Methods for Intrinsic Evaluation of Links in the Web of Data

Cristina Sarasua[1], Steffen Staab[1,2], and Matthias Thimm[1]

[1] Institute for Web Science and Technologies, University of Koblenz-Landau
[2] WAIS Research Group, University of Southampton
{csarasua,staab,thimm}@uni-koblenz.de

Abstract. The current Web of Data contains a large amount of interlinked data. However, there is still a limited understanding about the quality of the links connecting entities of different and distributed data sets. Our goal is to provide a collection of indicators that help assess existing interlinking. In this paper, we present a framework for the intrinsic evaluation of RDF links, based on core principles of Web data integration and foundations of Information Retrieval. We measure the extent to which links facilitate the discovery of an extended description of entities, and the discovery of other entities in other data sets. We also measure the use of different vocabularies. We analysed links extracted from a set of data sets from the Linked Data Crawl 2014 using these measures.

Keywords: Data Integration, Links, Quality, Monitoring, Semantic Web

1 Introduction

Linked Data principles encourage data publishers to connect the resources in their data sets to other resources "so that more things can be discovered"[3]. With the increasing number of available data sets and links between them [8, 12], it becomes highly important to observe the extent to which existing links have desirable properties, as we need to ensure high quality to encourage the usage of Linked Data. Links should (i) follow the recommendations that apply to high quality data [14] (i. e. links should be accessible, syntactically valid, and semantically accurate), and (ii) links should enable the discovery of "more things", facilitating new insights from the data. Established data-driven quality assurance methodologies [10, 14, 11] suggest that the key steps for improving the status quo are: the definition of measures, the analysis of measurements and the subsequent monitoring of updates. So, to be able to analyse the quality of links, we need measures that help us assess all relevant quality aspects, including (i) and (ii).

Previous empirical studies on the adoption of Linked Data principles [12, 6] report on the number of outgoing and incoming links of data sets, and the most frequently used predicates in RDF links. Recently, Hu et al. [7] studied degree distributions, as well as missing links in Bio2RDF based on symmetry and transitivity. Neto et al. [9] focused on the analysis of dead links in schema and entity link triples published in the Web of

[3] Berners-Lee, T. Linked Data Principles http://www.w3.org/DesignIssues/LinkedData.html

Data. While these studies, together with the findings provided by smaller evaluations of other link assessment methods focusing on (i) (e. g. Guéret et al. [4] and other quality dimensions like completeness [2] provide a characterization of existing links), they do not allow for assessing how many new things might be made discoverable thanks to the links (ii).

In this paper, we provide a framework for link analysis that takes into account principles of data integration in the Web of Data, addressing (ii). We suggest measures that focus on data quality dimensions inherent in the data, while extrinsic assessment would take into account the needs a user has in his specific context (cf. [14]). More specifically, our measures examine the effect that links have on entities (and consequently on data sets). We measure the extent to which links facilitate the discovery of an extended description of entities, and the discovery of other entities in other data sets. We also measure if they add different vocabularies (cf. Section 4.2) to the description of entities. Our measures are grounded on foundations of the field of Information Retrieval, as we acknowledge redundancy when we measure the gain in description, connectivity and number of used vocabularies. More precisely, the contributions of this paper are:

1. We identify a set of principles for data interlinking in the Web of Data (Section 3).
2. We define a set of measures to analyse available links in terms of these principles (Section 4).
3. We demonstrate the feasibility of the proposed framework with the implementation of the measures and carry out an empirical analysis of links extracted from the Linked Open Data Crawl [12] (Section 5).

2 Preliminaries

We introduce in this section the terminology and notation.

Definition 1. ***RDF Quadruple****: Given $\mathcal{U}$, a finite set of HTTP URIs, representing resources, $\mathcal{L}$ a finite set of literal values, and a finite set of blank nodes $\mathcal{B}$ where $U \cap \mathcal{L} = \mathcal{U} \cap \mathcal{B} = \mathcal{L} \cap \mathcal{B} = \emptyset$, a quadruple (s, p, o, c) is any element of the data space $Q = (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{L} \cup \mathcal{B}) \times \mathcal{U}$. s, p, o is a triple statement describing s, while c is the context (denoted by a URI) in which the triple is defined.*

Definition 2. ***RDF Data set****: An RDF data set D_c is a set of quadruples grouped by some context c $D_c \subseteq \{(s, p, o, c) \in Q\}$, where Q is the set of all quadruples.*

Definition 3. ***Home****: Given C the set of all contexts, and an entity (either a blank node or URI), $home : \mathcal{B} \cup \mathcal{U} \mapsto C$ is the function that maps the entity to the context c where the entity is defined. Note that when x is a vocabulary term (e. g. a class or a property), the c returned by $home(x)$ is the identifier of the vocabulary where the term x was defined.*

The $home$ function is customisable. For example, it can be defined to match the notion of data sets in the Linked Open Data literature [12], or it can be defined to match the graphs in data sets—the graphs in the SPARQL and N-Quads specifications. In this paper, we stick to the LOD cloud diagram[4] and analyse links on a data set basis.

For representing the relation between entities of different data sets, we define:

Definition 4. ***Link****: A link of D_c is a quadruple $(s, p, o, c) \in D_c$ such that $s \in \mathcal{U}, o \in \mathcal{U}, home(s) = c, home(o) \neq c$*

Definition 5. ***Interlinking****: The interlinking I_c of a data set D_c is the set of all links going out from D_c to any other data set: $I_c = \{(s, p, o, c) \in D_c | \; home(s) = c, home(o) \neq c\}$.*

To formally define our measures, we use a relational algebra-like notation. For this purpose we define selection σ, projection π and join $\bowtie$ as follows:

Definition 6. ***Selection****: Given $X \subseteq D_c$, a selection $\sigma_h(X)$ is the quadruples from X that satisfy a selection predicate h: $\sigma_h(X) = \{(s, p, o, c)|(s, p, o, c) \in X \wedge h(s, p, o, c) = true\}$*

Example 1. : We can select the quadruples of the data set D_c that are `owl:sameAs` links by $\sigma_{p=owl:sameAs}(D_c) = \{(s, p, o, c)|(s, p, o, c) \in D_c, p = owl : sameAs\}$

Definition 7. ***Projection****: Given $X \subseteq D_c$, and Y a subset of the elements in the quadruples in X, a projection $\pi_Y(X)$ on attributes Y is the subset of X including the elements Y: $\pi_Y(X) = \{(s, p, o, c)[Y]|(s, p, o, c) \in X\}$*

Example 2. : We can obtain the projection of all the entities appearing in the predicate and object positions of the quadruples of the data set D_c by $\pi_{p,o}(D_c) = \{(p, o)|(s, p, o, c) \in D_c\}$

Definition 8. ***EquiJoin****: Given $X_1 \subseteq D_1$ and $X_2 \subseteq D_2$, the Equi join of the two sets is the set of elements such that: $X_1 \bowtie_{X_1.o\theta X_2.s} X_2 = \{(X_1.s, X_1.p, X_1.o, X_2.s, X_2.p, X_2.o, c) \mid X_1.o = X2.s\}$*

Example 3. : In Table 1 case (I), the equijoin of the two quadruples on the name and the link is the 7-tuple “d1:nn owl:sameAs d2:nn d2:nn rdfs:label “Natasha” d1 .”.

Now, we may re-state **our task** at hand as follows: Given a data set D_c containing the interlinking I_c, our task is to compare D_c and $D_c \backslash I_c$ and analyse the value that I_c gives to the data in terms of the principles for data interlinking in the Web of Data described in the following section.

3 Principles for Data Interlinking in the Web of Data

The main reason to connect data sets is to enable their joint search, browsing or querying. As in any information system, when a user queries Linked Data it is important that she: (**n1**) finds all entities she is interested in (recall); (**n2**) finds only entities she is

[4] `http://lod-cloud.net`

	Source data set	Target data set(s)
	Entity Description	
(I)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation d1:stanford . d1:nn swrc:publication d1:p2012-1 . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natalya F. Noy"" . d2:nn dbo:affiliation d1:googleinc . d2:nn swrc:publication d2:p2015-1 .
(II)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn . d2:ms foaf:name "Mark Smith"	d2:nn foaf:name "Natalya F. Noy" . d2:nn cito:likes d2:sfo . d2:nn swc:holdsRole swc:Chair
(III)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natasha Noy" .
	Entity Connectivity	
(IV)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation dbr:Stanford_University . d1:nn owl:sameAs d2:p1 . d1:nn owl:sameAs d3:p5 . d1:nn owl:sameAs d4:p1 .	d2:p1 foaf:name "Natasha Noy" . d2:p1 dbo:affiliation dbr:Stanford_University . d3:p5 foaf:name "Natasha Noy" . d3:p5 dbo:affiliation dbr:Stanford_University . d4:p1 dbo:affiliation dbr:Stanford_University .
(V)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation dbr:Stanford_University . d1:nn owl:sameAs d2:p1 .	d2:p1 foaf:name "Natasha Noy" . d2:p1 dbo:affiliation dbr:Stanford_University . d3:p5 foaf:name "Natasha Noy" . d3:p5 dbo:affiliation dbr:Stanford_University .
	Vocabularies Involved in the Description	
(VI)	d1:nn foaf:name "Natasha Noy" . d2:nn rdf:type foaf:Person . d1:nn owl:sameAs d2:nn .	d2:nn sioc:creator_of d2:post2 . d1:nn rdf:type proton:Human . d2:nn vivo:teachingOverview "Natasha Noy was a tutor in the SSSW08 summer school" .
(VII)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natasha Noy" . d2:nn foaf:currentProject d2:bioportal . d2:nn foaf:pastProject d2:protege .

Table 1: Examples of different interlinking cases.

interested in (precision); (**n3**) is able to understand the relationship between entities in the Web; (**n4**) finds answers to all her questions no matter how heterogeneous in syntax, structure and semantics the questions are.

The existence of high quality links between entities can contribute to a better fulfilment of the aforementioned needs (n1-n4). In order to understand the way links can help, let us consider various interlinking examples (from (I) to (VII)) shown in Table 1. We analyse each of the examples, and derive from them desired properties for links (i. e. principles P1-P3).

Entity Description In case (I) we see two entities linked via an `owl:sameAs` link. The two connected entities have different names, but represent the same person (Natalya F. Noy, also known as Natasha Noy informally). The source data set contains the publications that Natasha wrote when she worked at Stanford, and the target data set contains publications she has written while working at Google Inc. If we search for the publications written by Natasha and only consider the source data set, we exclusively see her Stanford publications. If we consider the link connecting the two entities referring to Natasha, we are able to also find her Google publications, giving us higher recall (**n1**).

In case (II) the two entities are also connected via an `owl:sameAs`. The target data set contains data about conferences and program committees, while the source data set does not contain this kind of data. If we look for persons who have been chairs of scientific events, and we only take into account the source data set, we are not able to find any person because we lack the information about the chairs of the events. In an Information Retrieval scenario, we would use query relaxation techniques, and the search query would be reformulated as a search for persons. The result would include the entities for Natasha Noy and Mark Smith (who is a student assistant and was never a chair). Conversely, if we consider the link, we have relevant information for the query and only Natasha is retrieved in the results. Therefore, in this case the link enables us to have higher precision (**n2**).

Observation: These two cases, have something in common: the links (s, p, o, c) extend the description of entities s. The description of an entity is the set of quadruples with s as subject, and literals, URIs and blank nodes as objects (cf. Section 4.2). When the linked data sets provide redundant information, links do not help in recall, nor in precision. Example (III) is a clear example of a scenario where we have redundant information and the description is not extended. Therefore, we formulate the first principle as:

P1: Try to create an interlinking that extends the description of entities of the source data set.

Entity Connectivity Case (IV) connects the entity referring to Natasha in d1 to the corresponding entities representing Natasha in data sets d2, d3 and d4. While these links do not extend description of the entity in d1 (i. e. they do not follow the Principle P1), they help in understanding the relationship between the entities in the Web of Data (**n3**). This understanding is necessary when for example, a change in the affiliation of Natasha is materialised in d1 to update her affiliation. The descriptions in d2, d3 and d4 could be subsequently changed, in order to keep the data up-to-date.

Observation: In (IV), we can see the importance of creating multiple links from the same entity to different external entities and data sets, increasing its connectivity (cf. Section 4.2). In Case (V), which is similar to case (IV) but without the links to d3 and d4, we see that if the links from d1:nn to the entities in d3 and d4 do not exist (as in case (V)), it is harder to reach the entities in other data sets that would need to be updated. This is similar in cases where the links are created to group entities, or to enable the browsing of different types of entities. We formulate the second principle as:

P2: Try to create an interlinking that increases the number of entities and data sets that source entities are connected to.

Heterogeneity of Descriptions Case (VI) shows an example where the entity representing Natasha is connected via an `owl:sameAs` link to its corresponding entity in d2. The entity in d2 is described with vocabularies that are different from d1's vocabularies. In contrast, in case (VII) the entity in d2 contains a description that adds new information to the description of d1 (satisfies P1) but uses the same vocabulary as in d1 (i. e. FOAF).

Observation: in (VI), links help in answering a wider range of queries that might be formulated in different application contexts (**n4**). Using different vocabularies we are able to use and analyse entities from multiple perspectives. Hence, the third principle is:

P3: Try to create an interlinking that makes the source entities have a description with a higher number of vocabularies in their description.

These principles are not independent from each other. Principles P2 (entity connectivity) and P3 (vocabularies) are specializations of P1 (entity description). For some types of links (non-identity links), creating links to new entities in new data sets (P2), means that the description of the source entity is extended (P1). However, that does not necessarily happen the other way round. Analogously, if one uses further vocabularies in the links between entities (P3), the description of the source entity will be extended (P1). Therefore, when we analyse data in terms of these principles, we consider them as a three level test, in which having passed P1 is positive, but having passed P2 and P3, too, is even more positive. We do not claim that these principles are complete, and they may be extended.

4 Intrinsic Measures for Assessing the Quality of Links

The measures that we define do not provide an absolute assessment of the quality of links. That is, a particular measurement is not good or bad. Instead, we provide measures for a comparative assessment: we acknowledge that one interlinking is better than another in some dimension that we observe with regard to the principles in the previous section. It is up to the person or application inspecting the measurements to interpret its meaning, and make a decision based on it (e. g. a data publisher willing to improve her interlinking and using our measurements as a guide to decide where to start from).

We distinguish between descriptive statistics that give an overview of the size and the elements in I_c (see Section 4.1), and measures that assess the way the links in the interlinking I_c of the data set D_c follow the aforementioned principles (see Section 4.2).

4.1 Basic Descriptive Statistics

In order to describe basic properties of the interlinking of a data set, we use basic statistics proposed by related work (e. g. Void Vocabulary[5] and LOD Stats[6]), to compute the volume of the interlinking ($|I_c|$), and the distribution of linksets ($\{(x, |\sigma_{p=x}(I_c)|)\}$).

[5] `https://www.w3.org/TR/void/`

[6] `http://stats.lod2.eu/links`

4.2 Principles-based Measures

Since we would like to study the effect that links have on the entities of the source data set, our measures analyse links grouped by source entities. Note that in our analysis we focus on entities $e \in D_c$ such that $\nexists\ (e, rdf : type, rdfs : Class) \in D_c$. So, we look at the interlinking of individuals and not at vocabulary terms.

4.2.1 Two views of the quadruples about entities For each entity e, we distinguish two views of the set of quadruples that state something about e: the description view and the connectivity view of an entity.

Description view This view focuses on all the quadruples in X describing the entity e.

We define the description of an entity e in $X \subseteq D_c$ as the projection that selects the predicates and objects from the set of quadruples of X about e, and entities defined to be identical to e (usually defined via the predicates `owl:sameAs` or `skos:exactMatch`).

$$desc(e, X) = \pi_{(p,o)}(\sigma_{s=e}(X)) \cup \pi_{(Q.p,Q.o)}(\sigma_{X.p=identity}((X \bowtie_{X.o=Q.s} Q))) \quad (1)$$

In order to have a more detailed view of the description, we differentiate between the entity's classification (i. e. the quadruples referring to the `rdf:type` of the entity):

$$classif(e, X) = \sigma_{p=``rdf:type''}(desc(e, X)) \quad (2)$$

and the rest of the description:

$$descm(e, X) = desc(e, X) \backslash classif(e, X) \quad (3)$$

Example 4. In Table 1(VI), classif(d1:nn,D_1'= { (rdf:type, foaf:Person), (rdf:type, proton:Human)} and descm(d1:nn,D_1') = {(foaf:name, “Natasha Noy”), (owl:sameAS, d2:nn), (foaf:name, “Natasha Noy”),(sioc:creator_of, d2:post2),(vivo:teachingOverview, "...")}

Additionally, we make a specification of $descm(e, X)$ and define $descmp$ to project only the predicates (instead of the predicates and values as in $descm(e, X)$).

$$descmp(e, X) = \pi_{(p)}(descm(e, X)) \quad (4)$$

To identify the vocabularies used in the description of an entity we define:

$$vocabd(e, X) = \{home(p) | (p, o) \in desc(e, X)\} \quad (5)$$

Connectivity view This view focuses on the quadruples that state the connections between the entity e and other entities. Note that this view is a subview of the description view. Here, we ignore the quadruples about e, with literal values and quadruples describing identical entities to e.

We define the entity connectivity of an entity e in $X \subseteq D_c$ as the set containing the entities targeted from e:

$$econn(e, X) = \pi_o(\sigma_{s=e}(X)) \tag{6}$$

Analogously, we define the data set connectivity of an entity e in $X \subseteq D_c$ as the set containing the data sets targeted from e:

$$dconn(e, X) = \{home(o)|\ o \in econn(e, X)\} \tag{7}$$

Example 5. In Table 1(V), econn(d1:nn,D_1)={dbr:Google,d2:p1,d3:p5,d4:p1}, whereas dconn(d1:nn,D_1)={dbr,d2,d3,d4}

4.2.2 Measuring the principles at an entity and data set level Now that we have defined the sets for the description and the connectivity views (Section 4.2, let us look at the measures that are interesting to be applied on these sets, in order to state the extent to which the links going out of entity e follow principles P1, P2 and P3. We use the notation S to refer to any of the sets above.

Measure size Measuring the size of data is a standard way of characterizing data. We measure the size of each of the sets above by calculating the cardinality of the corresponding set (i. e. $|S|$).

Measure diversity When we observe if entities get their description (i. e. $classif(e, X)$ and $descm(e, X)$) extended when considering the links, we aim to identify redundancy. Furthermore, when we analyse the targeted entities and data sets, as well as the vocabularies used in the description and the links, we want to measure diversity both without and with links. In these two situations, we may encounter repetitions in the classification, the description, the entity connectivity, the data set connectivity, and the vocabularies used in the description. Therefore, we extend the notion of our sets and model multisets (allowing repeated elements), counting the number of times each element appears in the multiset: (S, n) where n is $n : S \mapsto \mathbb{N}_{\geq 1}$, a function that given an $s \in S$ tells the number of times that s appears in S.
Diversity is a measure that takes into account the number of different (and non redundant) types of elements in a set, and at the same time takes into account how equally distributed the elements of each type are present in the set. For these two purposes, we use the Shannon Entropy [13], a standard measure used in Information Theory to measure diversity.

$$H(ELS) = -\sum_{s \in S} prob(ELS = s) \times \log prob(ELS = s) \tag{8}$$

A low entropy value means that there is little diversity in the data. Note that $H(x) \geq 0$. In $classif(e, X)$, and $descm(e, X)$ repeated statements appear only when we consider the quadruples of the target data sets, because in one data set quadruples are supposed to be unique. Still, we calculate entropy to be able to signal redundancy when we compare the description with and without links.

Compare measurements In order to accomplish our task of comparing measurements considering the links vs. not considering the links, we differentiate between the total set of quadruples in D_c, and the set of internal quadruples defined as:

$$D_c^{\text{internal}} = D_c \backslash I_c$$

We compare a measurement on D_c vs. the measurement on D_c^{internal} by subtracting the latter to the former.

Based on these three rationales, we define the following list of measures (cf. Table 2) to analyse the way links follow the principles. To measure the extension in classification, description, entity connectivity, data set connectivity and the increase in the number of vocabularies employed, we use the difference in entropy. For example, to check if the classification is extended, we define two random variables CS (in D_c^{internal}) and CS' (in D_c) and calculate $H(CS') - H(CS)$. The difference is zero when there is no information gain, negative when redundant information is gained, and positive otherwise.

ID	Principle/Description	Vars.	Definition
m11a	P1 #classes	-	$\|classif(e, D_c^{\text{internal}})\|, \|classif(e, D_c)\|$
m11c	P1 Classification Extension (entropy)	CS, CS'	$H(CS') - H(CS)$
m12a	P1 #predicate-objects	-	$\|descm(e, D_c^{\text{internal}})\|, \|descm(e, D_c)\|$
m12c	P1 Description Extension	DE, DE'	$H(DE') - H(DE)$
m13a	P1 #predicates	-	$\|descmp(e, D_c^{\text{internal}})\|, \|descmp(e, D_c)\|$
m13c	P1 Predicate Description Extension	DEP, DEP'	$H(DEP') - H(DEP)$
m21a	P2 #targeted entities	-	$\|econn(e, D_c^{\text{internal}})\|, \|econn(e, D_c)\|$
m21c	P2 Entity connectivity Extension	EC, EC'	$H(EC') - H(EC)$
m22a	P2 #targeted data sets	-	$\|dconn(e, D_c^{\text{internal}})\|, \|dconn(e, D_c)\|$
m22c	P2 Data set connectivity Extension	DC, DC'	$H(DC') - H(DC)$
m31a	P3 #Vocabularies in desc.	-	$\|vocabd(e, D_c^{\text{internal}})\|, \|vocabd(e, D_c)\|$
m31c	P3 Increase #Vocabularies Used (entropy)	VD, VD'	$H(VD') - H(VD)$

Table 2: List of measures to analyse the fulfilment of data interlinking principles. Columns show: the name of the measure, the principle the measure belongs to, the random variables defined for the measure, and the formal definition of the measure.

5 Empirical Analysis

To demonstrate the feasibility of our approach for profiling the quality of links in the Linked Open Data cloud, we have implemented the measures in the SeaStar framework, which uses Java, the NxParser to parse N-Quads, and Jena for handling RDF data[7].

[7] Source code: `https://github.com/criscod/SeaStar`

5.1 Data

We use data from the Linked Open Data Crawl[8], as it has been recognised as a sound snapshot of the LOD cloud in 2014 [12]. First we extracted the links from the crawled data, by parsing the dump line by line, and identifying each quadruple containing a subject and an object with different graph provenance, and therefore a different $home(x)$. While parsing the dump file, we excluded all syntactically invalid quadruples to work with clean data. Second, in order to analyse the links on a data set basis, we split the data crawl into individual data sets, taking as contexts the data set identifiers provided by Schmachtenberg et al.[9]. We selected a set of 35 data sets from the LOD2014 crawl (from different domains and containing several types of links), analysing a total of 1+ million links.

5.2 Methodology

We computed each of the measures listed in Table 2 for each of the linked entities in the data sets, for all types of links in the 35 data sets. Once we had all the results, we first empirically validated the measures (Section 5.3). After that, we analysed the results on a data set basis (Section 5.4). We have published our experimental data and sources[10].

5.3 Measure validation

Following standard practices in the literature of quality measures [3], we validate our measures by (i) checking that they do not provide the same measurement for all data sets D_i; and (ii) verifying that our measures are not all correlated with each other – otherwise having multiple measures would be of limited utility.

Discriminative Measures We computed for each data set standard summary statistics such as the mean, standard deviation and quartiles considering all types of links simultaneously. As we see in the data files, the values for the measures vary across data sets, except for the classification extension (m11c) – where all data sets show a mean, standard deviation and quartiles of 0.0 for the difference in entropy. The other measures are discriminative.

Independent Measures We computed the Spearman correlation of all the measurements within each data set, putting all types of links together. Table 3 shows the correlation values. The first row contains NaN values because the standard deviation(s) are equal to zero. Measures m21 and m22 are highly correlated (0.96), which makes sense, since m21 looks at the number of target entities and m22 at the number of target data sets. In theory, one may link to many target entities within a few data sets and viceversa; but the empirical analysis suggests that having both might not particularly interesting. Having only m21 seems to be sufficient.

[8] Linked Data Crawl `http://goo.gl/lqxdgo`

[9] List of Data sets`http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/tables/datasetsAndCategories.tsv`

[10] Experimental data: extracted links `http://141.26.208.201/links/` Measurements `http://141.26.208.201/datameasures/` Python code and others `https://github.com/criscod/SeaStar/tree/master/data`

Measures	m11	m12	m13	m21	m22	m31
m11	NaN	NaN	NaN	NaN	NaN	NaN
m12		1.00	0.29	0.58	0.55	0.55
m13			1.00	-0.23	-0.22	0.76
m21				1.0	0.96	0.04
m22					1.0	0.02
m31						1.0

Table 3: Correlation between measures, for all data sets and all types of links.

Typelink	I	S	R	O	C	All
AEMET	0	0	96	0	57	153
BFS	1063	0	0	0	2862	3925
Bibbase	0	0	456	1401	0	1857
Bibsonomy	35646	0	2180	0	123080	160906
BNE	58	0	0	0	221	279
DNB	3577	0	8711	2278	55	14621
DWS Mannheim	71	0	296	39	926	1332
Eurostat	1182	0	2	0	1012	2196
Eye48	1	0	244	0	490	735
Fao	0	0	6	0	23	29
FigTrees	2	0	22	2	59	85
GeoVocab	11455	0	1759	113	7565	20892
GovWild	0	0	1998	0	0	1998
Harth	76	0	344	456	30	906
Icane	20	0	25	30	19	94
IMF	243	0	3	0	377	623
Korrekt	0	0	1174	0	7959	9133
L3S	1059	0	2478	1028	1089	5654
LinkedGeoData	634	0	12	0	254	900
LOD2	26	0	282	50	180	538
NDLJP	1	0	178	60	267	506
Ontologi	0	0	5686	0	736	6422
Openei	6	0	323	0	203	532
Reegle	327	0	432	0	135	894
Revyu	1402	0	2145	1806	39772	45125
RodEionet	9	0	981	0	0	990
SemanticWeb	161	0	783	0	576295	577239
Sheffield	121	0	2189	1	27064	29375
Simia	6691	0	25113	0	38069	69873
Soton	50	0	352	0	160	562
SWCompany	2023	0	13473	421	43136	59053
TomHeath	7	0	34	4	6	51
Torrez	0	0	266	0	493	759
TWRPI	2	0	12	0	65	79
UKPostCodes	1	0	7	0	1	9

Table 4: Different types of links in the 35 data sets that we analysed.

5.4 Results

Let us first look at the types of links that exist in the data sets and second, at the adoption of the 3 core principles. We focus on identity links (e. g. `owl:sameAs`), relationship links (e. g. `wgs84:location`), classification links (e. g. `rdf:type`), similarity links (e. g. `skos:closeMatch`), and other more general links (e. g. `rdfs:seeAlso`).

Basic Descriptive Statistics When we look at the type of links that is used the most in each of the data sets, in 17/35 data sets the type used at most is classification links (c), in 12/35 data sets it is relationship links (r), in 3/35 it is identity links (i) and in 3/35 it is other links (o). None of the data sets has similarity links (s). Table 4 shows the number of each type of link for each data set.

Principle-based Measurements Since our user is a data publisher willing to improve the interlinking, for each measure we analyse the inequalities among entities of the same data set. For that, we generate multiple box plots (one per entropy-based measure and type of link)[11]. If a box plot suggests that there are entities that get their description less extended than other entities in the data set, the data publisher could think of generating further links from those entities to new target data sets. The important features of these plots are the medians (in red), the range and interquartile range—which can show big differences among the measurements of different entities when they are big—and the outliers, which in our case are relevant as they can be one of the weak spots to be improved.

Classification: for all data sets and all types of links, the difference in cardinality (m11a) and entropy (m11c) has a median of 0.0 and the range of boxes is [0.0,0.0]. That means that there are no cases in the data where entities have been classified with classes defined in the source data set and the classification is inherited via identity links. However, given the number of links of type c, we see that data publishers do classify their entities with external classes.

Description: according to the m12a measurements, in all but two data sets the median of (p, o)-s gained is equal or below 2; the remaining two data sets show a median of 4 and 20. The median of new o-s gained instead (m13a) is 1 for 32 of the data sets (the other three have a median of 0). Observing the m12c measurements in the first row of box plots (Figure 2), we notice that in links of type c the medians of the difference in entropy stay between 0 and 1, while in links of type i the medians vary among data sets and go up to 8. Also, in identity links there are way more outliers than in classification links (see the case of Bibsonomy). It makes sense that entities are not described homogeneously, and often publishers do not have the resources to review each generated identity link. Both things motivate that SeaStar shows the user source entities and other data sets as more positive references. In the case of m13c measurements, and for all types of links, we find data sets that have negative values for the difference in entropy. That means that the links add some redundancy by adding statements with predicates that were already in the source entity. However, the positive thing is that only a few data sets have the box in the negative area, and that happens for links of type relationship (r) and others (o). For example, that occurs when the data publisher adds multiple `rdfs:seeAlso` internal and external links. The medians are between -0.4 and 0.7. Comparing the box plots for identity links (type i) of the m12 and m13 measurements, we notice that in the former the range of the boxes is larger than the boxes in m13 measurements; in m12 the distance between the min and max is around 4 where as in m13 is around 0.2.

Connectivity: the medians for the number of new entities targeted (m21a) for three data sets are 3,4, and 11, and for the rest these are all equal or below 2 new entities targeted. In the difference of entropy (m21c), the box plots do not show redundancy, which would only be possible if we compared D_c with a basis of previously generated links and new links were added over the same target entity. This would be a positive thing, if those links managed to extend the description (P1). M21 measurements show medians between 0 and 8 as for links of type i, between 0.0 and 2.0 for links of type r and o, and between 0.0 and 1.0 for links of type c. The box plot with links of type i, shows a more skewed box (either to the left or to the right) than m12 measurements of the same type of links.

[11] `https://github.com/criscod/SeaStar/tree/master/data/plots`

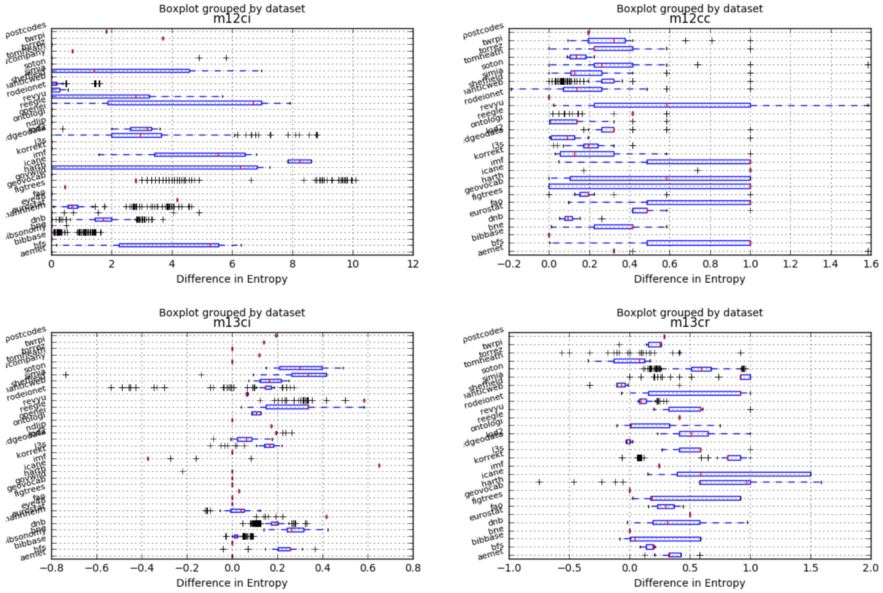

Fig. 1: Box plots showing m12c and m13c measurements for all data sets (m12c type c, m12c type, m13c type i, m13c type r).

Heterogeneity: measurements m31a show that 27 data sets gain 1 vocabulary in their description, while the rest do not gain any new. The difference in entropy (m31c) is in several data sets negative (in outliers and in the interquartile range). For links of type c the medians in measurements m31c are between 0.0 and 1.0, while for links of type r medians are between -0.1 and 1.0.

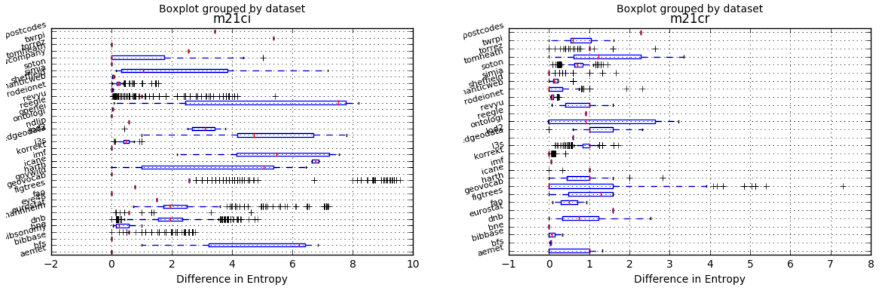

Fig. 2: Boxplots showing m21c measurements (links type i and type c) and box plots showing m31c measurements (links type r and type c).

6 Related Work

With the growth of Linked Data, there has been an increasing interest in assessing and monitoring the quality of available data [14].

Status of the Linked Data Web: while there were previous studies about the conformance of the Linked Data principles [6, **?**], the work by Schmachtenberg et al. [12] is the most recent study on the current adoption of Linked Data best practices. With regard to the linking principle, their analysis on data crawled from 1041 distinct data sets) showed descriptive statistics about the in- and out-degree of data sets (defined by the number of data sets pointing to / targeted by the data sets), and the most frequently used predicates.

Link Analysis: there are studies focusing exclusively on links. Halpin et al. [5] analyzed the usage of the `owl:sameAs` predicate in the links of the Linked Data space. They observed that sometimes the predicate was used with a meaning different from its original definition, and suggested to improve the quality of such links by using alternative and more suitable predicates (e. g. `skos:closeMatch` when not all properties of the target entity apply to the source entity; `foaf:primaryTopicOf` when the target entity represents but is not the same as the source entity). Hu et al. [7] empirically studied term and entity links in Biomedical Linked Data. Their findings include link and degree distributions, the analysis of symmetry and transitivity, and the evaluation of entity matching approaches over the links. Neto et al. [9] analysed the Linked Data crawl by Schmachtenberg et al., together with the set of Linked Open Vocabularies[12]. They examined the number of valid and dead links (i. e. in their work, links with an o that cannot be described in the target distribution), as well as the number of namespaces in link distributions and data sets. Albertoni et al. [2, 1] analysed the completeness of the interlinking of pairs of data sets and the extent to which data sets become more multilingual thanks to the links. These methods fail in stating the extent to which links add value to the source data set in terms of the principles that we mention in this paper.

Methods for Assessing Accuracy of Links: several methods have been developed to assess the semantic accuracy of links (e.g. to decide whether `ch:koblenz owl:sameAs de:koblenz` holds or not). Guéret et al. [4] defined a framework including three measures from the area of network theory: degree, clustering coefficient and betweeness centrality of the entities in links; as well as two measures that the authors define: number of unclosed same as chains and description enrichment defined as the raw number of new statements gained by the source entity. While Guéret's et al. notion of description enrichment is related to ours, the main differences are that we are able to observe further dimensions (e. g. how the classification of entities and the connectivity is extended by the links), our approach is not only restricted to `owl:sameAs` links (as it applies to any link) and we are able to signal redundancy.

7 Conclusions and Future Work

We have presented a collection of measures whose goal is to help in gaining insights into the quality of existing links, and understanding the effect that links produce in the

[12] LOV `http://lov.okfn.org/dataset/lov/`

source data set. After analysing 35 data sets of the LOD cloud with these measures our findings show that source entities are not classified with internal classes, but with external classes via links, and identity links do not contribute to inheriting new classes. We also observed that there is certain redundancy in the properties and vocabularies used as for extending the description. The differences between entities and data sets shown in the boxplots justify the need for our framework, which is able to pinpoint reference interlinked entities and data sets to data publishers.

As future work, we plan to extend our approach including mappings between classes and properties. We expect this add-on to help in identifying redundancy more precisely. Furthermore, we consider evaluating the usefulness of the measures with domain experts and observing the actions they take in data sets in response to the measurements.

Acknowledgements The research leading to these results has received funding from the European Union's FP7 under grant agreement no. 611242—Sense4Us project. We also thank Thomas Gottron for our discussions at an initial phase of this research, and Leon Kastler for his feedback.

References

1. Albertoni, R., De Martino, M., Podestà, P.: A linkset quality metric measuring multilingual gain in skos thesauri. In: Linked Data Quality co-located with ESWC 2015 (2015)
2. Albertoni, R., Pérez, A.G.: Assessing linkset quality for complementing third-party datasets. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops. EDBT '13 (2013)
3. Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A metrics-driven approach for quality assessment of linked open data. Journal of theoretical and applied electronic commerce research 9(2), 64–79 (2014)
4. Guéret, C., Groth, P.T., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Proc. ESWC 2012. pp. 87–102 (2012)
5. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isn't the same: An analysis of identity in linked data. In: Proc. ISWC 2010, pp. 305–320. Springer (2010)
6. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web 14, 14–44 (2012)
7. Hu, W., Qiu, H., Dumontier, M.: Link analysis of life science linked data. In: The Semantic Web-ISWC 2015, pp. 446–462. Springer (2015)
8. Max Schmachtenberg, Christian Bizer, A.J., Cyganiak, R.: Linking open data cloud diagram (2014), `http://lod-cloud.net/`
9. Neto, C.B., Kontokostas, D., Hellmann, S., Müller, K., Brümmer, M.: Assessing quantity and quality of links between linked data datasets (2016)
10. Pandian, C.R.: Software metrics: A guide to planning, analysis, and application. CRC Press (2003)
11. Rula, A., Zaveri, A.: Methodology for assessment of linked data quality. In: LDQ@ SEMANTICS (2014)
12. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Proc. ISWC 2014 - Part I. pp. 245–260 (2014)
13. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 3–55 (2001)
14. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked open data: A survey. Semantic Web Journal (2015)