	• http://www.spod.org/?m=20081014 • http://history.nasa.gov/SP-168/sect02b.htm#116 • http://history.nasa.gov/SP-168sp168.htm • Lunar ray crater Copernicus from Apollo 12, NASA photo.
<code>dbpprop:caption</code>	• 20 (xsd:integer)
<code>dbpprop:depth</code>	• 3.8
<code>dbpprop:diameter</code>	• 93.0
<code>dbpprop:eclw</code>	• W
<code>dbpprop:eponym</code>	• <code>dbpedia:Nicolaus_Copernicus</code>
<code>dbpprop:latitude</code>	• 9 (xsd:integer)
<code>dbpprop:longitude</code>	• 20 (xsd:integer)
<code>dbpprop:ncrs</code>	• N
<code>dbpprop:wikiPageUsesTemplate</code>	• <code>dbpedia:Template:Lunar_crater_data</code>
<code>oddmns:subject</code>	• <code>category:Impact_craters_on_the_Moon</code>
<code>geonss:point</code>	• 9.7,20.0
<code>rdf:type</code>	• <code>owl:Thing</code> • <code>owl:Feature</code> • <code>dbpedia-owl:Place</code> • <code>dbpedia-owl:LunarCrater</code> • <code>dbpedia-owl:NaturalPlace</code>
<code>rdfs:comment</code>	• Copernicus is a prominent lunar impact crater named after the astronomer Nicolaus Copernicus, located in eastern Oceanus Procellarum. It is estimated to be about 800 million years old, and typifies craters that formed during the Copernican period in that it has a prominent ray system.
<code>rdfs:label</code>	• Copernicus (lunar crater)
<code>geo:lat</code>	• 9.700000 (xsd:float)
<code>geo:long</code>	• 20.000000 (xsd:float)

Fig. 2 DBpedia data about the Copernicus crater.

Figure 2 shows DBpedia data concerning a lunar crater named after Copernicus. As one can see at the bottom, `geo:lat` and `geo:long` are used to represent the centroid of the crater. However, W3C Basic Geo uses WGS84 as a reference datum. Thus, and in contrast to the original Wikipedia data, the information

that the crater is not on Earth and that the coordinates use a different, selenographic reference system were lost in the triplification process. Consequently, and as depicted in Figure 3, systems such as the Fluidops Information Workbench render the crater on the Earth's surface near the city of Sarh, Chad. The same is true for the landing site of Apollo 11 – Tranquility Base – located in the Mare Tranquillitatis. In fact, the same problem occurs for all other locations on distant planets and their moons. Showcasing one consequence of such errors, the current DBpedia version (2015-04) indeed shows that the *moon landing* happened here on Earth, as is evident by the following SPARQL query which returns geographic coordinates in the southern part of Algeria.

```
SELECT ?lat ?long
WHERE { dbp:Tranquility_Base geo:lat ?lat ; geo:long ?long. }
```

lat	long
0.6875	23.4333
0.713889	23.4333
0.6875	23.7078
0.713889	23.7078

Listing 3.1 Query and results showing the location of the moon landing is in Algeria.

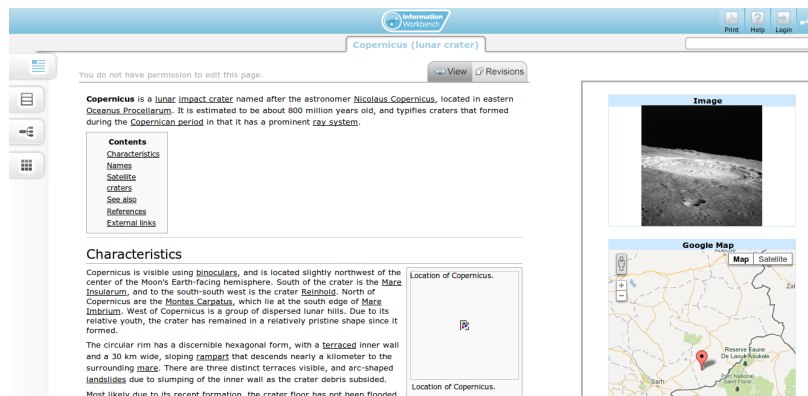


Fig. 3 Fluidops displays Linked Data about the Copernicus crater taken from DBpedia.

Three underlying issues contribute to the outlined problems. First, there is an ongoing debate on how to simplify data publishing on the Web and part of this discussion is about how to avoid burdening publishers through enforcing complex vocabularies and schema. However, the degree to which simplification results in oversimplification is largely context-dependent and while current proposals argue for not enforcing spatial reference system identifiers (SRID), the example above illustrated potential consequences. The counterargument made by the Web community is that for the majority of data published on the Web (that has some sort of geographic identifier), simple WGS84 point coordinates are indeed appropriate. The second issue is the lack of a clear *best practice* for publishing geo-data on the Linked Data cloud. While GeoSPARQL [12] is slowly gaining traction, there are various competing or complementary approaches such as the W3C Basic Geo vocabulary or SPARQL-ST [13] which can also handle

spatiotemporal data. The third issue lies in the nature of most vocabularies and ontologies themselves as well as a lack of domain expertise. Ontologies cannot fix meaning but only restrict the interpretation of domain terminology towards their intended meaning [10]. Consequently, while the W3C Basic Geo specs identify WGS84 as the reference coordinate system, this is not enforced through the axiomatization, and, thus, there is no way of preventing *geo:lat* and *geo:long* from being used to represent locations on celestial bodies other than the Earth. Finally, as discussed previously, most Linked Data today are created by data enthusiasts from existing data. This typically leads to lost expertise. We expect this problem to disappear with time as more domain experts adopt a Linked Data driven approach to publishing their (scientific) data.

The moon landing error mentioned above arose from using the wrong ontology to annotate data. There are also more subtle cases, however, with more dramatic consequences that arise from a lack of domain knowledge or an unclear scope. Consider, for example, the Gulf of Guinea which is one of the world's key oil exploration regions, recently gaining notoriety through frequent pirate attacks. Today's semantic search engines such as *Google's knowledge graph* or knowledge engines such as *Wolfram Alpha* can answer basic questions about countries bordering the Gulf of Guinea. For instance, both systems can handle a query such as '*What is the population of Nigeria?*'. However, no system can answer a query such as '*What is the total population of all countries bordering the Gulf of Guinea?*' or '*What are the major cities in this region ordered by population?*'. In principle, however, and leaving the natural language processing and comprehension of the underlying topological relations aside, such queries can be easily answered using SPARQL and Linked Data. To do so, one could, for instance, select a reference point in the gulf and use a buffer to query for all populated places and their population. Using PROTON's *populationCount* relation the query could be formulated as shown by the fragment in Listing 3.2.

```
SELECT (sum(?populationCount) as ?totalPopulation)
WHERE {
  [...] geo:lat ?lat ; geo:long ?long .
  ?place omgeo:nearby(?lat ?long "500mi");
  ptop:populationCount ?populationCount.}
  [...]
```

Listing 3.2 Fragment of a query for the total population of places within a radius of 500 miles around a location in the Gulf of Guinea.

This query, however, will return the population of cities, towns, countries, and so forth, and, thus, will not give a truthful estimate of the population (as citizens of a country and its cities will be counted multiple times). We will revisit the case of towns and cities later and for now will consider all types of geographic features that have a population value, e.g., to rank places by population. The Gulf of Guinea is also home to the intersection of the Equator with the Prime Meridian. Interestingly, and as shown by the results of Listing 3.3, this has surprising implications for the query discussed before. In GeoNames, the Earth, as such, is located in its own reference system at (0,0) together with the statement that its population is 6,814,400,000 and its feature type is *L parks,area*; see Figure 4. Hence, it is the most populated geographic feature in the Gulf of Guinea and thus

causes the gulf to have the world's highest population density. Moreover, these kinds of errors will propagate, e.g., via GeoNames' RDF *nearby* functionality. For instance, we can learn that the United States are nearby the Odessa Church.⁸

One could now argue that placing the Earth at (0,0) is an isolated case, and, thus, not a systematic error. However, this is not the case. Many existing mapping services return (0,0) to indicate geocoding failures. In fact, this is so common that the Natural Earth dataset has created a virtual island at the location called *Null Island* to better flag geocoding failures. Consequently, it is not surprising to find many features on the Linked Data cloud located to (0,0). The second problem, namely the population count, is also systematic. The Linked Data cloud is envisioned as a distributed global graph but it is not yet clear which data should be provided by linking to more authoritative sources and which data should be kept locally. Therefore, for instance, The New York Times Linked Data portal returns a population of 86,681 for Santa Barbara without providing detailed metadata, while GeoNames reports 88,410 (together with a change history). In contrast, DBpedia reports a population of 90,385 as well as corrected data for the latest update, namely 2014-01-01.

```
SELECT distinct ?lat ?long ?populationCount
WHERE {
  <http://sws.geonames.org/6295630/> geo:lat ?lat ; geo:long ?long ;
  ptop:populationCount ?populationCount .}
```

```
lat long populationCount
0 0 681440000
```

Listing 3.3 A query for the geographic coordinates of the Earth and its population.

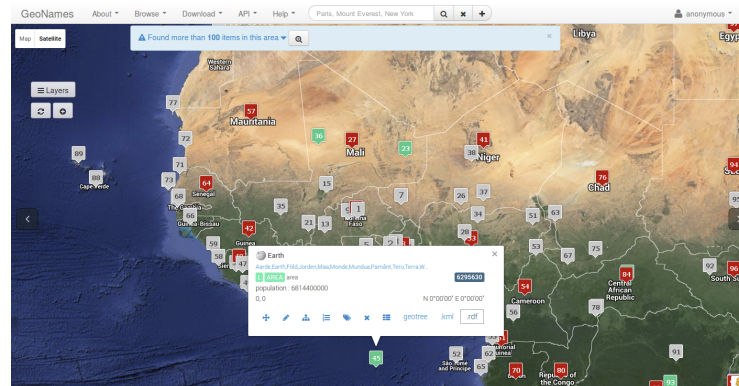


Fig. 4 The point-feature representation of the Earth.

Lessons Learned: Selecting or creating an appropriate ontology to semantically *lift* data is not trivial and the moon landing example shows some of the potential consequences. As most ontologies are lightweight and thus underspecified, it is important to check the documentation and intended use manually.

⁸E.g. via, wget <http://sws.geonames.org/6252001/nearby.rdf>

One proposal is to enforce the explicit specification of coordinate reference systems for all spatial data on the Linked Data cloud. This, however, has been controversially discussed in recent years as it would introduce another hurdle-to-entry for publishers and Web developers. Thus, it has been argued that a layered approach is needed. The second case, namely the population count for the Gulf of Guinea, highlights the need for tighter integration of different data sources based on their scope and authority. Today, a lot of data are published by providers that have limited expertise, cannot provide provenance records, or have no clear maintenance strategy. It is worth noting that the Web (and thus also Linked Data) follows the AAA slogan that **A**n **A**n **A**n about **A**n topic. While this strategy has enabled the Web we know today, it is a blessing and curse at the same time when it comes to scientific data and reliability. Future work will need to go beyond entity resolution (e.g., via *owl:SameAs*) by providing data conflation services (e.g., to merge/correct population data from different sources).

4 Modeling Errors

Another source of error is introduced by various modeling errors such as ontologies being overly simplistic or overly specific as well as errors that result from how data are semantically lifted using these ontologies. Many of these examples are related to how we assign locations to entities. Clearly, entities typed as *place* (and its subtypes) have a direct spatial footprint such as `dbr:Montreal geo:geometry POINT(-73.56 45.5)` even though this footprint may be contested, missing, or unknown, such as for the ancient city of Troy. A similar argument can be made for types that describe spatially fixed entities, e.g., statues. In some rare cases this is also true for otherwise mobile entities such as vessels. A common example for this is the HMS Victory that is located on a dry dock in Portsmouth, England. Wikipedia and thus DBpedia assign geographic coordinates to most places, many statues, and some other entities such as the HMS Victory.⁹ For many other types of entities, however, this is not an appropriate method for assigning locations. For instance, any living human has a (changing) position at any time. This position is not stable and thus not reported in a resource such as Wikipedia (although it may be stored in a trajectory database). In fact, one would be very surprised to find the up-to-date geographic coordinates for a specific person, car, ongoing event, and so forth in the Wikipedia.

From an ontological modeling perspective, one would expect entities of types such as *event* to be related to a place which in turn is related to a spatial footprint. In fact, the notion that events are located spatially via their physical participants and these participants are temporally located via events, is at the core of the DOLCE foundational ontology. One way of thinking about this is to consider the length of the *property path* that is expected between an entity of a given type and geographic coordinates. For example, Rene Descartes is related to Stockholm which has

⁹http://dbpedia.org/resource/HMS_Victory

a spatial footprint: `dbr:Rene.Descartes dbp:deathPlace dbr:Stockholm. dbr:Stockholm geo:geometry POINT(18.07 59.33)`. From this perspective, places are expected to be 0-degree spatial. Persons, events, and so forth, are expected to be 1-degree spatial, and information resources such as academic papers are expected to be 2-degree spatial (via the affiliations of their authors).

Interestingly, performing this experiment on DBpedia yields 1,893 0-degree persons, 371,655 1-degree persons, and 31,182 2-degree persons. Higher degree persons can easily be explained either by a lack of knowledge about their places of birth and death or by the many fictitious persons classified as *Person* in DBpedia. Zero degree persons, however, can be considered modeling errors and will appear in Figure 1. The same argument can be made for the 5,086 0-degree events, 1,507 0-degree sports teams, 448 0-degree biological species, and so forth.

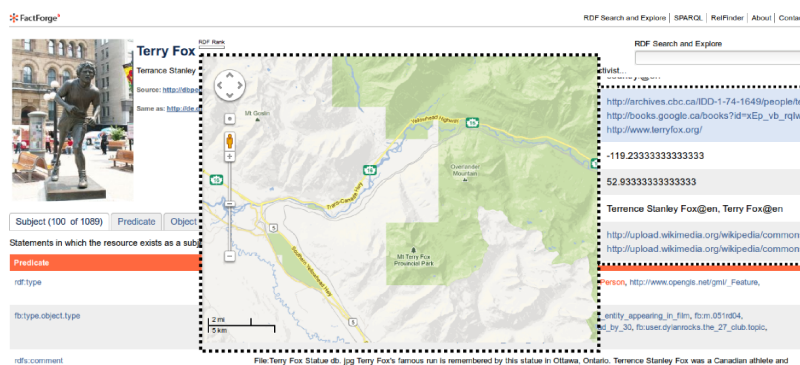


Fig. 5 The spatial footprint of the famous Canadian athlete Terry Fox.

Let us now illustrate the resulting problems using a concrete example. Figure 5 shows a query for *Terry Fox*. As can be seen on the right side of the figure, there are latitude/longitude coordinates assigned to him directly. The image on the left implies that the information about the *person* Terry Fox may have been accidentally conflated with the *statue* of Terry Fox which indeed may have a fixed location. Checking the geographic coordinates, however, reveals that they point to the Mt. Terry Fox Provincial Park (in the middle of Figure 5), thereby clearly revealing the modeling error and its consequences.

A second common example related to modeling is the mis-categorization of geographic features. These errors are difficult to quantify as there is no gold standard that would allow us to measure the *semantic accuracy* of type assignment. Nonetheless, some of the clear, e.g., legally defined, cases are worth discussing. For instance, there are 554 places in DBpedia that are classified as being a town while having a population over 100,000, e.g., Stuttgart, Germany, with a population over 500,000, and 3,694 cities with a population below 1,000 such as Eureka, Utah with a current estimated population of 667. The issue here is that the meanings of *city* and *town* varies greatly across countries and even between US states [7]. In Utah, for instance, every settlement with a population below 1,000 is legally considered a town. Hence, Eureka is a town and not a city. In contrast, the class *town* in Pennsylvania is a sin-

gletton class that contains Bloomsburg as its sole member. Nonetheless we can find triples such as `dbr:Bloomsburg_University_of_Pennsylvania dbp:city dbr:Bloomsburg,_Pennsylvania` in DBpedia. In both cases, the underlying problem is that the ontologies (which are often semi-automatically learned from data) are overly specific and introduce fine grained distinctions that are not supported through the data; see [1] for more details on feature types in DBpedia.

Lessons Learned: While there is sufficient theoretical work on how entities are located in space and time – namely by modeling location as a *relation* between objects and by spatially anchoring events via their physical participants – there seems to be a gap on how to apply these theoretical results to the practice of data publishing. The case of wrong or overly-specific type assignment is even more difficult to tackle as geographic feature types have spatial, temporal, and culturally indexed definitions as shown by the town and city example. Ongoing work investigates the role of spatial statistics for mining type characteristics *bottom-up* and may help to minimize categorization errors in the future [18].

5 Accuracy and Uncertainty Related Errors

DBpedia also stores 133,941 cardinal direction triples such as the statement, Ventura, CA is to the north of Oxnard, CA : `dbr:Ventura,_California dbp:south dbr:Oxnard,_California`.¹⁰ This leads to the interesting question of how accurate these triples are. Testing 100,000 of these triples reveals that 26% (26,420) of them are inaccurate when using the geometries provided by DBpedia. Our sample only includes triples where subject and object are both of type `dbo:Place` and have valid `geo:geometry` predicates. By considering all 133,941 cardinal triples in DBpedia, we find that 55,928 of them have a subject or object lacking `geo:geometry`, or are not of type `dbo:Place`. Of these, 17,957 triples list a cardinal direction relation to a RDF literal such as an `xsd:integer`, e. g., `dbr:Harrisburg,_Pennsylvania dbp:north 20 (xsd:integer)`.

More interesting, however, than discovering these (significant) data errors alone, is the question of how much uncertainty is introduced by using point-features to represent places and how this uncertainty is communicated [4]. Returning to the Ventura and Oxnard example, one can overlay the known administrative areas for both cities with a 1x1 kilometer grid and then pairwise compare all possible grid points. Figure 6 shows the spatial distribution of those grid points and an 1:n step out of this direction comparison. The directionality is determined by testing if the azimuth between two point geometries falls within ω (which is set to $\pi/8$) from the primary angle of the cardinal (N,S,E,W) or the intercardinal direction (NE,SE,SW,NW). For example, SE (*stko:southeast* here) covers the range $5\pi/8$ to $7\pi/8$ which is measured from the positive *y*-axis. Our results show that the cardinal direction S holds for 34.8% of the cases in which Ventura is located to the north of Oxnard, while the intercardinal direction SE holds for 50.5% cases in which Ventura is located to the northwest of Oxnard. In

¹⁰The way in which DBpedia uses cardinal directions can be easily misunderstood. The triple states that the entity south of Ventura is the city of Oxnard.

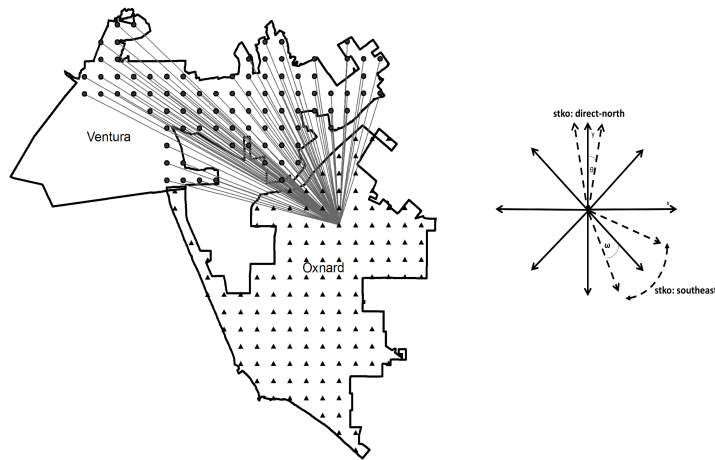


Fig. 6 A 1:n step in the direction computation for 1x1 km grids of Ventura (88 circles) and Oxnard (133 triangles). Grid points in the ocean were removed.

0.5% of the cases the correct direction is NW. This uncertainty (and the fact that SE seems to be the better choice), however, is not communicated by DBpedia.

A similar case, this time based on reporting coordinates and areas beyond a meaningful accuracy, can be found in many other examples. For instance, DBpedia states that the value for `dbo:PopulatedPlace/areaTotal` for Santa Barbara is $108.69662101458125 \text{ km}^2$. The location for Santa Barbara is given by the centroid `POINT(-119.71416473389 34.425834655762)` – thus indicating that the exact centroid of Santa Barbara is known at the sub-micron scale. This is not DBpedia specific and thus a systematic error. Similar cases can be found in the New York Times Linked Data hub that locates Santa Barbara at `geo:lat 34.4208305` and `geo:long -119.6981901`.¹¹ In contrast, the Taxon-Concept dataset uses the uncertainty parameter specified by RFC 5870, e.g., `geo:44.863876, -87.231892;u=10` for a sighting of the Danaus Plexippus butterfly, thereby presenting a possible solution to the problem.

Finally, it is worth noting that the lack of a clear uncertainty framework for Linked Data in general has dramatic consequences beyond location data alone. Listing 5.1, shows a query for regions in California and their population. Summing up the data for the South Coast and Central Coast would not yield a value of approximately 22,250,000 but merely 2,249,558. This surprising behavior is caused by the population of the South Coast being represented as a *string* instead of an *xsd:integer* which cannot be used (and is thus silently disregarded) by the SPARQL summation function.

```
SELECT ?region ?population
WHERE {
  ?region a yago:RegionsOfCalifornia;
    dbp:population ?population .}
```

¹¹<http://data.nytimes.com/N2261955445337191084>

```

region                population
[shortened results]  ...
South.Coast.(California) '~ 20million'@en //Not recognized as a (approximate) number
Central.Coast.(California) 2249558 //recognized as an xsd:integer

```

Listing 5.1 Population of (overlapping) regions in California.

Lessons Learned: The cardinal directions example shows the many and massive errors that exist in spatial information on the Linked Data cloud today. Blaming the datasets and their providers, however, is missing the more relevant and underlying problem – namely the effects of decontextualization on data [8] and their transformation into statements in triple form. Consider the following example: The sentence ‘Isla Vista, CA is the most populated municipality to the west of the Mississippi.’ is meaningful and partially correct. During natural language processing and triplification this sentence would be transferred to a triple such as `ex:Isla_Vista dbr:west ex:Mississippi`. This triple, however, is not only questionable but also leads to exactly those cardinal direction accuracy issues discussed before as the direction will depend on the point coordinates used to represent the Mississippi river. Finally, and as illustrated above, the lack of a general uncertainty framework for Linked Data requires urgent attention in future research.

6 Conclusions

Places and positioning information more broadly play a key role in interlinking data on the Web. Consequently, it is important to study the quality of these (geo-)data. Our work reveals that about 10% of all spatial data on the Linked Data cloud is erroneous to some degree. We identified major types of systematic errors, discussed their likely causes (some of which have been confirmed by the data providers), and pointed out lessons learned and directions for future research. Some of the identified problems can be easily addressed and prevented in the future, e.g., by unit testing against possible representational choices for geographic coordinates. Other cases remain more challenging such as proper ontological modeling or the representation of uncertainty. Those issues for which a clear best practice can be identified and agreed upon are currently being collected by the joint OGC/W3C *Spatial Data on the Web Working Group*.¹² Finding the right balance between simple models and data publishing processes on the one hand and preventing potentially harmful oversimplifications on the other hand remains the major challenge to be addressed in the future.

References

1. Adams, B., Janowicz, K.: Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science* **29**(4), 556–579 (2015)

¹²The views presented in this paper belong to the authors and do not necessarily represent the views or positions of the entire working group. A current draft of the best practice report is available at: <https://www.w3.org/TR/sdw-bp/>.

2. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: A uniform way of publishing other people's dirty data. In: *The Semantic Web - ISWC 2014*, pp. 213–228. Springer (2014)
3. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Welty, C.A.: Building watson: An overview of the deepqa project. *AI magazine* **31**(3), 59–79 (2010)
4. Fisher, P.F.: Models of uncertainty in spatial data. *Geographical information systems* **1**, 191–205 (1999)
5. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010* (2010)
6. Hogan, A., Hitzler, P., Janowicz, K.: Linked dataset description papers at the semantic web journal: A critical assessment. *Semantic Web* **7**(2), 105–116. (2016)
7. Janowicz, K.: Observation-Driven Geo-Ontology Engineering. *Transactions in GIS* **16**(3), 351–374 (2012)
8. Janowicz, K., Hitzler, P.: The digital earth as knowledge engine. *Semantic Web* **3**(3), 213–221 (2012)
9. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 747–758. *International World Wide Web Conferences Steering* (2014)
10. Kuhn, W., Kauppinen, T., Janowicz, K.: Linked data-a paradigm shift for geographic information science. In: *Geographic Information Science*, pp. 173–186. Springer (2014)
11. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* **6**(2), 167–195 (2015)
12. Perry, M., Herring, J.: Ogc geosparql-a geographic query language for rdf data. *Open Geospatial Consortium* (2012)
13. Perry, M., Jain, P., Sheth, A.P.: SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries. In: *Geospatial Semantics and the Semantic Web - Foundations, Algorithms, and Applications*, pp. 61–86. Springer (2011)
14. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pp. 114–129. Springer (2012)
15. Rietveld, L., Verborgh, R., Beek, W., Vander Sande, M., Schlobach, S.: Linked Data-as-a-Service: The Semantic Web Redeployed. In: *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia.*, pp. 471–487. Springer (2015)
16. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open phacts: semantic interoperability for drug discovery. *Drug Discovery Today* **17**(21), 1188–1198 (2012)
17. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. *Semantic Web* **7**(1), 63–93 (2015)
18. Zhu, R., Hu, Y., Janowicz, K., McKenzie, G.: Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS* (2016;forthcoming)