Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling

Johann Schaible[1], Thomas Gottron[2], and Ansgar Scherp[3]

[1] GESIS Leibniz-Institute for the Social Sciences, Cologne, Germany
johann.schaible@gesis.org
[2] Institute for Web Science and Technologies, University of Koblenz-Landau, Germany
gottron@uni-koblenz.de
[3] Kiel University and Leibniz Information Center for Economics, Kiel, Germany
mail@ansgarscherp.net

Abstract. The choice of which vocabulary to reuse when modeling and publishing Linked Open Data (LOD) is far from trivial. There is no study that investigates the different strategies of reusing vocabularies for LOD modeling and publishing. In this paper, we present the results of a survey with 79 participants that examines the most preferred vocabulary reuse strategies of LOD modeling. The participants, LOD publishers and practitioners, were asked to assess different vocabulary reuse strategies and explain their ranking decision. We found significant differences between the modeling strategies that range from reusing popular vocabularies, minimizing the number of vocabularies, and staying within one domain vocabulary. A very interesting insight is that the popularity in the meaning of how frequent a vocabulary is used in a data source is more important than how often individual classes and properties are used in the LOD cloud. Overall, the results of this survey help in better understanding the strategies how data engineers reuse vocabularies and may also be used to develop future vocabulary engineering tools.

Keywords: #eswc2014Schaible.

1 Introduction

With the increasing use of LOD, it becomes more and more important for data providers not only to publish their data as LOD, but also to model it in an easy to process way, i.e., make the data more human-readable and machine-processable. During the modeling process a data engineer has to—among many other tasks—decide with which vocabularies to express the data. Hereby, reusing vocabularies is clearly motivated by the best practices and recommendations for designing and publishing Linked Data [1]. Experienced Linked Data engineers decide which vocabularies to reuse based on their knowledge and "gut-feeling" in order to achieve several goals such as providing a clear structure of the data or making it easy to be consumed. Such goals, or aspects, lead to various vocabulary reuse strategies. For example, one might reuse only one domain specific vocabulary to provide a clear data structure, and the other might reuse popular vocabularies to make the data easier to be consumed. However, these strategies are quite

V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 457–472, 2014.

vague and not described in the literature in a formalized way. In fact, besides reusing "well-known" vocabularies, as it increases the probability that data can be consumed by applications [2], there are no established recommendations formulated on *how to choose* which vocabularies to reuse. This implies the challenge, especially for an unexperienced engineer, to decide on an appropriate mix of vocabularies optimally capturing the domain under investigation. More concrete, the Linked Data engineer needs to answer the question which vocabularies shall be used and how many shall be combined. There are various factors influencing the engineer's decision to reuse classes and properties from existing vocabularies. These factors include the popularity of a vocabulary, the match to the domain which is modeled, the maintenance of the vocabulary, the authority who has published the vocabulary, and others. Overall, deciding for which and how many vocabularies to reuse is a "non-trivial" task [3,4] and hardly addressed by today's research. Therefore, either the data engineer decides not to reuse vocabularies at all or the decision for which and how many vocabularies to reuse is solely based on the engineer's knowledge and experience. Thus, the main contribution of this paper is to condense and aggregate the knowledge and experience of Linked Data experts and practitioners regarding which reuse strategy to follow in a real-world scenario in order to achieve the previously stated goals.

Why this study? To the best of our knowledge, there is no study which empirically examines how to select vocabularies and vocabulary terms for reuse. More insights about the different factors and strategies that influence the engineers in their decision to select reusable classes and properties is needed. Such insights would provide guidance for the modeling process and aid the Linked Data engineer in deciding which vocabularies to reuse. In this study, we intend to identify these key factors and strategies.[1] To this end, we have conducted a survey among Linked Data practitioners and experts. The aim of the survey is to aggregate the knowledge and experience of these practitioners and experts to condense best practices and established approaches on how to select vocabularies for reuse.

We have asked the participants of the survey to rank several data models, which exemplify different vocabulary reuse strategies, from *most preferred* to *least preferred* with respect to the reuse of vocabularies. Such reuse strategies are "reuse mainly popular vocabularies", "reuse only domain specific vocabularies", or other. In addition, the participants had to answer different questions regarding their preferences when reusing vocabularies (cf. Section 2). We have obtained feedback from 79 participants acquired through public mailing lists (cf. Section 3). The main findings of our work are that reusing vocabularies directly is considered significantly better than defining proprietary terms and establishing links on a schema-level to other vocabulary terms. In addition, a trade-off should be made between reusing popular and domain specific vocabularies. Furthermore, additional meta-information on the domain of a vocabulary and on the number of LOD datasets using a vocabulary are considered the most helpful information for deciding which vocabulary to reuse (cf. Section 4 and Section 5). Overall, the results provide very valuable insights in how data engineers decide which vocabularies to reuse when modeling Linked Open Data (cf. Section 6). This may also lead to the

[1] An extended description of this study and a more detailed statistical analysis of its results is available as technical report in [5].

development of novel recommendation services for future vocabulary engineering tools (cf. Section 8).

2 The Survey

The survey consists of ranking tasks, where the participants have to decide which of the provided data models reuses vocabularies the best way, and explanations, where the participants have to rate different aspects why they have ranked the models the way they did. [2] Hereby, each data model represents a specific vocabulary reuse strategy such as reusing only popular or domain specific vocabularies. In Section 2.1, we define a set of features, which describes the data models and their underlying vocabulary reuse strategy, provide a detailed description of the survey design (Section 2.2), and finally illustrate and explain each of the data models in Section 2.3.

2.1 Features for LOD Modeling

To describe the differences of the data models that express the same example instance with different vocabularies and vocabulary terms, we make use of features such as the number of datasets using a vocabulary or the total occurrence of a vocabulary term. In general, such a set of features is based on datasets and vocabularies used in some LOD collection, e.g., a huge collection of RDF graphs that was crawled by a Linked Data crawler like the Billion Triple Challenge dataset.

Let $W = \{V_1, V_2, ..., V_n\}$ with $n \in \mathbb{N}$ be the set of all vocabularies used in the LOD cloud. Each vocabulary $V \in W$ consists of properties and type classes such that $V = P_V \cup T_V$. P_V is the set of all properties and T_V is the set of all classes in vocabulary V. Furthermore, let $\mathbb{DS} = \{DS_1, DS_2, ..., DS_m\}$ with $m \in \mathbb{N}$ be the set of all datasets in the LOD cloud. Each $DS \in \mathbb{DS}$ is a tuple $DS = (G, c)$ consisting of a context URI c of DS, where an RDF graph G can be found. G is a set of triples with

$$G = \{(s, p, o) | p \in URI, s \in URI, o \in (URI \cup LIT)\} \tag{1}$$

where URI is a set of URI's and LIT a set of literals. We define the function $\phi : \mathbb{DS} \rightarrow \mathcal{P}(W)$ that maps each dataset to the set of vocabularies used by the dataset

$$\phi((G, c)) = \{V \mid (\exists\,(s, p, o) \in G : p \in V) \vee (\exists\,(s, \text{rdf:type}, o) \in G : o \in V)\} \tag{2}$$

Hereby, $|\phi((G, c))|$ is the number of all used vocabularies in dataset DS. Accordingly, the function $\Phi : W \rightarrow \mathcal{P}(\mathbb{DS})$ specifies which datasets in the LOD cloud use a vocabulary $V \in W$

$$\Phi(V) = \{(G, c) \mid (\exists\,(s, p, o) \in G : p \in V) \vee (\exists\,(s, \text{rdf:type}, o) \in G : o \in V)\} \tag{3}$$

Therefore, $|\Phi(V)|$ is the number of datasets in the LOD cloud that use vocabulary V. To identify how often a vocabulary term $v \in V$ has occurred in the LOD cloud, we first

[2] The survey was designed with the online survey software *QuestBack Unipark* (`http://www.unipark.com/`) and is archived at the GESIS data repository service *datorium* (`http://dx.doi.org/10.7802/64`) including the raw result data in SPSS format.

define an auxiliary function $\psi : (V, \mathbb{DS}) \rightarrow \mathbb{N}$ that calculates the cardinality of the set of all triples $(s, p, o) \in G$ that include a vocabulary term $v \in V$ with

$$\psi(v, (G, c)) = |\{(s, p, o) \in G | v = p \vee (v = o \wedge p = \text{rdf:type})\}| \tag{4}$$

To finally calculate the overall occurrences of a vocabulary term $v \in V$ in the LOD cloud, we simply sum up the values $\psi(v, (G, c))$ over all $DS \in \mathbb{DS}$ with $\Psi : V \rightarrow \mathbb{N}$ that is defined as

$$\Psi(v) = \sum_{(G,c) \in \mathbb{DS}} \psi(v, (G, c)) \tag{5}$$

For the survey, we have retrieved metrics from LODStats [6] and the Linked Open Vocabulary index (LOV) [7] regarding the number of datasets using a specific vocabulary and vocab.cc [8] regarding the total occurrence of a vocabulary term.

2.2 Survey Design and Measurements

As mentioned, the survey consists of several ranking tasks and rating preferences regarding how much it influenced the ranking decision. For the ranking tasks, we provided several alternative data models, which exemplify different vocabulary reuse strategies, that had to be ranked from *most preferred* to *least preferred*. We let the participants rank such modeling examples instead of the reuse strategies directly in order to elude answers that are simply influenced by the theory of vocabulary reuse [1,2].

To illustrate the differences of the strategies, we use the previously defined features $\phi((G, c))$, $|\phi((G, c))|$, $|\Phi(V)|$, and $\Psi(v)$. The vocabularies in $\phi((G, c))$ provide information on which vocabularies have been used, e.g., some domain specific vocabularies, whereas the values of $|\Phi(V)|$ and $\Psi(v)$ provide information on the popularity of a vocabulary V and a vocabulary term v, respectively.

We consider the modeling examples and thus the underlying reuse strategies as different, if there is a difference in their features. For example, strategies like *minimize number of vocabularies* or *maximize number of vocabularies* are reflected by $|\phi((G, c))|$ that states the number of reused vocabularies. Listing 1.1 and Listing 1.2 provide two data models that describe the same example instance with different sets of vocabularies and different vocabulary terms.

```
<http://ex1.org/publ/01>
    rdf:type swrc:Publication;
    swrc:title "Title";
    swrc:author <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
    rdf:type swrc:Person;
    swrc:name "xyz".
```

Listing 1.1. Example data model M_a

- $\phi(M_a) = \{\text{swrc}\}$
- $|\phi(M_a)| = 1$
- $|\Phi(\text{swrc})| = 6$
- $\Psi(\text{swrc:Publication}) = 30$
- $\Psi(\text{swrc:title}) = 10,487$
- $\Psi(\text{swrc:author}) = 26,478$
- $\Psi(\text{swrc:Person}) = 30,510$
- $\Psi(\text{swrc:name}) = 35,756$

```
<http://ex1.org/pub/001>
    rdf:type swrc:Publication;
    dc:title "Title";
    dc:creator <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
    rdf:type foaf:Person;
    foaf:name "xyz".
```

Listing 1.2. Example data model M_b

- $\phi(M_b) = \{\text{swrc, dc, foaf}\}$
- $|\phi(M_b)| = 3$
- $|\Phi(\text{swrc})| = 6$
- $|\Phi(\text{dc})| = 287$
- $|\Phi(\text{foaf})| = 232$
- $\Psi(\text{swrc:Publication}) = 30$
- $\Psi(\text{dc:title}) = 3,605,629$
- $\Psi(\text{dc:creator}) = 1,653,155$
- $\Psi(\text{foaf:Person}) = 18,477,533$
- $\Psi(\text{foaf:name}) = 9,235,251$

Model M_a in Listing 1.1 follows the strategy to reuse only one domain specific vocabulary, namely the Semantic Web for Research Communities (SWRC[3]) vocabulary, and model M_b in Listing 1.2 follows the strategy to reuse popular vocabularies such as FOAF[4] and Dublin Core. [5] According to the features from Section 2.1, the FOAF and Dublin Core vocabularies are more popular than SWRC ($|\Phi(\text{foaf})| = 232 > 6 = |\Phi(\text{swrc})|$ and $|\Phi(\text{dc})| = 287 > 6 = |\Phi(\text{swrc})|$), which also applies to their classes and properties as indicated by the various values of Ψ. Nonetheless, the entire data model can be expressed with the SWRC vocabulary, and with $\Psi(\text{swrc:title}) = 10,487$ for example, SWRC is used in a few but quite large data sets. The central research question is to find out which vocabulary reuse strategies as the ones in M_a and M_b are considered better in a real-world scenario.

The different models and their strategies are based on several aspects of *preference* that we have identified from the state of the art about how to publish Linked Data [1,2]. In detail, they are: (A1) providing a clear structure of the data, (A2) making the data easier to be consumed, and (A3) establishing an ontological agreement in data representation. As part of our questionnaire, we asked the participants to rate these aspects on a 5-point-Likert scale at the beginning and after the first two ranking tasks, to investigate whether they have influenced the participant's ranking decision or not. Besides insights on the participant's answers, it allows us to make a qualitative correlation between the ratings and the rankings of the data models. For example, if aspect (A1) is considered the most important aspect and the ranking of the strategy which reuses only a minimum number of vocabularies is significantly the best, then this would suggest that in order to provide a clear data structure, one has to minimize the number of reused vocabularies instead of reusing popular vocabularies.

2.3 Ranking Tasks

The survey contains three ranking tasks, each covering a different aspect of the engineer's decision making process [3,9]. In the following, we will describe the different tasks, their motivation, and the used schema models (including the most decisive features). The models are fictive and prototypical for the different strategies. They are not real world schemas to prevent biased rankings as real-world schemas might be known

[3] `http://www.ontoware.org/index.html`, access 12/19/2013.
[4] `http://xmlns.com/foaf/spec/`, access 1/9/2014.
[5] `http://dublincore.org/documents/dcmi-terms/`, access 1/9/2014.

to some participants. The underlying strategies for the schemas are as follows: reuse popular vocabularies (*pop*), interlink proprietary terms with existing ones (*link*), minimize total number of vocabularies (*minV*), minimize number of vocabularies per concept (*minC*), confine to domain specific vocabularies (*minD*), and maximize number of vocabularies (*max*). Tables 1 to 3 illustrate for each ranking task the key features of the models and their underlying strategies. The upper section of the tables displays the reused vocabularies, and the lower section displays the most decisive vocabulary terms in the meaning of their total occurrence as in $\Psi(v)$. A "✓" in the table cells indicates whether the specific vocabulary V or vocabulary term v is used in the schema model, whereas a "$-$" indicates that the specific V or v is not used in the schema. The values in the last two columns show the features of the vocabularies ($|\Phi(V)|$) and their terms ($\Psi(v)$). Please note, meta-information such as $|\Phi(V)|$ and $\Psi(v)$ were provided to the participants only in the third ranking tasks for two reasons: (i) for the first two ranking tasks the goal was to aggregate and condense the participant's experience and "gut-feeling" without having these numbers at hand, and (ii) the third ranking task investigates how such meta-information influences the participant's ranking decision.

Furthermore, all data models within a ranking task describe data from the same domain (important for comparability). Between the ranking tasks, the models are from different domains (important to avoid domain-specific bias).

Ranking Task T_1: Reuse vs. Interlink. The first ranking task is about reusing vocabularies vs. establishing links on schema-level. We provided the participants with three schema models (displayed in Table 1). Each model expresses the same example instance, which represents an *Actor* who played in a certain *Movie*, with a different vocabulary reuse strategy. Model M_{1a} reuses vocabulary terms from the FOAF and Dublin Core vocabularies directly, which is considered very popular as indicated by the values $|\Phi(V)|$ and $\Psi(v)$, i.e., it follows the *pop* strategy. On the other hand, model M_{1b}, uses a self-defined vocabulary but links its classes and properties to the FOAF and Dublin Core vocabularies via rdfs:subClassOf and owl:equivalentProperty. It is arguable whether M_{1a} or M_{1b} is more likely to achieve such goals as provided in the aspects (A1), (A2), and (A3). Whereas M_{1a} reuses vocabulary terms directly and makes the data easier to read for humans, M_{1b} might be easier to be processed by Linked Data applications. Strategy *max*, exemplified by M_{1c}, pursues the same principle as M_{1a}, but maximizes the number of different vocabularies within one dataset by also using the MOVIE[6] and AWOL[7] vocabulary. We have set this strategy as a *lower boundary*, indicated by $|\Phi(\text{movie})| = 0$ and $|\Phi(\text{awol})| = 0$, to investigate whether other strategies are significantly different with respect to the quality of modeling and publishing LOD.

Ranking Task T_2: Appropriate Mix of Vocabularies. The second ranking task covers the topic of mixing an appropriate amount of different vocabularies. We provided the participants with four schema models $M_{2a} - M_{2d}$ described in Table 2. They all express the same example instance with different strategies about a *Publication* including a title, creation and publication date, as well as its *Author*, who has a name and a working place as properties. Model M_{2a} reuses only one vocabulary (strategy *minV*), which

[6] `http://data.linkedmdb.org/all`, access 1/12/2014.
[7] `http://bblfish.net/work/atom-owl/2006-06-06/`, access 1/12/2014.

Table 1. Ranking Task T_1: The models $M_{1a} - M_{1c}$, their reuse strategy, and features

	M_{1a}	M_{1b}	M_{1c}	$\lvert\Phi(V)\rvert$	$\Psi(v)$
Reuse Strategy	(*pop*)	(*link*)	(*max*)		
$\lvert\phi(M)\rvert$	2	4	3	/	/
$V =$ foaf	✓	✓	✓	232	/
$V =$ dc	✓	✓	–	287	/
$V =$ owl	–	✓	–	277	/
$V =$ rdfs	–	✓	–	533	/
$V =$ awol	–	–	✓	0	/
$V =$ movie	–	–	✓	0	/
$v =$ foaf:Person	✓	✓	✓	/	18, 477, 53
$v =$ dc:title	✓	✓	–	/	3, 605, 629
$v =$ foaf:made	✓	✓	–	/	57, 791
$v =$ rdfs:subClassOf	–	✓	–	/	12, 207
$v =$ owl:equivalentProperty	–	✓	–	/	127
$v =$ movie:performance	–	–	✓	/	0
$v =$ awol:title	–	–	✓	/	0

is neither used in very many dataset ($|\Phi(\text{swrc})| = 10$) nor are its vocabulary terms occurring frequently. However, it is highly domain specific and the entire data can be described by using terms from this vocabulary. Model M_{2b}reuses a maximum set of different vocabularies (strategy *max*) and is again the *lower boundary* in this ranking task. Most vocabularies are not used by many data sets, and with the exception of foaf:name and dcterms:title, the total occurrences of the remaining vocabulary terms is also quite low. Strategy *pop*, exemplified by M_{2c}, on the other hand reuses only the most popular vocabulary terms and vocabularies. The strategy *minC*, exemplified by M_{2d}, reuses one vocabulary per concept, i.e., the entity *Publication* is described via the popular Dublin Core vocabulary and the entity *Person* is described via the domain-specific SWRC vocabulary. Apart from M_{2b}, every other model and their underlying vocabulary reuse strategies in this ranking task is likely to comply with aspects (A1) to (A3). Reusing a minimum amount of vocabularies might provide a clear data structure, but it might also fail to capture the entire semantics of the data. Reusing mainly popular vocabularies might also fail to capture some domain specific semantics, but it is easy to understand by humans. In such case, M_{2d} might provide a well defined trade-off between M_{2a} and M_{2c}.

Ranking Task T_3: Vocabulary Reuse with Additional Meta-Information. This ranking task is different from the previous ones, as we wanted to investigate the influencing factors for vocabulary reuse by providing additional information about the vocabularies and vocabulary terms. Furthermore, by letting the respondents rank the given meta-information, we can also conclude whether it is helpful to provide additional information such as documentation on the semantics of a vocabulary term or pattern-based vocabulary term information. First, the participants were given an initial data model (IM), which represents an example instance of a *Music Artist*, who has a specific name and has published an *Album* having a title. The initial data model uses three vocabularies

Table 2. Ranking Task T_2: The models $M_{2a} - M_{2d}$, their reuse strategy, and features

	M_{2a}	M_{2b}	M_{2c}	M_{2d}	$\|\Phi(V)\|$	$\Psi(v)$
Reuse Strategy	*minV*	*max*	*pop*	*minC*		
$\|\phi(M)\|$	1	6	3	2	/	/
$V =$ swrc	✓	✓	✓	✓	10	/
$V =$ dc	–	✓	✓	✓	287	/
$V =$ foaf	–	✓	✓	–	232	/
$V =$ npg	–	✓	–	–	5	/
$V =$ umbc	–	✓	–	–	1	/
$v =$ swrc:author	✓	✓	–	–	/	16,754
$v =$ umbc:institution	–	✓	–	–	/	0
$v =$ npg:Contributor	–	✓	–	–	/	0
$v =$ foaf:name	–	✓	✓	–	/	3,287,920
$v =$ dc:title	–	✓	✓	✓	/	17,120,348
$v =$ foaf:Person	–	–	✓	–	/	2,333,589
$v =$ dc:creator	–	–	✓	✓	/	7,372,111

$\phi(DS) = \{\text{foaf, mo, rdfs}\}$, of which the MO[8] vocabulary is very specific for the domain of musical artists. Subsequently, the participants were provided the three schema models described in Table 3, each extending the IM with further properties such as the artist's homepage, the record's image, and others. Hereby, some vocabulary terms used in IM were updated with other vocabulary terms. Model M_{3a} extends the schema in IM with further properties from the MO ontology, but also replaces the other terms such as foaf:Agent with mo:MusicArtist or foaf:name with rdfs:label. Hereby, the *minD* strategy tries to express the data with (as few as possible) domain-specific vocabularies. The strategy *minV*, exemplified by M_{3b}, uses only one vocabulary, but the schema.org[9] vocabulary covers a broad range of different domains, including music artists. Thus, it is possible to express the entire dataset with this one vocabulary, although it is not quite popular as indicated by the features $|\phi|$ and Ψ. Model (M_{3c}) again follows the strategy to reuse popular vocabularies (*pop*). Their terms are very broad and not domain specific, but the popularity of the vocabularies and their terms is very high.

The additional meta-information, to which we will also refer to as "support type", on the provided data models contain the following information: ST_1- *Domain of a vocabulary:* domain of FOAF is people and relationships; domain of MO is musical work and artists; ST_2- *Statistics about vocabulary usage:* number of data providers in LOD cloud using FOAF: 500; number of data providers using MO: 50; ST_3- *Statistics about vocabulary term usage:* number of uses of foaf:homepage: 800; number of uses of mo:homepage: 200; ST_4- *Semantic information on vocabulary term:* foaf:homepage is used for the web page of a person, while mo:homepage is used for a fan/band page of an artist; and ST_5- *Statistics about vocabulary terms in triple context:* Most common object property between mo:MusicArtist and mo:Record is mo:published. Hereby, the data for ST_2, ST_3, and ST_5 is *fictive* and not retrieved from some web service.

[8] `http://purl.org/ontology/mo/`, access 1/4/2014.
[9] `http://schema.org/`, access 1/4/2014.

Table 3. Ranking Task T_3: The models $M_{3a} - M_{3c}$, their reuse strategy, and features

Model	M_{3a}	M_{3b}	M_{3c}	$\lvert\Phi(V)\rvert$	$\Psi(v)$
Reuse Strategy	*minD*	*minV*	*pop*		
$\lvert\phi(M)\rvert$	3	1	3	/	/
$V =$ foaf	✓	–	✓	232	/
$V =$ mo	✓	–	✓	4	/
$V =$ rdfs	✓	–	–	533	/
$V =$ schema	–	✓	–	3	/
$V =$ dc	–	–	✓	287	/
$v =$ foaf:Agent	✓	–	–	/	2,818,352
$v =$ mo:homepage	✓	–	–	/	0
$v =$ mo:MusicArtist	✓	–	✓	/	1,713,860
$v =$ schema:name	–	✓	–	/	0
$v =$ schema:Person	–	✓	–	/	375,277
$v =$ schema:MusicAlbum	–	✓	–	/	59,248
$v =$ dc:title	–	–	✓	/	3,605,629
$v =$ foaf:name	–	–	✓	/	9,235,251
$v =$ foaf:homepage	–	–	✓	/	8,244,952

3 Participants

Overall, $N = 79$ participants (16 female) took part in the survey. However, it was not mandatory to answer every question resulting in a participation range from minimum $N = 59$ to maximum $N = 79$. $N = 67$ finished the entire survey including demographic information. About 67% of these 67 participants work in academia, 23% work in industry, and 10% in both. The variety of the participants ranges from research associates (22) over post doctoral researchers (14) to professors (8) with an average age of $M = 34.6$ ($SD = 8.6$). On average the participants have worked for 4 years with Linked Open Data ($M = 4.07, SD = 2.64$), and rated their own expertise consuming and publishing LOD quite high ($M = 3.64, SD = 1$) on a 5-point-Likert scale from 1 (*none at all experienced*) to 5 (*expert*). Hereby, about $59,7\%$ of the participants consider themselves to be high experienced or above (4 or 5 on the Likert-scale) and $40,3\%$ consider themselves to have moderate knowledge or less. In total, we can say that our participants are quite experienced in the field of Linked Data. This makes the results of the survey very promising with respect to their validity for identifying the best strategy to choose appropriate vocabulary terms.

The participants were acquired using the following mailing lists: (a) public LOD mailing list,[10] (b) public Semantic Web mailing list,[11] and (c) EuropeanaTech-Community.[12] In addition, we contacted various authors and data maintainers of LOD datasets

[10] `http://lists.w3.org/Archives/Public/public-lod/`, access: 1/4/2014.
[11] `http://lists.w3.org/Archives/Public/semantic-web/`, access 1/4/2014.
[12] `http://pro.europeana.eu/web/network/europeana-tech`, access 1/4/2014.

Table 4. Results of the three ranking tasks $T_1 - T_3$

Ranking Task	Model	Strategy	Median Rank (Mdn)	Friedman test
T_1	M_{1a}	*pop*	1	$\chi^2(2, 78) = 11.521, p = .003$
	M_{1b}	*link*	2	
	M_{1c}	*max*	2	
T_2	M_{2a}	*minV*	3	$\chi^2(3, 63) = 40.536, p < .001$
	M_{2b}	*max*	4	
	M_{2c}	*pop*	1	
	M_{2d}	*minC*	2	
T_3	M_{3a}	*minD*	2	$\chi^2(2, 61) = 3.1$, **n.s.**, $p = .211$
	M_{3b}	*minV*	2	
	M_{3c}	*pop*	2	

on CKAN[13] as well as participants and lecturers from the Summer School for Ontological Engineering and Sematic Web (SSSW[14]) in person and asked them to participate in the survey and share their expertise.

4 Results of Ranking Tasks

We encode the obtained ranking position for the data models with numbers starting at 1, 2, and so on, i.e., the lower the ranking number the better rank position of a response option. For each ranking task, we performed a Friedman test to detect significant differences between the strategies (with $\alpha = .05$), as the answers are provided on an ordinal scale. Subsequently, we applied pairwise Wilcoxon signed-rank tests with Bonferroni correction, if significant differences have been found.

Table 4 summarizes the results of all three ranking tasks and gives a first insight into how the schema models and its underlying vocabulary reuse strategy have been ranked (including the significant differences between the rankings which are provided in the last column).

Ranking Task T_1. Regarding the task T_1, which was completed by $N = 78$ respondents, a significant difference of the three data models with respect to an appropriate reuse of vocabularies can be observed in Table 4. The Median (*Mdn*) ranks show that M_{1a} with the underlying strategy of reusing popular vocabulary terms is ranked better (*Mdn=1*) than the other two models and their strategy (*Mdn=2*). A post hoc analysis with Wilcoxon signed-rank tests, which were conducted with a Bonferroni correction applied (now $\alpha = .017$), provide final evidence that M_{1a} is significantly better than the other two models. However, there was no significant difference between the strategy to

[13] `http://datahub.io/group/lodcloud`, access 1/4/2014.
[14] `http://sssw.org/2013/`, access 1/11/2014.

Table 5. Results of the Support Types from Ranking Task T_3

Support Type	Support	*Mdn*	Friedman test
ST_1	Information on domain of vocabulary	2	$\chi^2(2, 78) = 11.521, p = .003$
ST_2	Number of LOD datasets using a vocabulary	2	
ST_3	Number of all occurrences of a vocabulary term in LOD cloud	3	
ST_4	Documentation of a vocabulary term	3	
ST_5	Information on most common use of an object property	4	

interlink self-defined vocabulary terms with external classes and properties and the *max* strategy that was merely provided as a lower boundary for vocabulary reuse.

Ranking Task T_2. The second ranking task, which was completed by $N = 63$ respondents, again shows a statistical significant difference between the four different reuse strategies and that the model with the strategy of reusing mainly popular vocabularies (M_{2c}) is ranked first (*Mdn=1*). A further post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied (now $\alpha = .008$), and it proves that M_{2c} is significantly better than M_{2a} and M_{2b}, but there was no significant difference to schema model M_{2d}. Furthermore, M_{2d} was significantly better than M_{2b} but no difference to M_{2a}, and schema model M_{2a} was significantly better than M_{2b}.

Ranking Task T_3. The last ranking task had two parts, and a total of $N = 61$ respondents have completed the first part and $N = 59$ completed the second part. In the first part, as shown in Table 4, the median ranks for the three model and their strategies are the same (*Mdn=2*). The results of the Friedman test to detect significant differences show that there is no difference between the strategies whatsoever (**n.s.**). In the second part, the participants had to rank which provided support type (the additional meta-information) was most helpful for making their ranking deciding. The five support types, their median ranks and whether there was a significant difference detected is displayed in Table 5. It can be observed that ST_1 and ST_2 are considered to be more helpful for making the right choice considering vocabulary reuse, whereas ST_4 seems not to be quite as helpful. Further post hoc analysis with Wilcoxon signed-rank tests show that ST_1 is significantly different to every other support type except ST_2. Support type ST_2 is again significantly different to all remaining support types but ST_4, and ST_3, ST_4, and ST_5 have no significant differences whatsoever.

5 Results of the Aspect Questions

We asked the participants to evaluate the different aspects regarding "why reuse vocabularies?", as introduced in Section 2.2, at the beginning of the survey and after the first and second ranking task. The median rating for the three aspects *A1:* provide a clear structure of the data, *A2*: make the data easier to be consumed, and *A3*: establish an ontological agreement was in general high (*Mdn≥4*). Applying Friedman test to measure whether there are significant differences to the second and third rating, shows that in each case, the respondents ranked the three aspects at the beginning of the survey significantly higher than after the two ranking tests, which is also proved by the post hoc analysis with Wilcoxon signed-rank tests. Basically, the median ratings for *A1* and *A2* was first *Mdn=5* and at the second and third rating it was *Mdn=4*. The aspect *A3* was asked only twice and the post hoc analysis showed that the first rating was significantly better that the second one despite the fact that the median rating for this aspect in both rating was *Mdn=4*. Furthermore, splitting the ratings into two groups with one group having an LOD experience of < 4 (moderate and below) and the other group being ≥ 4 (high to expert knowledge), shows that both groups have decreased the ratings of the aspects *A1* to *A3*.

6 Discussion

The results of analyzing the most important aspects to reuse vocabularies show that most participants have, in theory, the intention to publish Linked Open Data in an easy to process way, i. e., provide a clear structure of the data and make it as easy as possible to consume the data. However, it is very interesting to see that the theoretical intention to follow these best practices (*A1* to *A3*) seem to be higher than the intention to follow them in a real-life scenario. This is indicated by the ratings of *A1* to *A3* being high at the beginning ($Mdn = 5$) but not as high after asking the participants whether these aspects influenced the ranking decision ($Mdn = 4$). Nonetheless, each of these aspects was still rated with a median of $Mdn = 4$ on a 5-point-Likert scale, which still shows that these aspects are considered as "somewhat important". Therefore, the participant's goals to provide a clear structure and thereby increase the readability of the dataset can be considered as relatively consistent throughout the survey. Furthermore, there were no significant differences between the group of participants who have high to expert knowledge to the group with moderate LOD knowledge and below. This indicates that these goals are very genuine ones. Having these goals in mind, it is very interesting to look at the rankings of the three ranking tasks.

For **Ranking Task** T_1, the *pop* strategy is the significantly preferred choice. This is quite interesting, as theoretically, it is considered by the best practices to be important to establish links on schema level to other vocabulary terms. However, this *link* strategy was not significantly better than the *max* strategy (lower boundary). Furthermore, looking at the quite small total occurrence of properties such as owl:equivalentProperty indicates that other data providers do not follow this good practice either. In fact, looking at the total occurrence of the term owl:sameAs ($|\Phi(\text{owl:sameAs})| = 18,678,552$) indicates that for data providers it is more important to link Linked Open Data on instance level.

In **Ranking Task** T_2, the results showed that reusing widely-used vocabulary terms from widely-used vocabularies is considered better than reusing only domain specific vocabularies. This is quite interesting, as it is considered good practice to select the domain vocabulary first and use as many of its terms, if possible, as other vocabularies might not be needed. Apparently, this was not considered helpful in providing a clear data structure. In fact, correlating the ranking of the various aspects why vocabularies should be reused and the results of this ranking task, it seems that preferring widely-used vocabulary terms from widely used vocabularies serves the purpose more than reusing mainly the domain specific vocabulary. Despite this, both of these strategies were not significantly better than the strategy that uses a minimum amount of vocabularies per concept (*minC*). This *minC* strategy indeed seems to provide a good trade-off between reusing popular and domain specific vocabularies.

For **Ranking Task** T_3, no significant differences between the strategies were found in the first part of this task. The second part showed that the information on how many datasets use a specific vocabulary and the information on the domain of a vocabulary seem to be the most preferred additional meta-information. The results are interesting in a two-fold way: First, ranking task T_3 was very similar to ranking task T_2. Despite this similarity, the obtained results are very different. In detail, the information in ST_1 states that the MO vocabulary covers the domain of musical artists and their work as well as that the MO vocabulary is used by 50 data sets (fictive number; real number is $|\Phi(\text{MO})| = 3$). This might lead to believe that the MO vocabulary is a suitable candidate to express musical data, as it is used by many other data providers. Therefore, other vocabularies such as FOAF or Dublin Core are not needed, as MO is well-known and widely-used. Second, regarding the different support types, it is interesting to observe that the number of datasets using vocabulary V was considered more informative than the number of the total occurrences of vocabulary term $v \in V$. Particularly, to establish an ontological agreement in data representation, it seems to be better to reuse vocabulary terms from a vocabulary that is used by many, probably smaller datasets.

The results of our survey might have been influenced by several factors such as specific use cases, which were not considered in detail for ranking the LOD models, as well as the format in which we depicted the examples to the participants. Regarding different use cases, one might primarily use LOD for publishing the data on the web for automated consumption, but one might also define a LOD vocabulary to represent the domain knowledge for an own application. For example, the proprietary class myMov:Actor represents an actor. When modeling Linked Open Data and trying to provide a clear schema structure as well as to make the data easier to be consumed, the use of foaf:Person might be adequate. Whereas when defining an ontology, defining the proprietary vocabulary term and specifying a rdfs:subClassOf relationship might be considered better and more correct. As we did not specify the concrete application the Linked Data is created for, there are several other factors that might have influenced the results in a similar way. However, we did not focus on these factors as they are very difficult to grasp in a structured way and to simplify the study. The survey is addressing Linked Data practitioners, who work with Linked Open Data on a regular basis. Therefore, we showed the modeling examples in N3/Turtle syntax as this is the most common

way of representing data in a good human readable way. We might have excluded some participants, who might not be comfortable with N3/Turtle syntax.

7 Related Work

Previous studies regarding the datasets contained in the LOD cloud are mainly focused on investigating the compliance of LOD sources to different characteristics or best practices. Hogan et al. [3] performed an empirical analysis examining 4 million RDF/XML documents on their conformance to several best practices that were elaborated in [1], and in [9], the authors analyze LOD datasets and discuss common errors in the modeling and publishing process. In addition, Poveda Villalón et al. [10] performed a similar analysis of ontology reuse in the LOD context. As a result, reusing and mixing vocabularies is identified as an issue that is more difficult to resolve.

A study in the field of reusing *ontologies* was done by Simperl [4]. The author performs a feasibility study on reusing ontologies, where most prominent case studies on ontology reuse as well as methods and tools are enumerated. It is demonstrated that different methods for reusing ontologies are perfectly suitable to for a development of a new ontology, but in all case studies each reused ontology has to be found, evaluated, and chosen manually, which results in making the decision on which ontology to reuse based on personal experience.

There are also a couple of different methods that help the data engineer in deciding which vocabulary to reuse. However, these are focused on specific domains such as cultural heritage [11], governmental data,[15] bibliographic data,[16] and human resources [4]. These domain-specific methods provide valuable information on how to model and publish data as LOD in these domains, but may be too specific in order to apply it to the general case. The most recent work on the best practices about how to generally publish Linked Data is a tech report by the W3C [12]. It includes a basic checklist about what appropriate vocabularies must or should have, but besides the factor that one should reuse a vocabulary that is used by many other datasets, the other items on that checklist rather suggest to check whether a vocabulary is documented, self-descriptive, or is accessible for a long period. These aspects are not considered in our survey, but might be an interesting factor for future vocabulary recommendation tools.

The Linked Open Vocabulary index (LOV) [7] is an inspirational service to aid the Linked Data engineer in finding appropriate vocabulary terms for reuse. It provides the engineer with the most common and popular vocabularies as well as a lot of meta-information about each vocabulary and vocabulary term. This makes it possible to find the most suitable classes and properties to express data as LOD. However, it is solely based on a best string-match search and each vocabulary term has to be implemented in the engineering process manually. To alleviate this, a first implementation of a recommendation service for reusing ontologies is the Watson [13] plugin for the NeOn ontology engineering toolkit [14]. It uses semantic information from a number of ontologies and

[15] `http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook#Step_3_Re-use_Vocabularies_Whenever_Possible`, access: 5/16/2013.

[16] `http://aims.fao.org/lode/bd`, access: 5/16/2013.

other semantic documents published on the Web to recommend appropriate vocabulary terms, but it does consider the typical strategies for modeling Linked Data.

8 Conclusion

We presented a study that investigates which vocabulary reuse strategy is followed by Linked Data experts and practitioners in various real-life scenarios. It was examined via a survey consisting of ranking tasks, where the participants were asked to rank various modeling examples according to their understanding of good reuse of vocabularies, and rating assignments to explain which aspects most influenced the ranking decisions. The results of the ranking tasks illustrate that reusing vocabulary terms from widely-used as well as domain specific vocabularies directly is considered a better approach than defining proprietary terms and interlink them with external classes and properties. Furthermore, reusing popular vocabulary terms from frequently used vocabularies is more important than frequently used vocabulary terms from vocabularies that are not used by many data providers. To balance vocabulary terms from popular and domain specific vocabularies, it is considered to be important to maintain an appropriate mix, in order to provide a clear structure of the data and make it easier to be consumed. These findings of our survey can also be used for future vocabulary recommendation systems such as the LOVER approach [15] or implemented in existing tools such as Watson [13] for the NeOn ontology engineering toolkit [14].

Acknowledgement. We thank the participants for their time and effort. We additionally thank Natasha Noy, Asunción Gómez-Pérez, Laura Hollink, Jérôme Euzenat, and Richard Cyganiak for their valuable feedback on the survey and the modeling examples. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013), REVEAL (Grant agree no. 610928).

References

1. Bizer, C., Cyganiak, R., Heath, T.: How to publish linked data on the web. Web page (2007) (revised 2008) (access January 03, 2014)
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan Kaufmann (2011)
3. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web, 14–44 (2012)
4. Simperl, E.: Reusing ontologies on the semantic web: A feasibility study. Data Knowledge Engineering 68, 905–925 (2009)
5. Schaible, J., Gottron, T., Scherp, A.: Extended description of the survey on common strategies of vocabulary reuse in linked open data modeling. Technical Report 01/2014, Institute for Web Science and Technologies, Universität Koblenz-Landau (2014) `http://www.uni-koblenz.de/~fb4reports/2014/2014_01_Arbeitsberichte.pdf`

6. Auer, S., Demter, J., Martin, M., Lehmann, J.: LODStats – an extensible framework for high-performance dataset analytics. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 353–362. Springer, Heidelberg (2012), http://dblp.uni-trier.de/db/conf/ekaw/ekaw2012.html#AuerDML12
7. Vandenbussche, P.Y., Vatant, B., Rozat, L.: Linked open vocabularies: an initiative for the web of data. In: QetR Workshop, Chambery, France (2011)
8. Stadtmüller, S., Harth, A., Grobelnik, M.: Accessing information about linked data vocabularies with vocab.cc. In: Semantic Web and Web Science, 391–396 (2013)
9. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: Proceedings of the Linked Data on the Web Workshop, LDOW 2010 (2010)
10. Poveda Villalón, M., Suárez-Figueroa, M.C., Gómez-Pérez, A.: The landscape of ontology reuse in linked data. In: Proceedings Ontology Engineering in a Data-driven World, OEDW 2012 (2012)
11. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Synthesis Lectures on The Semantic Web: Theory and Technology (2012)
12. Atemezing, G.A., Villazón-Terrazas, B., Hyland, B.: Best practices for publishing linked data. W3C note, W3C (2014), http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/ (access January 03, 2014)
13. d'Aquin, M., Baldassarre, C., Gridinoc, L., Sabou, M., Angeletou, S., Motta, E.: Watson: Supporting next generation semantic web applications. In: Proceedings of the IADIS International Conference WWW/Internet 2007, pp. 363–371 (2007)
14. Haase, P., Lewen, H., Studer, R., Tran, D.T., Erdmann, M., d'Aquin, M., Motta, E.: The neon ontology engineering toolkit. In: Korn, J. (ed.) WWW 2008 Developers Track (2008)
15. Schaible, J., Gottron, T., Scheglmann, S., Scherp, A.: Lover: support for modeling data using linked open vocabularies. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops (2013)