# LinkGen: Multipurpose Linked Data Generator

Amit Krishna Joshi, Pascal Hitzler, and Guozhu Dong

Data Semantics Lab, Wright State University, Dayton, OH, U.S.A.
{joshi.35,pascal.hitzler,guozhu.dong}@wright.edu

**Abstract.** The paper presents LinkGen, a synthetic linked data generator that can generate a large amount of RDF data based on certain statistical distribution. Data generation is platform independent, supports streaming mode and produces output in N-Triples and N-Quad format. Different sets of output can be generated using various configuration parameters and the outputs are reproducible. Unlike existing generators, our generator accepts any vocabulary and can supplement the output with noisy and inconsistent data. The generator has an option to inter-link instances with real ones provided that the user supplies entities from real datasets.

## 1 Introduction

In recent years, we have seen a rapid adoption of semantic technologies by a number of large organizations such as BBC, Thomson Reuters, New York Times and Library of Congress [3]. Linked Open Data (LOD) cloud consists of more than 30+ billion triples and hundreds of datasets[1]. These datasets use a number of vocabularies to describe the group of related resources and relationships between them. According to [16], Linked Open Vocabularies (LOV) dataset now consists of more than 500 vocabularies, 20,000 classes and almost 30,000 properties. The vocabularies are modeled using either RDF Schema (RDFS) or richer ontology languages such as OWL [5].

Linking enterprise data is also gaining popularity and industries are perceiving semantic technologies as a key contributor for effective information and knowledge management [14, 18]. One of the major obstacles for building a linked data application is generating a synthetic dataset to test against specific vocabularies. In this paper, we present LinkGen, a synthetic linked data generator that generates arbitrarily large datasets for a given vocabulary. Generating synthetic data is not a new concept. It has been widely used in database field for testing database design and software applications as well as database benchmarking and data masking [2]. In the semantic web field, it has been primarily used for benchmarking Triplestores. Existing generators [9, 13, 4, 7] are designed for specific use cases and work well with certain vocabularies but cannot be re-purposed for other vocabularies. LinkGen, on the other hand, can work with widely available vocabularies and can be used in multiple scenarios including: 1)

[1] http://lod-cloud.net/

Testing new vocabulary 2) Querying datasets 3) Diagnosing data inconsistencies 4) Evaluating performance of datasets 5) Testing Linked Data aggregators 6) Evaluating various compression methods

Creating synthetic datasets that closely resemble real world datasets is very important. Numerous studies including [6, 15] found that URIs in real world linked datasets exhibit a power-law distribution. In order to automatically generate synthetic data that exhibit such power-law distribution, LinkGen employs random data generation based on various statistical distributions including Zipf's Law[2].

Real world linked datasets are by no means free of noise and redundancy. Linked Data quality and noise in Linked Data has been studied extensively in [10, 17, 11, 19]. The noise can be in the form of invalid data, syntactic errors, inconsistent data and wrong statements. LinkGen provides some of these options to add noise in the synthetic dataset. LinkGen also has the option to specify the number of triples to generate. It aids in testing existing linked data compression methods such as [6, 8] against varying database size and scenarios.

Specifically, the contribution of this work is a tool that can automatically generate synthetic datasets with the following properties:

- Dataset can be generated based on power-law distribution to resemble real world datasets
- Noise can be added to the synthetic dataset
- Dataset can be generated in both streaming and on-disk mode
- Synthetic instances can be linked to real-world entities if dictionary of real world entities is available.

The rest of this paper is organized as follows. Section 2 describes related work and existing generators. Section 3 describes the LinkGen generator with details on various parameters including data distribution and noisy data. Section 4 reports on experimental results and finally, Section 5 concludes the paper and identifies topics for further research. The tool is open source and available at GitHub[3] under GNU License[4].

## 2 Related Work

To the best of our knowledge, this is the first work that generates synthetic linked dataset for any vocabulary that can mimic real world datasets with features such as statistical distribution and noisy data. Quite a few synthetic generators exist that have been developed for benchmarking RDF stores using specific vocabularies. The Lehigh University Benchmark (LUBM) [7] consists of a data generator that produces repeatable and customizable synthetic dataset using Univ-Bench Ontology in the unit of a university. Different set of data can be generated by

[2] To review Zipf's and Pareto's Law, see [1]
[3] http://www.w3id.org/linkgen
[4] https://opensource.org/licenses/GPL-3.0

specifying the seed for random number generation, number of universities and the starting index of the universities.

Berlin SPARQL Benchmark (BSBM) [4] is built around an e-commerce use-case in which a set of products is offered by different vendors and consumers have posted reviews about products. BSBM constitutes a data generator that supports the creation of large datasets using number of products as the scale factor and can output in an RDF representation as well as relational representation.

SP$^2$Bench [13] has a data generator for creating DBLP[5]-like RDF triples and mimics correlations between entities using power law distributions and growth curves. The Social Intelligence Benchmark (SIB) [12] contains an S3G2 (Scalable Structure-correlated Social Graph Generator) that creates a synthetic social graph with correlations. Tontogen[6] is a protege-plugin that can create synthetic dataset using a uniform distribution of instances for relationships. WatDiv[7] and Sygenia[8] are two other tools that can generate data based on user supplied queries.

As noted above, none of the existing generators are suitable for creating synthetic data for different vocabularies. They have a little or no option to configure the output in regards to data distribution, noise and alignments.

## 3 Data Generator

In this section, we describe different concepts related to the data generator and provide details on how it works. At the core of data generation is a random data generator used for generating unique identifiers for each entity. In order to create different sets of output, LinkGen creates random data based on the seed value supplied by the user.

### 3.1 Entity Distribution

There are different statistical methods to generate and distribute entities in a dataset. LinkGen provides two statistical distribution techniques namely Gaussian distribution and Zipf's power-law distribution. Example of Gaussian distribution includes those in real life phenomena such as heights of people, errors in measurement and marks on a test. Examples of Power-law distributions include the frequencies of words and frequencies of family names. [6, 15] found that subject URIs in real world linked datasets exhibit a power-law distribution. LinkGen use zipf's law as a default option for entity distribution. Figure 1 taken from [6] shows the power-law distribution of subjects in a Wikipedia dataset.

[5] http://dblp.uni-trier.de/db/
[6] http://lsdis.cs.uga.edu/projects/semdis/tontogen/
[7] http://dsg.uwaterloo.ca/watdiv/download
[8] https://sourceforge.net/projects/sygenia/

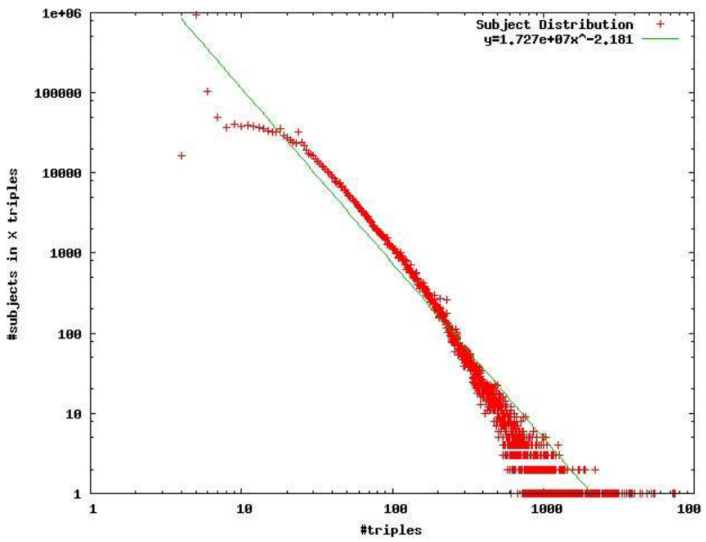


**Fig. 1.** Power-law distribution of subjects in Wikipedia

### 3.2 Noisy Data

Noisy data plays a critical role in applications that aggregate data from multiple sources and those that deal with semi-structured and unstructured data [10]. LinkGen creates noisy data by:

- Adding inconsistent data, for instance writing two conflicting values for a given dataType property
- Adding triples with syntactic errors, ex: typos in subjectURI or rdfs:Label
- Adding wrong statement by assigning invalid domain and range, ex: ns:PlaceInstance rdf:type ns:Person
- Creating instances with no type information

Users can specify a combination of parameters for generating noisy data. All parameters related to noise are prefixed with *noise.data* text in the configuration file ex: *noise.data.total* and *noise.data.num.notype*. If the output is in N-Quads format, the noisy data are added to a separate named graph.

### 3.3 Inter-linking real world entities

LinkGen allows mapping real world entities with automatically generated entities. For this, the user has to supply a set of real world entities expressed in RDF format: <ns:entityuri> rdf:type <ns:class>. LinkGen will then inter-link by using owl:sameAs triple, such as: <ns:entityuri> owl:sameAs <ns:classInstance>. This enables users to create a mixed dataset by combining synthetic dataset with the real dataset. This is important in scenarios where you would need to study the

effect of adding new triples in current live dataset. Existing SPARQL queries can be slightly modified to fetch additional results from test dataset by adding owl:sameAs statement in the query.

### 3.4 Output Data and Streaming mode

LinkGen creates a VoID[9] dump once the synthetic data is generated. VoID, the Vocabulary of Interlinked Datasets, is used to express metadata about RDF dataset and provides a number of properties for expressing numeric statistical characteristics about a dataset, such as the number of RDF triples, classes, properties or, the number of entities it describes.

LinkGen supports N-Triples and N-Quads format for output data. By default, the tool will save output to a file but it can be run in streaming mode, enabling users to pipe the output of RDF streams to other custom applications.

### 3.5 Config Parameters

There's an array of configuration parameters available to create unique synthetic datasets. The output is reproducible so running LinkGen multiple times with same set of input parameters will yield same output. Most useful configuration parameters include: a) distribution type which can be gaussian or zipf and b) seed values for creating different datasets

### 3.6 Data Generation Steps

The first step in data generation involves loading ontology and gathering statistics about all ontology components such as number of classes, datatype properties, object properties and properties for which domain and range are not defined. We also store the connectivity of each class and order the classes based on the frequency. Most connected class will lead to generation of larger number of corresponding entities.

The second step involves using statistical distribution to generate large number of entities and associating the weights for each one of them. Parameters for Zipf and Gaussian distribution are configurable and can be used to create different sets of output. For Zipf's distribution, sample size is equal to the size of maximum number of triples to be generated. For Gaussian distribution, two parameters viz. mean and standard deviation are required.

Next step involves going through each class and generating synthetic triples for associated properties using weighted entities. For each entity, at least two triples are added to denote its type. They are: *instance* rdf:type *Classs* and, *instance* rdf:type *owl:Thing*. It should be noted that not all properties have well defined domain and range. For instance, in DBpedia, more than 600 properties including the ones in Table 1 have either missing domain or range information

[9] https://www.w3.org/TR/void/

in the vocabulary. In such cases, RDF Semantics[10] permits using any resources as a domain of the property. Similarly, the range can be any Literal or resource depending on whether the property is datatypeProperty or objectProperty.

**Table 1.** Properties with no domain or range info in DBpedia ontology

| DataTypeProperty with no domain | ObjectProperty with no range |
|---|---|
| http://dbpedia.org/ontology/number | http://dbpedia.org/ontology/teachingStaff |
| http://dbpedia.org/ontology/width | http://dbpedia.org/ontology/daylightSavingTimeZone |
| http://dbpedia.org/ontology/distance | http://dbpedia.org/ontology/simcCode |
| http://dbpedia.org/ontology/fileSize | http://dbpedia.org/ontology/uRN |

For datatypeProperties which have range of XSD datatypes, we used a simple random generator to create literal values.

## 4 Evaluation

To evaluate our work, we generated varying number of synthetic datasets for two general purpose vocabularies: DBpedia[11] and schema.org[12]. For schema.org, we used an owl version available from TopBraid[13]. We built LinkGen using Apacha Jena[14], a widely used free and open source Java framework for building Semantic Web and Linked Data applications. At the current state, LinkGen supports only RDFS vocabularies. Although it can generate synthetic dataset for any vocabulary expressed in RDFS or OWL, it does not implement all class descriptions and property restrictions specified in the OWL ontology. Also, the support for blank nodes is not provided.

Table 2 shows the general characteristics of the dataset used for the experiment. For both DBpedia and Schema.org, the most connected classes were Person, Place and owl:Thing.

**Table 2.** Characteristics of the datasets used for evaluation

| | DBpedia | Schema.org |
|---|---|---|
| Number of distinct classes | 147 | 158 |
| Number of distinct properties | 2891 | 1002 |
| Number of distinct object properties | 1734 | 463 |
| Number of distinct data properties | 1100 | 490 |
| distinct properties without domain and/or range specification | 685 | 11 |

[10] https://www.w3.org/TR/2000/CR-rdf-schema-20000327/
[11] http://www.dbpedia.org
[12] http://www.schema.org
[13] http://topbraid.org/schema/
[14] http://jena.apache.org/

Figure 2 is the performance chart depicting the total time taken to create synthetic datasets of varying size for both vocabularies. There's a slight increase in time for DBpedia which may be due to the relatively high number of properties.

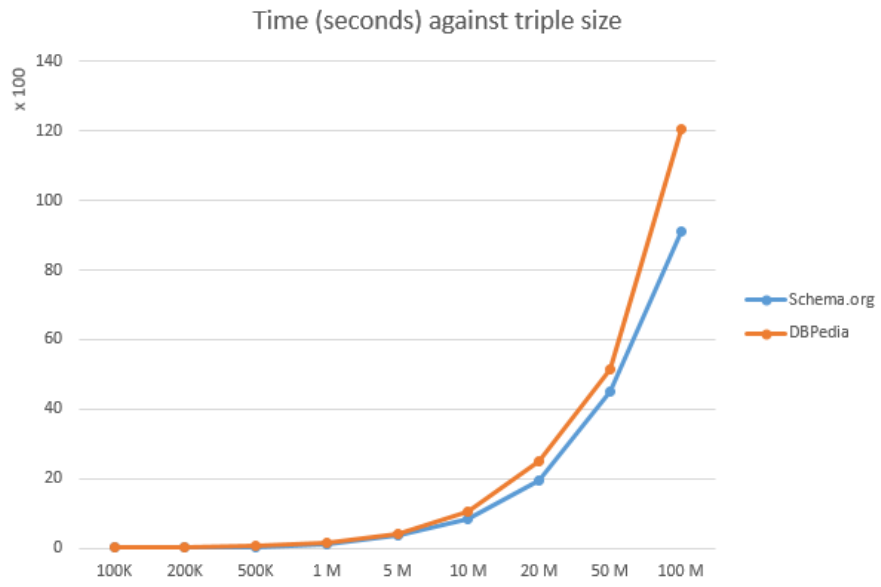


**Fig. 2.** Time taken for generating datasets of various sizes

## 5 Conclusion

In this paper, we have introduced a multipurpose synthetic linked data generator. The system can be configured to generate various sets of output to test semantic web applications under different scenarios. This includes defining a statistical distribution type for instances, adding inconsistent and noisy data, and integrating real world entities. The system supports streaming mode which can be used for evaluating applications that deal with streaming data. By generating a large amount of RDF data, it can aid in testing the performance of various applications that deal with querying, storage, visualization, compression and reporting. Experimental results show that out generator is highly performant and scalable. In the future, we will explore supporting OWL constraints as well as using parallel and distributed algorithms to generate massive dataset in short duration.

## References

1. Adamic, L.A.: Zipf, power-laws, and pareto - a ranking tutorial. Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html (2000)

2. Arasu, A., Kaushik, R., Li, J.: Data generation using declarative constraints. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. pp. 685–696. ACM (2011)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts pp. 205–227 (2009)
4. Bizer, C., Schultz, A.: Benchmarking the performance of storage systems that expose SPARQL endpoints. World Wide Web Internet And Web Information Systems (2008)
5. Brickley, D., Guha, R.V.: RDF vocabulary description language 1.0: RDF schema (2004)
6. Fernández, J.D., Martínez-Prieto, M.A., Gutierrez, C.: Compact representation of large RDF data sets for publishing and exchange. In: The Semantic Web–ISWC 2010, pp. 193–208. Springer (2010)
7. Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web 3(2), 158–182 (2005)
8. Joshi, A.K., Hitzler, P., Dong, G.: Logical linked data compression. In: The Semantic Web: Semantics and Big Data, pp. 170–184. Springer (2013)
9. Morsey, M., Lehmann, J., Auer, S., Ngonga Ngomo, A.C.: DBpedia SPARQL benchmark–performance assessment with real queries on real data. The Semantic Web–ISWC 2011 pp. 454–469 (2011)
10. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS) 10(2), 63–86 (2014)
11. Péron, Y., Raimbault, F., Ménier, G., Marteau, P.F.: On the detection of inconsistencies in RDF data sets and their correction at ontological level (2011)
12. Pham, M.D., Boncz, P., Erling, O.: S3g2: A scalable structure-correlated social graph generator. In: Selected Topics in Performance Evaluation and Benchmarking, pp. 156–172. Springer (2012)
13. Schmidt, M., Hornung, T., Lausen, G., Pinkel, C.: $SP^2$Bench: a SPARQL performance benchmark. In: Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on. pp. 222–233. IEEE (2009)
14. Semeraro, G., Basile, P., Basili, R., De Gemmis, M., Ghidini, C., Lenzerini, M., Lops, P., Moschitti, A., Musto, C., Narducci, F., et al.: Semantic technologies for industry: From knowledge modeling and integration to intelligent applications. Intelligenza Artificiale 7(2), 125–137 (2013)
15. Tummarello, G., Delbru, R., Oren, E.: Sindice. com: Weaving the open linked data. Springer (2007)
16. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semantic Web (Preprint), 1–16 (2015)
17. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in DBpedia. In: The Semantic Web: Trends and Challenges, pp. 504–518. Springer (2014)
18. Wood, D.: Linking Enterprise Data. Springer Science & Business Media (2010)
19. 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)