# A Collection of Benchmark Datasets for Systematic Evaluations of Machine Learning on the Semantic Web

Petar Ristoski<sup>1</sup>, Gerben Klaas Dirk de Vries<sup>2</sup>, and Heiko Paulheim<sup>1</sup>

<sup>1</sup> University of Mannheim, Germany Research Group Data and Web Science {petar.ristoski,heiko}@informatik.uni-mannheim.de <sup>2</sup> WizeNoze, Amsterdam, Netherlands g.k.d.devries@outlook.com

Abstract. Resource type: Datasets Permanent URL: http://w3id.org/sw4ml-datasets In the recent years, several approaches for machine learning on the Semantic Web have been proposed. However, no extensive comparisons between those approaches have been undertaken, in particular due to a lack of publicly available, acknowledged benchmark datasets. In this paper, we present a collection of 22 benchmark datasets of different sizes.Such a collection of datasets can be used to conduct quantitative performance testing and systematic comparisons of approaches.

Keywords: Linked Open Data, Machine Learning, Datasets, Benchmarking

## 1 Introduction

In the recent years, applying machine learning to Semantic Web data has drawn a lot of attention. Many approaches have been proposed for different tasks at hand, ranging from reformulating machine learning problems on the Semantic Web as traditional, propositional machine learning tasks to developing entirely novel algorithms. However, systematic comparative evaluations of different approaches are scarce; approaches are rather evaluated on a handful of often project-specific datasets, and compared to a baseline and/or one or two other systems.

In contrast, evaluations in the machine learning area are often more rigorous. Approaches are usually compared using a larger number of standard datasets, most often from the UCI repository<sup>3</sup>. With a larger set of datasets used in the evaluation, statements about statistical significance are possible as well [3].

At the same time, collections of benchmark datasets have become quite well accepted in other areas of Semantic Web research. Notable examples include the Ontology Alignment Evaluation Initiative (OAEI) for ontology matching<sup>4</sup>, the

<sup>&</sup>lt;sup>3</sup> http://archive.ics.uci.edu/ml/

<sup>&</sup>lt;sup>4</sup> http://oaei.ontologymatching.org/

Berlin SPARQL Benchmark<sup>5</sup> for triple store performance, the Lehigh University Benchmark (LUBM)<sup>6</sup> for reasoning, or the Question Answering over Linked Data (QALD) dataset<sup>7</sup> for natural language query systems.

In this paper, we introduce a collection of datasets for benchmarking machine learning approaches for the Semantic Web. Those datasets are either existing RDF datasets, or external classification or regression problems, for which the instances have been enriched with links to the Linked Open Data cloud [14]. Furthermore, by varying the number of instances for a dataset, scalability evaluations are also made possible.

#### 2 Related Work

Recent surveys on the use of Semantic Web for machine learning organize the proposed approaches in several categories, i.e., approaches that use Semantic Web data for machine learning [16], approaches that perform machine learning on the Semantic Web [11], and approaches that use machine learning techniques to create and improve Semantic Web data [8, 16]. Furthermore, there are some challenges, like the *Linked Data Mining Challenge*<sup>8</sup> or the *Semantic-Web enabled Recommender Systems Challenge*<sup>9</sup>, which usually focus on only a few datasets and a very specific problem setting.

### 3 Datasets

Our dataset collection has three categories: (i) existing datasets that are commonly used in machine learning experiments, (ii) datasets that were generated from official observations, and (iii) datasets generated from existing RDF datasets. Each of the datasets in the first two categories are initially linked to DBpedia<sup>10</sup>. This has two main reasons, (1) DBpedia being a cross-domain knowledge base usable in datasets from very different topical domains, and (2) tools like DBpedia Lookup and DBpedia Spotlight making it easy to link external datasets to DBpedia. However, DBpedia can be seen as an entry point to the Web of Linked Data, with many datasets linking to and from DBpedia. In fact, we use the RapidMiner Linked Open Data extension [9], to retrieve external links for each entity to YAGO<sup>11</sup> and Wikidata<sup>12</sup>. Such links could be exploited for systematic evaluation of the relevance of the data of different LOD dataset in different learning tasks.

In the dataset collection, there are four datasets that are commonly used for machine learning. For these datasets, we first enrich the instances with links to LOD datasets, and reuse the already defined target variable to perform machine learning experiments:

 $<sup>^{5}\</sup> http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/$ 

<sup>&</sup>lt;sup>6</sup> http://swat.cse.lehigh.edu/projects/lubm/

 $<sup>^7</sup>$  http://greententacle.techfak.uni-bielefeld.de/ $\sim$ cunger/qald/

 $<sup>^{8}\</sup> http://knowalod 2016. informatik.uni-mannheim.de/en/linked-data-mining-challenge/$ 

 $<sup>^9</sup>$  http://challenges.2014.eswc-conferences.org/index.php/RecSys

<sup>&</sup>lt;sup>10</sup> http://dbpedia.org

<sup>&</sup>lt;sup>11</sup> http://yago-knowledge.org/

<sup>&</sup>lt;sup>12</sup> http://www.wikidata.org

- The Auto MPG dataset<sup>13</sup> captures different characteristics of cars, and the target is to predict the fuel consumption (MPG) as a regression task.
- The AAUP (American Association of University Professors) dataset contains a list of universities, including eight target variables describing the salary of different staff at the universities<sup>14</sup>. We use the average salary as a target variable both for regression and classification, discretizing the target variable into "high", "medium" and "low", using equal frequency binning.
- The Auto 93 dataset<sup>15</sup> captures different characteristics of cars, and the target is to predict the price of the vehicles as a regression task.
- The Zoo dataset captures different characteristics of animals, and the target is to predict the type of the animals as a classification task.

For those datasets, cars, universities, and animals are linked to DBpedia based on their name.

The second category of datasets contains a list of datasets where the target variable is an observation from different real-world domains, as captured by official sources. Again, the instances were enriched with links to LOD datasets. There are thirteen datasets in this category:

- The Forbes dataset contains a list of companies including several features of the companies, which was generated from the Forbes list of leading companies 2015<sup>16</sup>. The target is to predict the company's market value as a classification and regression task. To use it for the task of classification we discretize the target variable into "high", "medium", and "low", using equal frequency binning.
- The *Cities* dataset contains a list of cities and their quality of living, as captured by Mercer [7]. We use the dataset both for regression and classification.
- The Endangered Species dataset classifies animals into endangered species<sup>17</sup>.
- The Facebook Movies dataset contains a list of movies and the number of Facebook likes for each movie<sup>18</sup>. We first selected 10,000 movies from DB-pedia, which were then linked to the corresponding Facebook page, based on the movie's name and the director. The final dataset contains 1,600 movies, which was created by first ordering the list of movies based on the number of Facebook likes, and then selecting the top 800 movies and the bottom 800 movies. We use the dataset for regression and classification.
- Similarly, the Facebook Books dataset contains a list of books and the number of Facebook likes. Each book was linked to the corresponding Facebook page using the book's title and the book's author. Again, we selected the top 800 books and the bottom 800 books, based on the number of Facebook likes.
- The Metacritic Movies dataset is retrieved from Metacritic.com<sup>19</sup>, which contains an average rating of all time reviews for a list of movies [12]. The initial

<sup>&</sup>lt;sup>13</sup> http://archive.ics.uci.edu/ml/datasets/Auto+MPG

<sup>&</sup>lt;sup>14</sup> http://www.amstat.org/publications/jse/jse\_data\_archive.htm

 $<sup>^{15}\ \</sup>rm http://www.amstat.org/publications/jse/v1n1/datasets.lock.html$ 

<sup>&</sup>lt;sup>16</sup> http://www.forbes.com/global2000/list/

<sup>&</sup>lt;sup>17</sup> http://a-z-animals.com/

 $<sup>^{18}</sup>$  We use the Facebook Graph API: https://developers.facebook.com/docs/graph-api

<sup>&</sup>lt;sup>19</sup> http://www.metacritic.com/browse/movies/score/metascore/all

dataset contained around 10,000 movies, from which we selected 1,000 movies from the top of the list, and 1,000 movies from the bottom of the list. We use the dataset both for regression and classification.

- Similarly, the *Metacritic Albums* dataset is retrieved from Metacritic.com<sup>20</sup>, which contains an average rating of all time reviews for a list of albums [13].
- The *HIV Deaths Country* dataset contains a list of countries with the number of deaths caused by HIV, as captured by the World Health Organization<sup>21</sup>. We use the dataset both for regression and classification.
- Similarly, the Traffic Accidents Deaths Country dataset contains a list of countries with the number of deaths caused by traffic accidents<sup>22</sup>.
- The Energy Savings Country dataset contains a list of countries with the total amount of energy savings of primary energy in 2010<sup>23</sup>, which was downloaded from WorldBank<sup>24</sup>. We use the dataset both for regression and classification.
- Similarly, the *Inflation Country* dataset contains a list of countries with the inflation rate for  $2011^{25}$ .
- The Scientific Journals Country dataset contains a list of countries with a number of scientific and technical journal articles published in 2011<sup>26</sup>.
- The Unemployment French Region dataset contains a list of regions in France with the unemployment rate, used in the SemStats 2013 challenge [10].

Again, for those datasets, the instances (cities, countries, etc.) are linked to DBpedia. For datasets which are used for classification and regression, the regression target was discretized using equal frequency binning, usually into a *high* and a *low* class.

The third, and final, category contains datasets that were generated from existing RDF datasets, where the value of a certain property is used as a classification target. There are five datasets in this category:

- The Drug-Food Interaction dataset contains a list of drug-recipe pairs and their interaction, i.e., "negative" and "neutral" [6]. The dataset was retrieved from FinkiLOD<sup>27</sup>. Furthermore, each drug is linked to DrugBank<sup>28</sup>. We drew a stratified random sample of 2,000 instances from the complete dataset. When generating the features, we ignore the foodInteraction property in DrugBank, since it highly correlates with the target variable.
- The AIFB dataset describes the AIFB research institute in terms of its staff, research group, and publications. In [1] the dataset was first used to predict the affiliation (i.e., research group) for people in the dataset. The dataset contains 178 members of a research group, however the smallest group contains only 4

 $<sup>^{20}\ \</sup>rm http://www.metacritic.com/browse/albums/score/metascore/all$ 

 $<sup>^{21}</sup>$  http://apps.who.int/gho/data/view.main.HIV1510

<sup>&</sup>lt;sup>22</sup> http://apps.who.int/gho/data/view.main.51310

<sup>&</sup>lt;sup>23</sup> http://data.worldbank.org/indicator/10.1\_ENERGY.SAVINGS

<sup>&</sup>lt;sup>24</sup> http://www.worldbank.org/

 $<sup>^{25}\ \</sup>rm http://data.worldbank.org/indicator/NY.GDP.DEFL.KD.ZG$ 

 $<sup>^{26}\ \</sup>rm http://data.worldbank.org/indicator/IP.JRN.ARTC.SC$ 

<sup>&</sup>lt;sup>27</sup> http://linkeddata.finki.ukim.mk/

<sup>&</sup>lt;sup>28</sup> http://wifo5-03.informatik.uni-mannheim.de/drugbank/

people, which is removed from the dataset, leaving 4 classes. Also, we remove the employs relation, which is the inverse of the *affiliation* relation.

- The AM dataset contains information about artifacts in the Amsterdam Museum [2]. Each artifact in the dataset is linked to other artifacts and details about its production, material, and content. It also has an artifact category, which serves as a prediction target. We have drawn a stratified random sample of 1,000 instances from the complete dataset. We also removed the material relation, since it highly correlates with the artifact category.
- The MUTAG dataset is distributed as an example dataset for the DL-Learner toolkit<sup>29</sup>. It contains information about complex molecules that are potentially carcinogenic, which is given by the isMutagenic property.
- The BGS dataset was created by the British Geological Survey and describes geological measurements in Great Britain<sup>30</sup>. It was used in [17] to predict the lithogenesis property of named rock units. The dataset contains 146 named rock units with a lithogenesis, from which we use the two largest classes.

An overview of the datasets is given in Tables 1, 2, and 3. For each dataset, we depict the number of instances, the machine learning tasks in which the dataset is used (C stands for classification and R stands for regression), the source of the dataset, and the LOD datasets to which the dataset is linked. For each dataset, we depict basic statistics of the properties of the LOD datasets, i.e., average, median, maximum and minimum number of *types*, *categories*, *outgoing relations* (rel out), *incoming relations* (rel in), outgoing relations including values (rel-vals out) and incoming relations including values (rel-vals in). The datasets, as well as a detailed description, a link quality evaluation, and licensing information, can be found online<sup>31</sup>.

From the given statistics, we can infer the following observations: (i) DBpedia contains significantly less *owl:sameAs* links to YAGO, compared to Wikidata; (ii) DBpedia provides the highest number of types and categories on average per entity; (iii) Wikidata contains the highest number of outgoing and incoming relations for most of the datasets; (iv) YAGO contains the highest number of outgoing and incoming relations values for most of the datasets.

### 4 Conclusion and Outlook

In this paper, we have introduced a collection of 22 benchmark datasets for machine learning on the Semantic Web. So far, we have concentrated on classification and regression tasks. There are methods to derive clustering and outlier detection benchmarks from classification and regression datasets [4,5], so that extending the dataset collection for such unsupervised tasks is possible as well. Furthermore, as many datasets on the Semantic Web use extensive hierarchies in the form of ontologies, building benchmark datasets for tasks like *hierarchical multi-label classification* [15] would also be an interesting extension.

 $<sup>^{29}</sup>$  http://dl-learner.org

 $<sup>^{30}</sup>$  http://data.bgs.ac.uk/

<sup>&</sup>lt;sup>31</sup> http://w3id.org/sw4ml-datasets

Me A	V	н		щ		A	А			
Metacritic Albums	FB Movies	FB Books	Cities	Forbes	Zoo	Auto 93	AAUP	Auto MPG	Name	
Metacritic	Facebook	Facebook	Mercer	Forbes	UCI ML	JSE	JSE	UCI ML	Source	
$c   \frac{R/C}{(c=2)}$	(c=2)	(c=2)	R/C (c=3)	$^{ m R/C}_{ m (c=2)}$	(c=3)	R	R/C (c=3)	R	Dataset • Task	
DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	DBpedia YAGO Wikidata	LOD	
	lia 1,600 O 1,339 ata 1,585		lia 212 O 187 ata 212	lia 1,585 O 1,003 ata 1,189	lia 101 O 8 ata 101	lia 93 O 80 ata 93	lia 960 O 889 ata 959	lia 371 O 331 ata 371	) #links	
								- 2		
$\begin{array}{ccc} 17.92 & 1\\ 7.22 & \\ 0.99 & \end{array}$	$\begin{array}{ccc} 24.90 & 27 \\ 12.08 & 14 \\ 1.01 & 1 \end{array}$	$\begin{array}{ccc} 19.08 & 20 \\ 8.37 & 10 \\ 1.00 & 1 \end{array}$	$\begin{array}{cccc} 31.28 & 35 \\ 16.66 & 19 \\ 2.11 & 2 \end{array}$	$\begin{array}{c ccc} 14.77 & 19 \\ 7.28 & 10 \\ 0.82 & 1 \end{array}$	8.61 1 0.74 1.00	$\begin{array}{ccc} 28.76 & 3 \\ 13.80 & 1 \\ 1.00 & \end{array}$	нN		ty avg me	
	$     \begin{array}{ccc}       27 & 55 \\       14 & 32 \\       1 & 4 \\       1 & 4     \end{array} $				$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{ccc} 31 & 43 \\ 16 & 19 \\ 1 & 2 \\ 1 & 2 \end{array}$	$     \begin{array}{ccc}       28 & 41 \\       2 & 17 \\       2 & 5     \end{array}   $	$\begin{array}{ccc} 31 & 46 \\ 16 & 21 \\ 1 & 3 \end{array}$	types med max	
000	0 0 1	000	100	000	000	005	000	005	min	H
4.27 3.22 0.00	$2.50 \\ 6.51 \\ 0.04$	$5.15 \\ 2.03 \\ 0.01$	$6.98 \\ 4.46 \\ 3.40$	4.87 2.35 0.22	$4.67 \\ 0.15 \\ 0.67 \\ $	1.13 9.09 0.12	9.38 3.31 0.88	$1.20 \\ 9.26 \\ 0.29$	cat avg m	able 1:
$ \begin{array}{c} 4 \\ 3 \\ 0 \\ 1 \end{array} $	$ \begin{array}{cccc} 11 & 60 \\ 6 & 27 \\ 0 & 1 \\ 1 \end{array} $	$     \begin{array}{c}       5 \\       2 \\       15 \\       15 \\       1     \end{array} $	$   \begin{array}{c}     7 \\     4 \\     4 \\     6   \end{array} $		$     \begin{array}{c}       3 \\       34 \\       1 \\       2     \end{array}   $	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	9 20 3 11 1 2	$     \begin{array}{ccc}       10 & 25 \\       9 & 23 \\       0 & 3 \\       \end{array} $	categories med max	Data
106 00	1 7 0 0 0	1 5 3 0 0	6 5 6 0 0 0	<u>3 2 2</u> 0 0 0	264	$     \begin{array}{ccc}       25 & 3 \\       18 & 0 \\       1 & 0   \end{array} $	2 0 2 0	00 N	x min	sets st
$10.85 \\ 8.05 \\ 17.64$	$12.43 \\ 8.39 \\ 48.75$	$     \begin{array}{r}       11.15 \\       8.41 \\       21.19     \end{array} $	$18.08 \\ 13.75 \\ 69.08$	$10.15 \\ 7.57 \\ 16.59$	$8.22 \\ 0.63 \\ 29.69$	$12.69 \\ 8.37 \\ 17.31$	$12.68 \\ 11.37 \\ 30.71$	$13.48 \\ 8.76 \\ 20.20$	avg	Table 1: Datasets statistics
$12 \\ 9 \\ 18$	48 J	$11 \\ 10 \\ 22 \\ 22 \\ 11 \\ 10 \\ 11 \\ 11 \\ $	19 15 67	11 16	35 1 9	12 10 17	$\frac{15}{29}$	$13 \\ 18$	rel out	01
17     10     45	21 17 107	$\frac{20}{55}$	$\frac{38}{153}$	27 21 137	$\frac{15}{57}$	22 11 26	28 15 83	27 18 61	rel out med max min	
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccc} 0 & 1.46 \\ 0 & 26.89 \\ 0 & 2.22 \end{array}$	$\begin{array}{c cccc} 0 & 1.64 \\ 0 & 25.32 \\ 0 & 3.15 \end{array}$	$\begin{array}{c c} 0 & 25.66 \\ 0 & 23.56 \\ 6 & 39.99 \end{array}$	$\begin{array}{c c} 0 & 2.76 \\ 0 & 52.07 \\ 0 & 5.00 \end{array}$	$\begin{array}{c c}3 & 3.54\\0 & 127.23\\3 & 8.28\end{array}$		$\begin{array}{ccc} 0 & 8.20 \\ 0 & 13.61 \\ 0 & 8.51 \end{array}$	$\begin{array}{ccc} 3 & 5.62 \\ 0 & 16.96 \\ 9 & 5.32 \end{array}$		
			$\begin{array}{ccc} 66 & 25 \\ 56 & 9 \\ 99 & 37 \\ 37 \end{array}$		5				rel j avg med	
	$     \begin{array}{ccc}       1 & 12 \\       6 & 138 \\       1 & 22 \\       1 & 22     \end{array} $	$     \begin{array}{ccc}       2 & 7 \\       3 & 138 \\       3 & 17 \\       3 & 17     \end{array} $		$\begin{array}{ccc} 2 & 27 \\ 2 & 138 \\ 5 & 52 \end{array}$	3 38 7 138 7 27	5 2 138 3 8	$\begin{array}{ccc} 7 & 36 \\ 3 & 138 \\ 7 & 44 \end{array}$	5 25 2 138 31	max	
<u>1 1 0</u>	0 1 0		μN	0 1 0	0 2 1	1 2 2	0 1 0		min	
$8.92 \\ 40.27 \\ 11.73 $	$11.65 \\ 55.01 \\ 56.37$	$7.04 \\ 24.37 \\ 16.41$	$16.26 \\ 222.54 \\ 105.29$	$\begin{array}{c} 10.44 \\ 34.42 \\ 12.69 \end{array}$	$13.26 \\ 5.39 \\ 18.20 $	$14.35 \\ 59.33 \\ 11.23$	$\begin{array}{c} 11.74 \\ 85.83 \\ 22.38 \end{array}$	$16.50 \\ 77.08 \\ 13.92$	rel	
$32 \\ 12 \\ 3$	$     \begin{array}{c}       12 \\       47 \\       23 \\       3     \end{array} $	$   \begin{array}{c}     7 \\     22 \\     16   \end{array} $	$\begin{array}{ccc} 13 & 1 \\ 214 & 6 \\ 89 & 3 \end{array}$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{c} 11 \\ 21 \\ 1 \end{array} $	$     \begin{array}{c}       11 \\       59 \\       11 \\       11   \end{array} $	$ \begin{array}{c} 11 \\ 68 \\ 21 \end{array} $	$   \begin{array}{c}     15 \\     70 \\     12   \end{array} $	rel-vals out	
$ \begin{array}{c} 63 \\ 361 \\ 49 \\ \end{array} $	$51 \\ 280 \\ 372 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ $	60 149 69	$ \begin{array}{c} 131\\ 681\\ 390\\ 2 \end{array} $	136 ( 510 ( 207 (		$\begin{array}{ccc} 64 & 4\\ 129 & 0\\ 25 & 4\end{array}$	66 ( 446 ( 97 (	70 278 54 4	l-vals out med max min	
$\begin{array}{c c} 0 \\ 0 \\ 2 \\ 2 \\ \end{array}$	0 0 4,6	$\begin{array}{c c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 $	$\begin{array}{c c} 0 & 1,4 \\ 0 & 8,0 \\ 2 & 5,2 \\ \end{array}$	010,5	$\begin{array}{c c}1 & 1\\0 & 26, 1\\1 & 1\\1 & 1\\1 & 1\end{array}$	4,0	$\begin{array}{c} 0 \\ 0 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{array}$	μų	-	
$5.28 \\ 2,749.90 \\ 8.77$	$4.96 \\ 4,682.42 \\ 20.75$	$2.80 \\ 4,735.50 \\ 7.47$		$14.30 \\ 10,531.37 \\ 30.14$	$\begin{array}{c ccccc} 146.28 & 24 & 3,686 \\ 26,173.23 & 28,418 & 28,418 \\ 125.82 & 92 & 785 \end{array}$	$22.60 \\ 4,025.90 \\ 19.91$	$62.18 \\ 2,455.27 \\ 296.92$	$36.65 \\ 3,236.24 \\ 59.33$	re	
7 23	$\frac{2}{12}$	4 8 2	678 19,810 3,555 72,320 1,599 99,865	107 8	24 98,418 92	18 46 19	23 110 20	$23 \\ 21 \\ 21 \\ 21 \\ 21 \\ 21 \\ 21 \\ 21 \\ $	rel-vals in med	
$\begin{array}{ccc} 3 & 50 \\ 10 & 28,418 \\ 7 & 54 \end{array}$	$\begin{array}{ccc} 2 & 110 \\ 43 & 28,418 \\ 12 & 230 \end{array}$	$\begin{array}{ccc} 2 & 42 \\ 8 & 28,418 \\ 4 & 165 \end{array}$	$\begin{array}{c} 678 \\ 555 \\ 555 \\ 72,320 \\ 599 \\ 99,865 \end{array}$	$\begin{array}{ccc} 4 & 1,925 \\ 107 & 28,418 \\ 8 & 2,881 \end{array}$	$3,686 \\28,418 \\785$	$\begin{array}{ccc} 18 & 64 \\ 46 & 28,418 \\ 19 & 57 \end{array}$	$\begin{array}{ccc} 23 & 2,488 \\ 110 & 28,418 \\ 20 & 31,777 \end{array}$	$\begin{array}{ccc} 23 & 509 \\ 60 & 28,418 \\ 21 & 755 \end{array}$	max min	
1 1 0	0 1 0	0 + 0	н <del>5</del> 0	0 - 0	0 2 12	3 4 2	0 1 0	<u>ω ο ο</u>	min	

Table	
÷	
Datasets	
statistics	

	min	0	-	0	24	550	66	77	693	66	77	757	66	77	693	66	77	693	66	37	88	137		1	0	0	1	4	0	0
	max n	31	18	117		,032 5			_			_		57			57	_		2,292	27	1,464 1	42	18	,930	158	,418	487	211	0
in	B		28,418	-	70,426	112,0	669, 273	218,957	423,559	557, 157	218,957	423,559	557, 157	218,957	423,559	557, 157	218,957	423,559	557, 157	$^{2,2}$	$^{4.5}$	1,4	114,742	28,418	460,9	-	28,4	4	2	
rel-vals i	med	2	37	11	,065	,879	71	,587	26	7,369 1,	,440	21	Ļ,	,440	21	÷.	40	21	Ļ,	969	24	93	15	418	70 6,	$^{24}$	418	26	50	0
rel-					1,0	4,8	4,971	Ξ	6,126	7,3	Η	5,821	5,810	1,4	5,821	5,810	1,440	5,821	5,810	6	1,424	-		$28 \\ -28 \\$			$2^{2}_{8}$			
	avg	3.56	,960.84	15.77	4,828.36	12,464.42	45,671.15	7,528.18	17,882.47	85,575.10	6,876.80	16,969.96	77,485.01	6,947.59	17,114.88	78,453.16	6,947.59	7,114.88	78,453.16	975.88	.,793.19	332.69	566.25	22,286.36	,909.94	34.71	127.08	32.25	62.49	0.00
			4,90		4,8;	12,40	45,6'	7,5;	17,8	85,5'	6,8,	16,90	77,48	6,9	17,1	78,43	6'.9	17,1	78,43	6	1,7	õ	ñ	22,26	21,90		20,4		-	
	min	0	0	0	10	0	2	0	0	2	0	0	0	0	0	2	0	0	2		$^{28}_{28}$	58	0	0	1	0	0	0	14	1
out	max	30	216	237	162	1,267	595	162	2,104	713	162	2,104	713	162	2,104	713	162	2,104	713	19	299	157	87	136	400	12	61	159	132	1
rel-vals	med	12	43	53	25	244	210	$^{24}$	290	213	23	279	210	24	279	210	$^{24}$	279	210	2	09	62	Π	0	22	ŝ	0	27	41	1
re	avg 1	1.42	48.84	52.70	27.75	302.34	229.46	27.44	345.03	242.38	26.72	329.28	238.69	26.85	331.16	237.94	26.85	331.16	237.94	7.19	81.12	83.12	12.65	9.53	22.04	3.63	7.28	34.99	41.96	1.00
	.я	0 1	14	0	6 2	5 30	2 22		5 34	2 24		5 32	2 23		5 33	2 23		5 33	2 23	7	т к	21 8	1	1	0	0	1	13	0	0
	ux min	2	138	13	89	138	148	94	138	148	94	138	148	94	138	148	94	138	148	24	9	51 2	55	38	78	5 L	38	17	2	0
l in	ed max	1	6 15		31 8	9 13	51 14	36 9	9 13	55 1	33	9 13	$55   1_4$	33	9 13	$55   1_4$	33	9 13	$55   1_4$	13	4	33	 ന	38 13	9	1	138 13	0	2	0
rel	avg med	35	22	.98		8			31			ŧ0			18			18			73		96	_	94	ŧ0	-	8	20	00
	aı	1.35	28.22	1:0	34.26	15.18	55.68	37.87	14.61	61.68	36.64	16.40	60.02	36.74	16.48	60.12	36.74	16.48	60.12	14.19	3.73	34.00	2.96	108.62	6.94	1.40	99.69	2.78	1.70	0.00
	min	c,	0	0	3	0	7	1	0	4	1	0	7	1	0	1-	-	0	7	e C	12	74	0	0	ŝ	0	0	0	41	3
t	max	19	11	99	28	35	173	28	35	191	28	35	191	28	35	191	28	35	191	10	14	119	25	6	137	15	8	79	71	3
rel out	med max	14	10	49	24	31	119	24	31	121	$^{24}$	31	119	24	31	119	24	31	119	6	12	84	11	0	37	14	0	47	64	ŝ
	avg	12.54	8.34	47.86	23.59	28.41	20.87	23.40	28.44	24.31	23.46	27.82	23.36	23.46	27.80	23.23	23.46	27.80	23.23	7.81	12.42	86.23	10.77	1.78	34.32	12.65	2.15	45.92	61.68	3.00
	mim	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		-	0	0	0	0	0	0	-	-
ies	max n	42	$^{28}$	1	23	18	9	23	16	9	23	16	9	23	16	9	23	16	9	15	x	4	34	12	9	18	9	-	~	~
categories	med n	11	9	0	13	6	ŋ	13	10	ŋ	13	10	ŋ	13	10	ŋ	13	10	ŋ	n	0	e	ю	0	0	5 L	0	0	/	~
Ca	avg n	1.87	6.43	0.03	2.61	9.28	1.83	3.12	9.62	1.94	3.12	9.52	4.92	3.11	9.44	4.88	3.11	9.44	4.88	3.73	2.77	2.58	6.32	0.76	0.44	5.46	0.68	0.01	~	~
F	min	0 1	0	0	0 1	0	-	0	0	-	0	0	-	0 1	0	-	0	0	-	0	x		0	0	0	0	0	0	2	1
	max m	45	19	1	52	24	x	53	27	10	53	27	10	53	27	10	53	27	10	32	14	e	89 89	16	4	38	31	ĉ	2	1
types	med m	27	14	1	37	15	4	38	15	4	38	15	4	38	15	4	38	15	4	21	x	-	12	0	-	4	0	0	2	-
	avg n	24.38	11.79	0.98	35.69	13.90	4.12	36.40	14.29	4.42	36.07	14.09	4.41	36.00	14.09	4.39	36.00	14.09	4.39	16.38	8.92	1.35	11.84	2.48	1.05	8.83	4.46	1.96	2.00	1.00
			8	_			_												_				Ξ							
	#links	2,000	1,588	1,981	114	108	114	146	139	146	162	152	162	160	150	160	160	150	160	26	26	26	301	65	301	1,989	588	1,908	2,00	2,000
	D	edia	00	data	edia	00	data	edia	00	data	edia	00	data	edia	00	lata	edia	00	data	edia	0	lata	edia	0	data	edia	0	data	3ank	LOD
	LOD	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DBpedia	YAGO	Wikidata	DrugI	FinkiLOD
L.	Task	U/ 0						U/ 0			0/0		(1)		2/2 10/2		U/ 0		(1=)	U/ 0							ζ		<sup>(C-2)</sup> DrugBank 2,000	_
Dataset				<u>ب</u>							mk /	<u>ب</u>		mk /	<u>ت</u>			<u>ب</u>	$ a-z-animals \begin{pmatrix} R/C \\ (c=2) \end{pmatrix} $						<u>ب</u>					
ñ	Source		Metacritic			OHW			OHW			WorldBank			WorldBank		WorldBank			SemStats				-anim				FinkiLOD		
	s		Me	_		_	_		_	_		Wo	_		Wo		Worl							a-z-	_			Fir		
	e	:+:	1010	S		2			nts	Control 1		ve vec	202		10	۲,	U	ole ole		ment	l e	1 5	bered	na la	ŝ		200	uou Hon	non	
	Name	otoor	Marci	MON	ΗΙΛ	Deaths	Country	Traffe	Accidents	Conntry	Enerov	Savinge	Conntry	Tuffation	Constant	IIII	Scientific Journals Country			Unemployment French Region			Endangered Species			Drug-Food Interaction				
		Metacritic Movies HIV Deaths			A A C			C S H				ЧŬ			Jo Sc			Unem F. R.			End SI			Ċ	1 <u>1</u>	=				

Table 2: Datasets statistics

Table 3: Datasets statistics

	types				rel out					re	in		I	el-va	ls out	;	rel-vals in					
Name	Task	#links	avg	med	$\max$	$\min$	avg	med	$\max$	$\min$	avg	med	$\max$	$\min$	avg	$\operatorname{med}$	$\max$	$\min$	avg	$\operatorname{med}$	$\max$	$\min$
AIFB	C $(c=4)$	176	1.4	1	2	1	7.1	7	9	5	2.0	2	5	0	18.2	7	219	2	19.8	9	246	0
AM	C(c=11)	1,000	1.0	1	1	1	19.8	20	29	9	0.6	1	3	0	21.9	20	283	7	3.2	1	273	0
MUTAG	C $(c=2)$	340	1.0	1	1	1	9.8	10	14	5			/	\	65.8	56	465	4		/	/	/
BGS	C $(c=2)$	146	1.0	1	1	1	29.7	31	- 36	21	1.4	2	4	0	25.2	24	54	15	2.7	2	12	0

Acknowledgements The work presented in this paper has been partly funded by the German Research Foundation (DFG) under grant number PA 2373/1-1 (Mine@LOD), and the Dutch national program COMMIT.

#### References

- 1. Bloehdorn, S., Sure, Y.: Kernel Methods for Mining Instance Data in Ontologies. The Semantic Web pp. 58–71 (2007)
- de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: The Semantic Web: Research and Applications, pp. 733–747. Springer (2012)
- Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)
- Emmott, A.F., Das, S., Dietterich, T., Fern, A., Wong, W.K.: Systematic construction of anomaly detection benchmarks from real data. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description. pp. 16–21. ACM (2013)
- Färber, I., Günnemann, S., Kriegel, H.P., Kröger, P., Müller, E., Schubert, E., Seidl, T., Zimek, A.: On using class-labels in evaluation of clusterings. In: MultiClust: Workshop on Discovering, Summarizing and Using Multiple Clusterings (2010)
- Jovanovik, M., Bogojeska, A., Trajanov, D., Kocarev, L.: Inferring cuisine-drug interactions using the linked data approach. Scientific reports 5 (2015)
- 7. Paulheim, H.: Generating possible interpretations for statistics from linked open data. In: 9th Extended Semantic Web Conference (ESWC) (2012)
- Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., Fanizzi, N.: Mining the semantic web. Data Mining and Knowledge Discovery pp. 613–662 (2012)
- 9. Ristoski, P., Bizer, C., Paulheim, H.: Mining the web of linked data with rapidminer. Web Semantics: Science, Services and Agents on the WWW (2015)
- 10. Ristoski, P., Paulheim, H.: Analyzing statistics with background knowledge from linked open data. In: Workshop on Semantic Statistics (2013)
- 11. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: A comprehensive survey. Web Semantics 36, 1–22 (2016)
- Ristoski, P., Paulheim, H., Svátek, V., Zeman, V.: The linked data mining challenge 2015. In: KNOW@ LOD (2015)
- Ristoski, P., Paulheim, H., Svátek, V., Zeman, V.: The linked data mining challenge 2016. In: KNOW@LOD (2016)
- 14. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: The Semantic Web–ISWC (2014)
- Silla, Jr., C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. pp. 31–72 (2011)
- 16. Tresp, V., Bundschus, M., Rettinger, A., Huang, Y.: Towards machine learning on the semantic web. In: Uncertainty Reasoning for the Semantic Web I (2008)
- 17. de Vries, G.K.D.: A fast approximation of the Weisfeiler-Lehman graph kernel for RDF data. In: ECML/PKDD (1). pp. 606–621 (2013)