Linkage Quality Analysis of GeoNames in the Semantic Web

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@ntnu.no

ABSTRACT

We examine the GeoNames gazetteer as a hub of Linked Geospatial Data. We survey quality and linkage characteristics to understand how well it supports the traversal of the Semantic Web and which limitations exist. We examine different traversal scenarios originating with GeoNames and present findings related to link availability and distribution along language and geospatial dimensions and discuss the role of cross-lingual issues.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; *Web mining*; *Semantic web description languages*;

KEYWORDS

Location-Aware Information Access; Geospatial Semantic Web; Linked Open Data; Gazetteer

ACM Reference Format:
Dirk Ahlers. 2017. Linkage Quality Analysis of GeoNames in the Semantic Web. In *GIR'17: 11th Workshop on Geographic Information Retrieval, November 30-December 1, 2017, Heidelberg, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3155902.3155904

1 INTRODUCTION

Gazetteers are the basis for most geospatial knowledge and are widely used in information retrieval to provide ground truth about places in the real world. They contain geographical features as well as cities and other populated places, along with coordinates and feature metadata. The most widely used freely available gazetteer is GeoNames.org, which has a worldwide coverage[1]. It is also the largest contributor to geospatial Linked Open Data (LOD) and is intensely crosslinked with DBpedia. This makes it an important hub to traverse geospatial data within the Semantic Web based on its roughly 11.5 million names. GeoNames provides its data as RDF[2] and provides linking and spatial search capabilities around existing places. While it has no own SPARQL endpoint, and thus no support for direct RDF query, it can be traversed as Linked Data from a given entry entity.

[1] http://www.geonames.org/
[2] GeoNames Ontology: http://www.geonames.org/ontology/

GIR'17, November 30-December 1, 2017, Heidelberg, Germany

ACM ISBN 978-1-4503-5338-0/17/11…$15.00
https://doi.org/10.1145/3155902.3155904

We examine GeoNames as part of the Semantic Web and examine linkage, idiosyncrasies, and potential limitations. After previous work looked at intrinsic geographic quality indicators for GeoNames [1], or at inconsistencies within Semantic Web resources [3], we are interested in its traversal as part of the Semantic Web to maintain broad coverage and linkage. We explore quality issues for traversing outlinks from GeoNames by *Linking GeoNames to Wikipedia*, *Linking GeoNames to DBpedia*, and *Linking GeoNames to other resources*.

2 ANALYSIS

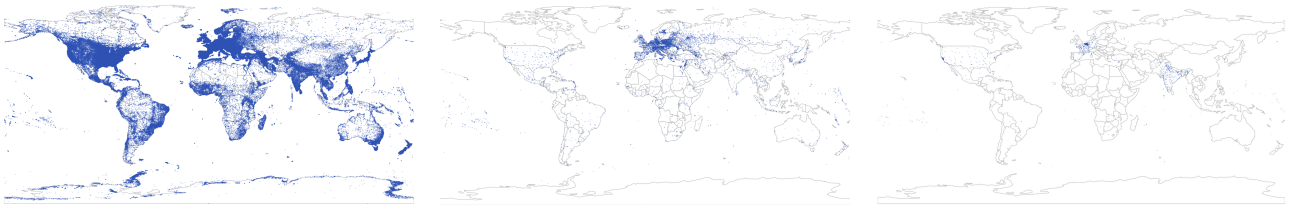
GeoNames provides different traversal mechanisms depending on the type of entity. Within its data it allows to traverse its hierarchy and perform radius searches for entities. For external linkage, *gn:wikipediaArticle* contains a link to Wikipedia and *rdfs:seeAlso* a link to DBpedia. Links from Wikipedia/DBpedia back to GeoNames are partially available in DBpedia, which has generated owl:sameAs links; there are usually no direct links from Wikipedia. These links can be discovered coming from GeoNames, the reverse does not always work. We leave the inlink examination for future work.

We find that about 5% of all GeoNames entities have a link; this is higher at about 13% for *populated places*. We only take this latter sample in this paper. First, we find an inconsistency between the full database dump[3] and the RDF endpoints regarding level of detail of non-semantic links: The RDF only contains links to Wikipedia; the database also contains a low number of (general) Web links. In many cases, these are the home pages of entities. For example, we find 590 links to .be domains within Belgium, linking to municipal sites. There is an additional non-explicit semantic linkage of 669 links to the Library of Congress Linked Data Service at id.loc.gov (cf. Table 2). In most cases, these co-occur with a Wikipedia link.

The linkage from GeoNames' RDF to DBpedia is straightforward. However, as DBpedia links are apparently derived inside the GeoNames "black box" only from the English Wikipedia links, other language versions than en.wikipedia.org are not available as RDF.

English is the dominating Wikipedia language. However, there are articles that only exist in language versions other than English (cf. [2, 4]). These are currently not findable by RDF traversal from GeoNames. Only in a few cases, GeoNames does link to a local Wikipedia language version even though an English article exists. In these instances, there is no direct link from a GeoNames entity to the DBpedia language version with the existing data. Then, it could be useful to try to construct the DBpedia URI oneself from the Wikipedia link. In the default case with an initial DBpedia reference, DBpedia can easily be traversed along its language versions. For example, we can derive a simple solution for semantic traversal that would, for a link to a non-english DBpedia version, check the

[3] http://download.geonames.org/export/dump/



(a) …with a link to the English Wikipedia (b) …with a link to a non-English Wikipedia (c) …with a link to non-Wikipedia sites

Figure 1: Mapping of GeoNames entities …

DBpedia property *owl:sameAs* for a link to the English version.

Normally, there is only one link to Wikipedia. In very few cases do we find multiple articles linked. For example, the Iraq GeoNames entity links to both Iraq and Mesopotamia in Wikipedia (and thus to DBpedia). In about 2% of entities with Wikipedia links, we find links to multiple Wikipedia language versions, which in most cases include an English link. We rarely find more than 2 links, so there is no direct exploration of the language versions.

For a larger perspective, we examine the distribution of Wikipedia outlink languages. The worldwide geographic distribution of entities is mapped in Fig. 1a for English links and Fig. 1b for all remaining languages. The former follows well-known expected density distributions of population and media use [6]. The latter is clustered interestingly in Europe and less dense in the rest of the world. This may hint at more local users adding links to GeoNames, but would need additional analyses to clarify underlying effects.

Also outlinks to non-English languages are heavily biased as shown in Table 1. Russian leads with 10042 links, followed by German with only 500. For all links to the Russian version, we find 8464 of them to be part of pairs, mostly with an English link, but 1578 entities with a Russian-only Wikipedia link.

Looking at cross-language distribution, we see for example around 2200 non-English links in Germany. Yet there are only 500 links to de.wikipedia.org in the whole dataset. The alternative languages are thus clearly not confined to their own spoken areas.

To follow up on the non-Wikipedia links, we show their classification by top level domain in Table 2 and map them in Fig. 1c. Again, there is no simply relation of country code to language. It shows a low overall number of non-Wikipedia links, a high number of Library of Congress links and Belgian municipalities, and other links distributed throughout the world, with higher density in the US, Europe, Afghanistan, and India.

Table 1: Frequency of top-10 outlink Wikipedia Languages

Language	Frequency
en	461728
ru	10042
de	500
fr	208
sv	173
es	139
lt	133
it	119
el	117
nl	96

Table 2: Frequency of top-10 outlink ccTLDs of non-Wikipedia links

TLD	Frequency
gov[4]	671
be	590
com	132
org	92
fr	33
de	29
pt	13
net	12
se	9
it	8

3 CONCLUSION AND FUTURE WORK

Our findings show certain issues with the GeoNames Semantic Web integration, some of which could be fixed. We uncovered interesting cross-dataset and cross-lingual issues and distribution biases and found additional linkage information not available in the RDF. It can be an important measure to understand how much information is currently hidden because it cannot be reached by *sameAs* and other semantic link traversal.

Previous work identified inconsistencies of properties and values between different language versions of DBpedia [3]. A bigger question would be about differences between language versions as a whole. There are open issues regarding link discovery mechanisms in matching GeoNames to other sources [2, 5]. Often Wikipedia URIs are used for comparison in entity matching, which we found to be partly unreliable.

Judging from our analyses, linking from GeoNames to the Semantic Web causes issues mainly to do with cross-lingual linking. The issues mostly manifest if there is no English version which can be used as a first traversal step as then subsequent exploration fails. Since DBpedia names are predictably constructible from Wikipedia titles this might be a good step to increase coverage. The listed issues can be useful in understanding limitations of the dataset and adapt integration accordingly. Our future work will concern in-link analysis from DBpedia, improved quality and quantitative assessment as well work towards automatic corrections and analysis and comparison of other sources such as TGN and local gazetteers.

REFERENCES

[1] Dirk Ahlers. 2013. Assessment of the Accuracy of GeoNames Gazetteer Data. In *7th Workshop on Geographic Information Retrieval (GIR '13)*. ACM.

[2] Dirk Ahlers. 2013. Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles. In *ECIR2013 – 35th European Conference on Information Retrieval*. Short paper.

[3] Elena Cabrio, Serena Villata, and Fabien Gandon. 2014. Classifying Inconsistencies in DBpedia Language Specific Chapters. In *International Conference on Language Resources and Evaluation LREC*.

[4] Mark Graham and Stefano De Sabbata. 2015. Mapping Information Wealth and Poverty: The Geography of Gazetteers. *Environment and Planning A* 47, 6 (2015), 1254–1264.

[5] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2012. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal* 194 (2012), 28–61.

[6] Vanessa Murdock. 2011. Your mileage may vary: on the limits of social media. *SIGSPATIAL Special* 3, 2 (2011), 62–66.

[4] including 669 links to id.loc.gov