A linkset quality metric measuring multilingual gain in 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 connections of information belonging to distinct datasets, that might be as important as dataset’s quality when consuming Linked Data. In this paper, we present a first linkset quality measure proposing a function able to estimate the new information gained through linksets among SKOS thesauri. A scoring function, the linkset importing is provided focusing on the multilingual gain, in terms of the new translated labels, obtained by complementing a SKOS thesaurus through skos:exactMatch links. We finally discuss how the linkset importing can be significantly used in the context of the EU project eENVplus.

1 Introduction

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 [14], data is only worth its quality. The research pertaining to Linked Data quality is especially focused on datasets [14]. However, one of the most interesting promises of Linked Data is “Linked Data will evolve the current web data into a Global Data Space” that implies 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 the way of linksets. Through linksets a Linked Data consumer can reach new information to complete and enrich data at hand, so, in order to keep the Linked Data promise, the quality of connections (hereinafter linkset quality) are as important as the quality of data. This paper proposes a method to shed light on this. It presents a measure, the linkset importing, estimating the linkset quality as the ability of a linkset to enrich a dataset with new properties values.
We are aware that quality is a multidimensional issue, and that, in analogy to the quality for dataset, even the quality of linkset might have different dimensions (e.g., correctness, completeness, trustworthiness). In fact, with the linkset importing we focus on an aspect of linkset quality the dimension completeness, more precisely, the completeness of a dataset obtained when complementing a dataset via its linkset. Linkset importing extends the linkset quality introduced in [2] focusing on skos:exactMatch linksets among thesauri exposed as Simple Knowledge Organization System (SKOS) Ontology in the Linked Data. This type of linksets and of datasets has been chosen considering the application scenarios we are facing in the EU funded project eENVplus (CIP-ICT-PSP grant No. 325232), where we deal with a remarkable number of environmental thesauri exposed as Linked Data [4] and with their skos:exactMatch linksets. Considerable efforts have been spent to interlink thesauri such as GEMET, EARTH, AGROVOC, EUROVOC, UNESCO, RAMEAU, TheSoz, but, currently, there is no way to assess the value of these interlinks in terms of usefulness and information gain. To this purpose, the linkset importing can be exploited to check the linkset complementation potential for any SKOS property; in particular, we focus on skos:prefLabel and skos:altLabel, in order to address the incomplete language coverage issue (see [12]), which affects many popular SKOS thesauri.

The organization of the paper is as follows: Section 2 introduces basic concepts such as 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 applies the linkset importing in an example which is grounded in the context of the EU project eENVplus. Finally, we discuss related work in Section 5 and the conclusions and future work in Section 6.

## 2 Basic Concepts

This paper considers resources on the Web represented using the Resource Description Framework (RDF)\(^2\). In particular, we use the notion of dataset and linkset provided in the Vocabulary of Interlinked Datasets (VoID)[7], an RDF vocabulary commonly adopted for expressing metadata about RDF datasets exposed as Linked Data. 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 between two datasets, defined the void:subjectsTarget and the void:objectsTarget, representing respectively the object and the subject of the linkset. Each RDF link is RDF triple (s,p,o), where s, p, o are generically indicated as RDF terms (hereafter, the set RDFTerms); more in detail, s and o, belonging respectively to the subject and the object datasets, may be RDF resources denoted by an IRI (hereafter, the set RDFRiri) (e.g.,

---

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.

2 [http://www.w3.org/TR/rdf11-primer/]
http://dbpedia.org/resource/Tectonics) and \( o \) may range also in RDF literals (hereafter, the set \( \text{RDFLit} \) (e.g., Dog) or RDFlit with ISO language tags\(^3\) (hereafter, the set \( \text{RDFLitLtag} \)) (e.g., Dog@en). While, \( p \) is a RDF property (hereafter, the set \( \text{RDFProp} \)) (e.g., skos:exactMatch) 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 with equivalent meaning.

This paper adapts the notion of complementation via a linkset introduced in [2] to SKOS thesauri. Given two thesauri \( X \), \( Y \) and a linkset \( L \) linking some concepts in \( X \) with some concepts in \( Y \), \( X \) can be complemented with \( Y \) via \( L \) resulting in a third thesaurus identified with \( X^L \). Informally, \( X^L \) contains all RDF triples of \( X \) and the SKOS/RDF triples reachable in \( Y \) via \( L \). Formally, let \( D \) be a dataset and \( t_D(s, p, o) \) be a predicate holding if and only if the RDF triple \( (s, p, o) \in D \), \( p \) a SKOS property, we define: 

\[
X^L = \{(s, p, o) | t_X(s, p, o) \vee t_L(s, \text{skos:exactMatch}, y) \wedge t_Y(y, p, o)\}.
\]

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 [5] 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. Scoring functions are functions evaluating quality indicators to measure the suitability of the data for some intended use. 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 the following sub-sections, we formalize two indicators and the importing scoring function. We do not provide explicitly any aggregate metrics.

3.1 Indicators

We present the indicator val4Prop that given an RDFRiri \( e \) of dataset \( X \) returns all the values associated to \( e \) for a specific RDF property, with the possibility to specify or not (using \( . \)) a language tag.

\(^3\) http://tools.ietf.org/html/bcp47#section-2.2.9
Definition 1 Let \( e \) be a RDF\R{iri} of dataset \( X \), \( p \) be a RDF\Prop, and \( lang \) be in RDF\Lit\Ltag \( \cup \{ \} \). We define:

\[
\text{val4Prop}_X(e,p,lang) = \begin{cases} 
\{ v | t_X(e,p,v) \} & \text{if } lang = \_ \\
\{ v@lang | t_X(e,p,v@lang) \} & \text{otherwise.}
\end{cases}
\]

Then, given an RDF\Terms \( e \) and a linkset \( L \), we define an operator \([ ]_L\) that returns \( e \) if \( e \) is not involved in any skos:exactMatch link or \( e \)'s skos:exactMatch-linked RDF\Terms otherwise.

Definition 2 Let \( L \) be a linkset, \( Z \) a set of RDF\Terms not including blank nodes. The operator \([ ]_L\) is defined as follows:

\[
[Z]_L = \{ y | z \in Z \land (t_L(z, \text{skos:exactMatch}, y) \lor (\neg t_L(z, \text{skos:exactMatch}, y) \lor z \in \text{RDF\Lit} \land y = z)) \}.
\]

Example 1. Considering datasets \( X \) and \( Y \) and linkset \( L \) in Fig. 1, \( \text{val4Prop}_X(x_2, \text{skos:prefLabel}, \text{en}) = \{ \text{Snake@en} \} \), whilst \( \text{val4Prop}_X(x_2, \text{skos:prefLabel}, \_ ) = \{ \text{Snake@en,Serpente@it} \} \), since in the latter there is no constraint on the language tag. Moreover, \( [\{ \text{Dog@en} \}]_L = \{ \text{Dog@en} \} \), since \( \{ \text{Dog@en} \} \subset \text{RDF\Lit\Ltag} \subset \text{RDF\Lit} \), and \( [\{ x_2, x_5 \}]_L = \{ x_2, y_5 \} \), since \( x_2 \) has no skos:exactMatch link, and \( y_5 \) is the skos:exactMatch-linked RDF\Term for \( x_5 \).

3.2 Scoring functions

Using the indicators presented in the previous section, we define now, the scoring functions characterizing our linkset quality measure. Linkset importing scoring function evaluates the percentage of “gained values” for a RDF property \( p \). “Gained values” are values not already present in the subject dataset \( X \), but reachable through the linkset \( L \) in object dataset \( Y \). Linkset importing assumes that the linkset correctness has been previously validated. In the following,

\footnote{RDF\Triples belonging to \( L \) are completely known. We assume the \( L \)'s RDF\Dump or SPARQL endpoint is specified in \( L \)'s VOID description. Thus \( \neg t_L(z, \text{skos:exactMatch}, y) \) can be verified under the close-world assumption.}
we present the importing scoring function for a single link, then, we generalize defining the average importing scoring function for the whole linkset $L$.

**Definition 3** Let $e \in RDFRiri$, $l \in L$ and $lang \in RDFLitLtag \cup \{\_\}$. The link importing for $e$ considering property $p$ through $l$ is defined as follows:

$$LinkImp_{4p_{L}}(e,p,l,lang) = \begin{cases} 0 & \text{if } den = 0 \\ LkImp_{4p_{L}}(e,p,l,lang) \times 100 & \text{otherwise} \end{cases}$$

where

$$LkImp_{4p_{L}}(e,p,l,lang) = 1 - \frac{|val_{Prop_{X}}(e,p,lang)|\bigcup|val_{Prop_{X}}(x_{3},l_{2},en)\bigcup|val_{Prop_{X}}(\{x_{3}\},\{l_{3}\},br,\_)|}{den}.$$  

Example 2. Considering the properties $pl = skos:prefLabel$, $al = skos:altLabel$ and $br = skos:broader$ showed in Fig. 1. $LinkImp_{4p_{L}}(x_{3},pl,l_{2},\_)$ = 100 * (1 - $\frac{|\{Dog@en\}|\bigcup|\{Dog@en\}|\bigcup|\{Dog@en,Cane@it\}|}{\text{den}}$) = 50% and $LinkImp_{4p_{L}}(x_{3},al,l_{2},en) = 0$% are, respectively, the percentage of new $pl$ in any language and new $al$ in English gained by $x_{3}$ via $l_{2}$. $LinkImp_{4p_{L}}(x_{5},br,l_{3},\_)$ = 100 * (1 - $\frac{|\{x_{3}\}|\bigcup|\{x_{5}\}|\bigcup|\{x_{5}\}|\bigcup|\{y_{6}\}|}{\text{den}}$) = 50% is the percentage of broader entities gained by $x_{5}$ via $l_{3}$. Only $y_{6}$ is gained, since $y_{3}$ is considered a duplication of $x_{3}$ ($\{x_{3}\}_{L} = \{y_{3}\}$).

The function measures the gain in completeness when complementing via a linkset, as a consequence, it returns 100% if and only if new values from the link object are imported for an empty subject. Generalizing to the entire linkset.

**Definition 4** Let $lang \in RDFLitLtag \cup \{\_\}$, the importing capability of $L$ with respect to $p$ is defined as the average importing of all links included in $L$.

$$AVGLinksetImp_{4p}(L,p,lang) = \frac{1}{|L|} * \sum_{e \in \{x | t_{L}(x,\_),t \in L \}} LinkImp_{4p_{L}}(e,p,l,lang)$$

4 Application

A prototype of the average importing scoring function has been implemented in JAVA/JENA, and applied to evaluate the quality of linksets, developed in the context of eENVplus project, among environmental SKOS thesauri. We focus on two linksets E2GEM (4365 links) and E2AGR (1436 links) which have both EARTH, a thesaurus of 14351 concepts, as subject dataset and respectively GEMET and AGROVOC [6] as object datasets. E2GEM and E2AGR have been created and validated in the context of eENVplus project [3]. Our purpose is to investigate which of two linksets imports in EARTH the greater number of $skos:prefLabel$ and $skos:altLabel$ in different languages.

The results are shown in Fig. 2, where radial axes include: (i) one axis for each considered language, and (ii) an axis “total” representing the average importing
Fig. 2. Average linkset importing for E2AGR and E2GEM.

for all languages. Focusing on skos:prefLabel, see Fig. 2(a), the linkset importing quality of E2GEM is better than E2AGR, in fact: (i) the average importing (axis “total”) in E2GEM is higher than in E2AGR; (ii) linkset E2GEM imports a greater number of languages with respect to E2AGR. On the other hand, when we consider the average linkset importing for skos:altLabel, see Fig. 2(b), the result is exactly the opposite, E2AGR performs better than E2GEM. In fact, (i) the axis “total” shows that, in average, E2AGR imports more skos:altLabel than E2GEM; and (ii) E2AGR imports skos:altLabel translated in more languages than E2GEM. So which linkset is the best has no trivial answer. In general, the choice of the best linkset depends on the specific languages in which we are interested. For example, considering the importing for skos:prefLabel (Fig. 2(a)) the set of languages imported from E2GEM largely differs from those importable via E2AGR. In fact, only 10 out of the 40 considered languages are importable from both linksets (i.e., ar, ru, es, tr, pl, de, fr, hu, cs), about 19 out of 40 (e.g., bg, ga, fi, sl, eu, ro) can be imported only considering E2GEM, and 8 out of 40 only from E2AGR. While, for skos:altLabel (Fig. 2(b)), we import about 20 out of 40 languages from E2AGR and about 4 of 40 from E2GEM. As already discussed the linkset quality evaluation performed using the average linkset importing score function, is partial and other indicators and measures should be defined in order to fully characterize the quality of a linkset. Nevertheless, the results showed in Fig. 2 allow a finer analysis of the linkset than the currently used measures based on the simple number of links. In fact, just considering the number of links, E2GEM drastically outperforms E2AGR.

5 Related work

A recent systematic review of quality assessment for linked data can be found in the SWJ submission [14]. This paper reviews quality dimensions traditionally
considered in data quality (e.g., availability, timeliness, completeness, relevancy, availability, consistency) and Linked Data specific dimensions, such as licensing and interlinking, considering, for the latter, the framework LINK-QA [9] and the works [13], [1]. LINK-QA defines two network measures specifically designed for Linked Data (i.e., Open SameAs chains, and Description Richness) and three classic network measures (i.e., degree, centrality, clustering coefficient) for determining whether a set of links improves the overall quality of linked data. Whilst, [13] and [1] detect the quality of interlinking via crowd-sourcing. The main differences with respect to our linkset importing scoring function are: (i) [9], [13], [1] work on links independently from the fact that links are part or not of the same linksets; (ii) [9], [13], [1] address the correctness of links, and not the completeness of the complemented. A set of scoring functions measuring completeness of the complemented are instead proposed in our previous work [2], for owl:sameAs linksets. We extend such work presenting a new measure based on skos:exactMatch linkset. The paper [12] defines a set of 26 quality issues for SKOS thesauri and shows how these can be detected and improved by deploying qSKOS [10], PoolParty checker, and Skosify [11]. Incomplete language coverage, arising when the set of language tags used by the literal values of concepts are not uniform for all concepts, is one of the considered issues and it is also one of the problem affecting most the environmental thesauri exploited in eENVplus project. Unfortunately, [12] uses linkset specific issues (i.e., missing out-links and in-links) as quality indicator for “stand-alone” SKOS thesauri. Thus, the power of linksets in the importing of new translated skos:prefLabel and skos:altLabel values to address incomplete language coverage, is not considered.

6 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 is twofold. On one hand, we want to draw the community 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. In our point of view, for the evolution of the Web of Data into the Global Data Space, linksets have the same importance of datasets. As a consequence, linksets quality should be considered as an independent branch of Linked Data quality, and not simply as one of the dataset quality dimensions. On the other hand, we contribute to the Linked Data quality assessment formalizing the linkset importing scoring function ables to evaluate linkset potential when complementing thesauri with their interlinked information. Although linkset importing is not sufficient for a complete linkset quality assessment, it has offered a starting point to evaluate the gain in term of translated labels on real linksets developed in the EU project eENVplus. As future works, we plan to evaluate the quality for other skos:exactMatch linksets in the context of eENVplus project and to encode the related results in DAQ [8], so that, quality results will be browsable with
third-party RDF CUBE visualizers. Moreover, we plan to investigate the linkset importing on \texttt{owl:sameAs} linksets and to define scoring functions for a larger set of linkset quality dimensions.

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

References


