Vocab.at – Automatic Linked Data Documentation And Vocabulary Usage Analysis

Miika Alonen, Tomi Kauppinen, Eero Hyvönen
Semantic Computing Research Group (SeCo)
Aalto University, Department of Media Technology
firstname.lastname@aalto.fi, http://www.seco.tkk.fi/

October 15, 2013

Abstract

A growing number of Linked Data is being published as RDF data dumps, as RDFa embedded in HTML pages, and via SPARQL endpoints. Unfortunately, the data available is often poorly documented and the consistency of the datasets is unknown. Gaining an understanding of whether a dataset qualifies for the intended use can then be very time consuming and impede the re-use of the data. When considering quality as fitness of use, documentation is a key component for assessing data quality. The common practice today is to document Linked Data vocabularies that are used by Linked Data. However, this approach neglects documenting the actual vocabulary usage in the datasets. In contrast, this paper presents an novel approach for assessing the vocabulary usage in Linked Data. The method generates missing documentation automatically and complements this by analysing the usage of vocabularies in the datasets. The resulted documentation shows the explicit vocabulary usage, which is invaluable when assessing the consistency and usefulness of the data. This method has been evaluated by developing a web service http://vocab.at and applying the analysis to selected datasets on the web.

1 Introduction

Linked Open Data (LOD)[1] is a way of promoting the re-use of data by publishing the data using the Web infrastructure and the Semantic Web standards[2]. The use of Linked Data breaches data silos by identifying the data with International Resource Identifiers (IRIs), describing the semantics using shared vocabularies and serving the data with standard protocols. The anticipated benefits of using the Linked Data approach are enhanced accessibility and increased usability of data through data harmonization, integration, and semantic enrichment by reasoning.

Large amounts of Linked Data are publicly available on the Web with diverse conventions in use in describing the datasets. The wast amount of data and absence of dataset descriptions impair the usefulness of Linked Data: it becomes *"merely more data"* as argued by Jain et al. (2010). To improve the usability of Linked Data it is important to document data for both human and machine users. For this purpose, a dataset can be described, e.g., with the Vocabulary of Interlinked Datasets (VoID) designed by Alexander et al. (2009). The use of dataset descriptions allows machines to determine how datasets can be accessed and whether a given dataset contains relevant information for a task at hand. However, assessing the usefulness and potential of re-using Linked Data

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://www.w3.org/standards/semanticweb/

from a human perspective depends on how easily the re-user can get an understanding of what the dataset actually contains and what is the quality of the data for the intended purpose.

The documentation of Linked Data is too often limited to just listing the vocabularies used to describe the data. The various ways in which vocabularies are used is not documented, which may impede understanding about the contents of the published data. We argue that without a proper documentation about the vocabulary usage, it is hard to evaluate the usability and usefulness of Linked Data. Linked Data publishers are often reluctant to document the datasets manually, as it is complicated, time consuming, and an error-prone task. Furthermore, dynamic changes and the evolution of Linked Data make the assessment even more difficult. For example, minor changes to the used vocabularies can affect the discoverability of the data.

Resource Description Framework (RDF)—the core of Linked Data—has been designed to be self-describing and to facilitate sharing and merging of the data (Cyganiak and Wood, 2013). Most significantly, the RDF model uses IRIs to identify things and relationships between them. Resources published as Linked Data and the used vocabulary terms should be accessible over the Web. In this way, the Web infrastructure makes the evaluation of Linked Data possible. However, in practice, the overall quality of the published datasets is far from perfect. Recent studies (Hogan et al., 2012; Hitzler, 2012; Auer et al., 2012) indicate that more than half of the published dataset have quality issues. One obvious reason to the low quality is the lack of know-how of Linked Data standards and best practices, and the fact that the data is often published without consideration of how to assess the data.

Different aspects of the Linked Data Quality have already been widely studied by several authors (Assaf and Senart, 2012; Hogan et al., 2012, 2010; Hartig and Zhao, 2009) and before them the principles for the data quality by many others Pipino et al. (2002); Juran et al. (1999); Redman and Blanton (1997); Wang et al. (1996). There are also tools available for Linked Data quality management (Mendes et al., 2012; Fürber and Hepp, 2011, 2010) and analytics (Auer et al., 2012; Langegger and Woss, 2009). However, there is no account of how widely these tools are used to improve the quality of Linked Data.

The Linked Data publication process should ideally include an assessment of the data in order to ensure the usefulness of the dataset for the task at hand, or for similar tasks, or for re-use of the data for different tasks. It is a common practice to use RDF validators for this. However, the RDF validation only ensures the validity of the dataset structure according to the RDF Specification. We argue that after the formal validation, the usage of RDF vocabularies should also be inspected. For example, in cases where RDF is automatically generated from legacy sources, an analysis of the resulted data and vocabulary usege can provide evidence that the transformation process was successful.

The contribution of this paper is to describe identified challenges in assessing the usability of Linked Data, and to present a new method for analysing the usage of vocabularies. Our approach emphasizes the importance of making Linked Data comprehensible to its re-user. For this, the method establishes a new kind of documentation that combines a vocabulary documentation and the statistics about how vocabulary terms are used in a given dataset at a certain time point.

The paper is structured as follows. Section 2 provides the motivation and describes the challenges of assessing the usability of Linked Data. Section 3 presents a method for vocabulary analysis and automatic documentation. In Section 4 we outline a service architecture implementing the method. Section 5 discusses a case study of assessing the quality of LOD datasets and presents the results. In Section 6 the related work is concerned, and Section 7 concludes the paper.

2 The Challenge of Assessing the Usability of Linked Data

One should be able to publish Linked Data of good quality by following the Semantic Web standards and the best practices, such as the Linked Data Design

Considerations (Heath and Bizer, 2011). However, it is very easy to lay aside some of the quality aspects in the process of creating and publishing Linked Data. The usefulness of the datasets should be the main priority when publishing the data, but this is often neglected even by the practitioners of the Linked Data (Hitzler, 2012).

One of the major obstacles in assessing the usability of Linked Data is the incomprehensibility of large datasets. There are methods for ranking datasets based on the interconnectivity (Toupikov et al., 2009) and means for assessing the data quality with approaches proposed by Mendes et al. (2012); Fürber and Hepp (2011, 2010). However, these methods do not necessarily improve the understandability of the dataset. The rating of the datasets based on the interlinkedness of the data is a good approach when looking for the related datasets, but links do not tell much about the usefulness of the data. In the other hand, the methods for assessing the quality of Linked Data already require understanding of the used data model.

The usability of the data depends on the requirements of the intended task. For Linked Data, it is impossible to know all of the requirements in advance. The dataset may qualify for a certain task, and in the same time fail in fulfilling the requirements of another. In that sense, the quality of Linked Data depends on the re-use of the data, and could be defined by rephrasing the common definition of quality as *fitness of re-use.*

The re-use of the Linked Data depends on the use case whose requirements can be based on many data quality principles, such as timeliness, consistency and accessibility, and other such criteria proposed for Linked Data quality (Assaf and Senart, 2012; Flemming and Hartig, 2010).

In a Linked Data re-use scenario, the user might not have any a priori information about the data. In order to assess the usability of the datasets, the user needs the information about the used data model. Linked Data is often created using various properties and classes from several vocabularies, and new vocabularies are formed for those information needs that are not in any vocabulary. The usual documentation of Linked Data, if any, is usually a list of used vocabularies and possibly an illustration of the used vocabulary terms. The problem with this kind of documentation is that it does not describe the dataset as it is, but only describes the vocabularies that are supposedly being used in the data.

To overcome this limitation we suggest a method for making the actual usage of vocabularies explicit by generating documentation automatically from the dataset. Automatically generated documentations are objective, up-to-date, and can aid to assess multiple data quality dimensions, such as the representational conciseness and consistency. This means that the user can evaluate if the data is complete and compatible with the current needs. The knowledge about the data structures should also help to improve the data quality with appropriate measures.

3 A Method for Assessing Vocabulary Usage in Linked Datatasets

In this section, we describe the method for automatically generating the documentation for assessing the vocabulary usage in Linked Data. The method can be applied to any dataset by acquiring the Linked Data from an arbitrary source, resolving the used vocabularies, creating a dataset description, and generating the dataset documentation.

Linked Data Sets (cf. Definition 1) are published on the web in a variety of formats. This requires different means for data processing. However, all formats use the same RDF data model, which allows us to define a generic method for assessing the vocabulary usage.

Definition 1 *Linked Data Set* $\mathbf{D}$ *is a set of triples constructed from URI references* $\mathbf{U}$*, Literals* $\mathbf{L}$ *and blank nodes* $\mathbf{B}$*. A single triple* $\mathbf{t}$ *is a construct that can be defined as* $\mathbf{t} = \langle \mathbf{s}, \mathbf{p}, \mathbf{o} \rangle$*, where* $\mathbf{s} \in \mathbf{U} \cup \mathbf{B}$*,* $\mathbf{p} \in \mathbf{U}$ *and* $\mathbf{o} \in \mathbf{U} \cup \mathbf{B} \cup \mathbf{L}$*.*

The best practice for describing the resources in Linked Data is to re-use existing terms from different vocabularies (Heath and Bizer, 2011). The machine processable version of the used vocabulary should be available on the Web following the Linked Data best practices (Berrueta and Phipps, 2008). The Linked Data may use multiple Linked Data Vocabularies (cf. Definition 2) to describe its classes and properties.

Definition 2 *Linked Data Vocabulary* $\mathbf{V}$ *is a set of triples* $\mathbf{T}$ *that describe the set of classes* $\mathbf{C}$ *and set of properties* $\mathbf{P}$ *which are identified with URIs meaning* $\mathbf{C} \cup \mathbf{P} \subset \mathbf{U}$. *Any dataset* $\mathbf{D}$ *may describe a triple* $\mathbf{t} = \langle \mathbf{s}, \mathbf{p}, \mathbf{o} \rangle$, *where class is* $\mathbf{o} \in \mathbf{C}$ *or property is* $\mathbf{p} \in \mathbf{P}$ *from the vocabulary* $\mathbf{V}$.

The use of Linked Data and vocabulary usage can be described with dataset description (Alexander et al., 2011) by analysing the dataset and collecting statistics about the property and class usage. We define the dataset description generally in Definition 3 as metadata created or collected from the data. The use of properties and classes can be described with separate partitions that represent the amount of the data and the usage of a certain vocabulary term. Property partitions are used to describe the total usage of properties and class partitions document the usage of classes. The property usage by a certain class can be also described in separate class partition.

Definition 3 *Dataset Description is a set of triples that describes metadata about certain Linked Data at a given time.*

The usage of vocabularies is resolved by requesting the vocabularies from the namespaces of used vocabulary terms, and the metadata about the unresolved vocabularies is stored to the dataset description. The unresolved vocabulary usage can then be calculated from the number of triples and the unresolved vocabularies. The unresolved vocabulary usage is then used to calculate a Vocabulary Score (cf. Definition 4), that is a metric that indicates how much of the used vocabulary is actually defined and dereferenceable.

Definition 4 *The Vocabulary Score is defined as* $\mathbf{V} = 1 - \frac{|\mathbf{Pu}|+|\mathbf{Cu}|}{|\mathbf{D}|}$ *where class usage is defined as* $\mathbf{Cu} = \{\langle \mathbf{s}, \mathbf{p}, \mathbf{o} \rangle | \mathbf{p} = \text{rdf:type}, \mathbf{o} \notin \mathbf{C}\}$ *and property usage as* $\mathbf{Pu} = \{\langle \mathbf{s}, \mathbf{p}, \mathbf{o} \rangle | \mathbf{p} = \text{rdf:type}, \mathbf{p} \notin \mathbf{P}\}$

The Vocabulary Score works as an indicator for the correct usage of Linked Data vocabularies. The score for a good quality dataset should always be 1, meaning the used vocabulary is accessible and properly published. However, the assessment of the rationality and usability of dataset needs more context dependent evaluation. The usefulness of the dataset can only be realized depending on the requirements placed for the data.

For assessing the consistency and usefulness of a dataset, its dataset description is used to automatically generate documentation. The documentation is generated from the class and property partitions, which describe the usage of vocabulary terms in the dataset. The resulted Dataset Documentation (cf. Definition 5) is a snapshot of the dataset and a mashup of multiple vocabularies.

Definition 5 *The Dataset Documentation is a depiction of the dataset description that represents the statistics and usage of vocabulary terms in more comprehendible human-readable format.*

Objective documentation is essential information for assessing the consistency of the datasets. Automatically generated dataset documentation is a platform for communicating the data model, actual usage of vocabulary terms and other statistics to the user. The used vocabularies are dereferenced from the original sources, and the documentation is enriched with corresponding definitions from the vocabularies. Unresolved namespaces, undefined properties and classes are described as Issues, which can be resolved by correcting the problems with the dataset. The cause of an issue in the dataset may also be a problem with the vocabulary, for example an invalid base URI in the schema.

4 The Design of vocab.at

The method for Linked Data vocabulary assessment and documentation was evaluated with our demonstrative service *vocab.at*. The system was implemented by using a light-weight architecture where the processing of the Linked Data and the interface is separated from each other. The backend of the system

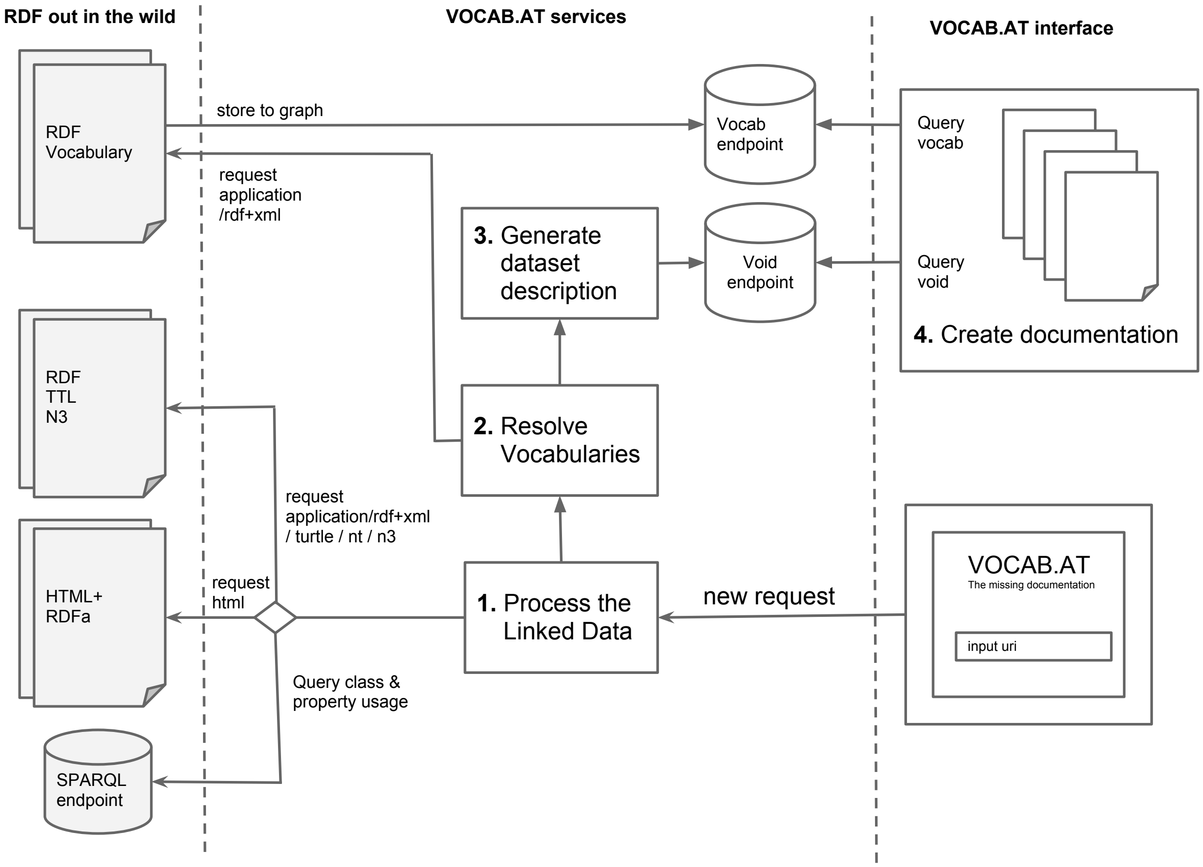


Figure 1: The design for creating Linked Data Documentations for vocabulary usage assessment

is built with Java and uses Jena for stream parsing[3] and to access[4] the Fuseki triplestore. The interface is implemented with HTML5 and JavaScript using AngularJS-framework[5]. The workflow of the system is illustrated in Figure 1, and each process step is explained further in following subsections.

4.1 Linked Data Processing

Linked Data is processed by the *vocab.at* service by first determining the type of the given RDF dataset. The content type referred to by the request URI is determined using content negotiation and SPARQL, and the corresponding parsing mechanism is selected. This is illustrated in step 1. in Figure 1. Once the right method for processing the dataset has been determined the statistics of the vocabulary usage is collected into a hashtable. The processing and querying is done by optimizing the performance so that large datasets are not loaded in the memory.

[3] http://jena.apache.org/documentation/io/rdf-input.html

[4] http://jena.apache.org/documentation/serving_data/

[5] http://angularjs.org/

An efficient way of processing RDF data dumps on the web is the statement-stream based approach, which was first applied to the Linked Data by Auer et al. (2012) to create statistics from the CKAN registry[6]. A similar approach is applied to the RDFa-annotated HTML by streaming the content and processing the found triples from the stream.

The SPARQL endpoint is the most challenging data source to process. It would be efficient to use SPARQL Graph Store HTTP Protocol (Ogbuji, 2013), and parse the resulted RDF with the statement-stream approach, but most of the SPARQL endpoints do not support or provide open access for the protocol. The viable way is to query the required statistics from the SPARQL endpoint with multiple explorative queries[7]. However, this may result in timeout when, for example, counting all of the triples in a large dataset.

The SPARQL queries for collecting the statistics from the endpoints need to be efficient. For better efficiency the queries are sent to selected graphs separately, by first querying all the class definitions. The class and property usage is then queried one class at a time. A drawback of this approach is that some of the datasets use only properties and blank nodes, and the usage of those properties needs to be queried separately.

4.2 Resolving Vocabularies

Once the Linked Data is processed and the vocabulary usage is stored into the hashtable, the URIs from the used classes and properties are dereferenced. The used vocabularies are resolved from the namespaces of the used terms. Used vocabularies should be dereferenceable according to the best practices of publishing the vocabularies. However, often the vocabulary terms are not dereferenceable and loading the whole vocabulary once is more efficient.

The vocabularies are loaded from the namespace URIs using *application/rdf+xml* content type, see step 2. in Figure 1. Each resolved vocabulary is stored into a SPARQL endpoint and into a separate graph. This way the vocabularies are easily accessible for querying definitions and determining, if a certain vocabulary term exists in the given vocabulary.

4.3 Generate Dataset Description

The usage of vocabularies is described as a dataset description that is generated from the hashtable (Cf. step 3. in Figure 1). The usage of vocabularies is expressed by describing the datasets using the VoiD vocabulary (Alexander et al., 2009). Dataset description is a way of formally describing the linked RDF datasets and the links between them. One of the original use cases of dataset descriptions was also to describe the vocabularies used by the dataset.

The generated dataset descriptions describe statistics about the class usage, property usage, and property usage of a certain class. The provenance of the generation process is expressed with the PROV vocabulary (Gil and Miles, 2013), and the errors and issues related to the vocabulary misuse are described using the VOCAB-vocabulary[8]. The vocabularies used by the *vocab.at* service is documented by using the system itself[9].

A dataset description is a snapshot of a vocabulary usage in the dataset at a given time. An example of a generated dataset description is presented in Example 1. The time of the generation and the generation process is described using the PROV vocabulary. The generation process is started by a user, by creating a new *prov:Activity* that is identified with a unique URI. The generated dataset description is described as a *void:Dataset*, which defines the used vocabulary with *void:propertyPartition* and *void:classPartition*.

Each partition is itself a *void:Dataset*[10], and class partitions are also being used to describe the property usage with the corresponding entities. The resulting dataset description is a set of blank nodes which is stored to a graph in the *vocab.at* SPARQL endpoint[11]. If there already is a dataset description created from the dataset, the equality to the earlier

[6] http://thedatahub.org

[7] http://code.google.com/p/void-impl/wiki/SPARQLQueriesForStatistics

[8] http://vocab.at/schema

[9] http://vocab.at?uri=http://vocab.at/sparql

[10] http://www.w3.org/TR/void/#class-property-partitions

[11] http://vocab.at/sparql

version is evaluated with the graph matching algorithm introduced by Carroll (2002). The new version of the description is stored to the endpoint only if the two dataset descriptions differ from each other, meaning that the vocabulary usage in the dataset has changed.

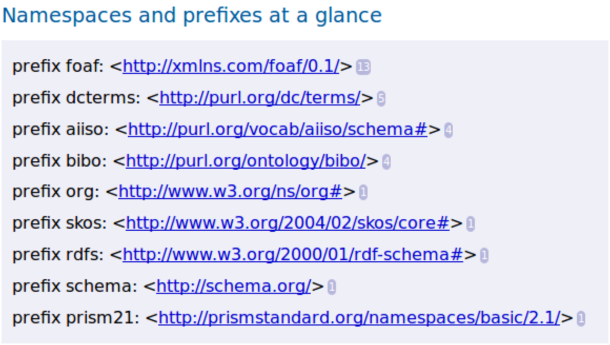
```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix void: <http://rdfs.org/ns/void#> .

<http://vocab.at/id/3Ivr> a <prov:Activity> ;
 <dct:source> <http://vocab.at/info> ;
 <prov:startedAtTime> "2013-09-24T08:00:46.122Z"^^
     xsd:dateTime ;
 <prov:endedAtTime> "2013-09-24T08:00:48.249Z"^^xsd
     :dateTime ;
 <prov:generated>
    [ a <void:Dataset> ;
        <dct:source> <http://vocab.at/info> ;
        <void:classPartition>
         [ <void:class> foaf:Document ;
           <void:distinctSubjects> 1 ;
           <void:propertyPartition>
             [ <void:entities> 1 ;
               <void:property> dct:subject ;
               <void:triples> 4
             ] ;
         ];
        <void:vocabulary>
         <http://purl.org/dc/elements/1.1/> ,
         <http://xmlns.com/foaf/0.1/> ;
         # continues, see http://vocab.at/data/300v
    ];
```

Example 1: Dataset description

4.4 Documentation Generation

The dataset descriptions stored into the *http://vocab.at/sparql* endpoint are used as an input to the document generator. The documentation generator is implemented as a dynamic *AngularJS* application that queries vocabulary usage from the SPARQL endpoint and generates the documentation on the fly. Dataset documentations created from the datasets are accessible from the main page of *vocab.at* by inputting the URI of the dataset to the main view, or directly with an *uri* parameter[12].

Each dataset documentation is identified with a permanent short URL, which can be used to directly to access a certain version of the dataset documentation. The used terms and the usage statistics are divided to different sections in the documentation, based on the used vocabularies, classes, and properties. The sections are combined with the related statistics from the dataset description.



Namespaces and prefixes at a glance

prefix foaf: <http://xmlns.com/foaf/0.1/> 13
prefix dcterms: <http://purl.org/dc/terms/> 5
prefix aiiso: <http://purl.org/vocab/aiiso/schema#> 4
prefix bibo: <http://purl.org/ontology/bibo/> 4
prefix org: <http://www.w3.org/ns/org#> 1
prefix skos: <http://www.w3.org/2004/02/skos/core#> 1
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> 1
prefix schema: <http://schema.org/> 1
prefix prism21: <http://prismstandard.org/namespaces/basic/2.1/> 1

Figure 2: Example of vocabulary usage

The vocabularies, used namespaces, and prefixes are defined in the vocabulary section (cf. Figure 2). The list of prefixes and namespaces describe the number of terms used from a certain vocabulary, and links to a more specific vocabulary section. For example, Figure 3 depicts a section on the use of *foaf* in a dataset.

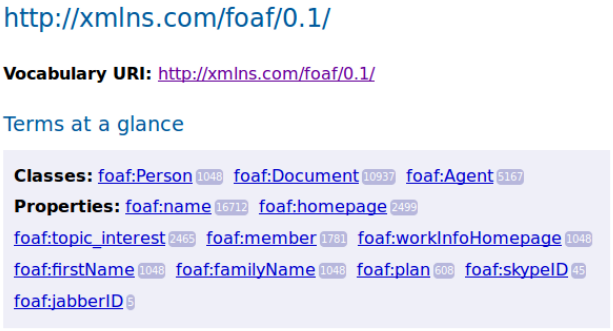
The vocabulary section describes the use of a certain vocabulary, and links to more specific documentation in class and property sections of the documentation (see Figure 4 for an example). The class section shows the number of entities using a certain class as well as the properties used with the class. The list of properties shows the distinct and the total number of property usage, which is helpful in debugging the datasets.

The generator resolves the used prefixes, using the *prefix.cc*[13] service, and creates templates for the SPARQL queries. The definitions from the vocabularies are queried from the vocabulary endpoint, and each resource is also linked to the original vocabulary documentation.

A dataset may have multiple dataset documentations, depending on the changes made in the dataset. The provenance information about the dataset is de-

[12] `http://vocab.at?uri=http://vocab.at/info`

[13] `http://prefix.cc`



http://xmlns.com/foaf/0.1/

Vocabulary URI: http://xmlns.com/foaf/0.1/

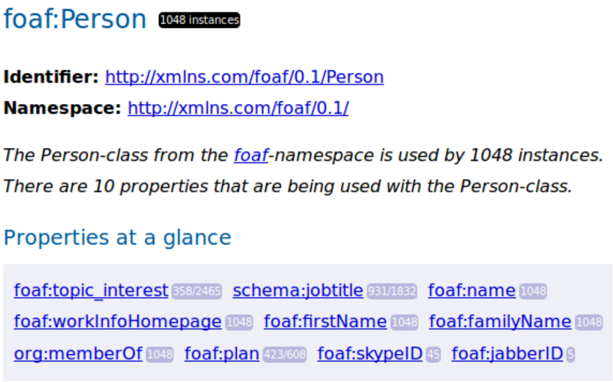
Terms at a glance

Classes: foaf:Person 1048 foaf:Document 10937 foaf:Agent 5167
Properties: foaf:name 16712 foaf:homepage 2499 foaf:topic_interest 2465 foaf:member 1781 foaf:workInfoHomepage 1048 foaf:firstName 1048 foaf:familyName 1048 foaf:plan 608 foaf:skypeID 45 foaf:jabberID 5

Figure 3: Example of the vocabulary section

scribed in the history section, as illustrated in Figure 5. The section lists the different versions of the documentations and changes in the vocabulary usage. The assessment section presents the statistics about the dataset and shows the formula and score of the vocabulary usage in percents for the better understandability.

In the end, the quality issues found in the datasets are listed and grouped into 1) unresolved vocabularies, 2) undefined properties and classes, and 3) general issues. Dataset users may also share and send new issues based on the observations they have made on the documentation. The reported issues are linked to the problematic resources by selecting the resource identifier and reporting the issue description.



foaf:Person 1048 instances

Identifier: http://xmlns.com/foaf/0.1/Person
Namespace: http://xmlns.com/foaf/0.1/

The Person-class from the foaf-namespace is used by 1048 instances. There are 10 properties that are being used with the Person-class.

Properties at a glance

foaf:topic_interest 358/2465 schema:jobtitle 931/1832 foaf:name 1048 foaf:workInfoHomepage 1048 foaf:firstName 1048 foaf:familyName 1048 org:memberOf 1048 foaf:plan 423/608 foaf:skypeID 45 foaf:jabberID 5

Figure 4: Example of the class section

5 Case Study: Assessing Linked Data Publications

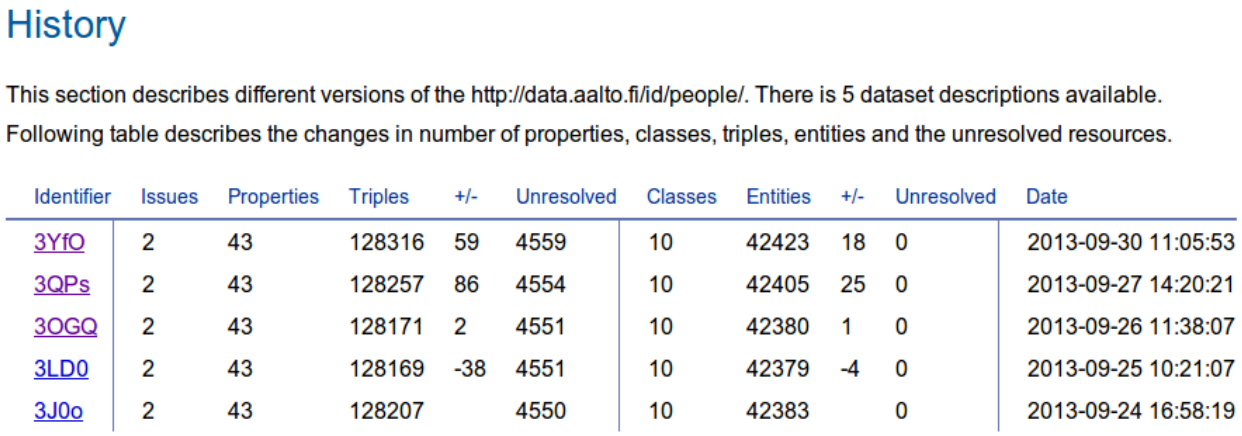
The *vocab.at* was evaluated by assessing a selection of presumably high quality datasets that have been peer reviewed: we evaluated LOD datasets whose descriptions were accepted for publication in a special issue of the Semantic Web journal[14]. Publishing dataset descriptions in a journal is a valuable asset to the research community as this promotes re-use of data, new research, and enables researchers to get credit from their data publishing work (Hitzler and Janowicz, 2013).

The assessment was made for the datasets which were directly accessible as data dumps or via a SPARQL endpoint. Only 8 of 20 datasets were accessible and automatically processable with *vocab.at* for various reasons. Some of the datasets were not available and VOID descriptions for these datasets did not provide access to the datasets. Many of the datasets were only available as packed data dumps containing random folder structure and multiple files. Some of the datasets were published only via SPARQL 1.0 endpoints which do not have required methods for counting the vocabulary usage.

Other datasets were accessible via REST-style interfaces providing access to only a singe resource at a time. Accessible datasets were selected based on the provided types of data. When the published dataset is divided multiple files, the default graph of the SPARQL endpoint was preferred, and if the data was published as one data dump, it was preferred. The results of the case study are presented in the Table 1, ordered by the Vocabulary Score of a dataset. The generated documentation and vocabulary assessment for the datasets can be found using the uri parameter *vocab.at?uri=dataset*.

The results from the case study reveal that most of the assessed datasets use properties and classes which are not defined or dereferenceable. The vocabulary score represents the dereferenceable term usage in the dataset. The number of issues is not directly affecting the vocabulary score, and even one issue can cause bad score depending on the use of the undefined

[14] http://www.semantic-web-journal.net/accepted-datasets



History

This section describes different versions of the http://data.aalto.fi/id/people/. There is 5 dataset descriptions available.
Following table describes the changes in number of properties, classes, triples, entities and the unresolved resources.

Identifier	Issues	Properties	Triples	+/-	Unresolved	Classes	Entities	+/-	Unresolved	Date
3YfO	2	43	128316	59	4559	10	42423	18	0	2013-09-30 11:05:53
3QPs	2	43	128257	86	4554	10	42405	25	0	2013-09-27 14:20:21
3OGQ	2	43	128171	2	4551	10	42380	1	0	2013-09-26 11:38:07
3LD0	2	43	128169	-38	4551	10	42379	-4	0	2013-09-25 10:21:07
3J0o	2	43	128207		4550	10	42383		0	2013-09-24 16:58:19

Figure 5: Example of the history section

vocabulary term.

Only two of the datasets got full scores on the vocabulary assessment. Notable is that one of the two—EARTh Dataset—only used the SKOS vocabulary. The other full scoring[15] dataset—The Linked Brazilian Amazon Rainforest Data (LBARD) (Kauppinen et al., 2013)—used several different vocabularies.

Dataset	Score	Issues
EARTh Dataset[16]	1	0
Linked Brazilian Amazon[17]	1	0
LOD EUScreen Dataset[18]	0.96	3
AEMET Dataset[19]	0.556	1
OGOLOD Dataset[20]	0.269	40
Kirjasampo Dataset[21]	0.173	13
TourMISLOD Dataset[22]	0.17	2

Table 1: Dataset Vocabulary Assessment

However, LBARD was not perfect when *vocab.at* was run first times against it late September 2013. But since it is a work of one of the authors of this paper, it was possible to improve the data and vocabularies based on the vocabulary usage analysis. In this sense the improvement of LBARD is an evidence of a successful use of *vocab.at* for pointing out the issues to be corrected.

The core cause for the errors before the improvement was that LBARD has been extended and linked to other data over time for various statistical analysis and visual analytics studies, but with the cost of introduced issues. Most notable issues at the start of the improvement were: 1) the terms introduced among the data itself were not deferenceable—this was corrected by configuring the server to support content negotiation, 2) many terms were added to data over time, but the plan of adding them also to deferenceable vocabularies was not realized—this was corrected by describing the terms, 3) spelling issues were observed in either data descriptions or in vocabularies—these were corrected.

The third best scored dataset—LOD EUScreen Dataset—had 3 issues, all due to undefined properties. Other datasets were far behind and most of the vocabularies used were not dereferenceable. It is notable that even if there are very few issues—like 1 in the case of AMETET Dataset—the score can be affected substantially if the issue(s) are dominating in the data. A small amount issues is likely easier to

[15] `http://vocab.at/page/20mm`

[16] `http://linkeddata.ge.imati.cnr.it:8890/sparql`

[17] `http://spatial.linkedscience.org/sparql`

[18] `http://lod.euscreen.eu/sparql`

[19] `http://aemet.linkeddata.es/sparql`

[20] `http://cu.atlas.bio2rdf.org/sparql`

[21] `http://saha.kirjastot.fi/dumps/`

[22] `http://tourmislod.modul.ac.at/openrdf-sesame/repositories/tourmis`

solve than a large amount.

The insight from this evaluation can be summarized as follows. The main reason for low vocabulary scores is usually the lack of the publication of vocabularies in use. Ideally Linked Data Vocabularies would be published following the best practices (Berrueta and Phipps, 2008). However, often the vocabularies are published just along with the data, i.e. without the consideration of neither vocabulary re-use nor data re-use. In few cases, there were also spelling errors in the vocabulary terms, either in data, or in vocabularies themselves.

We argue that—regardless of any score—the automatically generated documentation about the vocabulary usage is a valuable asset for improving the quality of data in the dataset. This was evidenced by the case of correcting LBARD to reach the maximum score 1 with the help of the documentation generated by *vocab.at*. We wish and expect that the community will similarly take the service into use to support improvement of the published other Linked Data.

Indeed, the correct usage of vocabularies should be the next step after a RDF validation. Our argument is that a documentation of the vocabulary usage can support correcting of many issues in the dataset. The most alarming scenario happens when the dataset is not accessible for machine processing. We state that Linked Data publication should be done either via an accessible RDF data dump, via RDFa tagged pages or by providing a SPARQL endpoint.

6 Related Work

The ability to create automatic documentation from RDF has already been used for creating the specifications of vocabularies. For example, the vocabulary specification of FOAF[23] is generated by SpecGen[24]. There are also other tools for publication and documentation of Linked Data Vocabularies, such as the Live OWL Documentation Environment[25] (LODE) by Peroni et al. (2012), Neologism (Basca et al., 2008), ldodds/dowl[26], parrot[27], OWLDoc[28], OntologyBrowser[29]. These kind of tools are strictly used to document the vocabulary developed in RDF or OWL, and do not consider the issue of vocabulary usage. Another form of automatic documentation is used by Linked Data Frontends (LDF), such as Pubby by Cyganiak and Bizer (2013) and URIBurner[30]. However, LDFs are not designed to give the user an overall view of the whole dataset, and focus on describing one resource at a time.

There is also research on Linked Data assessment (Auer et al., 2012; Langegger and Woss, 2009) that focuses on collecting the statistics of datasets and creating dataset descriptions. The dataset descriptions and the VoiD-vocabulary (Alexander et al., 2009) have also been used by Toupikov et al. (2009) for the rating of the datasets. Dataset descriptions by Lodstats[31] that make use of the statistics are extremely helpful when assessing the quality of a datasets.

Research on tracing the history of a dataset by extending the VoID vocabulary has been reported by Omitola et al. (2011). The Provenance vocabulary (PROV) of W3C (Gil and Miles, 2013) was just published as a recommendation. These contributions inspired our novel solution of describing the vocabulary usage in datasets, by using a provenance model and dataset descriptions for the automatic documentation of the dataset.

7 Conclusions

The vocabulary usage analysis and the Linked Data documentation are a novel way of assessing the usability of Linked Data. The automatically generated documentation from the data specifies the vocabulary usage explicitly and leaves no doubts about the used vocabulary terms. The presented approach improves the comprehensibility of Linked Data, which is essential for quality improvement and re-use of the data.

[23] `http://xmlns.com/foaf/spec/`

[24] `http://sioc-project.org/specgen`

[25] `http://www.essepuntato.it/lode`

[26] `https://github.com/ldodds/dowl`

[27] `http://ontorule-project.eu/parrot/parrot`

[28] `http://code.google.com/p/co-ode-owl-plugins/wiki/OWLDoc`

[29] `http://code.google.com/p/ontology-browser/`

[30] `http://linkeddata.uriburner.com/`

[31] `http://stats.lod2.eu/`

The vocabulary usage analysis and the generation of dataset documentation works also as an indicator for the accessibility of the dataset. Our case study showed that the accessibility of the published datasets can be a big issue even in highly ranked Linked Data publications. In our mind, best practices for the publication of Linked Data are too vague, which leads to using divergent approaches in data publication. We need a clearer set of guidelines for the publication format of the data dumps and recommendations to use SPARQL 1.1 endpoints that are accessible to the outside world.

In some cases—no matter of the tehniques used—documenting the vocabulary usage is doomed to fail. For example, DBpedia[32] misuses class and property definitions by generating additional semantics to the vocabulary terms, such as *dbyago:PeopleFromHelsinki*[33]. The impact of the additional semantics in the class definitions is realized as 63,553,605 classes that are impossible to document and comprehend. However, the problems with the massive vocabularies can be partly avoided by documenting the vocabulary usage of individual resources from dereferenceable URIs or limiting the documentation to certain classes.

The presented method for creating the vocabulary usage analysis and dataset documentation is generic and reproducible for other Linked Data implementations. The automatic documentation of vocabulary usage is a sensible approach to improve the comprehensibility of any Linked Data publication. The *vocab.at* service provides a preliminary solution for documenting Linked Data and is usable in most of the cases. Dataset publishers can also apply the presented method for creating dataset documentation within their own Linked Data Platforms; vocab.at is currently in use in this way in the Linked Data Finland service[34].

The *vocab.at* focuses currently only in the vocabulary usage analysis and ranks the dataset with the vocabulary score based on the dereferenceability of the vocabularies. The automatic documentation of Linked Data could also be extended for other cases, like the assessment of the incoming and outgoing links in the dataset. We also aim to improve the generation of the statistics and compare the implementation with other tools, such as LODstats. As for next steps in the future, we are planning to create a method for analysing the provenance from the different versions of the dataset descriptions. Detailed view to the history of the vocabulary term usage would improve the understanding in different versions of the data, and help in debugging the data flow from the data sources. The documentation could also inform the user about potentially related and useful datasets, based on the analysis of other dataset descriptions.

[32] `http://dbpedia.org/sparql`

[33] `http://dbpedia.org/class/yago/PeopleFromHelsinki`

[34] `http://ldf.fi/`

References

Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. (2009). Describing Linked Datasets. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*.

Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. (2011). Describing Linked Datasets with the VoID Vocabulary. Technical Report http://www.w3.org/TR/void/, W3C.

Assaf, A. and Senart, A. (2012). Data Quality Principles in the Semantic Web. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 226–229. IEEE.

Auer, S., Demter, J., Martin, M., and Lehmann, J. (2012). LODStats–an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management*, pages 353–362. Springer.

Basca, C., Corlosquet, S., Cyganiak, R., Fernández, S., and Schandl, T. (2008). Neologism: Easy Vocabulary Publishing. In *4th Workshop on Scripting for the Semantic Web*.

Berrueta, D. and Phipps, J. (2008). Best Practice Recipes for Publishing RDF Vocabularies. Technical Report http://www.w3.org/TR/swbp-vocab-pub/, W3C.

Carroll, J. (2002). Matching rdf graphs. In *The Semantic Web—ISWC 2002*, pages 5–15. Springer.

Cyganiak, R. and Bizer, C. (2013). Pubby - A Linked Data Frontend for SPARQL Endpoints. Online at http://wifo5-03.informatik.uni-mannheim.de/pubby/.

Cyganiak, R. and Wood, D. (2013). RDF 1.1 Concepts and Abstract Syntax. Technical Report http://www.w3.org/TR/rdf11-concepts/, W3C.

Flemming, A. and Hartig, O. (2010). Quality Criteria for Linked Data Sources. Online at http://bit.ly/ld-quality.

Fürber, C. and Hepp, M. (2010). Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In *Business Information Systems*, pages 35–46. Springer.

Fürber, C. and Hepp, M. (2011). SWIQA–A Semantic Web Information Quality Assessment Framework. In *Proceedings of the 19th European Conference on Information Systems (ECIS 2011)*.

Gil, Y. and Miles, S. (2013). PROV Model Primer. Technical Report http://www.w3.org/TR/prov-primer/, W3C.

Hartig, O. and Zhao, J. (2009). Using Web Data Provenance for Quality Assessment. In *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009)*, volume 526.

Heath, T. and Bizer, C. (2011). Linked data: Evolving The Web Into a Global Data Space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Hitzler, P. (2012). What's Wrong with Linked Data? Online at http://blog.semantic-web.at/2012/08/09/whats-wrong-with-linked-data/.

Hitzler, P. and Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web, Volume 4, Number 3*, pages 233–235.

Hogan, A., Harth, A., Passant, A., Decker, S., and Polleres, A. (2010). Weaving the Pedantic Web. In *Linked Data on the Web Workshop (LDOW2010) at WWW'2010*.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., and Decker, S. (2012). An Empirical Survey of Linked Data Conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:14–44.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., and Sheth, A. P. (2010). Linked Data Is Merely More Data. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*.

Juran, J. M., Godfrey, A. B., Hoogstoel, R. E., and Schilling, E. G. (1999). *Juran's Quality Handbook*, volume 2. McGraw Hill New York.

Kauppinen, T., de Espindola, G. M., Jones, J., Sánchez, A., Gräler, B., and Bartoschek, T. (2013). Linked Brazilian Amazon Rainforest Data. *Semantic Web Journal*. in press.

Langegger, A. and Woss, W. (2009). RDFStats-an Extensible RDF Statistics Generator and Library. In *Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on*, pages 79–83. IEEE.

Mendes, P. N., Mühleisen, H., and Bizer, C. (2012). Sieve: Linked Data Quality Assessment and Fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM.

Ogbuji, C. (2013). SPARQL 1.1 Graph Store HTTP Protocol. Technical Report http://www.w3.org/TR/sparql11-http-rdf-update/, W3C.

Omitola, T., Zuo, L., Gutteridge, C., Millard, I. C., Glaser, H., Gibbins, N., and Shadbolt, N. (2011). Tracing the Provenance of Linked Data Using voiD. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 17. ACM.

Peroni, S., Shotton, D., and Vitali, F. (2012). The Live OWL Documentation Environment: a Tool For the Automatic Generation of Ontology Documentation. In *Knowledge Engineering and Knowledge Management*, pages 398–412. Springer.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, 45(4):211–218.

Redman, T. C. and Blanton, A. (1997). *Data Quality For the Information Age*. Artech House, Inc.

Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., and Tummarello, G. (2009). DING! Dataset Ranking Using Formal Descriptions. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*.

Wang, R. Y., Strong, D. M., and Guarascio, L. M. (1996). Beyond accuracy: What Data Quality Means to Data Consumers. *J. of Management Information Systems*, 12(4):5–33.