Collaborative Project

GeoKnow - Making the Web an Exploratory for Geospatial Knowledge

Project Number: 318159 **Start Date of Project:** 2012/12/01 **Duration:** 36 months

Deliverable 3.4.2
Comparison with Other Data Sets

Dissemination Level	Public
Due Date of Deliverable	Month 34, 30/09/2015
Actual Submission Date	Month 36, 17/11/2015
Work Package	WP3, Spatial Knowledge Aggregation, Fusing and Quality Assessment
Task	T3.4
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	16
Filename	D3.4.2_Comparison_with other_ data_sets _Assessment.pdf

Abstract: This deliverable demonstrates how the developed quality measures are able to assess the quality of community-created geo-spatial datasets. This will allow for estimations on the progress of user mapping efforts.

Project funded by the European Commission within the Seventh Framework Programme (2007 - 2013)

History

Version	Date	Reason	Revised by
0.0	19/10/2015	Initialise the deliverable	Mohamed Ahmed Sherif
0.1	25/10/2015	Initial plots added	Muhammad Saleem
0.2	29/10/2015	Plots updated	Muhammad Saleem
0.3	02/11/2015	First deliverable draft	Mohamed Ahmed Sherif
0.4	04/11/2015	Second deliverable draft	Mohamed Ahmed Sherif
0.5	04/11/2015	Complete deliverable	Mohamed Ahmed Sherif
0.6	09/11/2015	Deliverable proofread	Muhammad Saleem
0.7	10/11/2015	Deliverable review	Axel-Cyrille Ngonga Ngomo
0.8	11/11/2015	Deliverable revise	Mohamed Ahmed Sherif
0.9	16/11/2015	Internal review	Giorgos Giannopoulos
1.0	17/11/2015	Final version	Mohamed Ahmed Sherif

Author List

Organization	Name	Contact Information
INFAI	Mohamed Ahmad Sherif	sherif@informatik.uni-leipzig.de
INFAI	Muhammad Saleem	saleem@informatik.uni-leipzig.de
INFAI	Axel-Cyrille Ngonga Ngomo	ngonga@informatik.uni-leipzig.de
INFAI	Giorgos Giannopoulos	giann@imis.athena-innovation.gr

Executive Summary

With the increasing availability of community-generated geo-spatial data sets comes the challenge of assessing the quality of such data. Community-provided data sets are especially prone to errors due to multiple autonomous data providers that use different assumptions about the structure and the semantics of data and the lack of uniform quality standards. This deliverable demonstrates how to assess the quality of various datasets using a set of geo-spatial data quality measures. We present the evaluation of 10 geo-spatial quality measures including datasets coverage, surface area and structuredness. We demonstrate the usefulness of our approach by using it on on three well-known Linked geo-spatial datasets – LinkedGeoData, NUTS, GeoLinkedData – and discuss the results obtained. Our results show that community-generated data sets differ clearly from manually curated data with respect to the proposed data quality metrics.

Contents

List of Figures

1 Introduction

The Linked Open Data (LOD) cloud hosts 9960[1] publicly available knowledge bases with many of them containing geo-spatial annotations. However, community-provided geo-spatial data sets can surpass the level of detail found in official and commercial maps. Assessing the quality of community provided data sets against reference data set(s) is one of the main challenges in the GeoKnow project. With GeoKnow aiming to address an extensive range of users including customers in an industrial environment, the quality of data sets resulting from different providers becomes a crucial factor for the acceptance and distribution of the project's results. With the constant growth of the community provided geo-spacial Linked Data sources over the last years comes with the need to develop measures for evaluating the quality of the existing data.

In this deliverable, we present and evaluate the measures for assessing the quality of user-contributed geo-spatial information. The goal is to compare different maps and different regions of a map represented in the RDF data model. To this end, we developed and evaluated measures for qualifying each data set pertaining to various aspects such as coverage, surface area and structuredness.

In the rest of this deliverable, we first introduce the selected geo-spatial data sets in Section 2. Then, in Section 3 we describe the different geo-spatial quality measures we used to perform our evaluation presented in Section 4. Finally, we conclude our work and propose some future work in Section 5.

[1] http://stats.lod2.eu/

2 Data Sets

We used three publicly available data sets for our experiments. The first data set, the *Nomenclature of Territorial Units for Statistics* or simply *NUTS*[2] was used as the core reference data set for our experiments. The NUTS data set is manually curated by http://statistics.data.gov.uk, where regions are described along with their temporal validity. We chose this data set because it contains fine-granular descriptions of 1,461 geo-spatial resources located in Europe. For example, Norway is described by 1981 points.

The second data set, *LinkedGeoData* or *LGD* for short, contains all the 332,494,353 geo-spatial resources from http://linkgeodata.org. LGD is an RDF conversion of the community provided information collected by the *OpenStreetMap* project[3]. For more details about LGD see [1, 4]. Finally, the third data set is *GeoLinkedData* or simple *GLD* from the open initiative of the Ontology Engineering Group (OEG). OEG collected GLD from diverse information sources belonging to the National Geographic Institute of Spain and made it available as RDF[4]. In our experiments, we used the all the 21,564,199 geo-spatial resources from GLD. For more details about GLD see [5]. In this deliverable, we compare the manually curated data set of NUTS against the two other community provided data sets of LGD and GLD for different quality measures.

[2] Version 0.91 available at http://nuts.geovocab.org/data/ is used in this work
[3] http://www.openstreetmap.org/
[4] http://geo.linkeddata.es

3 Measures

In this section, we describe the adapted measures that we used to qualify our aforementioned geo-spatial datasets. For a fully detailed description of the measures, please refer to GeoKnow's deliverable D3.5.2[5].

3.1 Properties per Class

This measure calculates how many distinct properties/predicates exist for each instance of a class. For each of the input data set classes, this measure counts the number of distinct predicates that are used in statements where the subject is an instance of that class.

3.2 Instances per Class

This measure calculates how many distinct instances exist for each class. This measure can be used to weigh the importance of a class in a data set.

3.3 Average Surface Area per Class

This measure calculates the average surface contained in polygons for each class. This measure is important to relativise the number of instances of some class. A class representing continents has only a few instances but the covered surface is much bigger than that of a class representing a city.

3.4 Number of Intersecting Classes per Instance

In this measure, for each instance of a given class, the measure calculates how many types this instance has. This is important to represent how specific the current class is. In a very specific class, outliers are more significant than in a general class.

3.5 Average Number of Points per Class

This measure represents the average of points per class. For each instance of the current class, the measure computes how many points are linked from this instance. This measure is important to differentiate between classes representing multi-point objects and those representing one point.

3.6 Average number of Polygons per Class

This measure reports the average number of polygons within instances of a given class.

3.7 Average Distance Between Point Sets which Represent the Same Resource

In this measure, we compute the average distance between polygons which represent the same resource in two linked data sets. This measure takes as input a source data set S, a target data set T and a set of point set distance functions D.

[5]Deliverable 3.5.2: Final report on spatial data quality assessment

We base the current work on an updated version of the set of point set distance functions first introduced in D3.4.1[6]. We will write $R = (p_1, \ldots, p_n)$ to denote that the vector description of the resource R comprises the points $p_1, \ldots, p_n$. A point p_i on the surface of the planet is fully described by two values: its latitude $lat(p_i) = \varphi_i$ and its longitude $lon(p_i) = \lambda_i$. We will denote points p_i as pairs (φ_i, λ_i). Then, the distance between two points p_1 and p_2 can be computed by using the *great elliptic arc distance* [2] dubbed as $\delta(p_1, p_2)$. Computing the distance between sets of points is yet a more difficult endeavour. Over the last years, several measures have been developed to achieve this task. Most of these approaches regard vector descriptions as ordered sets of points. The input for the distances consists of two point sets $s = (s_1, \ldots, s_n)$ and $t = (t_1, \ldots, t_m)$, where n resp. m stands for the number of distinct points in the description of s resp. t. W.l.o.g, we assume $n \geq m$. Here, we used 9 measures:

1. Mean measure
$$d_{mean}(s,t) = \delta\left(\frac{\sum_{s_i \in S} s_i}{n}, \frac{\sum_{t_j \in T} t_j}{m}\right). \tag{1}$$
2. Max measure
$$d_{max}(s,t) = \max_{s_i \in s, t_j \in t} \delta(s_i, t_j). \tag{2}$$
3. Min measure
$$d_{min}(s,t) = \min_{s_i \in s, t_j \in t} \delta(s_i, t_j). \tag{3}$$
4. Average measure
$$d_{average}(s,t) = \frac{1}{nm} \sum_{s_i \in S, t_j \in t} \delta(s_i, t_j). \tag{4}$$
5. Sum of Minimums measure (SOM)
$$d_{som}(s,t) = \frac{1}{2}\Big(\sum_{s_i \in s} \min_{t_j \in t} \delta(s_i, t_j) + \sum_{t_i \in t} \min_{s_j \in s} \delta(t_i, s_j)\Big). \tag{5}$$
6. Surjection measure
$$d_{surjection}(s,t) = \min_{\eta} \sum_{(e1,e2) \in \eta} \delta(e_1, e_2), \tag{6}$$
where η is the surjection from the larger of the point sets S and T to the smaller.
7. Fair Surjection measure (FS)
$$d_{fs}(s,t) = \min_{\eta'} \sum_{(e1,e2) \in \eta'} \delta(e_1, e_2), \tag{7}$$
where η' is the evenly mapped surjection from the larger of the sets s and t to the smaller.
8. Link measure
$$d_{link}(s,t) = \min_{R} \sum_{(s_i,t_j) \in R} \delta(s_i, t_j), \tag{8}$$
where minimum is computed from all relations R, where R is a linking between s and t satisfying the previous two conditions.

[6] Deliverable 3.4.1: Measures for Linked Geospatial Information

9. Hausdorff measure

$$d_{hausdorff}(s,t) = \max_{s_i \in S} \left\{ \min_{t_j \in T} \left\{ \delta(s_i, t_j) \right\} \right\}. \tag{9}$$

Given a source data set S, we start the measure computation procedure by querying S for the set of all classes C within S. For each source class c in C, we select the set of source class instances s for which the following apply: (1) they have a vector geometry; (2) they are linked with a target data set instance t in T; and (3) t also has a vector geometry representation. Afterwords, each of the aforementioned distance functions is applied against each pair of linked source and target instances, formally $d(s,t)$. Finally, for each d in D, we compute $averagePS(d)$ as the average all the computed point sets measures for each class, formally

$$averagePS(d) = \sum_{\forall s \in c} \frac{d(s,t)}{|s|}. \tag{10}$$

3.8 Class Coverage

This measure was introduced in [3]. the measure determines how well the instance data conform to `rdf:class` (class for short), i.e., how well a specific class is covered by the different instances of that class. The coverage of a class C, denoted by $Coverage(C)$, is defined as follow:

Definition 3.1 (Class Coverage)**.** For a data set D, let $P(C)$ denote the set of distinct properties having class C and $I(C)$ denote the set of distinct instances having class C. Let $I(p,C)$ denote the number of distinct instances having predicate p and class C. Then, the coverage of the class $CV(C)$ is

$$CV(C) = \frac{\sum_{\forall p \in P(C)} I(p,C)}{|P(C)| \times |I(C)|}$$

3.9 Weighted Class Coverage

Definition 3.1 considers the structuredness of a data set with respect to a single class. Obviously, a data set D has instances from multiple classes, with each instance belonging to at least one of these classes (if multiple instantiation is supported). It is possible that D might have a high structuredness for a class C, say $CV(C) = 0.8$, and a low structuredness for another class C', say $CV(C') = 0.15$. But then, what is the structuredness of the whole data set with respect to our class system (set of all classes)? Duan et al. [3] proposed a mechanism to compute this, by considering the weighted sum of the coverage $CV(C)$ of individual classes. In particular, for each class C, the weighted coverage is defined below.

Definition 3.2 (Weighted Class Coverage)**.** By using Definition 3.1, the weighted coverage for a class C denoted by $WT(CV(C))$ is calculated using the following formula:

$$WT(CV(C)) = \frac{|P(C)|+|I(C)|}{\sum_{\forall C' \in D} |P(C')|+|I(C')|}$$

3.10 Data set Structuredness

By using Definitions 3.1, 3.2, we are now ready to compute the structuredness, hereafter termed as coherence, of a whole data set D.

Definition 3.3 (Data Set Structuredness)**.** The over all structuredness or coherence of a data set D denoted by $CH(D)$ is defined as

$$CH(D)) = \sum_{\forall C \in D} CV(C) \times WT(CV(C))$$

4 Evaluation

We evaluated the measures described in Section 3 using the three data sets introduced in Section 2. We applied each measure against each data set and presented here an over all average, as well as standard deviation for each measure for each data set. All experiments were carried out on a 32-core 2.3 GHz server running *OpenJDK* 64-Bit Server 1.7.0_75 on *Ubuntu* 14.04.2 LTS. Each experiment was assigned 20 GB RAM.

4.1 Evaluating the Properties per Class Measure

Here, we evaluate the *properties per class* measure introduced in Section 3.1. As shown in Figure 1, GLD has (in average) almost the same number of properties per class as well as the same standard deviation as NUTS (the reference data set). On the other hand, in average, LGD have a higher number of properties per class than the other data sets, also with higher standard deviation.

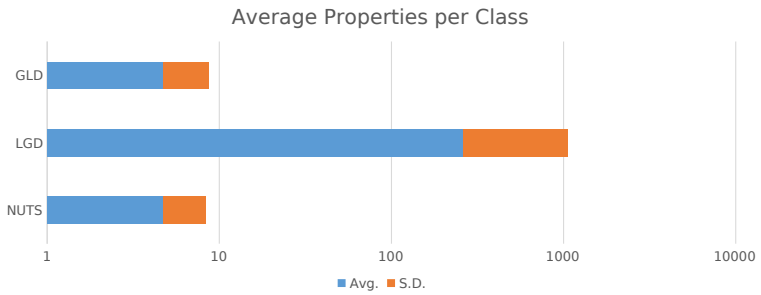

Figure 1: Comparison of the evaluation data sets on average properties per class over all classes.

4.2 Evaluating the Instances per Class Measure

Figure 2 shows the evaluation of the *instance per class measure* described in Section 3.2. The figure shows the average number of instances across all classes in each of the aforementioned data sets together with the relative standard deviations. LGD has higher number of instances per class over the other data sets. Yet, another observation is that the variation of number of instances per class is higher in the two test data sets of GLD and LGD over the reference of NUTS.

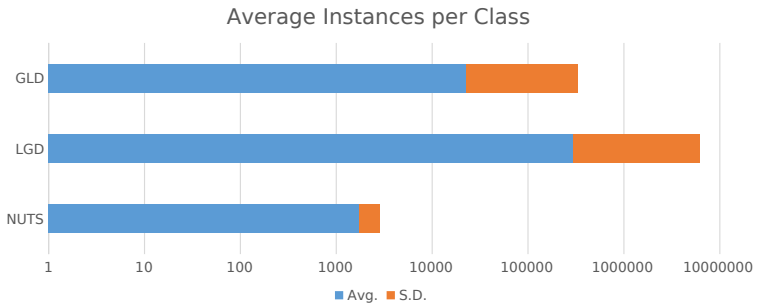

Figure 2: Comparison of the evaluation data sets on average instances per class over all classes.

4.3 Evaluating the Average Surface Area per Class Measure

Figure 3 shows the evaluation of the *average surface area per class* measure introduced in Section 3.3. The figure represents the average and the standard deviation of applying the measure against all classes in each data set. The zero area result for GLD and GLD come from the fact that each of the data sets represent each geo-spactial resource with at most one geo-location which simply have no surface area. For NUTS, one interesting observation is the big variance in the surface areas among different resources.

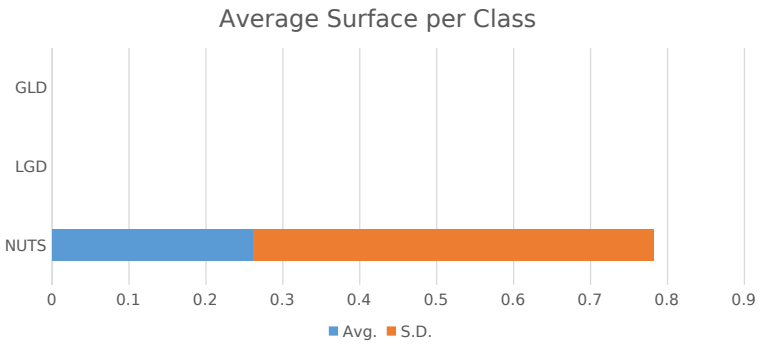

Figure 3: Comparison of the evaluation data sets on average surface per class over all classes.

4.4 Evaluating the Number of Intersecting Classes per Instance

Figure 4 shows the evaluation of the *average of intersecting classes per instance* measure introduced in Section 3.4. The zero result of the NUTS comes from the very specific classes of NUTS (i.e. each geo-spatioal resource belongs to at most one class). Moreover, LGD has less specific classes than GLD.

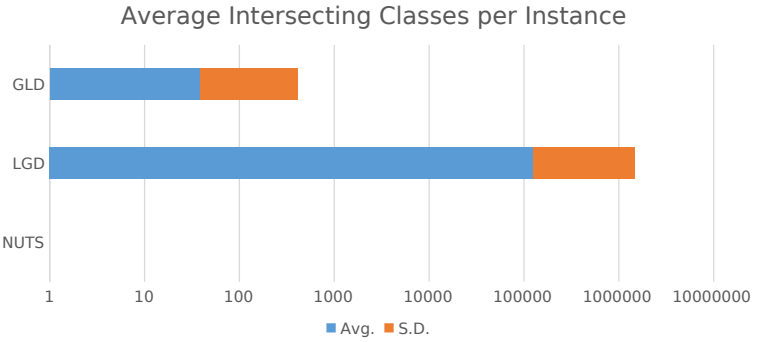

Figure 4: Comparison of the evaluation data sets on average external instances per class over all classes.

4.5 Evaluating the Average Number of Points per Class Measure

Figure 5 shows the evaluation of the *average number of points per classes* measure introduced in Section 3.5. The figure shows the average and the standard deviation of applying the measure against all classes in each evaluation data set. The reference data set (NUTS) has higher average of points per classes over the other data sets, which indicates NUTS to be more fine grained .

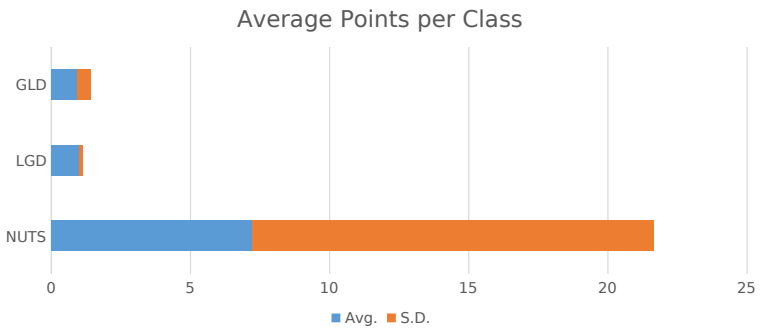

Figure 5: Comparison of the evaluation data sets on average points per class over all classes.

4.6 Evaluating the Average number of Polygons per Class Measure

Figure 6 shows the evaluation of the *average number of polygons per classes* measure introduced in Section 3.6. The figure represents the average and the standard deviation of applying the measure against all classes in each of the evaluation data set. The zero polygon result for GLD and GLD comes from the fact that each geo-spatial resource in GLD and LGD is represented with at most one geo-location in the from of one point not a polygon. On the other side, NUTS has in average 1.3 polygons per class.

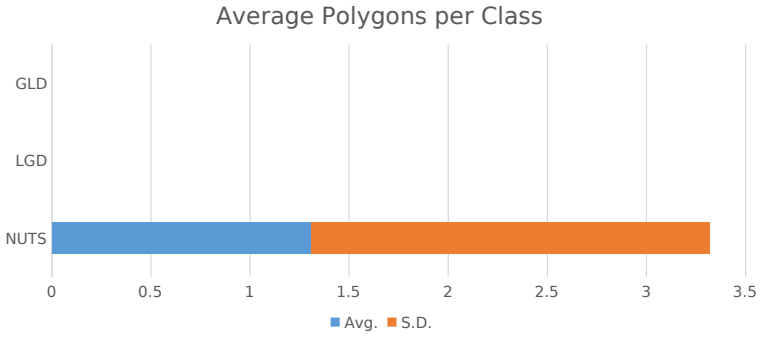

Figure 6: Comparison of the evaluation data sets on average polygons per class over all classes.

4.7 Evaluating the Average Distance Between Point Sets which Represent the Same Resource Measure

Figure 7 shows the evaluation of the *average distance between point sets which represent the same resource* measure introduced in Section 3.7, note that the figure shows the average and standard deviation of applying the nine aforementioned point set measure over all classes for each data set. For our experiments we used the three aforementioned data sets of NUTS, GLD and LGD as source data sets and DBpedia[7] as target data set for each of them. On average, the geo-spatial locations associated with NUTS resources have higher distance to the related DBpedia resources' geo-spatial locations than the other two test datasets.

4.8 Evaluating the Class Coverage Measure

Figure 8 shows the evaluation of the *average coverage of a class in a data set* measure introduced in Section 3.8. While both GLD and nuts have the same class coverage, LGD has a lower class coverage than the two

[7] http://wiki.dbpedia.org/

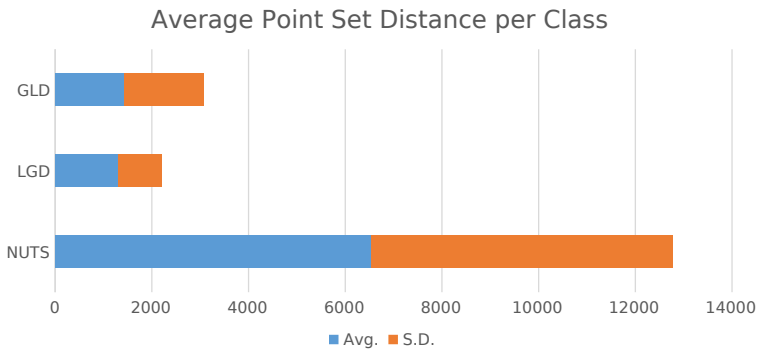

Figure 7: Comparison of the evaluation data sets on average point set distance per class over all point set distances over all classes.

of them.

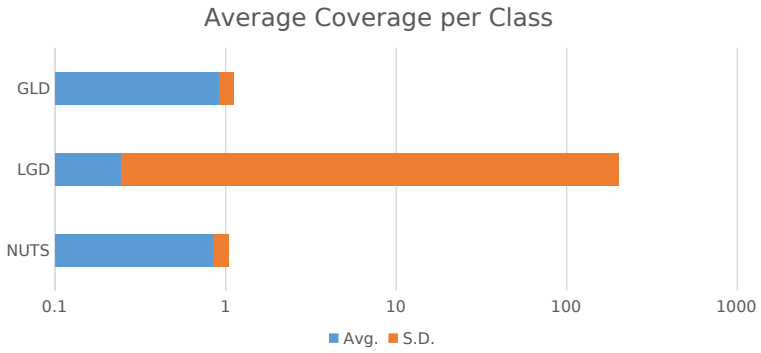

Figure 8: Comparison of the evaluation data sets on average class coverage over all classes.

4.9 Evaluating the Weighted Class Coverage Measure

Figure 9 shows the evaluation of the *average weighted coverage of a class in a data set* measure introduced in Section 3.9. While both GLD and nuts have the same weighted class coverage, LGD has a lower weighted class coverage.

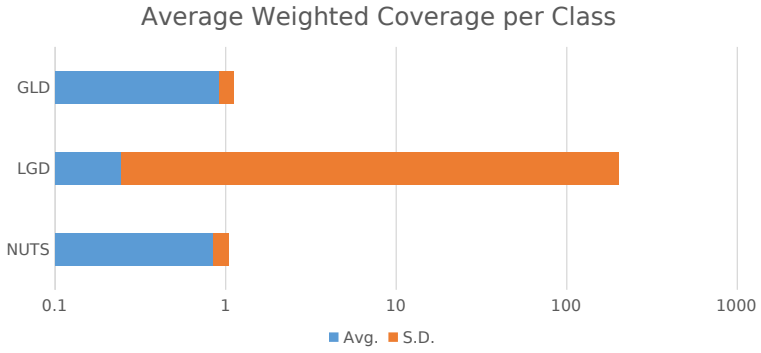

Figure 9: Comparison of the evaluation data sets on average weighted class coverage over all classes.

4.10 Evaluating the Data set Structuredness Measure

Figure 10 shows the evaluation of the *the coherence or structuredness of a data set* measure introduced in Section 3.10. According to the figure, GLD is the dataset with the highest structuredness, then come NUTS and LGD in the second and third places. Note that we do not have the standard deviations since each data set has a single value for this measure.

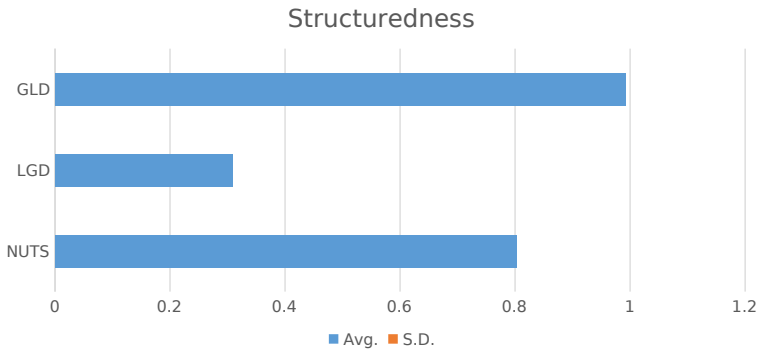

Figure 10: Comparison of the evaluation data sets on data set structuredness.

5 Conclusion and Future Work

In this deliverable, we demonstrated how the developed measures in D3.1.5 can be used to compare three different datasets with geo-spatial resources. We used NUTS as our reference dataset and LGD and GLD as two community-provided testing data sets. Pertaining to the number of instances per class, our evaluation showed that while LGD and GLD have – in average – higher number of instance per class over NUTS, NUTS has lower standard deviation. Also, GLD and LGD have lower average of point per class then our reference data set. Still, NUTS has a high variation of number of point per class between different classes. Some measures gave zero results when applied to specific data sets. For example, both *polygon per class* and *average surface per class* measures resulted in zero values when given GLD and LGD as input, this was because both data sets represent the geo-spatial resources using one geo-location. Moreover, the *number of intersecting classes per instance* measure gave zero result for NUTS because of the very specific classes of NUTS. Specking about structureness of our dataset, GLD came in the first place where NUTS and LGD came in the second and third places

In future work, we plan to implement more measures pertaining to geo-spatial dataset quality such as completeness and accuracy. Moreover, we aim to include more data sets with geo-spatial resources in our evaluation.

References

[1] Sören Auer, Jens Lehmann, and Sebastian Hellmann. LinkedGeoData - adding a spatial dimension to the web of data. In *Proc. of 8th International Semantic Web Conference (ISWC)*, 2009.

[2] RE Deakin. Great elliptic arc distance. *Lecture Notes. School of Mathematical & Geospatial Science, RNIT University, Melbourne, Australia, January*, 2012.

[3] Songyun Duan, Anastasios Kementsietsidis, Kavitha Srinivas, and Octavian Udrea. Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 145–156. ACM, 2011.

[4] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.

[5] LM. Vilches-Blázquez, B. Villazón-Terrazas, Oscar Corcho, and A. Gómez-Pérez. Geolinked data. an application case / un caso de aplicación. In *Proceedings of the I Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE 2010)*, October 2010. Ontology Engineering Group ? OEG.