EDBT '13 Proceedings of the Joint EDBT/ICDT 2013 Workshops
Pages 52-59
ACM New York, NY, USA ©2013
ISBN: 978-1-4503-1599-9
http://doi.acm.org/10.1145/2457317.2457327

Assessing Linkset Quality for Complementing Third-Party Datasets

Riccardo Albertoni*
OEG-DIA,
Facultad de Informática Universidad Politécnica de Madrid
Boadilla del Monte, Madrid, Spain
and
CNR-IMATI,
Via De Marini, 6, Torre di Francia, 16149
Genova, Italy
ralbertoni@fi.upm.es
albertoni@ge.imati.cnr.it

Asunción Gómez Pérez
OEG-DIA,
Facultad de Informática Universidad Politécnica de Madrid
Boadilla del Monte, Madrid, Spain
asun@fi.upm.es

ABSTRACT

Linked data best practices are getting extremely popular: various companies and public institutions have started taking advantage of linked data principles for exposing their datasets, and for relating their datasets to those served by third parties. Such enthusiasm is due to the linked data promise of evolving into a Global Data Space. Linksets are sets of links relating datasets and they surely play a fundamental role in this promise. However, a stable and well-accepted notion of linkset quality has not been yet defined. This paper contributes to overcome this lack by proposing a linkset quality measure. Among the different quality dimensions that can be addressed, the proposed measure focuses on completeness. The paper formally defines novel scoring functions and proposes an interpretation of these functions when maintaining and complementing third party datasets.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Database Management—*Heterogeneous databases*

General Terms

Linked data, Linksets, Quality

*This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. This Programme is supported by the Marie Curie Co-funding of Regional, National and International Programmes (COFUND) of the European Commission.

LWDM2013 '13 Genoa, Italy

Keywords

Quality assessment, Linksets

1. INTRODUCTION

Despite the adoption of linked data best practice has resulted in more accessible and structured data resources, the application of analysis tools on linked data collections is far from being a ready-to-go activity. Linked data consumption still have to deal with **data conditioning**, namely the process of getting data into a state in which data can be analysed and exploited for real applications. Data conditioning often includes error prone semi-automatic creation of integrated data caches to ensure efficiency and coherence consuming linked datasets independently provided. The problem is well known in the linked data community and in fact conceptual frameworks like the **Crawling Architectural Pattern**[8] have been developed with a mature technological stack to crawl, clean and integrated datasets before their data is exploited in linked data applications (e.g., Linked Data Integration Framework-LDIF[13]).

Unfortunately, the level of integration required depends on the kind of application is addressed. For example, applications visualizing mash-ups from different sources might be much less demanding in terms of integration and consistency than applications performing complex analysis. That because at some extent redundancies and inaccuracies can be filtered out by consumers in the former, whilst they might seriously affect the overall quality of the results in the latter[1]. For this reason, there is no guarantee that datasets integrated by a group of consumers will suit target applications of others. Neither it is possible to check the level of suitability of an integrated bunch of datasets being (i) the decisions taken during the conditioning process not finely documented and (ii) concepts for testing and describing the level of integration not yet consolidated in the linked data community.

In this paper, a notion of linkset quality is introduced. Linksets are pivotal for dataset integrations in the LOD. The proposed quality is specifically designed to estimate how linksets affect the dataset integration. Quality is often characterized in terms of "fitness for use" and it is defined with

respect to dimensions such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability, and verifiability[3]. This paper addresses linkset completeness as a mean to estimate possible losses in completeness when fusing two datasets via their linksets. So that the level of integration can be assessed and documented, and one step forward in the aforementioned long-distance trip can be taken. The proposed measures contribute in characterization of linksets and are intended for linked data publishers and consumers. Publishers can take advantages of them to check if a linkset they have provided is good enough or must be improved. Consumers are expected to consider these measures (a) to better understand whether they should or not rely on a linkset; (b) to have a first guess of what is the next action to complete the linkset; (c) to rank possible linkset alternative.

The rest of the paper is organized as follows: Section 2 introduces concepts upon which linkset quality is defined; Section 3 defines linkset quality measure in terms of quality indicators, scoring functions and quality assessment. Section 4 discusses the contribution with respect to the related work; Conclusions and future directions are drawn in the final section.

2. BASIC CONCEPTS

The proposed linkset quality measures are defined starting from the notion of dataset and linkset provided in the Vocabulary of Interlinked Datasets (VoID)[2]. VoID is a RDF vocabulary commonly adopted for expressing metadata about RDF datasets exposed in liked data. According to VoID,

- a **dataset**, more precisely a void:Dataset, is a set of RDF triples published, maintained or aggregated by a single provider. Typically, a linked data dataset exposes a set of RDF resources representing real world entities. RDF descriptions of the entities are associated to the resources and they are accessible on the Web, for example, resolving resources' HTTP dereferenceable URIs or through a SPARQL endpoint;

- a **linkset**, more precisely a void:Linkset, is a collection of RDF links, where each RDF link is a RDF triple whose subject and object are respectively in the subject and the object dataset. RDF links are typed by the RDF property deployed to connect objects to subjects. RDF links in a linkset should all have the same type, which implies that the linkset should be split in distinct linksets otherwise.

This paper considers **owl:sameAs linksets**, namely linksets whose RDF links are owl:sameAs. owl:sameAs is the RDF property adopted in the context of linked data to express the equivalence between RDF resources, namely, it is applied to RDF resources representing the same real entity.

Types for resources are given in terms of RDF and OWL classes. Typically, different providers adopt different RDF\ OWL Vocabularies to characterize the same kind of entities, so that, resources connected through a owl:sameAs in a linkset can have distinct types even if they are representing the same real entity. A relation of **type equivalence** $\sim$ is considered in this paper in order to capture when two types are just distinct concrete representations of the same kind of entities. $\sim$ holds between two types $type_a$ and $type_b$, (i.e., $type_a \sim type_b$) if and only if (i) there is a owl:sameAs RDF link between two resources a and b, so that a and b have respectively type $type_a$ and $type_b$ and a and b belong respectively to the subject and object dataset in the considered linkset or (ii) an alignment between $type_a$ and $type_b$ has been discovered in some ontology matching process.

One of the most interesting linked data promise is that "Linked data will evolve the web of data into a global data space". An implicit assumptions lies behind this promise: linksets can be exploited to complement and complete the information provided by independently served datasets. So that, analysts can run their experiments on richer data spaces resulting by complementing datasets with their interlinked datasets. For this reason, in this paper, we explicitly refer to a notion of **dataset complementation via a linkset**. Given two datasets X, Y; and a linkset L linking some of the entities exposed in X with some of the entities exposed in Y, we say that X can be complemented with Y via L and such as a complementation results in a third dataset hereafter indicated as X^L. In order to formally define X^L, we first introduce the predicate t, so that, given a dataset D and s, p, o, respectively subject, predicate and object in a RDF triple, $t_D(s, o, p)$ holds if and only if the triple (s, o, p) is in D. X^L, namely, X **complemented via** L is defined as $X^L = \{t \mid t \in t_X \vee [t = (s, p, o) \wedge t_L(s, \text{owl:sameAs}, y) \wedge t_Y(y, p, o)]\}$. It is worth noting that X^L and $X^L \cup Y$ usually differ: the former corresponds to X in which triples induced by the owl:sameAs have been materialized, whilst the latter also include all the triples from Y.

3. QUALITY MEASURE

The proposed linkset quality is structured coherently with the well-known quality terminology presented in [3], which adopts

Quality Indicators, namely, characteristics in datasets and linksets (e.g., pieces of dataset content, pieces of dataset meta-information, human ratings) which can give indication about the suitability of a dataset/linkset for some intended use. For example, in this paper, types of the entities involved in datasets and linksets as well as the cardinality of resources for types are considered as basic indicators;

Scoring Functions, namely, functions evaluating quality indicators to measure the suitability of the data for some intended use. In this paper, we have formalized three indicators which provide an assessment for type completeness and linkset type and entities coverage;

Aggregate Metrics, namely, user-specified assessment metric built upon scoring functions. These aggregations produce new assessment values through the average, sum, max, min or threshold functions applied to the set of scoring functions. In this paper, aggregate metric are provided in terms of the interpretations that can be drawn when combining the discussed scoring functions.

In this paper, linkset quality is designed to measure the suitableness of a linkset in a scenario of dataset complementation. Among the different quality dimensions that can

be considered (e.g., accuracy, timeliness, completeness, relevancy, consistency and interpretation), we address the dimension of completeness. Information completeness is usually defined on datasets in terms of "the degree to which information is not missing"[12]. So in coherence with this definition, we could define the linkset completeness as the degree to which links in the linksets are not missing. However, being biased by the aforementioned complementing perspective, we intuitively want to define the completeness of a linksets L in terms of the completeness of datasets obtained complementing via L. For this reason, we will consider complete a linkset L if X^L and Y^L maintain the same level of completeness of their source datasets X and Y. This means that if X and Y were complete when considering the two datasets disjointly, complementing X via a complete L we will still have a complete dataset. On the contrary, if we complement a datasets with an incomplete linkset the maintenance of the level of completeness is not granted.

Considering (i) two datasets SWDF[1] and DBLP[2], both about researchers and their publications. The former providing entities representing researchers, researchers' organizations, researchers' publications and related conferences whilst the latter is providing only entities for researchers, researchers' publications and related conference; (ii) a linkset L interlinking researchers from DBLP to the researchers in SWDF; (iii) a scenario in which data consumers want to build an enriched datasets DBLP^L by complementing DBLP with the information pertaining to organizations coming from SWDF; One of the basic question data consumers have to deal with is "whether or not complementing a dataset via a linkset makes sense". That decision depends on the dataset/linkset characteristics. For example, supposing DBLP and SWDF contain the same set of researchers but some are not yet included in the linkset L, no research organizations are imported by complementing DBLP via L for researchers not yet interlinked, and at the end, the percentage of researchers in $DBLP^L$ whose organizations are missing will be higher than in SWDF. Supposing SWDF contains a subset of the researchers provided in DBLP, a linkset L between SWDF and DBLP can at most consider the researchers the two datasets share. Depending on the number of shared researchers the DBLP complementation might result advantageous or not.

The following sections introduces measures to estimate losses and gains when complementing via linksets. The proposed quality formalization is build on top of sets defined in Table 1.

3.1 Quality Indicators

The definition of quality described in this paper relies on three basic **quality indicators**: Entities, Types and #E4Type.

Entities is the indicator returning the set of RDF resources exposed in a Dataset and it is defined as

$$\text{Entities} : \mathcal{D} \rightarrow 2^{\text{RDFEntities}}$$

$$\text{Entities}(X) = \{y \mid (t_X(y,*,*) \vee t_X(*,*,y)) \wedge y \notin \text{BlankNode}(X)\}$$

[1]http://data.semanticweb.org/
[2]http://dblp.l3s.de/d2r/

Table 1: Basic Notation

Set	*Definition*
$\mathcal{D}$	Set of void:Dataset.
$\mathcal{L} \subset \mathcal{D}$	Set of void:Linkset.
NativeTypes	Set of types natively defined in the RDF, OWL and SKOS specifications (e.g., owl:Class, owl:Restriction, rdfs:Resource, skos:Concept, skos:ConceptScheme).
UDTypes	Set of user defined types, namely resources in a collection of VOID datasets that are defined as instances of owl:Class or rdfs:Class.
RDFEntities	Set of entities exposed as RDF resources in the group of datasets considered.
BlankNode(X)	Set of blank nodes in the dataset X.

Types is the indicator returning the types of the entities exposed in a dataset or a linkset. It ignores types which are natively defined in the OWL, RDF, SKOS vocabularies and it is defined as

$$\text{Types} : \mathcal{D} \rightarrow 2^{\text{UDTypes}}$$

$$\text{Types}(X) = \{c \mid t_X(y, \text{rdf:type}, c) \wedge y \in \text{Entities}(X)\} \setminus \text{NativeTypes}$$

#E4Type is the indicator returning the number of entities for a given type exposed in a dataset or a linkset. It is defined as

$$\#\text{E4Type} : \mathcal{D} \times \text{UDTypes} \rightarrow \mathbb{N}$$

$$\#\text{E4Type}(X,T) = \mid \{z \mid t_X(z, \text{rdf:type}, T) \wedge z \notin \text{BlankNode}(X)\} \mid$$

It is worth noting that linksets are considered as a subset of datasets in VoID, so the indicators Entities, Types and #E4Type can be applied on datasets and linksets without distinction.
NLTypes and EquTypes are more complex indicators built on top of the previous which will be mentioned in some of the proposed scoring functions:

NLTypes returns the set of types in a dataset X that are not considered in a linkset L and it is defined as

$$\text{NLTypes} : \mathcal{D} \times \mathcal{L} \rightarrow 2^{\text{UDTypes}}$$

$$\text{NLTypes}(X, L) = \text{Types(X)} \setminus \text{Types(L)}$$

EquTypes relies on the type equivalence $\sim$ which we have introduced in Section 2. Given two dataset X and Y, EquTypes returns the subset of types in X that have an equivalent in Y according to $\sim$. It is defined as

$$\text{EquTypes}_\sim : \mathcal{D} \times \mathcal{D} \rightarrow 2^{\text{UDTypes}}$$

$$\text{EquTypes}_\sim(X,Y) = \{x \in \text{Types}(X) \mid \exists y \in \text{Types}(Y) \wedge x \sim y\}$$

3.2 Scoring Functions

The scoring functions proposed in this paper have been inspired by real attempts of consuming third party datasets. In particular, assessing similarity among entities exposed in different linked datasets[1] we have noticed that

- Linksets often cover only a minimal part of the entities exposed in the datasets that they involve;
- Linksets often include only a subset of the types of entities exposed in the datasets that they involve.

The proposed scoring functions ease in detecting the aforementioned situations, so that consumers can take the proper countermeasures and improve linkset completeness as well as the quality of datasets complemented via a linkset.

Three distinct scoring functions are introduced:

Linkset Type Coverage returns the percentage of types in a datasets that have been also considered in the linkset.

Linkset Type Completeness returns the percentage of mappable types in a datasets that have not yet been considered in the linksets when assuming an alignment among types.

Linkset Entity Coverage for Type returns what percentage of entities having a given type in a dataset are also involved in the analysed linkset.

A formalization for each of the above scoring functions is provided in the following. For each formalization, we provide examples discussing the score functions when applied on linksets involving the datasets DBLP and SWDF.

3.2.1 Linkset Type Coverage

Linkset Type Coverage measures at what extent a linkset covers the types involved in its subject or object datasets. Type coverage returns a value ranging in (0,1]. It is equal to 1 when the linkset covers all the types of entities exposed in a dataset (a.k.a. full coverage); when it is smaller than 1, it represents the percentage of types which are covered. Linkset Type Coverage is mathematically formalized in the equation *LTCov*.

$$\text{LTCov} : \mathcal{D} \times \mathcal{L} \rightarrow (0, 1]$$

$$\text{LTCov}(X, Y) = \frac{\mid \text{Types}(X) \cap \text{Types}(Y) \mid}{\mid \text{Types}(Y) \mid}$$

Figure 1 depicts some sets that can be obtained applying the quality indicator Types on the datasets DBLP, SWDF. L_1, L_2, L_3, L_4 are examples of possible linksets between DBLP and SWDF. In particular, entities of types foaf:Agent, swrc:Proceedings, foaf:Document are exposed in the dataset DBLP, whilst entities of types foaf:Person, swr:Proceedings, ro:FullPaper, ro:PosterPaper, ro:ShortPaper and foaf:Organization are provided by SWDF. Concerning the linksets, L_1 involves the types foaf:Agent and foaf:Person, L_2 involves the types foaf:Agent, foaf:Person, swrc:Proceedings and swr:Proceedings, and types for L_3 and L_4 can be listed in analogy with the previous examples.

The results applying LTCov on L_1, L_2, L_3, L_4 are reported in the first two rows of Table 2. Considering that results, it is worth noting that a full coverage can be achieved for DBLP but not for SWDF: L_3 and L_4 have a full coverage with respect to DBLP (i.e., LTCov(DBLP,L_3)=1 and LTCov(DBLP,L_4)=1), but being foaf:Organization a type of entity which is not provided by DBLP, we cannot enrich the interlinking L_4 more, and organizations cannot be included in L_4 and the full coverage cannot be obtained for SWDF.

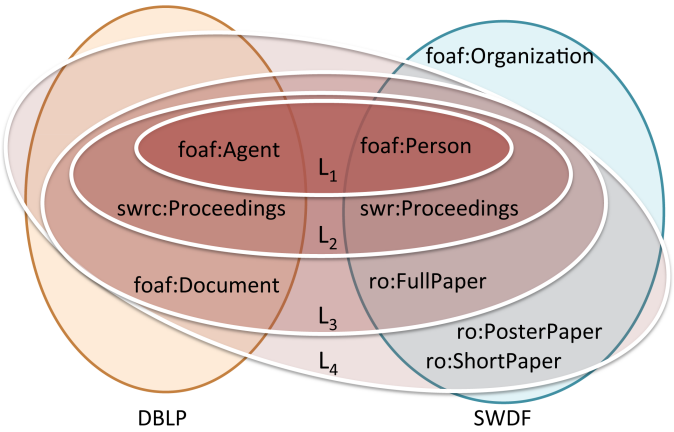

Figure 1: Diagram representing a subset of the types exposed in DBLP, SWDF and their linksets.

Table 2: LTCov and LTCom applied on the linksets depicted in Figure 1

	L_1	L_2	L_3	L_4
LTCov(DBLP, $L_?$)	1/3	2/3	3/3	3/3
LTCov(SWDF, $L_?$)	1/6	2/6	3/6	5/6
LTCom$_{\sim_1}$($L_?$,DBLP,SWDF)	1/1			
LTCom$_{\sim_2}$($L_?$,DBLP,SWDF)	1/2	2/2		
LTCom$_{\sim_3}$($L_?$,DBLP,SWDF)	1/3	2/3	3/3	
LTCom$_{\sim_4}$($L_?$,DBLP,SWDF)	1/3	2/3	3/3	3/3
LTCom$_{\sim_1}$($L_?$,SWDF,DBLP)	1/1			
LTCom$_{\sim_2}$($L_?$,SWDF,DBLP)	1/2	2/2		
LTCom$_{\sim_3}$($L_?$,SWDF,DBLP)	1/3	2/3	3/3	
LTCom$_{\sim_4}$($L_?$,SWDF,DBLP)	1/5	2/5	3/5	5/5

3.2.2 Linkset Type Completeness

Linkset Type Completeness measures the level of completeness of a linkset with respect to types involved in its datasets. If two datasets expose entities having compatible types, some interlinks among these entities could be established. Thus, if two equivalent/compatible types don't have yet any link in a linkset, this linkset is potentially incomplete. Type Completeness returns a value smaller than 1, whenever there is at least a type in the subject dataset which is equivalent to a type in the object dataset and has not yet been involved in the linkset. It is mathematically formalized in the equation LTCom.

$$\text{LTCom}_{\sim} : \mathcal{L} \times \mathcal{D} \times \mathcal{D} \rightarrow (0, 1]$$

$$\text{LTCom}_{\sim}(l, x, y) = 1 - \frac{\mid \text{NLTypes}(x, l) \cap \text{EquTypes}_{\sim}(x, y) \mid}{\mid \text{EquTypes}_{\sim}(x, y) \mid}$$

The ability of detecting incompleteness of LTCom strongly depends on correctness of the type equivalence relation. Type

equivalence always includes at least (i) the reflexives closure on types (i.e., $\{(\text{type}, \text{type}) \mid \text{type} \in \text{UDTypes}\} \subseteq \sim$); (ii) the mappings among types whose entities are actually interlinked in the considered linkset (i.e., if the linkset L is from a dataset X to a dataset Y and type_x, type_y are respectively the types of x and y then $\{(\text{type}_x, \text{type}_y) \mid x \in X, y \in Y, t_L(x, \text{owl:sameAs}, y)\} \subseteq \sim$). Besides the above, further mappings can be added to the relation $\sim$ by relying on (i) manually curated mappings or (ii) alignments provided by ontology matchers[5]. Although in general alignments among ontology schemas are not necessary symmetric, for the purposes of this paper, $\sim$ is considered a symmetric relation (i.e., $\{(\text{type}_x, \text{type}_y) \mid \exists \text{type}_x, \text{type}_y \in \text{UDTypes} \wedge \text{type}_y \sim \text{type}_x\} \subseteq \sim$).

This implies that, when we are assessing the type completeness for the linkset L_2, any type equivalence we want to consider must include the couples (foaf:Agent, foaf:Agent), (foaf:Person, foaf:Person), (swrc:Proceedings, swrc:Proceedings) and (swr:Proceedings, swr:Proceedings) for (i), (foaf:Agent, foaf:Person) and (swrc:Proceedings, swr:Proceedings) for (ii) and (foaf:Person, foaf:Agent) and (swr:Proceedings, swrc:Proceedings) for the symmetric closure. Table 3 shows some examples of mapping among types. Due to limitation in space, the reflexive and symmetric closures are not explicitly materialized in the table, but they will be considered when assessing LTCom.

Table 3: Examples of type equivalence relations.

$\sim_1$	{(foaf:Agent, foaf:Person)}
$\sim_2$	{(foaf:Agent, foaf:Person), (swrc:Proceedings, swr:Proceedings)}
$\sim_3$	{(foaf:Agent, foaf:Person), (swrc:Proceedings, swr:Proceedings), (swrc:Document, ro:FullPaper)}
$\sim_4$	{(foaf:Agent, foaf:Person), (swrc:Proceedings, swr:Proceedings), (swrc:Document, ro:FullPaper), (swrc:Document, ro:PosterPaper), (swrc:Document, ro:ShortPaper)}

The results of LTCom applied on linksets L_1, L_2, L_3, L_4 depicted in Figure 1 are shown in the last eight rows of Table 2. LTCom largely depends on the mapping relation adopted: the more precise $\sim$ captures equivalences on types, the more LTCom is informative. For example, considering an incomplete mapping such as $\sim_1$, linksets L_2, L_3, L_4 appear to be equivalent in terms of type completeness, but when we move on more precise mapping such as $\sim_4$ it turns out that they largely differs.

Focusing on results obtained when exploiting the mapping relation $\sim_4$, further remarks can be outlined:

(i) when assessing the completeness of a linkset L which involves the dataset X and Y, both completeness directions LTCom(L,X,Y) and LTCom(L,Y,X) must be considered. For example, L_3 results complete for DBLP, but not for SWDF (i.e., LTCom(L_4, DBLP, SWDF)=1 and LTCom(L_3, SWDF, DBLP)<1), whilst L_4 results complete for both (i.e., LTCom(L_4, SWDF, DBLP)= LTCom(L_4, DBLP, SWDF)=1). That implies L_4 is better for complementing a dataset than L_3 at least when L_4 and L_3 result equivalent with respect to the other indicators.

(ii) Type Coverage and Completeness detect different situations and they are somehow independent each other. For example, L_4 is complete even if it does not provide a full coverage on SWDF (i.e., LTCov(SWDF,L_4)<1 but LTCom(L_4, SWDF, DBLP)= LTCom(L_4, DBLP, SWDF)=1).

3.2.3 *Linkset Entity Coverage for Type*

Entity Coverage measures the percentage of entities of a selected type considered in the linkset. Selecting a type, Entity Coverage returns 1 when all the entities for that type in a dataset are also considered in the linkset. Otherwise it returns the percentage of entities in the dataset that are already considered in the linkset. Entity Coverage is mathematically formalized by ECov4T, which is specified as follows:

$$\text{ECov4T} : \mathcal{D} \times \mathcal{L} \times \text{UDTypes} \rightarrow [0, 1]$$

$$\text{ECov4T}(X, L, T) = \frac{\#\text{E4Type}(L, T)}{\#\text{E4Type}(X, T)}$$

Table 4 provides examples of entities cardinality (i.e., values for the indicator #E4Type): the first column shows the types exposed in the datasets and linksets we are considering; the second, third, forth and fifth columns provide #E4Type values respectively DBLP, SWDF, L_3 and L_4. Tables 5 and 6 show the results applying ECov4T on DBLP and SWDF considering the indicators in Table 4 and varying on the linksets L_3, L_4.

Table 4: Examples of #E4Type indicators: number of entities for types.

	DBLP	SWDF	L_3	L_4
foaf:Document	1984087		2784	4500
foaf:Agent	1000000		5320	9002
swrc:Proceedings	1108400		225	225
ro:FullPaper		2784	2784	2784
ro:ShortPaper		1201		1201
ro:PosterPaper		602		515
foaf:Person		9223	5320	9002
swr:Proceedings		225	225	225

Table 5: Results applying Entity Coverage on linksets L_3 and L_4 for types in DBLP.

	L_3	L_4
foaf:Document	0.140%	0.227%
foaf:Agent	0.532%	0.9%
swrc:Proceedings	0.020%	0.020%

Focusing on Table 5, it is worth noting that a very low percentage of researches in DBLP (i.e., entities of type foaf:Agent) are covered by L_3 and L_4. That kind of information is precious when complementing a dataset via linksets. Let us suppose we want to complement DBLP via L_4 in order to include in DBLP the researchers' organizations (i.e., foaf:Organization) served by SWDF. If we consider L_4 as linkset we obtain DBLP^{L_4} in which organizations are missing for at least all the researchers that are not

Table 6: Results applying Entity Coverage on linksets L_3 and L_4 for types in SWDF.

	L_3	L_4
ro:FullPaper	100%	100%
ro:ShortPaper		100%
ro:PosterPaper		86%
foaf:Person	58%	98%
swr:Proceedings	100%	100%

linked to SWDF. Since L_4 involves only the 0.9% of the researchers exposed in DBLP, we can estimate the number of researchers without organizations in at least the 91% (i.e., 100% - ECov4T(DBLP, L_4, foaf:Agent)) of the researchers provided in DBLP^{L_4}.

On the contrary, considering Table 6, we can note that all the entities for ro:FullPaper, ro:ShortPaper and swr:Proceedings in SWDF are included in the linkset L_4. That suggests that complementing SWDF via L_4 does not produce data missing. Unfortunately, in this specific example, it is not very useful to complement SWDF via L_4 because there are no complementary information to researchers in DBLP (i.e., ETCov(L_4,DBLP)=1), but in general, we might have complementaries in type of entities and the complementation could make sense.

3.3 Quality Assessment

The quality assessment is defined in terms of an interpretation upon the aforementioned score functions. The interpretation is illustrated in Table 7 and Table 8. Linked data providers and consumers are expected to exploit the proposed interpretation in order to (i) detect flaws/facts that might affect completeness; (ii) obtain suggestions about how to deal with detected flaws/facts; (iii) estimate consequences deriving from the linkset adoption. In particular, Table 7 presents facts and suggestions that can be detected considering the scoring function on types (i.e., Linkset Type Coverage and Completeness), whilst Table 8 presents facts and suggestions that can be derived considering also results from Entity Coverage. Linked data providers and consumers must consider the tables in the aforementioned order: Table 7 always before Table 8. They should check first for Type Completeness (Table 7, first row), because Type Coverage is not really meaningful when considering type incomplete linksets. Once Type Completeness is reached, namely when all the types for linkable entities are included in the linkset, they can start worrying about type complementaries, so that they can discern (i) if new types are included in the complemented dataset (Table 7, second and third rows); (ii) if complementation lacks of some types for the imported entities (Table 7, first and forth row); (iii) if the complementation is just contributing at the extensional level and no types are added (Table 7, fifth row). Afterwards, considering Table 8 they have an estimation of the number of instances that will be affected by the complementation.

For example, if we assume the equivalence type relation $\sim_4$ and we apply the interpretation on linksets L_1, L_2, L_3, L_4, all the linksets but L_4 will result type incomplete, so they should be integrated with new links among the entities having equivalent types in DBLP and SWDF.

In the case of L_4, we deal with two distinct situations depending on which dataset we are complementing. It does not make much sense to complement SWDF via L_4: according to Table 7 third row, DBLP is not providing additional types of entities for SWDF. The only attempt we can make is to jump in one of the situations where LTCov(DBLP,L_4)=1, for example, by finding a third dataset to complement DBLP with organizations, and after that, we can check if we have ended up with LTCov(SWDF,L_4)<1 or =1.

Complementing DBLP via L_4 makes more sense, instead. We end up with the situation described in Table 7 fourth row, then checking Table 8, we can estimate the percentage of researchers in DBLP^{L_4} getting /not getting their organizations. If we are not fully satisfied of what we have obtained in DBLP^{L_4} we might try to enrich it more looking for linksets complementing DBLP^{L_4} instead of DBLP.

4. RELATED WORK

In the context of web of data some efforts to define quality measures have been already carried out. For example, WIQA is a Information Quality Assessment Framework[3] which applies different filtering policies relying on complex metadata such as provenance chains and background information about providers. WIQA proposes a policy language WIQA-PL, it deploys an engine for interpreting such policies and it provides explanations of why information satisfies a specific policy. It makes information filtering decisions based on quality criteria comprehensible and traceable, but basically, scoring functions for quality are seen as parameters of the system and a definition of novel quality measures specifically designed for linked data is out of the WIQA scope.

On the contrary, EU funded projects such as PlanetData and LOD2 have specifically addressed quality metrics for linked data: LOD2 reviews quality dimensions which are traditionally considered in data and information quality (e.g., accuracy, timeliness, completeness, relevancy, availability, representational consistency). It lists some linked data specific indicators and some criteria upon which linked data quality can be defined[9, 6]. Unfortunately, it does not propose any indicator or criteria for completeness. PlanetData formalizes scoring functions for some of the aforementioned quality dimensions[10]. It discusses completeness of a dataset in terms of intensional, extensional and LDS completeness. In particular, a dataset is considered intensionally complete, when at the schema level, it contains all of the necessary attributes for a given task from. It is considered extensionally complete, when at the data (instance) level, it contains all of necessary objects for a given task[4]. Finally, a dataset is considers LDS complete, when properties considered relevant for a task have values. LDS completeness has been deployed in SIEVE[11], a framework adapting the ideas proposed in WIQA to fit into Linked Data Integration Framework (LDIF)[13]. All these efforts mainly cover quality of datasets and even if they consider quality in a data fusion context, they substantially overlook the issues pertaining to linkset quality.

At the best of our knowledge, LINK-QA[7] is the work getting closer to our linkset quality. It deploys two network measures specifically designed for Linked Data (Open SameAs chains, and Description Richness) and three classic network measures (degree, centrality, clustering coefficient) for determining whether a set of links improves the overall quality of linked data. However, our quality substantially differs from LINK-QA: (i) LINK-QA works on links independently of they are part or not of the same linksets; (ii)

Table 7: Quality assessment on a linkset L when interpreting results only from scoring functions on types

LTCov (X,L)	LTCov (Y,L)	$LTCom_{\sim}$	Facts/Suggestions
		one of the two directions <1	Linksets L results incomplete looking at types. Namely, there are equivalent types between X and Y which have not yet been considered in the linkset. **Next Action:** check if entities with types in $(NLTypes(X, L) \cap EquTypes_{\sim}(X, Y))$ can be interlinked with entities in Y whose type is ∼related.
<1	<1	both directions =1	Linkset L is complete w.r.t. types, the two datasets provide complementary types of entities, in principle, it is possible to complement both X via L and Y via L. **Next Action:** check entity coverage to know about data missing in X^L and Y^L.
<1	=1	both directions =1	Y is a subset of X in terms of type of entities provided. L can be employed to enlarge the set of entities of X, but entities imported from Y will be incomplete: types offered in X's schema which are not included in Y's will be missing. **Next Action:** it doesn't make much sense to complement X via L, unless LTCov(X,L) can become 1 by complementing X^L or Y with a third linkset providing the types missing in Y.
=1	<1	both directions =1	Y is providing types of entities not provided in X. **Next Action:** check entity coverage to know about data missing in X^L.
=1	=1	both directions =1	X and Y are providing the same types of entities. L can be employed to complement the dataset at the extensional level considering $X^L \cup Y$. **Next Action:** check Entity Coverage to know if datasets expose different entities (i.e., ECov4T(Y,L,t)<1 for some t).

Table 8: Quality assessment on a linkset L interpreting results from the scoring function on types and entity coverage

LTCov (X,L)	LTCov (Y,L)	LTCom (L,X,Y)	ECov4T (X,L,t)	Facts/Suggestions
<1	<1	=1	=1	The two datasets provide exactly the same entities for types t. If (i) t is the only type in the linkset or (ii) ECov4T(X,L, t)=1 for every type t included in the linkset, the linkset can be used jointly without introduce any data missing.
<1	<1	=1	=k<1	X^L will have (1-k)% of entities of type t incomplete.
<1	=1	=1	=k<1	X^L will have (1-k)% entities which will miss info type not included in L.

LINK-QA addresses correctness and it does not deal with completeness[3]; (iii) LINK-QA is for ranking sets of links, it can be used to say a linkset is better than another, but it does not suggest what is the next move a consumer should take to improve his linkset.

5. CONCLUSIONS AND FUTURE DIRECTIONS

This paper addresses linkset quality introducing novel quality measures in the context of linked data. The proposed quality measures aim at providing concepts to increase the consumers' awareness about risks and advantages they take when exploiting linked datasets. The paper defines linkset completeness in terms of different indicators and scoring functions. It demonstrates these score functions on a set of realistic examples and proposes interpretations upon them. These interpretations allow to estimate the consequence of incompleteness and to suggest actions which can be undertaken to improve the linksets. So far, a first JAVA-JENA prototype implementing the proposed measures has been developed and tested on few in-house linksets. However, a detailed validation is foreseen as part of the future work. In particular, in the next stages, measures will be tested on datasets included in the LOD cloud. This with the aim of (i) verifying the hypothesis under which indicators and scoring functions are applicable in the LOD cloud; (ii) providing

[3]The quality dimensions addressed by LINK-QA are not explicitly stated, but we exclude LINK-QA copes with completeness considering that it tries to correlate network measures and bad link detection, and the gold standard adopted in experimentation, i.e. the LATC Linkset-specification available at https://github.com/LATC/24-7-platform/tree/master/link-specifications, provides examples of correct and wrong links but it does not provide information about linkset completeness.

a statistical evidence about the relevance of the kinds of incompleteness that can be detect by mean of the proposed measures; (iii) characterizing more the linksets provided in the LOD, so that indicators and score functions can be included as part of linkset descriptions.

6. ACKNOWLEDGEMENTS

The authors would like to thank Jérôme Euzenat for his suggestions about ontology alignment and mapping relations, and Mari Carmen Suáerez Figueroa for having provided some in-house linksets on which the preliminary testing has been performed.

7. REFERENCES

[1] R. Albertoni and M. D. Martino. Semantic similarity and selection of resources published according to linked data best practice. In R. Meersman, T. S. Dillon, and P. Herrero, editors, *OTM Workshops*, volume 6428 of *Lecture Notes in Computer Science*, pages 378–383. Springer, 2010.

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary, 2011, `http://www.w3.org/TR/2011/NOTE-void-20110303/`.

[3] C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. *J. Web Sem.*, 7(1):1–10, 2009.

[4] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1), 2008.

[5] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[6] A. Flemming and O. Hartig. Quality criteria for linked data. 2010, `http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources`

[7] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing linked data mappings using network measures. In E. Simperl, P. Cimiano, A. Polleres, Ó. Corcho, and V. Presutti, editors, *ESWC*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2012.

[8] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[9] P. N. Mendes and C. Bizer. Survey report state of the art in mapping, quality assessment and data fusion. Technical report, LOD2- Creating Knowledge out of Interlinked data, Deliverable 4.3.1, 2011, `http://static.lod2.eu/Deliverables/Deliverable_4.3.1_FP7_LOD2_20110131.pdf`.

[10] P. N. Mendes, C. Bizer, J. H. Young, Z. Miklos, J.-P. Calbimonte, and A. Moraru. Conceptual model and best practices for high-quality metadata publishing. Technical report, PlanetData, Deliverable 2.1, 2012, `http://planet-data-wiki.sti2.at/web/File:D2.1.pdf`.

[11] P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In D. Srivastava and I. Ari, editors, *EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.

[12] L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002.

[13] A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker. LDIF - Linked Data Integration Framework. In O. Hartig, A. Harth, and J. Sequeda, editors, *COLD*, volume 782 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.