IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-16, May 2019 E-ISSN: 2347-2693

A Survey on Interlinking in Linked Open Data

Shweta S A [1*], Shreyas Suresh Rao[2]

[1,2]Dept. of Computer Science and Engineering, SOE, Presidency University, Bengaluru, India

Corresponding Author: shwetasa3@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si16.6974 | Available online at: www.ijcseonline.org

Abstract:- Semantic Web interconnects diverse data sources on the Web, thereby presenting a global database of Web resources. Interlinking is an important activity in establishing semantic links between applications on the World Wide Web. During the interlinking activity, first, link discovery is done to identify the datasets to be linked, and subsequently, link generation is done to generate the matching links, based on appropriate comparator algorithms. The paper presents a literature review on linked open data, focusing on the interlinking activity. Furthermore, the paper presents a novel map-reduce based approach for comprehensively presenting the interlinking algorithms. Lastly, the paper throws some insights on the LOD datasets available in the LOD Cloud 2018.

Keywords:- Semantic Web, Linked open data, Resource description framework, LOD Cloud, Hadoop Map-Reduce

I. INTRODUCTION

1.1 Semantic Web

The semantic Web is an extension of the World Wide Web, term invented by "Sir Tim Berners-Lee" in 1989. Semantic Web represents an effective means of data representation in the form of a global linked database [1]. The Semantic Web is completely managed by the World Wide Web Consortium (W3C). It expands on W3C's Resource Description Framework (RDF), and is usually designed with syntaxes that use Uniform Resource Identifiers (URIs) to represent data. The key objective of the Semantic Web is to trigger the development of the current Web to empower clients to seek, find, discover and share information across geographical and organizational boundaries without any effort [1]. The semantic Web basically takes into consideration the association of data utilizing a system that can be effectively read by machines – regardless of the kind, namely, PCs, IoT gadgets, cell phones etc.

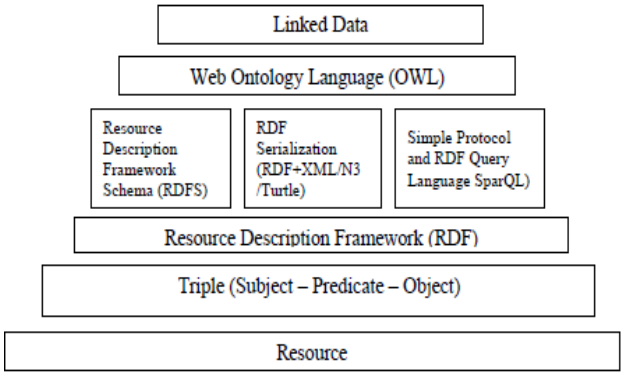

Figure 1: Structure of Semantic Web

Semantic Web has been decomposed into six layers where in each layer utilizes the services offered by the subsequent lower layers. Each layer is described below:

Resource (first layer). The base layer in semantic Web is a resource which refers to any conceptual or physical entity on the Web identified with a unique Uniform Resource Identifier (URI).

Triple (second layer). The relationships between the resources are represented as a collection of triples. A triple is a statement consisting of three constituent parts: subject, predicate and object.

Resource Description Framework (RDF) (third layer). Represents a data interchange format that encodes data in the triple form of subject, predicate and object.

The fourth layer comprises of three components namely RDF schema (RDFS), RDF serialization and SPARQL which operate on RDF data of the third layer. The different operations performed on RDF data include structuring, translating, retrieving and manipulating. While RDFS structures the RDF data, RDF serialization provides formats for translating RDF data and SPARQL provides a query language for retrieving and manipulating RDF data.

Web Ontology Language (OWL) (fifth layer). OWL adds semantics to the schema described by RDFS.

Linked Data (sixth layer). Linked data represents the top most layer in the semantic Web structure depicted in Figure 1. Linked data layer utilizes the services provided by all the

five other lower layers. The RDF data needs to be serialized into RDF supported formats that is queryable through SPARQL [10]. Structure of the semantic web gives strong base for understanding linked data concepts.

1.2 Linked Open Data (LOD)

Linked data refers to machine-readable RDF data present on the Web, which is linked to other data sources on the Web through RDF links, thereby contributing towards a giant global data space of data [4]. The linked data is also referred as Linked Open Data (LOD) since data is available not only to the publishing organization producing the data but also to any other organization or Web users accessing the data across the world. Linked data also offers lot of best practices for publishing and interlinking organized information on the Web. Publishing organizational data openly on Web provides an opportunity for other organizations, academicians and researchers to analyze and use that data in their work environment. The LOD datasets will use the given vocabulary standards (RDF, RDFS and SPARQL) [5].

In this paper we provide a comprehensive review of the importance of interlinking as a key LOD concepts. The paper reviews many algorithms for comparing the strings, date and integer. There is no comprehensive application for comparing the all algorithms performance together, so we propose a hadoop-based map-reduce technique to solve the problem. The other issues will be handled as future research work.

The next part of the paper is organized as follows. Section 2 explains the literature review of semantic web, LOD and interlinking. Section 3 presents an insight on the LOD structure, datasets in RDF format and provides an overview on the future research direction for LOD. Section 4 concludes the paper and presents some future work.

II. LITERATURE REVIEW

An important concept within the semantic web is linked data (LD). The main goal of LD is to relate data described using RDF model. So that machine can browse the data and then can interlink the related source of data by using many types of OWL (web ontology language). A best example for LOD is DBpedia. It will extract structured information from Wikipedia articles. This converts all Wikipedia content into RDF and provides link to this data to other databases like Geonames. The information provided by DBpedia is machine-readable, hence convenient for system-system interaction [13].

The term linked open data refers to a dataset which is open to Internet users, typically in RDF format [11]. LOD is public data and can be reused and republished without any restrictions. W3C created the RDF framework to describe the data on the web in structured format. RDF format expresses the data in triple format with different URI's like subject, predicate and object. RDF data can be stored in different format such as Turtle, JSON, N-Triples and XML.

The main aim of linked open data and semantic web technologies is to interlink existing data, available across multiple data sources. Interlinking of data in semantic web defines more about sharing and linking related data. Interlinking supports data reconciliation. It is provided by RDF triples that set up a link between the entity identified by the subject with the entity identified by the object and we represent the predicate using appropriate ontologies. Quality of interlinking data helps to access the data easily and it will increase the efficiency of LOD. It helps to establish the semantic links between the source dataset (enterprise data) with other target datasets on the World Wide Web [2].

The SPARQL is a query language to link data between diverse data sources. The OWL is considered as the base ontology language for semantic web and LOD. "Ontology is an explicit specification of a conceptualization". The different types of ontologies are Dublin Core, FOAF (Friend of a Friend), SKOS (Simple Knowledge Organization System). Linked data in the education field is solving several problems in educational technologies.

2.1 Semantic Web Tools

1. SILK Framework: - SILK is an open source framework used for integrating the different categories of data sources. This framework is used to generate the links between related data items. This framework specifies the different types of RDF links to interconnect all the related types of datasets.

2. Google Refine / Open Refine:- This tool is used for cleaning, transforming and format conversion of messy data, achieved through Web services.

3. D2R Server: - D2R Server (D2R – Database to Relational) is a tool that publishes relational database content in RDF form. This tool also provides a SPARQL endpoint for querying the RDF data through SPARQL queries and displays the results in various visualization formats [2].

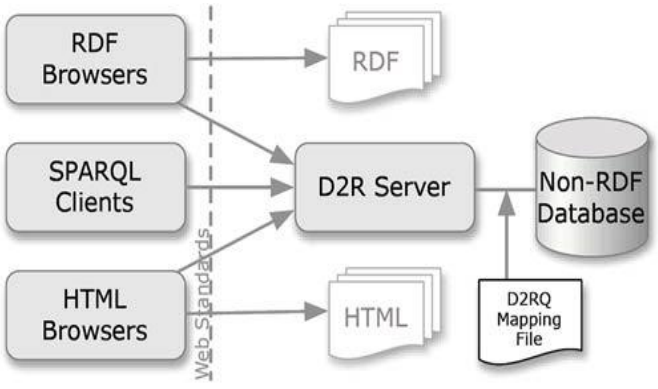

Figure 1: Design Diagram for D2R Server

During the time of interlinking from source data set to target data sets, we are using different data types such as "String", "Date", and "Integer". For "Date" and "Integer" types direct comparison can be performed. For the String type we use character based comparator algorithm and token based distance measures. Character based comparator algorithms are Levenshtein, Jaro-winkler, Smith-waterman, Jaro, Levenshtein Distance etc. These algorithms will compare strings at the character level [2].

Token based distance measures algorithms are Jaccard, Dice, Soft Jaccard etc. The character and token based algorithms are used in many domains like educational, banking and E-Government for interlinking on the Web [13].

2.2 LOD Standards for publishing the open data

Tim Berners-Lee has sketched out the accompanying standards for publishing linked data on the Web. These tenets help to publish the data to any one of the single worldwide information space. The Linked data standards, firmly identified with the five-star open model for distributing information [4], can be outlined as distributing organized, interlinked information, in non-exclusive organizations, utilizing URIs.

The tenets for publishing open data are given beneath:
1) Using URIs for names. URIs meaning resources must be resolvablee at http:// or https://
2) The RDF data model always contains the structured data. The RDF serialization formats include: RDF/XML, Turtle, N-Triples, N3 notation, JavaScript Object Notation (JSON).
3) A URI must give accommodating information using the benchmarks, for instance, RDF, RDFS, and OWL. RDFS check and comment properties can be used to give useful information.
4) Include associations with various URIs, with the objective that enables end-users to perform research.
5) Interlinking the data from other data sources using RDF links.

2.3 Linked Data publishing workflow:-

When publishing linked data, we have to follow the principles explained in Section 2.2. Figure 2 demonstrates the step by step design for how linked data can be published in global space. Linked data always operates with structured data; linking unstructured and semi-structured data is cumbersome requiring more research. Data is represented in RDF format (Subject, Predicate and Object). The linked data can be stored as relational database, Triple store or RDF store, prior to publication on the Web [29].

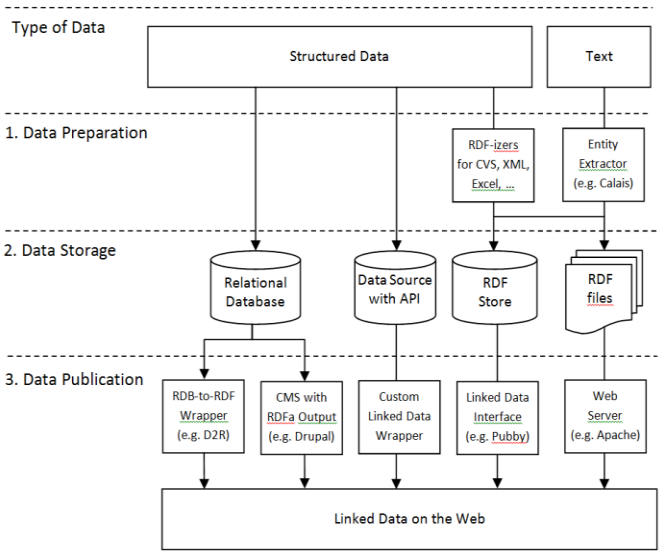

Figure 2: Workflow for Linked Data Publishing

III. INSIGHTS

3.1 LOD Cloud 2018

Linked data is a machine discernible RDF data which is published on the web. Since the data is available not only to the publishing organization it can be available throughout the global space.

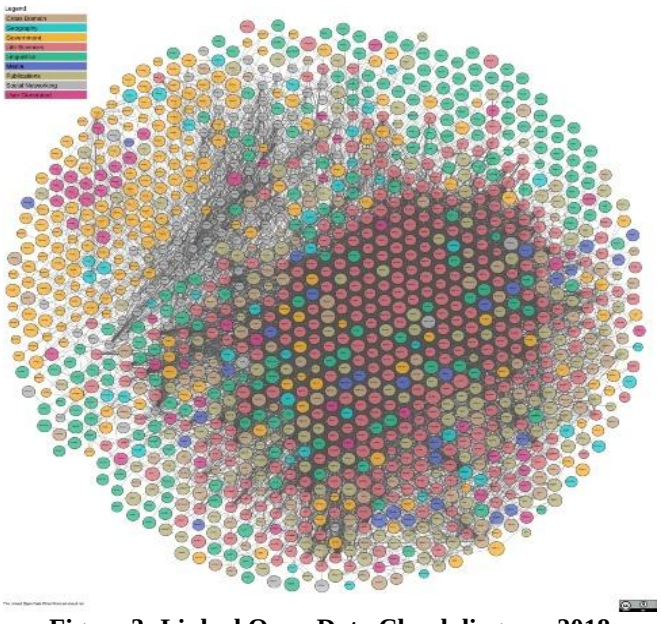

Figure 3: Linked Open Data Cloud diagram 2018

The main reason for motivation of interlinking concept is LOD cloud which is depicted in below Figure 3.

The figure shows datasets published by various organizations in linked data format. As of now LOD cloud

has **1,224** datasets with **16,113** links (as of June 2018). LOD cloud represents the largest open linked data collection in the world and enables any organization, researcher or academician to access the linked data for research or utility purposes [3]. In the center of the cloud we can see the biggest datasets like DBpedia and GeoNames, trailed by the W3C.

3.2 LOD Datasets

Dataset is an accumulation of related data. We are comparing the datasets from past 4 years from LOD cloud as shown in the figure 4 to know about the improvement and usage of the LOD cloud year-wise. When we start comparing the datasets, the numbers of datasets are increased year-wise. The biggest connected datasets are "DBpedia, Wikidata, Geonames, LinkedGeoData and YAGO" [16].

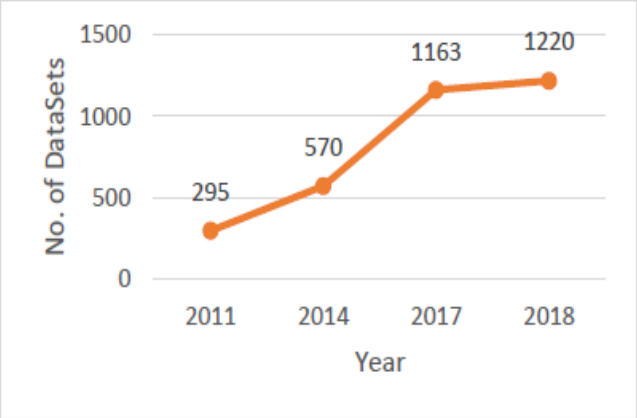

Figure 4: Increment in datasets year-wise in LOD Cloud

3.3 Limitations of SILK Framework tool:- SILK framework is an open source tool for interlinking the heterogeneous RDF datasets. SILK tool uses the graphical editor for linking the datasets and algorithms. In earlier days everyone used SILK workbench for link discovery technique. In this framework we use different character comparator algorithms and token based distance measures for comparing between datasets. The main defect of this tool is we have to check each algorithm separately during execution. We do not have options to check all the algorithms together. We should select one algorithm for execution and source-target datasets details for generating the RDF links between different data sources. We can take multiple source and target datasets but we cannot choose multiple algorithms at the same time for generating the links.

In Figure 5, levenshtein distance algorithm is used for comparing the date for all related datasets. This is the main drawback for SILK framework tool. To overcome this problem we can use hadoop Map-Reduce technology in big data.

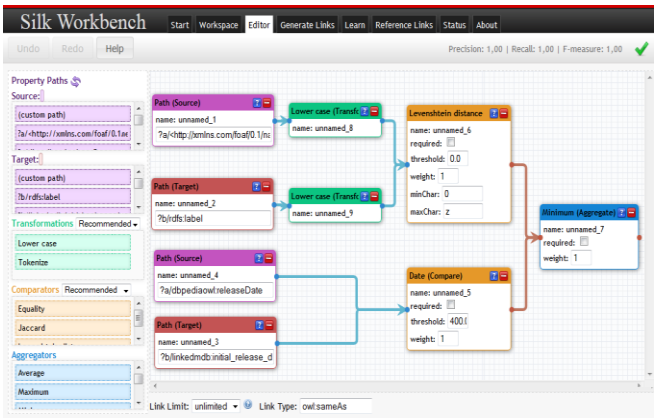

Figure 5: Example for SILK workbench tool

3.4 Big Data: - Now a days, Big data is exceptionally imperative for Organizations or Companies. Since it empowers them to accumulate, store, oversee, and control "Very Large Amounts of Data, Extremely High Velocity of Data and Extremely Wide Variety of Data". Big data stores structured, semi-structured and unstructured data [25]. Some of the examples for big data are social media, stock exchange data, transport data etc.

Hadoop Map Reduce

Hadoop is an open source framework which helps to store and process the big data in a dispersed manner. It is intended to scale up from single server to multiple number of machines. We can share huge amount of data into thousands of machines for better quality of output. Hadoop map reduce is one of the good technology to handle large amounts of datasets. The map-reduce algorithms divides into two sub-groups, such as Map and Reduce. Map takes a set of data and changes over into another set of data, where single components are separated into tuples. Reduce task will take output from map as an input and it will combines all the data into small set of tuples. We can split the data parallel to consume the time and it will increase the efficiency of output. The reduce job will process after the map job. Figure 6 explains how the map-reduce algorithm will take input and produces the output for the given dataset.

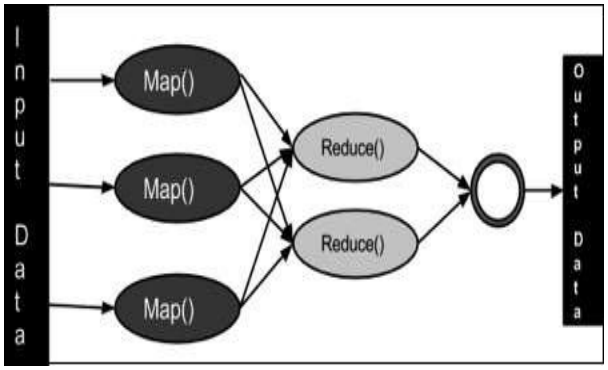

Figure 6: Overview of Hadoop Map-Reduce algorithm

The hadoop technology can run in two categories such as Hadoop single-node cluster and hadoop multi-node cluster.
In the single-node cluster both master and slave will act on the same machine. But, in multi-node cluster the one system will act as master and rest all machines will act as slaves, so we can store large datasets in master machine and distribute to all other slaves machines to parallelize the link discovery process. It will increase the scalability of data [26].

However, there are no applications that can automate the collective comparisons between the algorithms, to enable the end-user to visualize the comparisons comprehensively. If we use hadoop map reduce technology for interlinking the datasets, we can resolve the SILK framework limitation and build a comprehensive application that addresses the lacuna effectively.

IV. CONCLUSION AND FUTURE RESEARCH WORK

The paper provides a literature review on the concepts of Semantic Web, Linked Open Data and interlinking in LOD. Furthermore, the paper proposes a Hadoop map-reduce approach to solve the interlinking lacuna of SILK framework.

Future work includes: 1. Implement character-based distance algorithms viz., Levenstein, Jaro-Winkler and Smith-Waterman for link discovery in RDF dump datasets. 2. Implement token-based distance algorithms i.e., 'Jaccard' for link discovery in RDF dump datasets. 3. Implement a hadoop map-reduce algorithm to parallelize the link discovery process. 4. Implement a dashboard to visualize/ analyze the collective comparisons between the algorithms.

REFERENCES

[1]. Semantic Web techopedia [Online] Available: https://www.techopedia.com/definition/27961/semantic-web

[2]. S. S. Rao and A. Nayak, LinkED: A Novel Methodology for Publishing Linked Enterprise Data", CIT. Journal of Computing and Information Technology, Vol. 25, No. 3, 191–209 191 doi:10.20532/cit.2017.1003477 September 2017.

[3]. State of the LOD Cloud. [Online; accessed June-2018]. https://lod-cloud.net/#diagram

[4]. C. Bizer , T. Heath, and T. Berners-lee, "Linked Data – The Story So Far", *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009. http://dx.doi.org/10.4018/jswis.200908191

[5]. G. Antoniou and F. Van Harmelen."A Semantic Web Primer" MIT Press second edition 2008.

[6]. T. Berners-Lee, " Linked data",2006. [Online] Available: http://www.w3.org/DesignIssues/LinkedData.html

[7]. L. Jie, L. Wei and L. Liming, "Linked Open Data Query Based on Natural Language", Chinese Journal of Electronics Vol.26, No.2, March 2017

[8]. https://www.researchgate.net/publication/315323934_Linked_Open_Data_Query_Based_on_Natural_Language

.

[9]. C. K. Pereira, S. W. Matsui Siqueira et al., "Linked data in Education: a survey and a synthesis of actual research and future challenges", IEEE Transaction Journal Citation information: DOI 10.1109/TLT.2017.2787659, 1939-1382, 2017 https://ieeexplore.ieee.org/document/8240662

[10]. W. Beek, L. Rietveld, S Schlobach and F. V. Harmelen, "LOD Laundromat *Why the Semantic Web Needs Centralization (Even If We Don't Like It)", IEEE Journals and magazines Systems*, vol. 20, no. 2, pp. 78-81, 2016. https://ieeexplore.ieee.org/document/7420493

[11]. D2R Server: Accessing databases with SPARQL and as Linked Data. [Online] Available: http://d2rq.org/d2r-server#features

[12]. A. Zaveri *et al*., "Quality Assessment for Linked Data: A Survey", *Semantic Web Journal*, vol. 7, no. 1, pp. 63–93, 2015. http://dx.doi.org/10.3233/SW-150175

[13]. https://www.completesoftware.net/2012/02/interlinking-of-data-what-is-it-and-why-is-it-important/

[14]. M. Rani H G et al, "An Investigate Study on the Quality Aspects of Linked Open Data", ACM ISBN 978-1-4503-6576-5/18/10 DOI: http://doi.org/10.1145/3291064.3291074

[15]. E. Rajabi et al, "Interlinking Educational Data: An Experiment with GLOBE Resources", 2013 ACM 978-1-4503-2345-1/13/11 http://dx.doi.org/10.1145/2536536.2536588

[16]. W. Leyh et al,"Interlinking Standardized OpenStreetMap Data and Citizen Science Data in the OpenData Cloud", Springer International Publishing AG 2018 DOI 10.1007/978-3-319-60366-7_9

[17]. https://www.lesliesikos.com/lod-datasets/

[18]. Z. Zhang et al, "An unsupervised data-driven method to discover equivalent relations in large linked datasets", Semantic Web 8 2017, 197-223, DOI 10.3233/SW-150193.

[19]. J. Debattista et al, " Luzzu- A Methodology and Framework for Linked Data Quality Assessment", ACM Journal of Data and Information Quality, Vol. 8, No. 1, Article 4, October 2016.

[20]. T. Kawamura, A ohsuga, "Applying Linked Open Data to Green Design", IEEE Journal, Volume 30, issue 1, page 28-35 2015. https://ieeexplore.ieee.org/document/6916497

[21]. R. Stojanov and S Gramatikov et al, " Linked Data Authorization Platform", IEEE Journal, Volume 6, page s :1189-1213, Year 2018. https://ieeexplore.ieee.org/document/8120116

[22]. S. Subhashree, R Irny, et al, "Review of Approaches for Linked Data Ontology Enrichment", Springer ICDCIT 2018, LNCS 10722, pp. 27–49, 2018 https://doi.org/10.1007/978-3-319-72344-0_2

[23]. A. Subramanian and S. Srinivasa, "Semantic Integration of Structured Data Powered by Linked Open Data", ACM. ISBN 978-1-4503-3293-4/15/07 July 13 - 14, 2015, http://dx.doi.org/10.1145/2797115.2797130

[24]. A. Torre-Bastida et al, "A Rule Based Transducer for Querying Incompletely Aligned Datasets", ACM transactions on the web, vol. 12, No. 4, Article 23. Date September 2018.

[25]. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., &Gómez-Pérez, A. 2018. A comprehensive quality model for Linked Data. Semantic Web, 9, 3-24. DOI=10.3233/SW- 170267

[26]. https://www.journaldev.com/8734/introduction-to-bigdata

[27]. https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[28]. Stravoskoufos, K., Petrakis, E.G., Mainas, N., Batsakis, S., & Samoladas, V. 2016. SOWL QL: Querying Spatio-Temporal Ontologies in OWL. *Journal on Data Semantics*, 5, https://link.springer.com/article/10.1007/s13740-016-0064-5

[29]. http://linkeddatabook.com/editions/1.0/

Authors Profile

Shweta S A is a research scholar in the department of Computer Science and Engineering at Presidency University, Bengaluru, India. She Completed her M.Tech in CSE from SJBIT in 2017 Bengaluru. Her research interests include Semantic Web, Linked Open Data.

Dr. Shreyas Suresh Rao is an Assistant Professor in the department of Computer Science and Engineering at Presidency University, Bengaluru, India. He obtained his Ph.D from Manipal Institute of Technology, Manipal Academy of Higher Education, India. His previous experience includes seven years of working at SLK Software Services Pvt. Ltd., Bangalore where he played various roles such as Business Analyst, Team Leader, Analyst/Designer, and had been involved in the execution of several end-to-end enterprise projects in banking, manufacturing and automobile domains. His technical expertise lies in ASP.NET technologies such as Web services, Web applications, Silverlight, and SharePoint. His research interests include Semantic Web, Linked Open Data, crowdsourcing, knowledge engineering and Web services.

