Towards Linkset Quality for Complementing SKOS Thesauri

Riccardo Albertoni, Monica De Martino, and Paola Podestà

Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche,
Via De Marini, 6, 16149 Genova, Italy
{albertoni,demartino,podesta}@ge.imati.cnr.it

Abstract. Linked Data is largely adopted to share and make data more accessible on the web. A quite impressive number of datasets has been exposed and interlinked according to the Linked Data paradigm but the quality of these datasets is still a big challenge in the consuming process. Measures for quality of linked data datasets have been proposed, mainly by adapting concepts defined in the research field of information systems. However, very limited attention has been dedicated to the quality of linksets, the result of which might be important as dataset's quality in consuming data coming from distinct sources. In this paper, we address linkset quality proposing the *linkset importing*, a novel measure which estimates the completeness of dataset obtained by complementing SKOS thesauri with their skos:exactMatch-related information. We validate the proposed measure with an in-house developed synthetic benchmark: experiments demonstrate that our measure may be adopted as a predictor of the gain that is obtained when complementing thesauri.

1 Introduction

Linked Data is largely adopted by data producers such as European Environment Agency, US and some EU Governs, whose first ambition is to share (meta)data making their processes more effective and transparent. The increasing interest and involvement of data providers surely represents a genuine witness of the Web of Data success, but in a longer perspective, the quality of the exposed data will be one of the most critical issues in the data consumption process. After all, as discussed in [13], data is only worth its quality.

The research pertaining to Linked Data quality is especially focused on datasets [13]. However, one of the most interesting promises that Linked Data makes is "Linked Data will evolve the current web data into a Global Data Space", which implicitly assumes the exploitation of data items coming from different sources as a whole. In the Linked Data context, this is possible by connecting information belonging to different sources by way of linksets. It is using connections in linksets that a Linked Data consumer can complete and enrich data to hand, and as a consequence, the quality of connections (hereinafter linkset quality) is as critical as the quality of data if we want to keep the Linked Data

promise. This paper addresses linkset quality proposing the *linkset importing*, a measure which assesses linksets as good as they improve a dataset with its interlinked entities' properties. It extends the linkset quality introduced in [1] considering a specific kind of linkset: `skos:exactMatch` linkset, which connects thesauri exposed as Simple Knowledge Organization System (SKOS) Ontology in the Linked Data. This type of linkset has been chosen considering the application scenarios we are facing in the EU funded project eENVplus (CIP-ICT-PSP grant No. 325232), where in order to maintain a framework of thesauri for the environment, we have to deal with a remarkable number of thesauri exposed as Linked Data[2] and their `skos:exactMatch` linksets. In that context, we realized that conspicuous efforts have been spent to interlink thesauri such as GEMET, EARTh, AGROVOC, EUROVOC, UNESCO, RAMEAU, TheSoz. However, currently, there is no way to assess what is the real *value* of these interlinks: how useful and enriching are the provided linksets?

The paper proposes a method to shed light on this. It formalizes a measure to estimate how good a linkset is to enrich a thesaurus with the properties values reachable from its `skos:exactMatch`-related entities. Although the formalized linkset quality can be exploited to check the linkset complementation potential in respect to any property, in this paper, we especially focus on `skos:prefLabel` and `skos:altLabel`. That is because, complementation with respect to these two properties can be deployed to ease multilingual issues such as incomplete language coverage[1], an issue that is discussed in [12] which affects many of the most popular SKOS thesauri. As example of application of our *linkset importing*, we suggest its adoption in the metadata of published linksets: the assessment of *linkset importing* can be provided for every language of interest in order to inform about how fitting a linkset is to improve the thesaurus multilingualism.

The organization of the paper is as follows: Section 2 introduces concepts on which the paper relies on (i.e., dataset, linkset and complementation of a dataset via its linkset). Section 3 formalizes the *linkset importing* quality providing related indicators and score functions. Section 4 introduces the goal of our experimentation and explains the methodological and architectural setting adopted to validate the proposed quality measure. Section 5 discusses the results of experimentation showing the proposed measure as an effective predictor for the `skos:prefLabel` and `skos:altLabel` gain which may be obtained by complementing thesauri via its `skos:exactMatch` linksets. Finally, we discuss related work in Section 6 and the conclusions and future work in Section 7.

2 Basic Concepts

The proposed linkset quality measure is defined starting from the notion of dataset and linkset provided in the Vocabulary of Interlinked Datasets (VoID)[3]. VoID is a RDF vocabulary commonly adopted for expressing metadata about

[1] Incomplete language coverage arises when `skos:prefLabel` and `skos:altLabel` are provided in all the expected languages only for a subset of the thesaurus concepts.

RDF datasets exposed as Linked Data. According to VoID, we consider two concepts: dataset and linkset.
A **dataset (D)**, more precisely a `void:Dataset`, is a set of RDF triples published, maintained or aggregated by a single provider.
A **linkset (L)**, more precisely a `void:Linkset`, is a special kind of dataset containing only RDF links. Each RDF link is an RDF triple *(s,p,o)*, where s, o are concepts respectively in the subject and object RDF dataset, while p is property linking s and o, that indicates the type of the link.
RDF links in a linkset should all have the same type, otherwise, the linkset should be split in distinct linksets. This paper considers `skos:exactMatch` **linksets**, namely linksets made by RDF `skos:exactMatch` links. In the context of SKOS thesauri, `skos:exactMatch` binds SKOS concepts to have the equivalent meaning.

In this paper we refer to the notion of **thesaurus complementation via a linkset**. Given two thesauri X, Y and a linkset L linking some of the concepts in X with some of the concepts in Y, we say that X can be complemented with Y via L and that such complementation results in a third thesaurus identified with X^L. Informally, X^L contains all RDF triples of X and the RDF triples reachable in Y via L. In order to formally define X^L, we introduce the predicate **t**. Given a dataset D and the RDF triple *(s,p,o)* the predicate $t_D(s,p,o)$ holds if and only if the triple *(s,p,o)* is in D. Thus X^L is defined as $X^L = \{t \mid [t = (s,p,o) \wedge t_X(s,p,o)] \vee [t = (s,\bar{p},\bar{o}) \wedge t_L(s,\texttt{skos:exactMatch},y) \wedge t_Y(y,\bar{p},\bar{o})]\}$. Notice that, X^L and $X^L \cup Y$ usually differ. The former corresponds to X in which triples induced by the `skos:exactMatch` have been materialized, while the latter also include all the triples from Y.

3 Linkset Importing Quality

This section formalizes the *linkset importing*, a quality measure which assesses linksets as good as they improve a dataset with its interlinked entities' properties. *Linkset importing* is structured coherently with the well-known quality terminology presented in [4] including **quality indicators**, **scoring functions** and **aggregate metrics**. **Quality indicators** are 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. In this paper, we define a set of **VoID-inspired** indicators which provide metadata pertaining to the linkset under analysis (e.g., its related subject/object datasets). **Scoring functions** are functions evaluating quality indicators to measure the suitability of the data for some intended use. We have formalized *linkset importing* which measures the percentage of values that can be imported in subject dataset, from the object dataset, when complementing via a linkset. **Aggregate metrics** are user-specified 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, we do not provide any aggregate metrics.

3.1 Indicators

Taking the sets and notations defined in Table 1, we define the following quality indicators which are exploited in the formalization of the score functions.

Table 1. Basic Sets deployed in the formalization

Set	*Definition*
$\mathcal{D}$	set of void:Dataset
$\mathcal{L} \subset \mathcal{D}$	set of void:Linkset
RDFEntities	set of entities exposed as RDF resources in the group of datasets considered
RDFProperties	set of property that can be defined in a RDF/OWL
RDFValues	set of typed values that can be defined in a RDF/OWL
RDFTriples	set of RDF triples in the form (s,p,o) where s ∈ RDFEntities, p ∈ Properties, o ∈ RDFEntities ∪ RDFValues
Languages	set of language tag adopted in RDF/OWL
Links(L)	set of triples $t_L(*,*,*)$, belonging to the linkset L
Entities(X)	set of entities e belonging to the dataset X (i.e. $\{e \mid t_X(e,*,*)\}$)
VoID(L)	set of VoID triples describing the linkset L

We present **VoID-inspired indicator**, a set of indicators that, given a linkset L, returns respectively: (i) the subject dataset (**LSubject(L)**); (ii) the object dataset (**LObject(L)**).

Definition 1. *Let L be a linkset. We define:*
LSubject: $\mathcal{L} \rightarrow \mathcal{D}$; *LSubject(L)*= $\{X \mid t_{VoID(L)}(L, (void : subject), X)\}$
LObject: $\mathcal{L} \rightarrow \mathcal{D}$; *LObject(L)*= $\{Y \mid t_{VoID(L)}(L, (void : object), Y)\}$

We present the indicator **val4Prop** which, given a datasets X, a property p and an entity e in X, returns the property values associated to e. We, then, specialize the **val4Prop** indicators for specific languages when p ranges in RDF literals with language tags.

Definition 2. *Let X be an dataset, p be an RDFProperties, and e be an Entities(X). We define:*
val4Prop:$\mathcal{D} \times$ *Entities(X)* $\times$ *RDFProperties* $\rightarrow 2^{RDFValues \cup RDFEntities}$;
$val4Prop_X$ *(e,p)*= $\{v \mid t_X(e,p,v)\}$
When p ranges in RDF Literals with language tag, it can be specialized as follows:
val4Prop:$\mathcal{D} \times$ *Entities(X)* $\times$ *RDFProperties* $\times$ *Language* $\rightarrow 2^{RDFValues}$;
$val4Prop_X$ *(e,p, lang)*=$\{v@lang \mid t_X(e,p,v@lang)\}$

Now, we define an operator which given a set of entities returns either the entities itself or its mapping with respect to a linkset L. When it is applied to RDF literals it returns the literals without modifications.

Definition 3. *Let L be a linkset, X be the LSubject(L), and Y be the LObject(L), Z a set of RDFValues and RDFEntities. The operator* $[\ ]_L$ *is defined as follows:*
$[\ ]_L$*:RDFValues* $\cup$ *RDFEntities* $\times \mathcal{L} \rightarrow 2^{RDFValues \cup RDFEntities}$*;*
$[Z]_L$=$\{y|(\exists l \in Links(L)$ *t.c.* l=$(x,$ *skos:exactMatch,*$y) \wedge x \in X) \vee$
$((\neg \exists l \in Links(L)$ *t.c.* l=$(x,$ *skos:exactMatch,*$y) \wedge x \in X) \vee x \in RDFValues) \wedge y = x)\}$

3.2 Scoring functions

In this section, using the indicators presented in the previous section, we are able to define the importing scoring functions that characterize our linkset quality measure. The aim of the importing scoring function is to calculate the importing potential of a linkset L for a property p. Informally, this function evaluates how many new values for a property p, are distinct from those already existing in the subject dataset X and can be reachable through L. First of all, we present the importing scoring function for a single link and then we generalize defining the average importing scoring function for the whole linkset. In the following we consider a linkset L, the datasets X and Y, respectively, the subject and object dataset of L and the property of interest p.

Definition 4. ***Link importing for property*** *Let* $e \in Entities(X)$ *and* $l \in Links(L)$*. The importing measures the percentage of values for p that can be imported to e through the link l is defined as follows:*

LinkImp4p: $\mathcal{L} \times Entities(X) \times RDFProperties \times Links(L) \rightarrow \mathbb{R}^+ \cup \{0\}$*;*

$$LinkImp4p_L(e,p,l) = \begin{cases} 0 & \text{if } den \neq 0 \\ \overline{LinkImp4p_L}(e,p,l) & \text{otherwise} \end{cases}$$

where

$$\overline{LinkImp4p_L}(e,p,l) = 1 - \frac{|val4Prop_X(e,p)|}{\underbrace{|[val4Prop_X(e,p)]_L \cup val4Prop_{X^L}([e]_{\{l\}},p)|}_{den}}$$

Now, we consider the entire linkset.

Definition 5. ***Average Linkset importing for property.*** *Let L a linkset, the importing capability of L with respect to p is defined as the average importing of all links included in L.*
AVGLinksetImp4p: $\mathcal{L} \times RDFProperties \rightarrow \mathbb{R}^+ \cup \{0\}$*;*
AVGLinksetImp4p(L,p) $= \frac{1}{|Links(L)|} * \sum_{e \in \{x | t_L(x,*,*)\} l \in Links(L)} LinkImp4p_L(e,p,l)$

Link importing for multilingual literals Multilingualism in literals can become pivotal depending on the properties considered. For example, when analysing `skos:exactMatch` linksets, importing can be applied to `skos:prefLabel`, `skos:altLabel` properties independently from the languages as well as for a specific set of languages. In order to consider specific languages and to measure the impact of a property for each of these languages the above formulas can be rewritten as LinksetImp4p(L,p,lang) by replacing val4Prop$_X$(e,p) with val4Prop$_X$(e,p,lang) in Definition 4.

4 Importing Validation

The validation aims at demonstrating the ability of AVGLinksetImp4p, proposed in Section 3 as measure of quality for linksets, to evaluate the improvement in completeness of a dataset when this is complemented via linksets related information. Validation addresses the following research questions:

RQ 1 Does our measure detect linksets that do not bring advantages in term of completeness of the complemented dataset?

RQ 2 Does our measure detect linksets that bring advantages in term of completeness of the complemented dataset?

RQ 3 What about the reliability of the our measure results ? When does our measure provide reliable information for completeness? When is it not reliable?

In the following, we introduce the basic concepts of the methodology adopted to validate the scoring function AVGLinksetImp4p. Then, we present the modular validation architecture discussing in detail each components and the choices made.

4.1 Methodology Motivation and Principles

Due to the novelty of the research field of Linked Data quality, there exist, as far as we know, only a few benchmarks to validate aggregated quality measures and quality score functions (e.g., lodqa [2] and the LACT link specification [3]), and, unfortunately, none of them can be exploited to answer our research questions.

The validation methodology we have adopted is inspired by the settings used in the Ontology Alignment Evaluation Initiative (OAEI)[4] and further improved in [5] to evaluate ontology matching systems, since it faces similar issues. In fact, OAEI develops a flexible test generator with an extensible set of alterators which may be used programmatically for generating different test sets from different starting ontologies (i.e., seed ontologies) in order to provide a full-coverage of all the possible different situations that ontology matchers have to face, that is, the problem space. Test sets in OAEI are developed starting from a seed ontology, and then creating pairs of *altered* seed ontology. For each pair of *altered* seed ontologies, a reference alignment is provided (a.k.a., the ground truth) and compared with the alignment provided in output by the matching systems.

Analogously, we define a *Test Sets Generator* that provides an extensive collection of test sets. Test sets include a seed dataset, several pairs of *altered* seed dataset and a linkset for each pair. Thus, we are able to create a significant number of situations to analyse the problem space. Our research questions investigate the relation between AVGLinksetImp4p and the completeness of the complemented dataset, thus, our *ground truth*, is based on the completeness gain

[2] `http://lodqa.wbsg.de/`

[3] `https://github.com/LATC/24-7-platform/tree/master/link-specifications`

[4] `http://oaei.ontologymatching.org/`

reached in the complemented dataset. In order to estimate the completeness, we consider the seed dataset, which is the most *complete* dataset at hand, as the *gold standard*.

This methodology, as pointed out in [5], might suffer of the following drawbacks: (i) **lack of realism:** tests are artificially created to cover a problem space thus they are not necessary representative of all the issues that can be encountered in the reality; (ii) **lack of variability:** it is not possible to vary the seed and the applied transformations; (iii) **lack of discriminability:** tests are not able to really discriminate between matchers, since they are not enough difficult.

Concerning the lack of realism, the goal of our system is to validate our AVGLinksetImp4p (a.k.a., the linkset importing) analysing its behaviour in the critical situations which might lead to complementation incompleteness with respect to some specific properties and considering a specific linkset. Thus, we limit the problem space to cover the following linkset issues: (i) the linkset does not provide any importing for the considered properties; (ii) the linkset imports few properties values; (iii) the linkset imports enough properties values to fill up the gold standard; (iv) the linkset covers a very limited number of entities exposed in the datasets.

Lack of variability and discriminability are addressed, as suggested in [5], developing a flexible, extensible, open architecture[5]. Our architecture provides a *Test Sets Generator* module that, through a fine tuning of parameters, ensures the possibility to vary the seed dataset and to perform random alteration on the seed dataset with different precision, with the aim of fully-cover the problem space. The framework can be extended by third parties to provide further alterators or seed datasets to enlarge the problem space.

As discussed in Section 1, we focus on SKOS thesauri interlinked with `skos:exactMatch` relation, considering the `skos:prefLabel` and `skos:altLabel` properties of the SKOS concepts. Since, we want to deal with several SKOS thesauri and `skos:exactMatch` linksets within the EU project eENVplus [6].

4.2 A Modular Validation Architecture

In this section we present the validation architecture, shown in Figure 1. It has been implemented using Java, and in particular the technology provided by Jena API to manage RDF datasets, and it is made by two main modules: the *Test Sets Generator* and the *Importing and Completeness Assessment* modules.

The *Test Sets Generator* module performs two essential tasks. First, the creation, from the seed thesaurus T of the subject and object thesauri and of the `skos:exactMatch` linkset between them. T may be real or artificially generated and it represents the gold standard. Second, the alteration of the subject, object thesauri and of the linkset to generate several test sets and to possibly provide a full-coverage of the problem space.

[5] The framework is available at `http://purl.org/net/linksetq`
[6] http://www.eenvplus.eu/

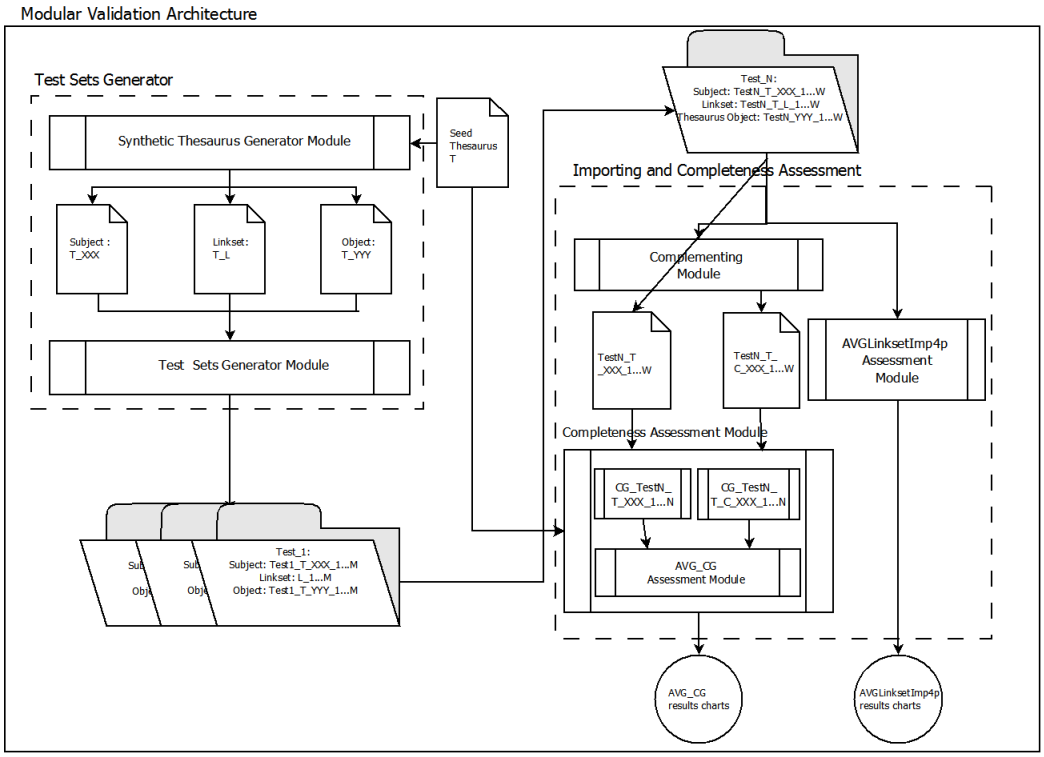

Fig. 1. Modular Validation Architecture

The *Importing and Completeness Assessment* module provides the assessment of AVGLinksetImp4p and of the completeness gain of the complemented thesaurus with respect to the gold standard.

Test Sets Generation. The goal of these components is to provide an extensive collection of test sets representative enough to possibly fully-cover the problem space. This component takes in input a seed thesaurus T, that is elaborated by the *synthetic thesaurus generator* module. Such a component duplicates T in two datasets T_XXX and T_YYY changing the original namespace in two different namespaces, in order to have the same concepts with the same properties in both thesauri. A linkset T_L is then generated between them. The datasets T_XXX, T_YYY and T_L are taken in input by the *Test Set Generator* module that, applying some modifications on each of the three creates different test sets.

Importing and Completeness Assessment. This component evaluates the AVGLinksetImp4p for the linksets in the generated test sets and the completeness gain of the complemented thesaurus with respect to these linksets. It relies on the notions of (i) *Thesaurus values restricted to property p*, (ii) *Completeness*

Table 2. Description of the Test sets generated.

Test 1. Alteration of *T_XXX*. *T_YYY* and *T_L* do not change. Deletion in subject thesauri: *Test 1.1*: 10% of `skos:prefLabel` and `skos:altLabel`; *Test 1.2:* 30% of `skos:prefLabel` and 10% `skos:altLabel`; *Test 1.3:* 60% of prefLabel and 50% `skos:altLabel`; *Test 1.4:* 100% of `skos:prefLabel` and 0% `skos:altLabel`;	**Test 2.** Alteration of *T_YYY*. *T_XXX* and *T_L* do not change. Deletion in object thesauri: *Test 2.1:* 10% of `skos:prefLabel` and `skos:altLabel`; *Test 2.2:* 30% of `skos:prefLabel` and 10% `skos:altLabel`; *Test 2.3:* 60% of prefLabel and 50% `skos:altLabel`; *Test 2.4:* 100% of `skos:prefLabel` and 0% `skos:altLabel`;
Test 3. Alteration of *T_L*, while, *T_XXX* and *T_YYY* do not change. Creation of different linkset deleting the 10% (Test 3.1), 30% (Test 3.2), 50% (Test 3.3), 90% (Test 3.4), 99% (Test 3.5) and 99.9% (Test 3.6) of `skos:exactMatch`	
Test 4. Alteration of *T_XXX* and *T_YYY*; *T_L* does not change. 8 different combinations of 2 subjects, 4 objects and one linkset. Deletions in *T_XXX*: *T_XXX_1*: 90% of `skos:prefLabel` and 50% `skos:altLabel`; *T_XXX_2*: 100% of `skos:prefLabel` and 90% `skos:altLabel`; and for *T_YYY*: 10% of `skos:prefLabel` and 10%`skos:altLabel`; 30% of `skos:prefLabel` and 10% `skos:altLabel`; 60% of `skos:prefLabel` and 50% `skos:altLabel`; 100% of `skos:prefLabel` and 90% `skos:altLabel`; The test sets: (i) Test 4.1/4.2/4.3/4.4 has *T_XXX_1* as fixed subject thesaurus and change the object thesauri; (ii) Test 4.5/4.6/4.7/4.8 has *T_XXX_2* as fixed subject thesaurus and change the object thesauri; the linkset *T_L* is the same for all the tests.	
Test 5. Alteration: *T_XXX* and *T_YYY* and *T_L*, 48 different combinations of 2 subjects, 4 objects 6 and linksets. Modification for the subject *T_XXX*: (i) *T_XXX_1*: 90% of `skos:prefLabel` and 50% `skos:altLabel`; (ii) *T_XXX_2* 100% of `skos:prefLabel` and 90% `skos:altLabel`. Modification for the object *T_YYY*: (i) *T_YYY_1*: 10% of `skos:prefLabel` and 10%`skos:altLabel`; (ii) *T_YYY_2*: 30% of `skos:prefLabel` and 10% `skos:altLabel`; (iii) *T_YYY_3*: 60% of `skos:prefLabel` and 50% `skos:altLabel`; (iv) *T_YYY_4*: 100% of `skos:prefLabel` and 90% `skos:altLabel`. Modification for the linkset *T_L* deleting: (i) *T_L_1*: 10% of `skos:exactMatch` ; (ii) *T_L_2*: 30% of `skos:exactMatch` ; (iii) *T_L_3*: 50% of `skos:exactMatch` ; (iv) *T_L_4*: 90% of `skos:exactMatch` ; (v) *T_L_5*: 99% of `skos:exactMatch`; (vi) *T_L_6*: 99.9% of `skos:exactMatch`. Test sets organized in groups for X=1, ..., 5 as follows $Test_5.X$ groups a sub-set of tests where only the linkset *T_L_X* is fixed, while from $Test_5.X_5.1$ to $Test_5.X_5.4$ the subject is fixed as *T_XXX_1* and the object change from *T_L_1* to *T_L_6*; and, from $Test_5.X_5.5$ to $Test_5.X_5.8$ the subject is fixed as *T_XXX_2* and the object change from *T_L_1* to *T_L_6*.	

wrt Gold Standard and (iii) *Average Completeness Gain wrt Gold Standard*. In the following, we consider the thesauri $T, G \in \mathcal{D}$, where T is a subject thesaurus in one of the test sets, and G is the gold standard. Moreover, the property p belongs to $TPrp$=$\{$`skos:prefLabel`, `skos:altLabel`$\}$.

The concepts of restriction of a thesaurus T to a specific property p is the set of `skos:Concept` in T having the property p.

Definition 6. ***(T restricted to property p)*** *The restriction of T wrt the property p is defined as follows:*
$T|_p = \{c \mid c \text{ is a } \texttt{skos:Concept} \wedge t_T(c,p,*)\}$

The notion of completeness of a thesaurus with respect to the gold standard, is derived by property completeness for datasets [13][6], and corresponds to the comparison between the number of values for the property p in the considered thesaurus and the number of values for p in the gold standard.

Definition 7. ***(Completeness wrt Gold Standard)*** *The completeness of T wrt G for p is defined as follows:*
$CG(T,G,p) = \frac{1}{|G|_p|} \sum_{t \in Entities(T)} \frac{|val4Prop_T(t,p,*)|}{|val4Prop_G(t,p,*)|}$

Using the notion of completeness wrt the gold standard it is immediate to define the average completeness gain wrt gold standard. The average completeness gain calculates the increment of completeness of a considered thesaurus after its complementation using a specific linkset L.

Definition 8. ***(Average Completeness Gain wrt Gold standard)*** *Let L be a linkset, T^L the complemented thesaurus. The average completeness gain of T wrt G for the property p is defined as follows:*
$AVG\_CG(G, T, T^L, p) = CG(T, G, p) - CG(T^L, G, p)$

The *Importing and Completeness Assessment* module takes in input all the test sets created by the *Test Sets Generator*. Thus, let consider a test set *TestN*, let $TestN\_i$=< *TestN_T_XXX_i, TestN_T_L_i, TestN_T_YYY_i*> (i.e., <subject_thesaurus, linkset, object_thesaurus>), where each of its triples is generated as described in Table 2. The AVGLinksetImp4p is evaluated directly on the linksets, considering the triple $TestN\_i$. On the other side, the completeness evaluation requires a further step, the generation of the complementation of the subject thesaurus *TestN_T_XXX_i* with the object thesaurus *TestN_T_YYY_i* via *TestN_T_L_i*. We identify with *TestN_T_C_XXX* the result of the *Complementing Module*. Then, the *Completeness Assessment Module* takes in input the subject thesaurus *TestN_T_XXX_i* and its complemented *TestN_T_C_XXX* and calculates the average completeness gain (AVG_CG).

5 Experimental Results

The goal of the experimental evaluation is to investigate the effectiveness of AVGLinksetImp4p in the evaluation of the average completeness gain (AVG_CG) of a thesaurus complemented via a specific linkset. We analyse the behaviour of these two functions on the synthetic test sets presented in Table 2. In the set up of the validation, we have considered the GEneral Multilingual Environment Thesaurus (GEMET) as seed thesaurus. GEMET is a `cc-by` licensed thesaurus which includes 5209 `skos:Concepts` with `skos:prefLabel` and `skos:altLabel` in more than 30 languages. Thus, we have performed the experiments considering both `skos:prefLabel` and `skos:altLabel`. However, due to space limitations and considering that the results obtained for the two properties are consistent, we present only the result for p =`skos:prefLabel`.

Figure 2 shows on the x axis all the test sets considered and on the y axis the values for AVGLinksetImp4p, AVG_CG and $|Links(L)|/|G|_p|$. The last function $|Links(L)|/|G|_p|$ is based on the cardinality of the linkset ($|Links(L)|$) and on the cardinality of the gold standard restricted to property p ($|G|_p|$). $|Links(L)|/|G|_p|$ derives from linkset completeness [7] and linkset coverage [1], in fact, it represents the *coverage* of the linkset wrt the entities in the gold standard with property values for p.

We can observe that:

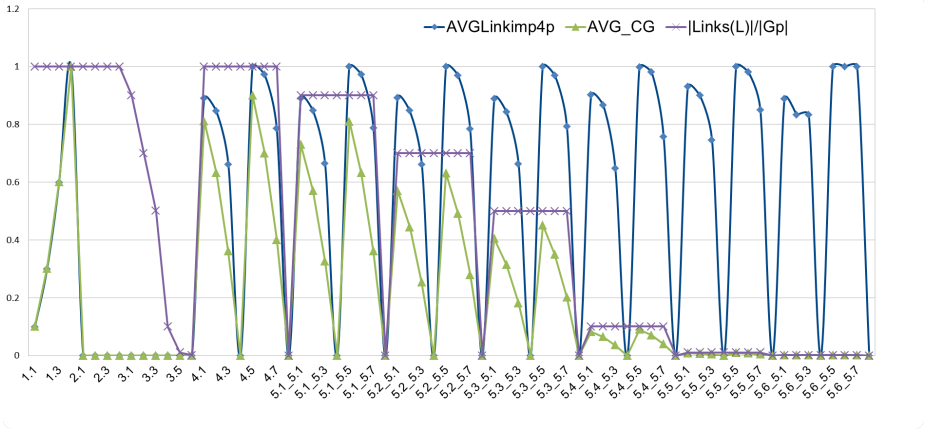

Fig. 2. AVGLinksetImp4p, AVG_CG and $|Links(L)|/|G|_p|$ considering `skos:prefLabel`

- the average importing function (i.e., AVGLinksetImp4p) is an upper bound of the average gain (AVG_CG);
- for the whole test sets 1, 2 and 3, AVGLinksetImp4p and AVG_CG have exactly the same value. In particular, in test set 1 we delete information only in the subject, thus, through the interlinking, that is complete, we capture in the object dataset all the information necessary to complete the subject. For test sets 2 and 3, both AVGLinksetImp4p and AVG_CG are zero, in fact, the subject is complete, consequently we do not import any information. This is an interesting example showing that an high number of links, 5220 in this case, is not necessarily synonymous with good quality linkset. In fact, you can have a linkset with a great number of links that do not bring any advantages in terms of completeness gain;
- AVGLinksetImp4p has an unexpected behaviour for the group of test 5.5_$*$ and 5.6_$*$. In both situations even if the AVG_CG is zero the AVGLinksetImp4p value is high.

To explain the unexpected situation we can look further at Figure 2. It is clear that the AVG_CG is affected by the coverage. A low coverage means few links import `skos:prefLabel` values, the consequences are twofolds: (i) a low AVG_CG since for the majority of the considered entities, AVG_CG is zero, because, they are not involved in the interlinking; (ii) on the other hand, these few links might import a significant percentage of `skos:prefLabel` values for single link. In this case, AVGLinksetImp4p is high, but, it is representative only of a small subset of the entities considered, so it might be misleading for the completeness gain of the overall complemented thesaurus.

Following these considerations, we decide to normalize AVGLinksetImp4p using the coverage as coefficient, and to adopt this normalized version as the *linkset importing* function, identifying it with **NormAVGLinksetImp4p**. In Figure 3 NormAVGLinksetImp4p is compared with AVG_CG (with test sets on

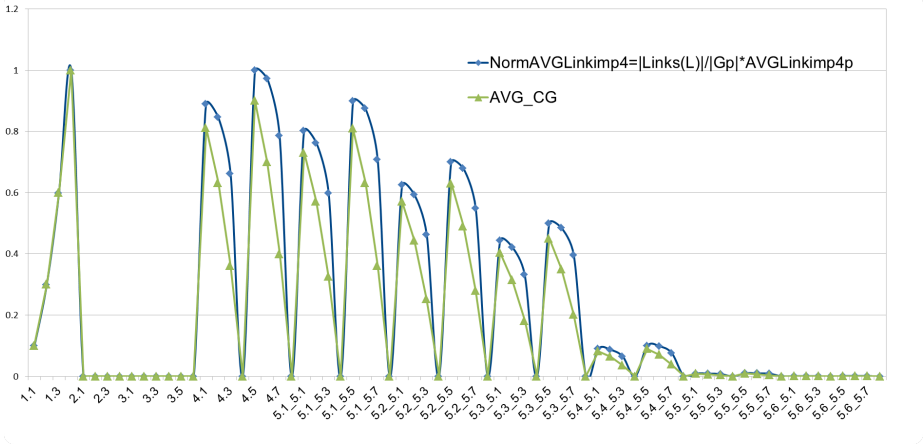

Fig. 3. Normalized AVGLinksetImp4p and AVG_CG considering `skos:prefLabel`

the x axis and on the y axis the function values). NormAVGLinksetImp4p is still an upper bound for AVG_CG; it still corresponds exactly to the AVG_CG in test sets 1, 2 and 3, where the complemented thesaurus corresponds to the gold standard. Besides, referring to the research questions discussed in Section 4: **(RQ1)** NormAVGLinksetImp4p is able to foresee when a linkset contributes to a completeness gain of the complemented thesauri (NormAVGLinksetImp4p>0) and **(RQ2)** when it does not (NormAVGLinksetImp4p=0). Concerning the reliability of the NormAVGLinksetImp4p results **(RQ3)**, we can say that the precision of the NormAVGLinksetImp4p depends on how much the complemented thesaurus property values are near to the gold standard property values. If they are nearly the same, the precision of NormAVGLinksetImp4p is high, if they are very different, the precision of NormAVGLinksetImp4p is low.

6 Related work

A recent systematic review of quality assessment for linked data can be found in the SWJ submission [13] and in the deliverable produced by EU funded project PlanetData [9]. Both the works review quality dimensions which are traditionally considered in data and information quality (e.g., availability, timeliness, completeness, relevancy, availability, consistency) as well as more Linked Data specific dimensions such as licensing and interlinking. Among the measure reviewed in these two works, we discuss in this section the measures for completeness and interlinking which are those more closely related to the contribution of this paper. In [13], completeness is measured in terms of (i) *schema completeness*, the degree to which the classes and properties of an ontology are represented [6] [10], (ii) *property completeness*, the measure of the missing values for a specific property [6], (iii) *population completeness*, the percentage of all real-world objects of a particular type that are represented in the datasets [6, 10], These measures

basically correspond to the notions of intensional, extensional and LDS completeness discussed in [9] and are defined for datasets, not for measuring Linkset quality. We have considered the *property completeness* to calculate the quality gain of the complemented thesaurus. Nevertheless, dataset quality and linkset quality still measure different things.

More related to our contribution is the framework LINK-QA [7], which is also discussed in [13] under the interlinking quality dimension. It defines 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 importing substantially differs from the scoring functions proposed in LINK-QA: (i) LINK-QA works on links independently from the fact that they are part or not of the same linksets; (ii) LINK-QA addresses correctness of links, it does not deal with gain in completeness of the complemented[7]. As pointed out in [13], LINK-QA also proposes an interlinking completeness which determines the degree to which entities in the dataset are interlinked. This completeness measure is closely related to our previous work [1]. [1] proposes a set of scoring functions for assessing `owl:sameAs` linkset quality. It strongly relies on the notion of types, namely classes of the entities exposed in a dataset and its related Linksets, and includes (i) linkset type coverage, which returns the percentage of types in a datasets that have been also considered in the linkset; (ii) linkset type completeness, which returns the percentage of mappable types in a datasets that have not yet been considered in the linksets; (iii) linkset entity coverage for type, which returns what percentage of entities having a given type in a dataset are also involved in the analysed linkset. In particular, linkset entity coverage for type specializes the interlinking completeness proposed in [7], it works out the completeness but grouping entities according to their type. A measure ($|Links(L)|/|G|_p|$) derived by this coverage has been applied as multiplicative coefficient to normalize the importing measure proposed in this paper. Besides, as discussed in the experimentation the importing measure goes beyond linkset completeness and linkset entity coverage for type, it measures the contribution in complementation that can carried by links, not only the extent to which links are available for entities.

A set of quality measures specific for SKOS thesauri relevant for this paper have been proposed in [12]. The paper summarizes a set of 26 quality issues for SKOS thesauri and shows how these can be detected and improved by deploying qSKOS [8], PoolParty checker, and Skosify [11]. Among the mentioned issues, incomplete language coverage is particularly worth for our work. Incomplete language coverage arises when the set of language tags used by the literal

[7] The quality dimensions addressed by LINK-QA are not explicitly stated. We exclude that LINK-QA considers completeness, since, it tries to correlate network measures and bad link detection. Moreover, 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.

values linked with a concept are not the same for all concepts. Our measure assesses the goodness of a linkset when complementing a thesauri to import further `skos:altLabel` and `skos:prefLabel`, and might represent a shortcut to address incomplete language coverage. Unfortunately, an analysis on linksets among thesauri is not included in [12]: missing out-links and in-links are adopted as indicator of SKOS thesaurus quality, but, their potential for thesaurus complementation and importing of `skos:prefLabel` and `skos:altLabel` values, when dealing with incomplete language coverage, is not considered.

7 Conclusions and Future Work

In this paper, we make a step towards the Linked Data quality assessment, a still open and critical research issue. Our contributions can be considered from two different points of view. On one hand, we draw attention to the critical issue of the linkset quality. In fact, we directly address the definition and the assessment of linkset quality measures, while, the majority of existing works focus on dataset quality. We provocatively state, that, in the evolution of the Web of Data into the Global Data Space, linksets should have the same importance of datasets thus, linksets quality should be considered as an independent branch of Linked Data quality, and not simply as one of the quality dimensions in the dataset quality assessment. On the other hand, we take a step towards in the Linked Data quality assessment. We formalize a linkset quality measure, the *linkset importing* function, which evaluates linkset potential when complementing datasets with their interlinked information. We validate our measure on `skos:exactMatch` linksets by considering the properties `skos:prefLabel` and `skos:altLabel`. The validation shows that the normalized version of *linkset importing* is a good predictor for the `skos:prefLabel` and `skos:altLabel` completeness gain in the complemented thesaurus. We provide a modular validation architecture which is a small open in-house benchmark, that might be extended in future for evaluating other measures (also provided by third parties) addressing a larger problem space. Future work includes the application of our scoring function on a set of real linksets, i.e., the linksets among environmental thesauri developed by the EU project eENVplus to evaluate the potential of role these linksets to improve their multilingual support. Furthermore, a future interesting issue is the investigation of the behaviour of the normalized average importing on other kind of linksets (e.g., `owl:sameAs`).

Acknowledgements. This research activity has been partially carried out within the EU funded project eENVplus (CIP-ICT-PSP grant No. 325232).

References

1. R. Albertoni and A. Gómez-Pérez. Assessing linkset quality for complementing third-party datasets. In G. Guerrini, editor, *EDBT/ICDT Workshops*, pages 52–59. ACM, 2013.

2. R. Albertoni, M. D. Martino, and P. Podestà. Environmental thesauri under the lens of reusability. In *EGOVIS 2014*, volume 8465 of *Lecture Notes in Computer Science*, pages 222–236. Springer, 2014. to appear.
3. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets with the VoID Vocabulary, 2011.
4. C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. *J. Web Sem.*, 7(1):1–10, 2009.
5. J. Euzenat, M.-E. Rosoiu, and C. T. dos Santos. Ontology matching benchmarks: Generation, stability, and discriminability. *J. Web Sem.*, 21:30–48, 2013.
6. C. Fürber and M. Hepp. Swiqa - a semantic web information quality assessment framework. In V. K. Tuunainen, M. Rossi, and J. Nandhakumar, editors, *ECIS*, 2011.
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. C. Mader, B. Haslhofer, and A. Isaac. Finding quality issues in skos vocabularies. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *TPDL*, volume 7489 of *Lecture Notes in Computer Science*, pages 222–233. Springer, 2012.
9. 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.
10. 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.
11. O. Suominen and E. Hyvönen. Improving the quality of skos vocabularies with skosify. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 383–397. Springer, 2012.
12. O. Suominen and C. Mader. Assessing and improving the quality of skos vocabularies. *J. Data Semantics*, 3(1):47–73, 2014.
13. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked open data: A survey. *Submitted to Semantic Web Journal*, 2014.