A Metric-driven Approach for Interlinking Assessment of RDF Graphs

Najme Yaghouti[1], Mohsen Kahani[2], Behshid Behkamal[3]
Web Technology Lab
Ferdowsi University of Mashad
Mashhad, Iran
[1]*najme.yaghouti@stu.um.ac.ir*
[2]*kahani@um.ac.ir*
[3]*behkamal@um.ac.ir*

Abstract— **In recent years the web has evolved from a global information space of linked documents to one where both documents and data are linked. What supports this evolution is a set of best practices in publishing and connecting structured data on the web that is called linked data. The usefulness of linked data relies on how much related concepts are linked together. The aim of this research is to propose a metric-driven approach for interlinking assessment of a single dataset. The proposed metrics are categorized into three groups called internal linking, external linking and link-ability from other datasets. These metrics consider both graph structure (topology) and schema of datasets (semantic information) to evaluate interlinking with appropriate accuracy.**

Keywords— Interlinking; Linked Data; Metrics; RDF graph

I. Introduction

In recent years the web has evolved from a global information space of linked documents to one where both documents and data are linked. What supports this evolution is a set of best practices for publishing and linking structured data on the web that is called linked data. Linked Open Data (LOD) provides a distributed model for the semantic Web which allows any data provider to publish its publicly available data and meaningfully link them with other information sources over the Web. Existing information on semantic web are useful if and only if there are appropriate links between related concepts. Since human decisions become less important in web of data and smart agents should traverse links to retrieve suitable information for the users, one of the most important issues in linked data domain is having appropriate links between different datasets [17]. In order to connect different datasets, it is useful to evaluate the interlinking potential of a dataset with LOD cloud. To the best of our knowledge there are no automatic tools for the assessment of interlinking of a single dataset. So far, this can only be done using expert opinion and with manual and semi-automatic approaches. In this study, we propose a metric-driven approach for evaluating interlinking of a single dataset. If such automatic assessment tools are developed, then it becomes possible to identify datasets with low interlinking before publishing them, and thus we can first improve such datasets and then publish them on the web.

978-1-4673-9181-8/15/$31.00 ©2015 IEEE

The rest of this paper is organized as follows: Section 2 describes the general terms. Section 3 reviews the works about link quality assessment, and also graph theory in other domains. Section 4 describes the problem and our contribution. Section 5 introduces the proposed approach for interlinking assessment of a dataset. In Section 6 intermediate results are presented. Evaluation strategy and some general discussions are provided in Section 7. Finally, the paper is concluded in Section 8.

II. Terminology

In this section, we will define key concepts and terms used throughout this paper.

Interlinking: refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources [32]. Interlinking can be defined from two perspectives, internal linking and external linking. Internal linking refers to relating concepts inside a dataset [17], while external linking examine how much a dataset links to other datasets [19]. In this research we extend this definition by proposing a new concept, i.e. link-ability. Link-ability is the potential of a dataset to be linked from other datasets. Thus, our notion of interlinking is based on three concepts: internal linking, external linking and Link-ability. The reason we consider link-ability as new impressive factor in interlinking is because once we publish a dataset, it is important for us to be linked from other data sources to increase our dataset's popularity and having a significant role in providing information to the LOD cloud.

Measurement: is a mapping from the real world to the formal world in predefined certain conditions [9, 15].

Metric: is a symbol or number which (based on the mapping that was referred in measurement) is assigned to an entity in a real world to quantify one of the entity's attribute. Metrics can be subjective or objective [4, 14].

RDF: A graph-based data model which is used to describe web resources and relations between them. In this data model information is represented as <subject, predicate, object> triples [7].

RDF graph: An RDF graph is a collection of RDF triples [29].

Complete graph: In the mathematical field of graph theory, a complete graph is a simple graph in which every pair of distinct vertices is connected by a unique edge.

Node: Each entity or literal which is placed as a subject or an object of an RDF triple is represented by a node in a graph.

Edge: A line which connects two nodes inside a graph.

Link: A line which connects two nodes from two different graphs.

Internal edge degree: No. of edges which enter a specific node.

External edge degree: No. of edges which get out of the node. Since LOD graph is a directed graph, we distinguish between internal edge degree and external edge degree.

Degree Centrality: The number of edges incident upon a node. In the case of a directed graph, we define two separate measures of degree centrality, namely internal degree centrality and external degree centrality. Accordingly, the first one is a count of the number of ties directed to the node and the second one is the number of ties that the node directs to others.

Clustering Coefficient: The local clustering coefficient of a node in a graph quantifies how close its neighbours are to being a clique (complete graph).

Dereferenceable URI: is a resource retrieval mechanism that uses any of the internet protocols (e.g. HTTP) to obtain a copy or representation of the resource it identifies**.**

III. State of the Art

Interlinking is a new issue in linked data domain. Despite its importance, it has only recently received some attention from the Semantic Web community. Since our proposed approach for interlinking assessment is based on graph theory, we primarily focus on link analysis and assessment with respect to Semantic Web and LOD, but also briefly touch upon graph based metrics in similar applications that are relevant to our work.

A. Link Analysis and Assessment in Linked Data Domain

Many attempts have been made to analyze and assess the quality of links, with different goals in mind. Usually link structure analysis is done with the goal of obtaining information about semantic web structure. These analyses are done with different purposes including: interlinking assessment in semantic web, which shows that interlinking between resources is improving thorough the years [12], scalable data processing which represents information about most used subjects, predicates, objects, n-grams and similar information [23]. Also, Guéret and his colleagues have measured the robustness of network by evaluating the effect of link distribution in network robustness against attacks [16].

Link quality assessment methods can be divided into two groups. The first group assesses all types of links, while the latter one evaluates specific link types. Type-independent evaluation can be done in different ways. Some of the works focus on quality evaluation of links using a metric-based approach. For example, the study done by [17, 28] represent a platform called LATC to increase the quality and quantity of links on the web. The aim of this platform is adding links to the network to make it more robust against attacks. The accuracy of the evaluation method is not sufficient due to lack of considering link types and using sampling techniques. The authors of [19] presented a set of guidelines for publishing linked data on the web as well as some metrics for evaluating data publishers' conformance with these guidelines. They presented two metrics for interlinking assessment.

There are also a number of works focusing on link quality assessment without using metrics. For example, LiQuate is an automatic tool which evaluates the quality of data and links in LOD cloud using Bayesian networks. This system takes advantage of statistical reasoning to return the probability of ambiguity or incompleteness of data or links [27]. Another work is an unsupervised approach for finding wrong links. In this work, each link is represented as a feature vector in an upper dimension space. Then the wrong links, which are far from global distribution of other links, can be identified using multidimensional outlier detection methods [25].

In recent years, there are several reported works done on the evaluation of specific link types. Most of these works have focused on evaluating the quality of owl:sameAs links. Also, in [24] a methodology is represented for the analysis and evaluation of owl:sameAs links in the web of data and ranking their reliability. Here, quality evaluation is done according to functional properties of involved instances.

A domain independent approach is represented in [8] to investigate if a link between two entities is correct or not. Their approach focuses on neighbors of those entities and compares their features together. The authors of [1] evaluated the quality of a linkset by measuring its completeness. In this work completeness of a linkset is a way for estimating possible losses in the completeness of combining two datasets.

B. Graph Theory Applications in other Domains

Graph theory has been used in different domains such as computer networks, software engineering and neural networks. In computer networks, graph theory has been used for different purposes. For example, modelling attacks against computer networks and choosing reactions as responses to these attacks [22], modelling topological structures of networks and studying problems such as routing, resource reservation and administration [33], using traffic dispersion graphs as an approach for monitoring, analyzing and visualizing network traffic [20] and evaluating robustness and connectivity in different topologies [6] are some of the studies which have used graph theoretical metrics in their works. In the software engineering domain a set of graph theoretical metrics have been defined on feature models in software product lines to prevent the manufacturing of low quality products [2]. In neural network domain, graph theoretical metrics have been used for analyzing human brain's MRI data to recognize some diseases and cure them [10, 11].

As mentioned above, there are few works which focus on link quality assessment between different datasets. However, none of them focuses on interlinking evaluation of a **single** dataset. One of the biggest challenges in type-independent evaluations in the past works is their low accuracy due to the lack of consideration of semantics. Another reason for the low accuracy is due to the sampling techniques that are used since this is inevitable due to the large volume of data in different datasets. On the other hand, most interlinking assessment methods are manual or semi-automatic and they cannot be applied to large datasets. They also cover few important factors in interlinking. In this paper, we discuss the importance of assessing the interlinking of a given dataset before its publication as a part of the linked open data cloud. The main objective of this study is to present a scalable approach to evaluate the interlinking of a single dataset with a desirable accuracy.

IV. Problem Statement and Contribution

Although interlinking is an important issue for the successful organic growth of the Web of Data, there are only a very limited number of research initiatives that focus on interlinking assessment of a dataset in the Web of Data.

Therefore, the main objective of this work is to propose a metric-driven framework that evaluates the interlinking of a single dataset automatically, before it is publicly published. For this purpose, we should first identify the main factors that affect interlinking such as link type, class type, restrictions and other semantic metadata. In other words, an attempt is made to observe and clearly formulate a set of metrics that are quantitatively measurable for a given dataset.

Then, we shall try to find meaningful statistical correlations between the proposed metrics and interlinking (which is not directly measurable) through empirical observational studies. Based on these correlations, a framework will be created that will be able to predict the interlinking of datasets by just observing their measurable metrics.

The fact is that none of the previous works has offered a solution for interlinking assessment of a **single** dataset. In this paper, we argue the importance of applying a metric based approach for assessing the interlinking of a given dataset before its publication as part of the linked open data cloud. The novel contributions of this work can be summarized as follows:

- We extend the definition of interlinking as an important quality dimension of any dataset in the context of LOD by proposing three sub-factors including: internal linking, external linking and link-ability;
- We systematically propose and validate a set of automatically measurable metrics for measuring the interlinking of a dataset;
- We introduce a novel approach for the interlinking assessment of a given dataset on LOD, which can be developed as an automatic tool.

V. The Proposed Approach

Our approach for interlinking assessment involves the measurement of internal linking, external linking and link-ability. To achieve this goal, the following steps must be performed:

1) Exploratory analysis of the previous and current works on link analysis and assessment, and also the application of graph theory in different domains;
2) Recognition of effective factors in interlinking assessment such as different types of links;
3) Devising a set of metrics for interlinking assessment of a single dataset;
4) Theoretical validation of the proposed metrics;
5) Selection of appropriate RDF datasets for evaluating the proposed approach;
6) Converting RDF datasets to RDF graphs;
7) Implementing an automated tool for measuring the proposed metrics;
8) Empirical evaluation of the proposed metrics by developing a questionnaire for subjective evaluation of the selected datasets;
9) Developing predictive (learning-based) techniques to find a correlation between the proposed metrics and interlinking;
10) Applying the final framework to predict the interlinking of datasets by measuring the metrics.

According to these steps, we have undertaken an exploratory analysis of the previous and current well-known works related to our research. We have systematically reviewed the works which evaluate the quality of links and classified them into two groups: link-independent assessment methods and specific link type evaluation methods. Studying related works, we realized that the existing approaches are not accurate enough.

In order to cover this problem, we tried to extract effective factors on interlinking of a dataset and we defined a set of metrics to measure these factors. The employed approach for metric definition process is Goal-Question-Metric (GQM) [3]. We have applied this approach for assessing the inherent quality dimensions of a dataset in our prior research [5]. In GQM, the goals are gradually refined into several questions and each question is then refined into metrics. Also, one metric can be used to answer multiple questions. Based on this approach, twenty different metrics are proposed in order to be used as measurement references for interlinking of a linked open dataset. Since there has been a few metrics focusing on interlinking evaluation in related works [17, 19, 28] we had to study metrics proposed in other areas in an attempt to benefit from them and adapt them for our work or define new metrics.

The main idea behind the design of these metrics has been comprehensiveness and simplicity. To achieve comprehensiveness, we have tried to cover as many influential factors in interlinking of a dataset as possible. Therefore, the metrics are proposed in three main groups: internal linking metrics, external linking metrics and link-ability metrics. We have also defined metrics by considering the graph structure of a dataset, types of links, number of nodes, and schema of a dataset (semantic metadata of a dataset) in order to achieve a high accuracy. Thus, proposed metrics are classified into graph-based metrics and semantic-based metrics. Graph-based metrics focus on the graph structure (topology) of a dataset while semantic-based metrics consider some semantic metadata such as link type, class type, restrictions and other semantic information. Furthermore, we have tried to define all the metrics in simple ratio scale. Taking into account the proposed metrics, it is understood that developing simple metrics is our secondary objective.

Table 1 provides the description of each of the proposed metrics. As mentioned before, some of the metrics are graph-based ones while others are semantic-based metrics. To better distinguish these two groups of metrics, the names of graph-based metrics are highlighted with Bold, and semantic-based metrics are Italic.

VI. Intermediate Results

An important outcome of this work is the evaluation of the assumption made in this study that a dataset within LOD can be automatically processed using metrics and evaluated for measuring its interlinking before release. It is expected that given the values of the metrics computed for a given dataset, the developed predictive models are able to estimate/predict the value of interlinking which is not directly measurable.

In order to put the proposed metrics into practical use, we have devised and implemented a tool that is able to automatically compute the metric values for any given input dataset. The code is available online at [31].

For better observation of metric behavior, different datasets from a variety of LOD domains are selected. These datasets are selected from NeOn project[1], which is under EU-FP6 program[2]. Table 2 presents the details of the selected datasets. Here, we have reported the results of our experiments over five datasets. Table 3 summarizes the results of the experiments conducted on them. As presented in Table 3, we were able to compute the values of all metrics using our automated metric computation tool, which shows that all the metrics have an objective nature.

Given the fact that the role of a metric is to measure an aspect of a dataset that would distinguish it from other datasets as much as possible and in light of the fact that all of the proposed metrics, except three of them, have different values on different datasets, it can be concluded that these are discriminatory metrics. Thus, these metrics are suitable for evaluating the interlinking of a dataset.

Only three of the metrics (M7, M13, M17) take on the same values on different datasets. Since, we need the schema of a dataset for computing semantic metrics and many of the available datasets do not present their schema, there are limitations in selecting datasets for evaluation. Thus, if we were able to find other datasets with schema, we would probably obtain different values.

VII. Evaluation Strategy

In the previous section, the proposed approach was presented. Here, we explain how to theoretically validate the metrics. The proposed metrics are validated from a measurement-theoretic perspective. Furthermore, we should subjectively evaluate the proposed model using experts' opinions, although we do not cover it in this research.

Generally, any kind of measure is a homomorphism from an empirical relational system to a numerical relational system [14]. Therefore, it is necessary to theoretically analyze these measures within the framework of measurement theory. There are two main groups of frameworks for the theoretical validation of metrics. The Frameworks in the first group are directly based on measurement theory principles [26]; while the ones in the second group present a set of properties which have to be satisfied by the metrics [9].

A metric is theoretically valid, if it could be validated by either measurement theory-based frameworks or property-based frameworks. In this work, we have validated our metrics according to a well-known framework in the second group, i.e. the property-based measurement framework [9]. This framework provides five types of metrics including size, length, complexity, coupling and cohesion and offers a set of desirable properties which a metric of that type should satisfy. Since, all of the proposed metrics are of the size type, according to the property-based measurement framework [9], they are expected to possess three main properties, namely, non-negativity (size cannot be negative), null value (size is expected to be null when a system does not contain any elements) and additivity (when modules of a system do not have any elements in common, we expect their size to be additive). These properties have been analyzed for the proposed metrics and it has been noted that all of the metrics satisfy these three properties.

In this step of this study, we have received the experts' subjective perceptions about interlinking of experimented datasets. Currently we are using machine learning techniques to find the correlations between the measured values of the metrics and interlinking.

VIII. Conclusion and Future Works

In this paper, one of the quality dimensions in linked data domain, i.e. interlinking is described. A new definition is presented for interlinking using three concepts including internal linking, external linking and link-ability and a set of metrics are proposed to assess them. In order to cover the challenges that exist in past works, we have focused on the graph structure of a dataset, types of links, nodes and schema of a dataset (semantic information) to assess the interlinking of a dataset with desired accuracy. To put the proposed metrics into practice, we have implemented an automated tool and computed the metric values for five datasets across a variety of LOD domains with different sizes. Finally, the suitability of these metrics has been discussed.

We are currently focusing on the empirical evaluation of the proposed metrics by capturing experts' opinions about interlinking for all of the datasets used in this experiment. We also consider developing statistical models that would take the values of the metrics proposed in this paper for each dataset into account and predict the interlinking of the dataset once it is integrated into the LOD.

Improving the proposed metrics, implementing the automatic tool in a parallel manner, defining new metrics by considering other issues which affect interlinking, considering other quality dimensions besides interlinking, such as security, accuracy, etc. can be considered as future works.

TABLE I. DEFINITION OF PROPOSED METRICS FOR INTERLINKING.

Dimension	Metric Name	Formula	Description
Internal-Linking	**Graph-Completeness**	$\frac{\textbf{No. of edges}}{\textbf{No. of edges in a complete graph}}$	The ratio of the number of edges over the number of edges in a clique. It is an estimation of internal linking. The higher the value, the more internally connected it is.
	Node-Degree	$\frac{\textbf{No. of edges}}{\textbf{No. of nodes}}$	The ratio of the number of edges over the number of nodes in a graph. This metric is an estimation of nodes' degree. The higher the value, the more connected are the nodes to each other.
	Entity-In	$\frac{\textbf{No. of entity nodes with internal edges} > \theta}{\textbf{No. of all entity nodes}}$ $(\boldsymbol{\theta} = \textbf{Mean internal edges of entity nodes})$	The ratio of the number of entity nodes with high internal edges over the number of all entity nodes in a graph. It counts the entity nodes which have a high internal edge degree. This metric is an estimation of internal linking and the higher the value, the more internally linked the graph is.
	Entity-Out	$\frac{\textbf{No. of entity nodes with external edges} > \theta}{\textbf{No. of all entity nodes}}$ $(\boldsymbol{\theta} = \textbf{Mean external edges of entity nodes})$	The ratio of the number of entity nodes with high external edges over the number of all entity nodes in a graph. This metric counts the entity nodes which have a high external edge degree. The higher the value, the more internally connected the nodes are
External-Linking	**External-Links**	$\frac{\textbf{No. of external links}}{\textbf{No. of all edges and links}}$	The ratio of the number of external links over the number of all links and edges in a graph. The higher the value, the more externally linked it is.
	External-Nodes	$\frac{\textbf{No. of entitiy nodes with external links}}{\textbf{No. of all entity nodes}}$	The ratio of the number of nodes with external links out of graph over the number of all nodes in a graph. This metric counts the nodes which have external links to other datasets. This can be an estimation of external linking.
	External-SameAs	$\frac{\textbf{External sameAs links to other resources}}{\textbf{Total No. of external links}}$	The ratio of the number of external SameAs links to other datasets over the number of all external links. Out of all external links, owl:sameAs links have a more important role in connecting related concepts. So the higher the value, the more external linking we have.
Link-Ability	*Parent-Class*	$\frac{\textbf{No. of parent classes}}{\textbf{Total No. of classes}}$	The ratio of the number of parent classes over the number of all classes. Parent classes presents more generic concepts, and generic concepts have more potential to be linked from other datasets.
	Restriction-Number	$\frac{\textbf{No. of properties without restriction}}{\textbf{No. of all properties}}$	The ratio of the number of properties without restrictions over the number of all properties. The less we have restrictions on a dataset, the more potential it has to be linked from other datasets.
	High-CC	$\frac{\textbf{No. of nodes with clustering coefficient} > \theta}{\textbf{No. of all nodes}}$ $(\boldsymbol{\theta} = \textbf{Mean Clustering Coefficient})$	The ratio of the number of nodes with high clustering coefficient over the number of all nodes in a graph. Nodes with a high clustering coefficient are the key nodes in a graph. The more we have key nodes in a dataset, the more potential it has to be linked from other datasets.
	URI-Length	$\frac{\textbf{No. of nodes with short URIs}}{\textbf{No. of all nodes}}$	The ratio of the number of nodes with short URIs over the number of all nodes in a graph. Short URIs are more readable. The more we have short URIs in a dataset, the more potential it has to be linked from other datasets.

[1] Networked Ontology

[2] Europian Sixth Framework Programme: http://ec.europa.eu/research/fp6/index_en.cfm

Dimension	Metric Name	Formula	Description
	High-In-DC	$\frac{\textbf{No. of nodes with internal degree centrality} > \theta}{\textbf{Total No. of nodes}}$ **(θ = Mean Internal Degree Centrality)**	The ratio of the number of nodes with high internal degree centrality over the number of all nodes in a graph. This metric count the nodes which have a high internal degree centrality. (We distinguish internal and external degree centrality because the LOD graph is a directed graph). Nodes with high degree centrality are popular and important nodes in a graph. The more we have important nodes in a dataset, the more potential it has to be linked from other datasets.
	High-Out-DC	$\frac{\textbf{No. of nodes with external degree centrality} > \theta}{\textbf{Total No. of nodes}}$ **(θ = Mean External Degree Centrality)**	The ratio of the number of nodes with high external degree centrality over the number of all nodes in a graph. It is similar to the previous metric, but it focuses on external degree centrality.
	ObjPrp-Links	$\frac{\textbf{No. of object property edges}}{\textbf{Total No. of edges}}$	The ratio of the number of object property edges over the number of all edges in a graph. Object property edges connect two entities to each other, while datatype property edges connect an entity and a literal. The more we have object property edges in a dataset, the more potential it has to be linked from other datasets.
	Entity-Number	$\frac{\textbf{No. of entity nodes}}{\textbf{No. of all nodes}}$	The ratio of the number of entity nodes over the number of all nodes in a graph. This metric counts the entity nodes. Entity nodes are the ones which can be linked from other datasets, while literal nodes and blank nodes cannot be linked from other datasets. The more we have entity nodes in a dataset, the more potential it has to be linked from other datasets.
	Distinct-Objects	$\frac{\textbf{No. of edges with the label OWL: differentFrom}}{\textbf{No. of all edges}}$	The ratio of the number of edges with the label owl:differentFrom over the number of all edges in a graph. owl:differentFrom edges are an estimation of diversity of instances. The more different instances we have in a dataset, the more potential it has to be linked from other datasets.
	Class-Size	$\frac{\textbf{No. of entities}}{\textbf{No. of classes}}$	The ratio of the number of entities over the number of classes. This metric is an estimation of class size. The bigger a class is, the more instances it has and so the more potential it has to be linked from other datasets.
	URI-Type	$\frac{\textbf{No. of nodes with derefrencable URIs}}{\textbf{No. of all nodes}}$	The ratio of the number of nodes with dereferenceable URIs over the number of all nodes in a graph. This metric counts the nodes with dereferenceable URIs. The more we have dereferenceable URIs in a dataset, the more potential it has to be linked from other datasets.
	Vocab-Importance	$\frac{\textbf{No. of important used vocabs}}{\textbf{No. of all used vocabs}}$	The ratio of the number of important used vocabularies over the number of all used vocabularies. This metric counts the important vocabularies which are used in the schema of a dataset. Widespread vocabularies are important in relating concepts in different datasets. The more we have common vocabularies in a dataset, the more potential it has to be linked from other datasets.
	Node-Description	$\frac{\textbf{No. of descriptive nodes}}{\textbf{No. of all nodes}}$	The ratio of the number of nodes which provides some descriptive information over the number of all nodes in a graph. This metric counts nodes which contain descriptive information. These descriptions increase readability, so the more we have descriptive nodes in a dataset, the more potential it has to be linked from other datasets.

TABLE II. DETAILS OF THE DATASETS USED IN OUR EXPERIMENTS.

Dataset	No. of triples	No. of instances	No. of classes	No. of properties
FAO Water Areas	5,365	293	7	19
Water Economic Zones	25,959	693	22	127
Large Marine Ecosystems	6,006	358	9	31
Geopolitical Entities	22,725	312	11	101
ISSCAAP Species Classification	368,619	23,856	22	93

TABLE III. RESULTS OF THE EXPERIMENTS.

No	Metrics	FAO Water Areas	Water Economic Zones	Large Marine Ecosystems	Geopolitical Entities	ISSCAAP Species Classification	Mean	STDEV
1	Graph-Completeness	0.004	0.001	0.006	0.001	0.00	0.002	0.002
2	Node-Degree	3.317	1.519	1.521	2.540	1.982	2.175	0.683
3	Entity-In	0.090	0.079	0.204	0.349	0.001	0.144	0.121
4	Entity-Out	0.416	0.290	0.295	0.031	0.005	0.207	0.161
5	External-Links	0.487	0.411	0.443	0.997	0.001	0.467	0.316
6	External-Nodes	0.196	0.186	0.189	0.625	0.002	0.239	0.282
7	External-SameAs	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	High-In-DC	0.211	0.034	0.164	0.156	0.002	0.113	0.080
9	High-Out-DC	0.180	0.179	0.151	0.044	0.005	0.111	0.073
10	High-CC	0.211	0.121	0.160	0.181	0.006	0.135	0.071
11	ObjPrp-Links	0.403	0.492	0.453	0.003	0.980	0.466	0.310
12	Entity-Number	0.197	0.188	0.201	0.047	0.971	0.320	0.330
13	Distinct-Objects	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	Class-Size	46.000	80.600	29.333	40.363	150.800	69.419	40.865
15	Parent-Class	0.684	0.333	0.333	0.920	0.333	0.520	0.241
16	Restriction-Number	0.294	0.909	1.00	0.189	0.954	0.669	0.351
17	URI-Type	1.00	1.00	1.00	1.00	1.00	1.00	0.000
18	URI-Length	0.105	0.781	0.215	0.612	0.161	0.374	0.270
19	Vocab-Importance	0.714	0.454	0.555	0.800	0.555	0.615	0.124
20	Node-Description	0.163	0.000	0.076	0.000	0.000	0.047	0.064

REFERENCES

[1] R. Albertoni and A.G. Pérez, "Assessing linkset quality for complementing third-party datasets," Proceedings of the Joint EDBT/ICDT 2013 Workshops. ACM, 2013.

[2] E. Bagheri and D. Gasevic, "Assessing the maintainability of software product line feature models using structural metrics," Software Quality Journal 19.3, pp. 579-612, 2011.

[3] V. R. Bassili, G. Caldiera, and H.G. Robach, "The goal question metric approach," In Encyclopedia of software engineering, ed: John Wiley & Sons, pp. 528-532, 1994.

[4] C. Batini and M. Scannapieca, "Data quality: concepts, methodologies and techniques," Springer, 2006.

[5] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," Journal of Theoretical and Applied Electronic Commerce Research, Vol. 9, Issue 2, pp. 64-79, 2014.

[6] A. Bigdeli, A. Tizghadam, and A. Leon-Garcia, "Comparison of network criticality, algebraic connectivity, and other graph metrics," Proceedings of the 1st Annual Workshop on Simplifying Complex Network for Practitioners. ACM, 2009.

[7] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," International Journal on Semantic Web and Information Systems (IJSWIS) 5.3, pp. 1-22, 2009.

[8] C. Böhm, Enriching the Web of Data with topics and links, PhD Thesis, Universitätsbibliothek, Hamburg, Germany, 2013.

[9] L. C. Briand, S. Morasca, and V.R. Basili, "Property-based software engineering measurement," Software Engineering, IEEE Transactions, 22(1): pp. 68-86, 1996.

[10] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," Nature Reviews Neuroscience10.3, pp. 186-198, 2009.

[11] L. Deuker, "Reproducibility of graph metrics of human brain functional networks," Neuroimage 47.4, pp. 1460-1468, 2009.

[12] W. Ge, J. Chen, W. Hu, and Y. Qu, "Object Link Structure in the Semantic Web," In ESWC (2), volume 6089 of Lecture Notes in Computer Science, pp. 257–271, 2010.

[13] M.J. Eppler and D. Wittig, "Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years," 5th International Conference on Information Quality, 2009.

[14] N.E. Fenton and S.L. Pfleeger, Software metrics: a rigorous and practical approach, PWS Publishing Co, 1998.

[15] L. Finkelstein, "Widely, strongly and weakly defined measurement," Measurement, 34(1), pp. 39-48, 2003.

[16] C. Guéret, P.T. Groth, F.V. Harmelen, and S. Schlobach, "Finding the Achilles Heel of the Web of Data: Using Network Analysis for Link-Recommendation," 9th International Semantic Web Conference, pp. 289–304, 2010.

[17] C. Gueret, P. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," The Semantic Web: Research and Applications, 2012.

[18] O. Hartig, "Trustworthiness of data on the web," In Proceedings of the STI Berlin & CSW PhD Workshop. Citeseer, 2008.

[19] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, "An empirical survey of Linked Data conformance," 2012.

[20] M. Iliofotou, "Network monitoring using traffic dispersion graphs (tdgs)," Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007.

[21] ISO, ISO/IEC 25012- Software engineering - Software product Quality Requirements and Evaluation (SQuaRE), in Data quality model, 2008.

[22] M. Jahnke, C. Thul, and P. Martini, "Graph based metrics for intrusion response measures in computer networks," Local Computer Networks, 32nd IEEE Conference on. IEEE, 2007.

[23] C. Joslyn, B. Adolf, S.A. Saffar, J. Feo, E. Goodman, D. Haglin, and D. Mizell, "High performance semantic factoring of giga-scale semantic graph databases," Contribution to Semantic Web Challenge at ISWC, 2010.

[24] L. Papaleo, N. Pernelle, and F. Saıs, "On evaluating the quality of RDF identity links in the LOD," In proceedings of IC'2014 Workshop "From Open Sources to Web of Data, 2014.

[25] H. Paulheim, "Identifying wrong links between datasets by multi-dimensional outlier detection," In proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings co-located with 11th Extended Semantic Web Conference, 2014.

[26] G. Poels and G. Dedene, "Distance-based software measurement: necessary and sufficient properties for software measures," Information and Software Technology, 42(1), pp. 35-46, 2000.

[27] E. Ruckhaus, O. Baldizán, and M. Vidal, "Analyzing linked data quality with LiQuate," On the Move to Meaningful Internet Systems: OTM Workshop, 2013.

[28] C. Stadler, C. Gueret, J. Lehmann, P. Groth, and A. Jentzsch, 2012. "LATC: Linking open data around the clock," http://latc-project.eu/

[29] W3C, 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-rdf-graph

[30] Y. Wand and R.Y. Wang, "Anchoring data quality dimensions in ontological foundations," Communications of the ACM, 39.11, pp. 86-95, 1996.

[31] N. Yaghouti, The code of metrics calculation tool with the datasets. Available: https://sourceforge.net/projects/interlinkingassessment/

[32] A. Zaveri, A. Rula, A. Maurino, R.P.J. Lehmann, S. Auer, and P. Hitzler, "Quality Assessment for Linked Data: A Survey," Accepted in Semantic Web Journal: http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey, in press.

[33] E.W. Zegura, K.L. Calvert, and M.J. Donahoo, "A quantitative comparison of graph-based models for Internet topology," IEEE/ACM Transactions on Networking (TON) 5.6, pp. 770-783, 1997.