UNIVERSITÄT
KOBLENZ · LANDAU
Institut WeST

FB 4
Informatik

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

Johann Schaible
Thomas Gottron
Ansgar Scherp

Nr. 1/2014

Arbeitsberichte aus dem Fachbereich Informatik

Arbeitsberichte des Fachbereichs Informatik

ISSN (Print): 1864-0346
ISSN (Online): 1864-0850

Herausgeber / Edited by:

Der Dekan:
Prof. Dr. Lämmel

Kontaktdaten der Verfasser

Johann Schaible, Thomas Gottron, Ansgar Scherp
Institut WeST
Fachbereich Informatik
Universität Koblenz-Landau
Universitätsstraße 1
D-56070 Koblenz
E-Mail: johann.schaible@gesis.org, gottron@uni-koblenz.de, mail@ansgarscherp.net

Extended Description of the 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. Modeling and publishing Linked Open Data (LOD) involves the choice of which vocabulary to use. This choice is far from trivial and poses a challenge to a Linked Data engineer. It covers the search for appropriate vocabulary terms, making decisions regarding the number of vocabularies to consider in the design process, as well as the way of selecting and combining vocabularies. Until today, 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. Participants of our survey are LOD publishers and practitioners. Their task was 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 understanding the strategies how data engineers reuse vocabularies, and they may also be used to develop future vocabulary engineering tools.

1 Introduction

Publishing data as Linked Open Data (LOD) makes it possible to interlink it with other external data sources in a two-fold way: (i) on instance level, for example via `owl:sameAs` links, or (ii) on schema level via properties like `owl:equivalentProperty`. This enables the data provider to enrich the data with further information such as knowledge from related domains or data sets comprising additional meta-information about the same resources. However, 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 classes and properties from existing vocabularies, rather than reinventing them, is clearly motivated by the best practices and recommendations for designing and publishing Linked Data [1]. Experienced Linked Data engineers follow these recommendations 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 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. Hereby, engineers decide which vocabularies to reuse based simply on their knowledge, experience, and "gut-feeling".

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. Thus, the main contribution of this paper is to condense and aggregate the knowledge, the experience, and the "gut-feeling" 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 in the real world 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.[4] To this end, and to aggregate the expert's knowledge and experience, we have conducted a survey among Linked Data practitioners and experts.

We have obtained feedback from 79 participants acquired through public mailing lists. We have asked the participants of the survey to rank several data

[4] This is an extended description of the paper published in the ESWC 2014 proceedings [5]

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. 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 external vocabulary terms.[5] 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. Overall, the results provide very valuable insights in how data engineers decide which vocabularies to reuse when modeling Linked Open Data. They might also provide valuable requirements for developing novel vocabulary recommendation services in existing and future vocabulary engineering tools.

In Section 2, we describe the survey and the vocabulary reuse strategies that were implemented in the examples used in the survey. Subsequently, we give first information on how we collected the data and on the participants of the survey in Section 3. We present the results of the ranking tasks in Section 4 and the results of which aspects are considered as most important to reuse vocabularies in Section 5, before we discuss the results in Section 6. Section 7 comprises the related work, including other studies that cover the topic of vocabulary reuse as well as already existing tools and services that support the data engineer in modeling Linked Open Data, before we conclude our work in Section 8.

2 The Survey

The survey[6] 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. Each data model represents a specific vocabulary reuse strategy such as reusing only popular or domain specific vocabularies. This way, we intend to find out which of the strategies is the most preferred one by Linked Data practitioners in a real-world scenario. 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 using the defined set of features.

[5] Raw result data in SPSS format can be found here: `http://dx.doi.org/10.7802/64`

[6] 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`)

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 define a generic set of features for LOD data models 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.[7]

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. Here, we do not regard blank nodes, since using blank nodes is not considered good practice. 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 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

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

[7] `http://challenge.semanticweb.org/`, access 03/10/2014

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

Listing (1.1) Example data model M_a

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

Listing (1.2) Example data model M_b

Fig. 1: Examples of data models which have to be ranked in the ranking tasks

As mentioned, this set of features can be based on various collections of RDF graphs. However, for the survey, we have retrieved the metrics for these features 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

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 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 stated in [1, 2]. To illustrate the differences of such strategies, we use the previously defined features $\phi((G,c))$, $|\phi((G,c))|$, $|\Phi(V)|$, and $\Psi(v)$. The vocabularies in $\phi(DS)$ provide information on which vocabularies have been used and whether they are domain specific or not. The number $|\phi((G,c))|$ indicates how many different vocabularies have been used to describe the data, and $|\Phi(V)|$ 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 that were defined in Section 2.1. 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 in Figure 1 provide two example data models that describe the same data item with different sets of vocabularies and different vocabulary terms. Table 1 illustrates the data models, their underlying vocabulary reuse strategy, and their features regarding $|\Phi(V)|$ and $\Psi(v)$. We can see that model M_a uses a minimum amount of vocabularies (reuse strategy "*min V*" with $|\phi(M_a)| = 1$) and M_b reuses primarily popular vocabularies (reuse strategy "*pop*" with $|\phi(M_1)| = 3$). Model M_a expresses the data using the Semantic Web for Research Communities (SWRC) vocabulary,[8] i.e., with a highly domain specific vocabulary $\phi(M_a) = \{\text{swrc}\}$, whereas M_b makes use of

[8] `http://www.ontoware.org/index.html`, access 12/19/2013

Table 1: The models M_a and M_b, their reuse strategy, and features.

	M_a	M_b	$\lvert\Phi(V)\rvert$	$\Psi(v)$
Reuse Strategy	$(minV)$	(pop)		
$\lvert\phi(M)\rvert$	1	3	/	/
$V =$ foaf	–	✓	232	/
$V =$ dc	–	✓	287	/
$V =$ swrc	✓	✓	10	/
$v =$ swrc:Publication	✓	✓	/	30
$v =$ swrc:title	✓	–	/	10,487
$v =$ dc:title	–	✓	/	17,120,348
$v =$ swrc:author	✓	–	/	16,754
$v =$ dc:creator	–	✓	/	7,372,111
$v =$ swrc:Person	✓	–	/	30,510
$v =$ foaf:Person	–	✓	/	2,333,589
$v =$ swrc:name	✓	–	/	35,756
$v =$ foaf:name	–	✓	/	3,287,920

more popular, i.e., well-known and widely-used, vocabularies such as FOAF[9] and Dublin Core[10] $\phi(M_b) = \{\text{foaf, dc}\}$. Hereby, the vocabularies FOAF and Dublin Core are considered more popular as SWRC as the values of $|\Phi(V)|$ are indicating ($|\Phi(\text{foaf})| = 232 > 6 = |\Phi(\text{swrc})|$ and $|\Phi(\text{dc})| = 287 > 6 = |\Phi(\text{swrc})|$). In addition, M_b makes also use of more popular vocabulary terms than M_a as indicated by the various values of Ψ. Nonetheless, the total numbers of occurrences of the SWRC vocabulary terms such as $\Psi(\text{swrc:title}) = 10,487$, are still quite high and thus indicate a highly domain specific use by a few but large datasets. 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. In other words, which of the features that represent a model is considered more important?

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 correlation between the user ratings of the aspects and the rankings of the data models. For example, if aspect (A1) is significantly considered the most important aspect and the ranking of the strategy which reuses only a minimum number of vocabularies is significantly

[9] `http://xmlns.com/foaf/spec/`, access 1/9/2014
[10] `http://dublincore.org/documents/dcmi-terms/`, access 1/9/2014

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 maximizing them.

2.3 Ranking Tasks

The survey contains three ranking tasks, each covering a different aspect of the engineer's decision making process [3, 9]. For example, should the Linked Data engineer focus on reusing vocabularies directly, or on defining proprietary classes and properties and afterwards establishing links on schema-level to external vocabularies? In the following, we will describe the different tasks and their motivation along the used schema models including their features. The schema 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 to some participants. The different vocabulary reuse strategies that we investigate 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*)
- maximize number of vocabularies (*max*)

Figure 2 to Figure 4 display the different data models that had to be ranked by the participants, and Table 2 to Table 4 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 $\Psi(v)$ Hereby, 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 this vocabulary or vocabulary term 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 though, the models are from different domains (important to avoid domain-specific bias). The ranking tasks are not linked to each other, as each task answers a different research question such as which amount of mixed vocabularies is considered best. We used only $3-4$ data models per task, as only some strategies were important for each task (and its goal), e.g., the first ranking task covers reusing vocabularies vs. establishing links. Differentiating between

```
<http://ex1.org/actor/1661/>
  rdf:type foaf:Person;
  foaf:name "Jack Nickolson";
  foaf:made
    <http://ex1.org/6354/>.
<http://ex1.org/6354/>
  rdf:type myMov:Film;
  dc:title "Batman".
```

Listing (1.3) Data model M_{1a}

```
<http://ex1.org/actor/1661/>
  rdf:type foaf:Person;
  awol:name "Jack Nickolson";
  movie:performance
    <http://ex1.org/6354/>.
<http://ex1.org/6354/>
  rdf:type movie:Film;
  awol:title "Batman".
```

Listing (1.4) Example data model M_{1c}

```
<http://ex1.org/actor/1661/>
  rdf:type myMov:Actor;
  myMov:name "Jack Nickolson";
  myMov:made <http://ex1.org/6354/>.
<http://ex1.org/6354/>
  rdf:type myMov:Film;
  myMov:title "Batman".
myMov:Actor a rdf:Class;
  rdfs:subClassOf foaf:Person.
myMov:name a rdf:Property;
  owl:equivalentProperty foaf:name.
myMov:made a owl:ObjectProperty;
  owl:equivalentProperty foaf:made.
myMov:title a rdf:Property;
  owl:equivalentProperty dc:title.
```

Listing (1.5) Example data model M_{1b}

Fig. 2: Data Models that had to be ranked for ranking task T_1

reusing minimal number of vocabularies and minimal number of vocabularies per concept is not important in this case. It also kept the survey understandable and manageable for participants.

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 the Listings in Figure 2) that had to be ranked based on the decision whether it is better to reuse vocabulary terms directly or use self-defined terms and establish links on schema level to external vocabulary terms via properties such as `rdfs:subClassOf` or `owl:equivalentProperty`. Each model expresses the same example instance that represents an *Actor* who played in a certain *Movie* with a different strategy, i.e., different vocabulary terms form different vocabularies. Hereby, the vocabulary exemplified by the namespace **myMov** is the self-defined vocabulary. Table 2 illustrates the vocabulary reuse strategies of these data models and their features. Model M_{1a} reuses vocabulary terms from the FOAF and Dublin Core vocabularies directly, which seem to be quite popular as indicated by the values $|\Phi(V)|$ and $\Psi(v)$, i.e., it follows the *pop* strategy. In detail, it has a fair amount of reused vocabularies ($|\phi(M_{1a})| = 2$) and the popularity of the vocabularies ($|\Phi(\text{foaf})| = 232$, $|\Phi(\text{dc})| = 287$) as well as the total occurrence of their vocabu-

Table 2: 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
$v =$ foaf:name	✓	✓	–	/	9,235,251
$v =$ awol:name	–	–	✓	/	0
$v =$ foaf:made	✓	✓	–	/	57,791
$v =$ movie:performance	–	–	✓	/	0
$v =$ movie:Film	–	–	✓	/	12,494
$v =$ dc:title	✓	✓	–	/	3,605,629
$v =$ awol:title	–	–	✓	/	0
$v =$ rdfs:subClassOf	–	✓	–	/	12,207
$v =$ owl:equivalentProperty	–	✓	–	/	127

lary terms, such as `foaf:Person` (Ψ(foaf:Person) = 18,477,533) and `dc:title` (Ψ(dc:title) = 3,605,629) is very high. 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` properties. However, with Ψ(rdfs:subClassOf) = 12,107 and Ψ(owl:equivalentProperty) = 127 it can be observed that this strategy, namely strategy *link*, is not used very often. 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}, is similar to M_{1a}, but instead of reusing well-known vocabularies it maximizes the number of different vocabularies within one dataset by also using the MOVIE[11] and AWOL[12] vocabulary. We have set this strategy as a *lower boundary*, indicated by $|\Phi(\text{movie})| = 0$ and $|\Phi(\text{awol})| = 0$, to investigate whether the other two strategies are significantly different to M_{1c} with respect to the quality of modeling and publishing Linked Open Data.

Ranking Task T_2: Appropriate Mix of Vocabularies The second ranking task covers the topic of mixing an appropriate amount of different vocabularies.

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

```
<http://ex1.org/lod/publ/001>
  rdf:type swrc:Publication;
  swrc:title "Example Title";
  swrc:creationDate
    "Example Issued Date";
  swrc:startDate
    "Example Available Date";
  swrc:author
    <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
  rdf:type swrc:Person;
  swrc:name "xyz";
  swrc:institution
    "Example Institution".
```

Listing (1.6) Data model M_{2a}

```
<http://ex1.org/lod/publ/001>
  rdf:type swrc:Publication;
  dcterms:title "Example Title";
  xfoaf:issueDate
    "Example Issued Date";
  dcterms:available
    "Example Available Date";
  swrc:author
    <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
  rdf:type npg:Contributor;
  foaf:name "xyz";
  umbc:institution
    "Example Institution".
```

Listing (1.7) Example data model M_{2b}

```
<http://ex1.org/lod/publ/001>
  rdf:type swrc:Publication;
  dcterms:title "Example Title";
  dcterms:issued
    "Example Issued Date";
  dcterms:available
    "Example Available Date";
  dcterms:creator
    <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
  rdf:type foaf:Person;
  foaf:name "xyz";
  swrc:institution
    "Example Institution".
```

Listing (1.8) Example data model M_{2c}

```
<http://ex1.org/lod/publ/001>
  rdf:type dcterms:BibliographicResource;
  dcterms:title "Example Title";
  dcterms:issued
    "Example Issued Date";
  dcterms:available
    "Example Available Date";
  dcterms:creator
    <http://ex1.org/pers/xyz>.
<http://ex1.org/pers/xyz>
  rdf:type swrc:Person;
  swrc:name "xyz";
  swrc:institution
    "Example Institution".
```

Listing (1.9) Example data model M_{2d}

Fig. 3: Data Models that had to be ranked for ranking task T_2

We provided the participants with the four schema models $M_{2a} - M_{2d}$, illustrated in Figure 3 and described in Table 3, and let them decide about finding the appropriate number of different vocabularies in a data set. The schema models 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 working place as properties. With the SWRC vocabulary, model M_{2a} reuses only one vocabulary (strategy $minV$), which is neither used in very many dataset ($|\Phi(\text{swrc})| = 10$) nor are its vocabulary terms occurring frequently; $\Psi(\text{swrc:author}) = 16,754$ is the vocabulary term with the highest number of occurrences in this schema model. However, it is highly domain specific and the entire data can be described by using classes and properties from this vocabulary such as `Publication, Person, author` and `institution`. 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 the most popular vocabulary terms and vocabularies, as it can be observed in the metrics. Regrading the vocabular-

Table 3: Ranking Task T_2: The models $M_{2a} - M_{2d}$, their reuse strategy, and features

	M_{2a}	M_{2b}	M_{2c}	M_{2d}	$\lvert\Phi(V)\rvert$	$\Psi(v)$
Reuse Strategy	*minV*	*max*	*pop*	*minC*		
$\lvert\phi(M)\rvert$	1	6	3	2	/	/
V = swrc	✓	✓	✓	✓	10	/
V = dc	–	✓	✓	✓	287	/
V = xfoaf	–	✓	–	–	0	/
V = foaf	–	✓	✓	–	232	/
V = npg	–	✓	–	–	5	/
V = umbc	–	✓	–	–	1	/
v = swrc:Publication	✓	✓	✓	–	/	30
v = dc:BibliographicResource;	–	–	–	✓	/	0
v = swrc:title	✓	–	–	–	/	10,487
v = dc:title	–	✓	✓	✓	/	17,120,348
v = swrc:creationDate	✓	–	–	–	/	0
v = xfoaf:issueDate	–	✓	–	–	/	0
v = dc:issued	–	–	✓	✓	/	232,329
v = swrc:startdate	✓	–	–	–	/	0
v = dc:available	–	✓	✓	✓	/	1,308
v = swrc:author	✓	✓	–	–	/	16,754
v = dc:creator	–	–	✓	✓	/	7,372,111
v = swrc:Person	✓	–	–	✓	/	30,510
v = npg:Contributor	–	✓	–	–	/	0
v = foaf:Person	–	–	✓	–	/	2,333,589
v = swrc:name	✓	–	–	✓	/	35,756
v = foaf:name	–	✓	✓	–	/	3,287,920
v = swrc:institution	✓	–	✓	✓	/	241
v = umbc:institution	–	✓	–	–	/	0

ies they are: $|\Phi(\text{foaf})| = 256$, $|\Phi(\text{dc})| = 378$), and regarding the vocabulary terms they are: $\Psi(\text{dc:creator} = 7,372,111), \Psi(\text{foaf:Person}) = 2,333,589$). 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}. We intend to investigate whether it is better to use as few vocabularies as possible or to use several different vocabularies. Reusing as few vocabularies as possible is more likely to increase the readability of the data, but there is also a risk of fitting an entity into a less suitable vocabulary term.

Reusing several vocabularies is more likely to provide better fitting vocabulary terms, but it might also decrease the readability of the data.

Ranking Task T_3: Vocabulary Reuse with Additional Meta-Information This ranking task is different from the previous ones, as we want 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 published an *Album* having a title. The initial data model uses three vocabularies $\phi(DS) = \{\text{foaf, mo, rdfs}\}$, of which the MO[13] vocabulary is very specific for the domain of musical artists. Subsequently, the participants were provided the three schema models, which are illustrated in Figure 4 and described in Table 4 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. For example, it might occur that the vocabulary term `foaf:Agent` might be updated with the term `mo:MusicArtist` to describe the musician entity.

Model M_{3a} extends the schema in IM with further properties from the MO ontology, but also updates 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 domain specific vocabularies as possible as well as to use generic vocabulary terms such as `rdfs:label` for entities that cannot be expressed with the domain specific vocabulary. The number of datasets using the MO ontology is not very high ($|\Phi(\text{mo})| = 4$) but the total occurrences of `mo:MusicArtist` ($\Psi(\text{mo:MusicArtist}) = 1,713,860$) indicates that there is a large dataset on musical artists. The strategy *minV*, exemplified by M_{3b}, uses only one vocabulary, but the schema.org[14] vocabulary covers a broad range of different domains, including music artists. Therefore, it possesses some specific classes and properties for music artists such as `schema:MusicAlbum` or `schema:album`, but also general vocabulary terms such as `schema:Person` or `schema:name` to cover general data entities. 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*) such as FOAF and Dublin Core, even though some vocabulary terms are not quite domain specific for describing the data. For example, the property `dc:title` describes the name of an album, but is not considered specific for the music domain. Thus, such vocabulary terms are very broad, but their popularity, and the popularity of the whole vocabularies, is very high.

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

```
<http://ex1.org/artist/artist_01>
  rdf:type mo:MusicArtist;
  rdfs:label "Joe␣Somebody";
  mo:published
    <http://ex1.org/record_01/>;
  mo:homepage <http://www.jsb.org>;
  mo:activity_start "2002-09-24";
  mo:label
    <http://ex1.org/label_xyz/>.
<http://ex1.org/record_01/>
  rdf:type mo:Record;
  rdfs:label "Example␣Record␣Title";
  mo:track_count 20;
  mo:image <http://ex1.org/image01>.
```

Listing (1.10) Data model M_{3a}

```
<http://ex1.org/artist/artist_01>
  rdf:type schema:Person;
  schema:name "Joe␣Somebody";
  schema:album
    <http://ex1.org/record_01/>;
  schema:url <http://www.jsb.org>;
  schema:foundingDate "2002-09-24";
  schema:acountablePerson
    <http://ex1.org/label_xyz/>.
<http://ex1.org/record_01/>
  rdf:type schema:MusicAlbum;
  schema:name "Example␣Record␣Title";
  schema:numTracks 20;
  schema:image <http://ex1.org/image01>.
```

Listing (1.11) Example data model M_{3b}

```
<http://ex1.org/artist/artist_01>
  rdf:type mo:MusicArtist;
  foaf:name "Joe␣Somebody";
  mo:published
    <http://ex1.org/record_01/>;
  foaf:homepage <http://www.jsb.org>;
  mo:activity_start "2002-09-24";
  mo:label
    <http://ex1.org/label_xyz/>.
<http://ex1.org/record_01/>
  rdf:type mo:Record;
  dc:title "Example␣Record␣Title";
  mo:track_count 20;
  foaf:img <http://ex1.org/image01>.
```

Listing (1.12) Example data model M_{3c}

Fig. 4: Data Models that had to be ranked for ranking task T_3

The additional meta-information, to which we will also refer to as "support types", on the provided data models contain the following information:

1. ST_1 - *Domain of a vocabulary:* domain of FOAF is people and relationships; domain of MO is musical work and artists.
2. ST_2 - *Statistics about vocabulary usage:* number of data providers in LOD cloud using FOAF: 500; number of data providers using MO: 50.
3. ST_3 - *Statistics about vocabulary term usage:* number of uses of `foaf:homepage`: 800; number of uses of `mo:homepage`: 200.
4. 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.
5. 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.

Table 4: Ranking Task T_3: The models $M_{3a} - M_{3c}$, their reuse strategy, and features

Model	M_{3a}	M_{3b}	M_{3c}	$\|\Phi(V)\|$	$\Psi(v)$
Reuse Strategy	*minD*	*minV*	*pop*		
$\|\phi(M)\|$	3	1	3	/	/
V = foaf	✓	✓	✓	232	/
V = mo	✓	–	✓	4	/
V = rdfs	✓	–	–	533	/
V = schema	–	✓	–	3	/
V = dc	–	–	✓	287	/
v = mo:MusicArtist	✓	–	✓	/	1,713,860
v = schema:Person	–	✓	–	/	375,277
v = rdfs:label	✓	–	–	/	91,521,315
v = schema:name	–	✓	–	/	0
v = foaf:name	–	–	✓	/	9,235,251
v = mo:published	✓	–	✓	/	0
v = schema:album	–	✓	–	/	0
v = mo:homepage	✓	–	–	/	0
v = schema:url	–	✓	–	/	0
v = foaf:homepage	–	–	✓	/	8,244,952
v = mo:activity_start	✓	–	✓	/	0
v = schema:foundingDate	–	✓	–	/	0
v = mo:label	✓	–	✓	/	0
v = schema:acountablePerson	–	✓	–	/	0
v = mo:Record	✓	–	✓	/	5,770
v = schema:MusicAlbum	–	✓	–	/	59,248
v = dc:title	–	–	✓	/	3,605,629
v = mo:track_count	✓	–	✓	/	0
v = schema:numTracks	–	✓	–	/	0
v = mo:image	✓	–	–	/	23,065
v = schema:image	–	✓	–	/	3
v = foaf:img	–	–	✓	/	11,004,064

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 on a 5-point-Likert scale from 1 (*none at all experienced*) to 5 (*expert*). Considering

consuming Linked Data the expertise was about $M = 3.67$ ($SD = 0.99$) and considering publishing Linked Data it was $M = 3.61$ ($SD = 1.1$). 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,[15] (b) public Semantic Web mailing list,[16] and (c) EuropeanaTech-Community.[17] In addition, we contacted various authors and data maintainers of LOD datasets on CKAN[18] as well as participants and lecturers from the Summer School for Ontological Engineering and Sematic Web (SSSW[19]) 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 5 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 first ranking task, which was completed by $N = 78$ respondents, there was a significant difference of the three data models with respect to an appropriate reuse of vocabularies, $\chi^2(2, 78) = 11.521$, $p = .003$. The Median (Mdn) ranks show that M_{1a} with the underlying strategy of reusing popular vocabulary terms directly 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. The tests showed that strategy *pop* compared to strategy *link* was considered better by 48 respondents and as worse by 25 respondents (rank of strategy *pop* $<$ rank of strategy *link* $= 48$ and rank of strategy *pop* $>$

[15] `http://lists.w3.org/Archives/Public/public-lod/2013Apr/0120.html`, access: 1/4/2014
[16] `http://lists.w3.org/Archives/Public/semantic-web/`, access 1/4/2014
[17] `http://pro.europeana.eu/web/network/europeana-tech`, access 1/4/2014
[18] `http://datahub.io/group/lodcloud`, 1/4/2014
[19] `http://sssw.org/2013/`, access 1/11/2014

Table 5: Results of the three ranking tasks $T_1 - T_3$

Ranking Task	Model	Strategy	Median Rank	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	$\chi^2(2, 61) = 3.1, \textbf{n.s.}, p = .211$
	M_{3b}	*minV*	2	
	M_{3c}	*pop*	2	

rank of strategy *link* = 25). Compared to strategy *max*, strategy *pop* show similar numbers (rank of strategy *pop* $<$ rank of strategy *max* $=$ 48 and rank of strategy *pop* $>$ rank of strategy *max* = 24). Evenly spread was the ranking between strategy *link* and strategy *max* (rank of strategy *link* $<$ rank of strategy *max* $=$ 35 and rank of strategy *link* $>$ rank of strategy *max* $=$ 34). In detail, there were no significant differences between strategy *max* and strategy *link* as the Wilcoxon signed-rank test shows ($Z = -0.181, \textbf{n.s.}, p = .856$). However, there was a statistically significant better vocabulary reuse strategy regarding strategy *pop* vs. strategy *link* ($Z = -3.214, p < .001$) and strategy *pop* vs. strategy *max* ($Z = -3.197, p < .001$).

Ranking Task T_2. The second ranking task, which was completed by $N = 63$ respondents, again shows that the model with the strategy of reusing mainly popular vocabularies (M_{2c}) is ranked first (*Mdn = 1*). The Friedman test also shows a significant difference between the four data models ($\chi^2(3, 63) = 40.536, p < .001$). A further post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied (now $\alpha = .008$). There was no significant difference between strategy *minV* and strategy *minC* ($Z = -0.602, n.s., p = .551$) as well as no significant difference between strategy *pop* and strategy *minC* ($Z = -2.292, n.s., p = .021$), despite the different median ranks. However, the other comparisons show significant differences, i.e., strategy *minV* to strategy *pop* ($Z = -2.616, p < .008$), strategy *minV* to strategy *max* ($Z = -3.902, p < .001$), strategy *pop* to strategy *max* ($Z = -5.632, p < .001$), and strategy *minC* to strategy *max* ($Z = -3.926, p < .001$).

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 5, the median ranks for the three model and their strategies are the same (*Mdn = 2*). Indeed, the results of the Friedman test to detect significant differences show that there is no significant difference between the strategies whatsoever ($\chi^2(2, 61) = 3.1, \textbf{n.s.}, p = .211$).

Table 6: 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	

In the second part , the participants had to rank which provided support type (the additional meta-information) was most helpful for making their ranking decision. The median ranks for the five support types and whether there was a significant difference detected is displayed in Table 6 The results of the Friedman test show that there was a significant difference between the five different types of support ($\chi^2(3, 59) = 36.165, p < .001$). 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 with a Bonferroni correction applied (now $\alpha = 0.005$) show that $ST1$ is not significant different to $ST2$ ($Z = -0.586, n.s., p = .564$), but is indeed significantly different to all other support types: $ST1$ to $ST3$ ($Z = -3.788, p < .001$), $ST1$ to $ST4$ ($Z = -3.493, p < .001$), and $ST1$ to $ST5$ ($Z = -4.333, p < .001$). The second support type $ST2$ is significantly different to $ST3$ ($Z = -4.547, p < .001$) and to $ST5$ ($Z = -3.804, p < .001$), but not to $ST4$ ($Z = -2.555, \mathbf{n.s.}$, $p = .01$). Finally, there are no significant differences between $ST3$ and $ST4$ ($Z = -0.581, \mathbf{n.s.}, p = .565$), $ST3$ and $ST5$ ($Z = -1.289, \mathbf{n.s.}, p = .199$), and $ST4$ and $ST5$ ($Z = -2.118, \mathbf{n.s.}, p = .034$).

5 Results of the Aspect Questions

We asked the participants to evaluate the different aspects regarding why reusing vocabularies, which was introduced in Section 2.2, at three points in time, namely at the beginning of the survey and after the first and second ranking task. Basically, the majority of the respondents rated each aspect quite high. 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 ($Mdn \geq 4$). 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 ((A1): $\chi^2(2, 63) = 6.881, p = .031$; (A2): $\chi^2(2, 63) = 34.889, p < .001$; (A3): $\chi^2(1, 63) = 6.429, p = .017$). Post hoc analysis with a Bonferroni correction applied (now for (A1) and (A2): $\alpha = 0.017$), showed that regarding (A1) there is a significant difference between the first rating and the second rating ($Z = -2.523, p = .011$), as well as between the first rating and the third rating ($Z = -2.511, p = .011$). However, the second rating was not significantly different to the third one ($Z = -.146, p = .909$). Regarding (A2), post hoc analysis showed the first rating is significantly different to the second ($Z = -3.778, p < .001$) and third rating ($Z = -4.805, p < .001$). No differences were found between the second and the third rating though ($Z = -1.937, n.s., p = .065$). The median ratings dropped for both (A1) and (A2) from $Mdn = 5$ to $Mdn = 4$. The aspect (A3) was asked only twice and the post hoc analysis with a Bonferroni correction applied (now $\alpha = 0.025$) showed that the first rating was significantly better that the second one ($Z = -2.155$, $p = .032$), despite the fact that the median rating for this aspect is $Mdn = 4$. 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). Between the answers of these two groups, there are no significant differences in the rating before and after the ranking tasks.

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 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), which reuses simply a maximum of vocabulary terms. 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` on the other hand ($|\Phi(\text{owl:sameAs})| = 18,678,552$) 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 as the best option. This time, all strategies were significantly better than the lower boundary of reusing a maximum amount of different vocabularies (the *max* strategy). The *pop* strategy, which reuses popular (in the meaning of widely-used) vocabulary terms, was significantly better compared to the *minV* strategy that used only one domain specific vocabulary. 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. In fact, if the data can be described with one domain vocabulary, which can be considered well-known by users working in this domain, it may seem odd to reuse another (popular) vocabulary. Apparently, this was not considered to help to provide a clear data structure. 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 ranking 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. The additional information in part one of T_3 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 (the additional meta-information), 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

though, to reuse vocabulary terms from a vocabulary that is used by many, probably smaller datasets. One might be more familiar with these vocabularies and Linked Data applications might have tailored support for these vocabulary terms.

The results of our survey might have been influenced by several factors such as the 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.

The results of the survey can also be used for defining requirements for future vocabulary recommendation services. From the first ranking task, we can derive the requirement to filter vocabulary terms for alignment on schema level and rank widely used vocabulary terms from widely used vocabularies higher than others. The second ranking task underlines such a requirement. However, we can also derive that vocabulary terms from already reused vocabularies should be ranked high, in order to maintain an appropriate mix of different vocabularies and thereby provide a clear structure of the data. As modeling is an iterative process, in each modeling iteration, a new recommendation is computed. With every new vocabulary term that is added to the model, various other rules are created or updated based on this new set of vocabularies and vocabulary terms [10]. Therefore this effects the recommendations in every step of the modeling process, i. e., the filtering and ranking of the result set, and it becomes more important to maintain an appropriate mix of vocabularies. Finally, providing explicit additional metadata about the use of vocabularies and their domain, as shown by the results from the third ranking task, is very likely to help the engineer with the decision of reusing the best possible vocabulary term.

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. [11] 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 [12], governmental data,[20] bibliographic data,[21] 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 [13]. It includes information on how to find relevant vocabularies for reuse and 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 [14] plugin for the NeOn ontology engineering toolkit [15]. It uses semantic information from a number of ontologies and other semantic documents

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

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

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 [16] or implemented in existing tools such as Watson [14] for the NeOn ontology engineering toolkit [15].

Acknowledgement We thank the participants of the survey 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 Communitys 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 03/10/2014.
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (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 (2012) 14 – 44
4. Simperl, E.: Reusing ontologies on the semantic web: A feasibility study. In: Data Knowledge Engineering **68**(10) (2009) 905 – 925
5. Schaible, J., Gottron, T., Scherp, A.: Survey on common strategies of vocabulary reuse in linked open data modeling. In: Proceeding of the Extended Semantic Web Conference (ESWC'14), Springer (2014)
6. Auer, S., Demter, J., Martin, M., Lehmann, J.: Lodstats - an extensible framework for high-performance dataset analytics. In: EKAW. Volume 7603 of Lecture Notes in Computer Science., Springer (2012) 353–362

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. (2013) 391–396
9. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. Proceedings of the Linked Data on the Web Workshop (LDOW2010) (2010)
10. Zangerle, E., Gassler, W., Specht, G.: Recommending structure in collaborative semistructured information systems. In: Proceedings of the Fourth ACM Conference on Recommender Systems. RecSys '10, New York, NY, USA, ACM (2010) 261–264
11. 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)
12. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Synthesis Lectures on The Semantic Web: Theory and Technology (1) (2012)
13. Atemezing, G.A., Villazón-Terrazas, B., Hyland, B.: Best practices for publishing linked data. W3C note, W3C (January 2014) http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/.
14. 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 (2007) 363–371
15. Haase, P., Lewen, H., Studer, R., Tran, D.T., Erdmann, M., d'Aquin, M., Motta, E.: The neon ontology engineering toolkit. In J. Korn, editor, WWW 2008 Developers Track (2008)
16. 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)

Bisher erschienen

Arbeitsberichte aus dem Fachbereich Informatik

(http://www.uni-koblenz-landau.de/koblenz/fb4/publications/Reports/arbeitsberichte)

Johann Schaible, Thomas Gottron, Ansgar Scherp, Extended Description oft he Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling, Arbeitsberichte aus dem Fachbereich Informatik 1/2014

Ulrich Furbach, Claudia Schon, Sementically Guided Evolution of SHI ABoxes, Arbeitsberichte aus dem Fachbereich Informatik 4/2013

Andreas Kasten, Ansgar Scherp, Iterative Signing of RDF(S) Graphs, Named Graphs, and OWL Graphs: Formalization and Application, Arbeitsberichte aus dem Fachbereich Informatik 3/2013

Thomas Gottron, Johann Schaible, Stefan Scheglmann, Ansgar Scherp, LOVER: Support for Modeling Data Using Linked Open Vocabularies, Arbeitsberichte aus dem Fachbereich Informatik 2/2013

Markus Bender, E-Hyper Tableaux with Distinct Objects Identifiers, Arbeitsberichte aus dem Fachbereich Informatik 1/2013

Kurt Lautenbach, Kerstin Susewind, Probability Propagation Nets and Duality, Arbeitsberichte aus dem Fachbereich Informatik 11/2012

Kurt Lautenbach, Kerstin Susewind, Applying Probability Propagation Nets, Arbeitsberichte aus dem Fachbereich Informatik 10/2012

Kurt Lautenbach, The Quaternality of Simulation: An Event/Non-Event Approach, Arbeitsberichte aus dem Fachbereich Informatik 9/2012

Horst Kutsch, Matthias Bertram, Harald F.O. von Kortzfleisch, Entwicklung eines Dienstleistungsproduktivitätsmodells (DLPMM) am Beispiel von B2b Software-Customizing, Fachbereich Informatik 8/2012

Rüdiger Grimm, Jean-Noël Colin, Virtual Goods + ODRL 2012, Arbeitsberichte aus dem Fachbereich Informatik 7/2012

Ansgar Scherp, Thomas Gottron, Malte Knauf, Stefan Scheglmann, Explicit and Implicit Schema Information on the Linked Open Data Cloud: Joined Forces or Antagonists? Arbeitsberichte aus dem Fachbereich Informatik 6/2012

Harald von Kortzfleisch, Ilias Mokanis, Dorothée Zerwas, Introducing Entrepreneurial Design Thinking, Arbeitsberichte aus dem Fachbereich Informatik 5/2012

Ansgar Scherp, Daniel Eißing, Carsten Saathoff, Integrating Multimedia Metadata Standarts and Metadata Formats with the Multimedia Metadata Ontology: Method and Examples, Arbeitsberichte aus dem Fachbereich Informatik 4/2012

Martin Surrey,Björn Lilge, Ludwig Paulsen, Marco Wolf, Markus Aldenhövel, Mike Reuthel, Roland Diehl, Integration von CRM-Systemen mit Kollaborations-Systemen am Beispiel von DocHouse und Lotus Quickr, Arbeitsberichte aus dem Fachbereich Informatik 3/2012

Martin Surrey, Roland Diehl, DOCHOUSE: Opportunity Management im Partnerkanal (IBM Lotus Quickr), Arbeitsberichte aus dem Fachbereich Informatik 2/2012

Mark Schneider, Ansgar Scherp, Comparing a Grid-based vs. List-based Approach for Faceted Search of Social Media Data on Mobile Devices, Arbeitsberichte aus dem Fachbereich Informatik 1/2012

Petra Schubert, Femi Adisa, Cloud Computing for Standard ERP Systems: Reference Framework and Research Agenda, Arbeitsberichte aus dem Fachbereich Informatik 16/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Natalia A. Zenkova, Simulating social objects with an artificial network using a computer cluster, Arbeitsberichte aus dem Fachbereich Informatik 15/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Simulating medical objects using an artificial network whose structure is based on adaptive resonance theory, Arbeitsberichte aus dem Fachbereich Informatik 14/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Comparing the efficiency of serial and parallel algorithms for training artificial neural networks using computer clusters, Arbeitsberichte aus dem Fachbereich Informatik, 13/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, A parallel algorithm for selecting activation functions of an artificial network, Arbeitsberichte aus dem Fachbereich Informatik 12/2011

Katharina Bräunlich, Rüdiger Grimm, Andreas Kasten, Sven Vowé, Nico Jahn, Der neue Personalausweis zur Authentifizierung von Wählern bei Onlinewahlen, Arbeitsberichte aus dem Fachbereich Informatik 11/2011

Daniel Eißing, Ansgar Scherp, Steffen Staab, Formal Integration of Individual Knowledge Work and Organizational Knowledge Work with the Core Ontology *strukt*, Arbeitsberichte aus dem Fachbereich Informatik 10/2011

Bernhard Reinert, Martin Schumann, Stefan Müller, Combined Non-Linear Pose Estimation from Points and Lines, Arbeitsberichte aus dem Fachbereich Informatik 9/2011

Tina Walber, Ansgar Scherp, Steffen Staab, Towards the Understanding of Image Semantics by Gaze-based Tag-to-Region Assignments, Arbeitsberichte aus dem Fachbereich Informatik 8/2011

Alexander Kleinen, Ansgar Scherp, Steffen Staab, Mobile Facets – Faceted Search and Exploration of Open Social Media Data on a Touchscreen Mobile Phone, Arbeitsberichte aus dem Fachbereich Informatik 7/2011

Anna Lantsberg, Klaus G. Troitzsch, Towards A Methodology of Developing Models of E-Service Quality Assessment in Healthcare, Arbeitsberichte aus dem Fachbereich Informatik 6/2011

Ansgar Scherp, Carsten Saathoff, Thomas Franz, Steffen Staab, Designing Core Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 5/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, The prediction of currency exchange rates using artificial neural networks, Arbeitsberichte aus dem Fachbereich Informatik 4/2011

Klaus G. Troitzsch, Anna Lantsberg, Requirements for Health Care Related Websites in Russia: Results from an Analysis of American, British and German Examples, Arbeitsberichte aus dem Fachbereich Informatik 3/2011

Klaus G. Troitzsch, Oleg Kryuchin, Alexander Arzamastsev, A universal simulator based on artificial neural networks for computer clusters, Arbeitsberichte aus dem Fachbereich Informatik 2/2011

Klaus G. Troitzsch, Natalia Zenkova, Alexander Arzamastsev, Development of a technology of designing intelligent information systems for the estimation of social objects, Arbeitsberichte aus dem Fachbereich Informatik 1/2011

Kurt Lautenbach, A Petri Net Approach for Propagating Probabilities and Mass Functions, Arbeitsberichte aus dem Fachbereich Informatik 13/2010

Claudia Schon, Linkless Normal Form for ALC Concepts, Arbeitsberichte aus dem Fachbereich Informatik 12/2010

Alexander Hug, Informatik hautnah erleben, Arbeitsberichte aus dem Fachbereich Informatik 11/2010

Marc Santos, Harald F.O. von Kortzfleisch, Shared Annotation Model – Ein Datenmodell für kollaborative Annotationen, Arbeitsberichte aus dem Fachbereich Informatik 10/2010

Gerd Gröner, Steffen Staab, Categorization and Recognition of Ontology Refactoring Pattern, Arbeitsberichte aus dem Fachbereich Informatik 9/2010

Daniel Eißing, Ansgar Scherp, Carsten Saathoff, Integration of Existing Multimedia Metadata Formats and Metadata Standards in the M3O, Arbeitsberichte aus dem Fachbereich Informatik 8/2010

Stefan Scheglmann, Ansgar Scherp, Steffen Staab, Model-driven Generation of APIs for OWL-based Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 7/2010

Daniel Schmeiß, Ansgar Scherp, Steffen Staab, Integrated Mobile Visualization and Interaction of Events and POIs, Arbeitsberichte aus dem Fachbereich Informatik 6/2010

Rüdiger Grimm, Daniel Pähler, E-Mail-Forensik – IP-Adressen und ihre Zuordnung zu Internet-Teilnehmern und ihren Standorten, Arbeitsberichte aus dem Fachbereich Informatik 5/2010

Christoph Ringelstein, Steffen Staab, PAPEL: Syntax and Semantics for Provenance-Aware Policy Definition, Arbeitsberichte aus dem Fachbereich Informatik 4/2010

Nadine Lindermann, Sylvia Valcárcel, Harald F.O. von Kortzfleisch, Ein Stufenmodell für kollaborative offene Innovationsprozesse in Netzwerken kleiner und mittlerer Unternehmen mit Web 2.0, Arbeitsberichte aus dem Fachbereich Informatik 3/2010

Maria Wimmer, Dagmar Lück-Schneider, Uwe Brinkhoff, Erich Schweighofer, Siegfried Kaiser, Andreas Wieber, Fachtagung Verwaltungsinformatik FTVI Fachtagung Rechtsinformatik FTRI 2010, Arbeitsberichte aus dem Fachbereich Informatik 2/2010

Max Braun, Ansgar Scherp, Steffen Staab, Collaborative Creation of Semantic Points of Interest as Linked Data on the Mobile Phone, Arbeitsberichte aus dem Fachbereich Informatik 1/2010

Marc Santos, Einsatz von „Shared In-situ Problem Solving“ Annotationen in kollaborativen Lern- und Arbeitsszenarien, Arbeitsberichte aus dem Fachbereich Informatik 20/2009

Carsten Saathoff, Ansgar Scherp, Unlocking the Semantics of Multimedia Presentations in the Web with the Multimedia Metadata Ontology, Arbeitsberichte aus dem Fachbereich Informatik 19/2009

Christoph Kahle, Mario Schaarschmidt, Harald F.O. von Kortzfleisch, Open Innovation: Kundenintegration am Beispiel von IPTV, Arbeitsberichte aus dem Fachbereich Informatik 18/2009

Dietrich Paulus, Lutz Priese, Peter Decker, Frank Schmitt, Pose-Tracking Forschungsbericht, Arbeitsberichte aus dem Fachbereich Informatik 17/2009

Andreas Fuhr, Tassilo Horn, Andreas Winter, Model-Driven Software Migration Extending SOMA, Arbeitsberichte aus dem Fachbereich Informatik 16/2009

Eckhard Großmann, Sascha Strauß, Tassilo Horn, Volker Riediger, Abbildung von grUML nach XSD soamig, Arbeitsberichte aus dem Fachbereich Informatik 15/2009

Kerstin Falkowski, Jürgen Ebert, The STOR Component System Interim Report, Arbeitsberichte aus dem Fachbereicht Informatik 14/2009

Sebastian Magnus, Markus Maron, An Empirical Study to Evaluate the Location of Advertisement Panels by Using a Mobile Marketing Tool, Arbeitsberichte aus dem Fachbereich Informatik 13/2009

Sebastian Magnus, Markus Maron, Konzept einer Public Key Infrastruktur in iCity, Arbeitsberichte aus dem Fachbereich Informatik 12/2009

Sebastian Magnus, Markus Maron, A Public Key Infrastructure in Ambient Information and Transaction Systems, Arbeitsberichte aus dem Fachbereich Informatik 11/2009

Ammar Mohammed, Ulrich Furbach, Multi-agent systems: Modeling and Virification using Hybrid Automata, Arbeitsberichte aus dem Fachbereich Informatik 10/2009

Andreas Sprotte, Performance Measurement auf der Basis von Kennzahlen aus betrieblichen Anwendungssystemen: Entwurf eines kennzahlengestützten Informationssystems für einen Logistikdienstleister, Arbeitsberichte aus dem Fachbereich Informatik 9/2009

Gwendolin Garbe, Tobias Hausen, Process Commodities: Entwicklung eines Reifegradmodells als Basis für Outsourcingentscheidungen, Arbeitsberichte aus dem Fachbereich Informatik 8/2009

Petra Schubert et. al., Open-Source-Software für das Enterprise Resource Planning, Arbeitsberichte aus dem Fachbereich Informatik 7/2009

Ammar Mohammed, Frieder Stolzenburg, Using Constraint Logic Programming for Modeling and Verifying Hierarchical Hybrid Automata, Arbeitsberichte aus dem Fachbereich Informatik 6/2009

Tobias Kippert, Anastasia Meletiadou, Rüdiger Grimm, Entwurf eines Common Criteria-Schutzprofils für Router zur Abwehr von Online-Überwachung, Arbeitsberichte aus dem Fachbereich Informatik 5/2009

Hannes Schwarz, Jürgen Ebert, Andreas Winter, Graph-based Traceability – A Comprehensive Approach. Arbeitsberichte aus dem Fachbereich Informatik 4/2009

Anastasia Meletiadou, Simone Müller, Rüdiger Grimm, Anforderungsanalyse für Risk-Management-Informationssysteme (RMIS), Arbeitsberichte aus dem Fachbereich Informatik 3/2009

Ansgar Scherp, Thomas Franz, Carsten Saathoff, Steffen Staab, A Model of Events based on a Foundational Ontology, Arbeitsberichte aus dem Fachbereich Informatik 2/2009

Frank Bohdanovicz, Harald Dickel, Christoph Steigner, Avoidance of Routing Loops, Arbeitsberichte aus dem Fachbereich Informatik 1/2009

Stefan Ameling, Stephan Wirth, Dietrich Paulus, Methods for Polyp Detection in Colonoscopy Videos: A Review, Arbeitsberichte aus dem Fachbereich Informatik 14/2008

Tassilo Horn, Jürgen Ebert, Ein Referenzschema für die Sprachen der IEC 61131-3, Arbeitsberichte aus dem Fachbereich Informatik 13/2008

Thomas Franz, Ansgar Scherp, Steffen Staab, Does a Semantic Web Facilitate Your Daily Tasks?, Arbeitsberichte aus dem Fachbereich Informatik 12/2008

Norbert Frick, Künftige Anfordeungen an ERP-Systeme: Deutsche Anbieter im Fokus, Arbeitsberichte aus dem Fachbereicht Informatik 11/2008

Jürgen Ebert, Rüdiger Grimm, Alexander Hug, Lehramtsbezogene Bachelor- und Masterstudiengänge im Fach Informatik an der Universität Koblenz-Landau, Campus Koblenz, Arbeitsberichte aus dem Fachbereich Informatik 10/2008

Mario Schaarschmidt, Harald von Kortzfleisch, Social Networking Platforms as Creativity Fostering Systems: Research Model and Exploratory Study, Arbeitsberichte aus dem Fachbereich Informatik 9/2008

Bernhard Schueler, Sergej Sizov, Steffen Staab, Querying for Meta Knowledge, Arbeitsberichte aus dem Fachbereich Informatik 8/2008

Stefan Stein, Entwicklung einer Architektur für komplexe kontextbezogene Dienste im mobilen Umfeld, Arbeitsberichte aus dem Fachbereich Informatik 7/2008

Matthias Bohnen, Lina Brühl, Sebastian Bzdak, RoboCup 2008 Mixed Reality League Team Description, Arbeitsberichte aus dem Fachbereich Informatik 6/2008

Bernhard Beckert, Reiner Hähnle, Tests and Proofs: Papers Presented at the Second International Conference, TAP 2008, Prato, Italy, April 2008, Arbeitsberichte aus dem Fachbereich Informatik 5/2008

Klaas Dellschaft, Steffen Staab, Unterstützung und Dokumentation kollaborativer Entwurfs- und Entscheidungsprozesse, Arbeitsberichte aus dem Fachbereich Informatik 4/2008

Rüdiger Grimm: IT-Sicherheitsmodelle, Arbeitsberichte aus dem Fachbereich Informatik 3/2008

Rüdiger Grimm, Helge Hundacker, Anastasia Meletiadou: Anwendungsbeispiele für Kryptographie, Arbeitsberichte aus dem Fachbereich Informatik 2/2008

Markus Maron, Kevin Read, Michael Schulze: CAMPUS NEWS – Artificial Intelligence Methods Combined for an Intelligent Information Network, Arbeitsberichte aus dem Fachbereich Informatik 1/2008

Lutz Priese,Frank Schmitt, Patrick Sturm, Haojun Wang: BMBF-Verbundprojekt 3D-RETISEG Abschlussbericht des Labors Bilderkennen der Universität Koblenz-Landau, Arbeitsberichte aus dem Fachbereich Informatik 26/2007

Stephan Philippi, Alexander Pinl: Proceedings 14. Workshop 20.-21. September 2007 Algorithmen und Werkzeuge für Petrinetze, Arbeitsberichte aus dem Fachbereich Informatik 25/2007

Ulrich Furbach, Markus Maron, Kevin Read: CAMPUS NEWS – an Intelligent Bluetooth-based Mobile Information Network, Arbeitsberichte aus dem Fachbereich Informatik 24/2007

Ulrich Furbach, Markus Maron, Kevin Read: CAMPUS NEWS - an Information Network for Pervasive Universities, Arbeitsberichte aus dem Fachbereich Informatik 23/2007

Lutz Priese: Finite Automata on Unranked and Unordered DAGs Extented Version, Arbeitsberichte aus dem Fachbereich Informatik 22/2007

Mario Schaarschmidt, Harald F.O. von Kortzfleisch: Modularität als alternative Technologie- und Innovationsstrategie, Arbeitsberichte aus dem Fachbereich Informatik 21/2007

Kurt Lautenbach, Alexander Pinl: Probability Propagation Nets, Arbeitsberichte aus dem Fachbereich Informatik 20/2007

Rüdiger Grimm, Farid Mehr, Anastasia Meletiadou, Daniel Pähler, Ilka Uerz: SOA-Security, Arbeitsberichte aus dem Fachbereich Informatik 19/2007

Christoph Wernhard: Tableaux Between Proving, Projection and Compilation, Arbeitsberichte aus dem Fachbereich Informatik 18/2007

Ulrich Furbach, Claudia Obermaier: Knowledge Compilation for Description Logics, Arbeitsberichte aus dem Fachbereich Informatik 17/2007

Fernando Silva Parreiras, Steffen Staab, Andreas Winter: TwoUse: Integrating UML Models and OWL Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 16/2007

Rüdiger Grimm, Anastasia Meletiadou: Rollenbasierte Zugriffskontrolle (RBAC) im Gesundheitswesen, Arbeitsberichte aud dem Fachbereich Informatik 15/2007

Ulrich Furbach, Jan Murray, Falk Schmidsberger, Frieder Stolzenburg: Hybrid Multiagent Systems with Timed Synchronization-Specification and Model Checking, Arbeitsberichte aus dem Fachbereich Informatik 14/2007

Björn Pelzer, Christoph Wernhard: System Description:“E-KRHyper“, Arbeitsberichte aus dem Fachbereich Informatik, 13/2007

Ulrich Furbach, Peter Baumgartner, Björn Pelzer: Hyper Tableaux with Equality, Arbeitsberichte aus dem Fachbereich Informatik, 12/2007

Ulrich Furbach, Markus Maron, Kevin Read: Location based Informationsystems, Arbeitsberichte aus dem Fachbereich Informatik, 11/2007

Philipp Schaer, Marco Thum: State-of-the-Art: Interaktion in erweiterten Realitäten, Arbeitsberichte aus dem Fachbereich Informatik, 10/2007

Ulrich Furbach, Claudia Obermaier: Applications of Automated Reasoning, Arbeitsberichte aus dem Fachbereich Informatik, 9/2007

Jürgen Ebert, Kerstin Falkowski: A First Proposal for an Overall Structure of an Enhanced Reality Framework, Arbeitsberichte aus dem Fachbereich Informatik, 8/2007

Lutz Priese, Frank Schmitt, Paul Lemke: Automatische See-Through Kalibrierung, Arbeitsberichte aus dem Fachbereich Informatik, 7/2007

Rüdiger Grimm, Robert Krimmer, Nils Meißner, Kai Reinhard, Melanie Volkamer, Marcel Weinand, Jörg Helbach: Security Requirements for Non-political Internet Voting, Arbeitsberichte aus dem Fachbereich Informatik, 6/2007

Daniel Bildhauer, Volker Riediger, Hannes Schwarz, Sascha Strauß, „grUML – Eine UML-basierte Modellierungssprache für T-Graphen“, Arbeitsberichte aus dem Fachbereich Informatik, 5/2007

Richard Arndt, Steffen Staab, Raphaël Troncy, Lynda Hardman: Adding Formal Semantics to MPEG-7: Designing a Well Founded Multimedia Ontology for the Web, Arbeitsberichte aus dem Fachbereich Informatik, 4/2007

Simon Schenk, Steffen Staab: Networked RDF Graphs, Arbeitsberichte aus dem Fachbereich Informatik, 3/2007

Rüdiger Grimm, Helge Hundacker, Anastasia Meletiadou: Anwendungsbeispiele für Kryptographie, Arbeitsberichte aus dem Fachbereich Informatik, 2/2007

Anastasia Meletiadou, J. Felix Hampe: Begriffsbestimmung und erwartete Trends im IT-Risk-Management, Arbeitsberichte aus dem Fachbereich Informatik, 1/2007

„Gelbe Reihe“
(http://www.uni-koblenz.de/fb4/publikationen/gelbereihe)

Lutz Priese: Some Examples of Semi-rational and Non-semi-rational DAG Languages. Extended Version, Fachberichte Informatik 3-2006

Kurt Lautenbach, Stephan Philippi, and Alexander Pinl: Bayesian Networks and Petri Nets, Fachberichte Informatik 2-2006

Rainer Gimnich and Andreas Winter: Workshop Software-Reengineering und Services, Fachberichte Informatik 1-2006

Kurt Lautenbach and Alexander Pinl: Probability Propagation in Petri Nets, Fachberichte Informatik 16-2005

Rainer Gimnich, Uwe Kaiser, and Andreas Winter: 2. Workshop "Reengineering Prozesse" – Software Migration, Fachberichte Informatik 15-2005

Jan Murray, Frieder Stolzenburg, and Toshiaki Arai: Hybrid State Machines with Timed Synchronization for Multi-Robot System Specification, Fachberichte Informatik 14-2005

Reinhold Letz: FTP 2005 – Fifth International Workshop on First-Order Theorem Proving, Fachberichte Informatik 13-2005

Bernhard Beckert: TABLEAUX 2005 – Position Papers and Tutorial Descriptions, Fachberichte Informatik 12-2005

Dietrich Paulus and Detlev Droege: Mixed-reality as a challenge to image understanding and artificial intelligence, Fachberichte Informatik 11-2005

Jürgen Sauer: 19. Workshop Planen, Scheduling und Konfigurieren / Entwerfen, Fachberichte Informatik 10-2005

Pascal Hitzler, Carsten Lutz, and Gerd Stumme: Foundational Aspects of Ontologies, Fachberichte Informatik 9-2005

Joachim Baumeister and Dietmar Seipel: Knowledge Engineering and Software Engineering, Fachberichte Informatik 8-2005

Benno Stein and Sven Meier zu Eißen: Proceedings of the Second International Workshop on Text-Based Information Retrieval, Fachberichte Informatik 7-2005

Andreas Winter and Jürgen Ebert: Metamodel-driven Service Interoperability, Fachberichte Informatik 6-2005

Joschka Boedecker, Norbert Michael Mayer, Masaki Ogino, Rodrigo da Silva Guerra, Masaaki Kikuchi, and Minoru Asada: Getting closer: How Simulation and Humanoid League can benefit from each other, Fachberichte Informatik 5-2005

Torsten Gipp and Jürgen Ebert: Web Engineering does profit from a Functional Approach, Fachberichte Informatik 4-2005

Oliver Obst, Anita Maas, and Joschka Boedecker: HTN Planning for Flexible Coordination Of Multiagent Team Behavior, Fachberichte Informatik 3-2005

Andreas von Hessling, Thomas Kleemann, and Alex Sinner: Semantic User Profiles and their Applications in a Mobile Environment, Fachberichte Informatik 2-2005

Heni Ben Amor and Achim Rettinger: Intelligent Exploration for Genetic Algorithms – Using Self-Organizing Maps in Evolutionary Computation, Fachberichte Informatik 1-2005