Semantic Web 9 (2018) 207–220
DOI 10.3233/SW-180290
IOS Press

Editorial

On the quality of vocabularies for linked dataset papers published in the Semantic Web journal

Stella Sam [a], Pascal Hitzler [a,*] and Krzysztof Janowicz [b]
[a] *DaSe Lab, Wright State University, Dayton, OH, USA*
[b] *STKO Lab, University of California, Santa Barbara, USA*

1. Introduction

Linked Data and knowledge graphs more generally[1] have been a main driver of research and technology transfer in our field. Linked data serves as a testbed for Semantic Web technologies, as an outreach to application communities interested in data sharing, and contributes more generally to the big data effort by making a large variety of structured data available for all kinds of purposes in various communities [24]. As of December 2017, the LOD Laundromat aggregator [38] alone shows over 38,000,000,000 Linked Data triples which is only a fraction of the data that has been published as Linked Data to date.

Much of Linked Data is generated by researchers as a contribution to the community effort. Due to the importance of this type of data for the advance of research and applications in our field, the Semantic Web journal began in 2012 to solicit papers containing linked dataset descriptions.[2] The idea behind this was to provide the broader researchers and practitioners community with concise summaries of the data, the different tools and endpoints that have been made available to work with the data, worked examples, best practice, and so on, to foster the usage of Linked Data beyond the Semantic Web community. In addition, dataset descriptions allow authors to record their efforts in a citeable paper publication, since this type of credit is still much more important than alternative measures of productivity and impact which take the reuse of other artefacts such as datasets and software into account. The initiative quickly became very popular,[3] and the first linked dataset papers then appeared in 2013 in the journal [24].[4]

More recently, the issue of Linked Data *quality* came into focus [55]. This followed on the heels of application interest and the rise of questions regarding issues such as trustworthiness of content, reliability of resources, or ease of reuse. Indeed in the last few years a significant number of publications have laid out possible quality measures and corresponding algorithms for quality assessment. Truth be told, though, the final verdict is still outstanding regarding the question what dimensions or measures are in fact the most relevant for assessing Linked Data quality.

[*] Corresponding author. E-mail: pascal.hitzler@wright.edu.

[1] There are some notable differences, such as that Linked Data is usually expressed in RDF, perhaps with an OWL ontology, is supposed to be available on the Web, and "linked" with other datasets. Knowledge graphs more broadly speaking would not necessarily conform with any particular syntax, may not be linked to other external sources, and may not even be (easily) accessible on the Web. Nonetheless, many of these differences relate to technologies and not fundamentally different paradigms.

[2] http://www.semantic-web-journal.net/blog/semantic-web-journal-special-call-linked-dataset-descriptions

[3] See http://www.semantic-web-journal.net/accepted-datasets for a list of all accepted and/or published SWJ dataset papers.

[4] It should be noted that we have in the meantime increased the bar for such papers, i.e., we have become more selective, in particular regarding tangible evidence for usefulness of the presented datasets. The reason for this is that both community and technology have advanced, and the journal needs to keep up with such developments.

One of the aspects which has so far not received sufficient attention, is the question of relevance of a quality schema or vocabulary for the quality of a linked dataset. If it is indeed relevant, as has been argued [39], then this in turn raises the question how to assess the quality of such schema or vocabulary, in particular since such an evaluation may not be identical to a quality assessment of ontologies independent of a Linked Data context.

To this end, some of us proposed a relatively simple schema – the *5-star LD vocabulary principles*, or *5-star principles* in short – for assessing Linked Data vocabulary quality, in an editorial published in this journal a few years ago [29], as an early contribution to the topic. A majority of the linked dataset papers which the Semantic Web journal has published so far had already been published or submitted, and we did not require adherence to the ideas in that editorial for future submissions, though since the publication of the editorial we have encouraged authors to consider them.

So far, however, there hasn't been any coordinated assessment of the 5-star LD vocabulary principles and how they pan out for real existing datasets. Therefore, in this work, we will look at all linked dataset papers from the perspective of the 5-star vocabulary principles. This serves both as a partial assessment of the quality of linked dataset papers in the Semantic Web journal, and as assessment of the 5-star principles and their applicability in practice.[5]

The paper is structured as follows. Section 2 recalls the 5-star principles and discusses ambiguities and other issues arising when attempting to apply them, as well as explanations on how we resolved these for our present analysis. Section 3 presents the data we have collected, for all linked dataset papers which have so far been published in the Semantic Web journal. Section 4 contains a discussion of the data and what conclusions we can draw from this as we go forward.

2. The 5-star linked data vocabulary quality categorization and how to apply it

2.1. Background and motivation

In 2010, Tim Berners-Lee published a non-technical schema [5] of awarding anywhere between 0 to 5 stars in rating (Linked) Data depending on how easy it was to discover, use and understand it. However, the 5 Star Linked Data appears to be just a necessary precondition to what is really needed and does not necessarily make the resulting Linked Data more reusable to humans or machines. It does not make any assumptions about the use of vocabularies either. In practice, querying Linked Data that do not refer to a vocabulary is difficult and ascertaining that the results reflect the intended query is impossible. A good vocabulary must restrict potential interpretations of the used classes and roles towards their intended meaning.

In March of 2014, the *5-star LD vocabulary principles* [29] were introduced, to encourage data owners, engineers and practitioners to publish and use vocabularies on the Web. Similar to Tim Berners-Lee's stars, which do not directly refer to the quality of the data, this star rating is also not directly concerned with the quality of the vocabularies themselves. Instead, it rates the vocabulary *use* of a linked dataset following the intuition that data that utilizes well-established, documented, maintained, and interconnected vocabularies is easier to (re)use than Linked Data that may be 5-star data in Berners-Lee's sense but does not utilize such vocabularies.

2.2. The model

According to the model, a linked dataset would also be awarded anywhere between 0 to 5 stars as follows:

0 Stars – No web-accessible information on the vocabulary used in the generation of the dataset is available.

1 Star – There exists web-accessible, de-referenceable, human readable information about the vocabulary used.

2 Stars – It is a 1 Star dataset. Also information on the vocabulary used, exists in a machine readable form with explicit axiomatization of it. Ideally RDFS, OWL, RIF or related W3C standards on the Semantic Web stack can be used. This is the level in which automatic reasoning can come into play.

3 Stars – It is a 2 Star dataset. The related vocabulary is linked to other vocabularies. The vocabulary level links must be between classes and properties and not just between individuals. Explicit alignments (e.g. via subClassOf and equivalentClass axioms) are considered better than direct reuse of external vocabularies, though both are acceptable.

4 Stars – It is a 3 Star dataset. Metadata about the related vocabulary such as license model, contact person, last modification date, the ontology used, knowl-

[5] In a sense, this paper also complements the earlier editorial [26] in which other quality aspects were studied.

edge management methodology, etc. used to arrive at the vocabulary is available in a de-referenceable, machine readable format.

5 Stars – It is a 4 Star dataset. Its related vocabulary is linked to by other vocabularies. It is this star that reflects on the external usage and thus usefulness of the vocabulary and thereby ultimately the dataset using it.

We followed this model to manually rate the linked datasets accepted by, described and published in the Semantic Web journal.

2.3. *The rating*

The following steps were followed to rate the datasets:

1. If a vocabulary was developed and used for the generation of the dataset, with the clear presence of a web-accessible, de-referenceable and in human readable form, description of the vocabulary, the 1st star was awarded to the dataset.
2. If the vocabulary was well-defined with all of the various axioms that comprised it, the 2nd star was awarded.
3. The 3rd star was awarded, if other standard, well-established vocabularies were re-used in a supplemental nature along with (linked to by) this vocabulary, through the use of properties such as subClassOf, subPropertyOf and unionOf. It showed that this vocabulary defines only the required classes and properties that do not already exist in some other standard, well-established vocabulary. If the classes and properties do exist in some other vocabulary, they were re-used by establishing links at the vocabulary level and not at the individual level.
4. The 4th star was awarded, if there was evidence of availability of web-accessible metadata regarding the vocabulary. This ensured that the metadata was available in a standard format such as VoID, VOAF, etc.
5. If it has been clearly shown, that the developed vocabulary has been linked to by other vocabularies to be used in a manner consistent with point 3 above, then the 5th star was awarded.

A total of 49 datasets have been published so far in the Semantic Web journal since 2012 and our analysis, given below, shows that most of them are at least at the 4-star level.

2.4. *Ambiguities & issues*

The following ambiguities and issues came to light while applying the 5-star principles to the practical reality of the ways vocabularies were used in the generation of datasets.

- For several datasets, the reference links mentioned in the related description paper, do not work and there is no mention on how to find the most recent, updated ones. However, there was ample evidence in the description paper that necessary, web-accessible information regarding the vocabulary used, was present, which resulted in the dataset earning the star rating that it did.
- For the generation of some datasets, no new vocabulary was developed. However, other standard, already well-established vocabularies were reused. This was sufficient for the dataset to earn 4 stars. If the reused vocabulary was, in turn, reused in the generation of a different dataset, the dataset automatically earned 5 stars.
- In some cases, there was no need for the developed vocabulary to be supplemented by other standard, well-established ones, and hence was not linked to other vocabularies, the dataset thus not earning the 3rd star. However, the metadata regarding the vocabulary was present, the dataset thus earning the 4th star without having earned the 3rd star. This was resolved by awarding the dataset 3 stars.

3. Vocabulary quality for SWJ linked datasets

Table 1 presents the data from our analysis for the 49 papers and the star rating of vocabulary use.

4. Discussion

At this stage, our collected data is mainly descriptive. The editorial introducing the 5-star vocabulary categories was about the presentation of a position on the importance of Linked Data vocabularies and a anticipation on what aspects may be indicative of quality while at the same time providing simple guidelines and rules. As such, it was meant as a starting point for discussions, rather than as an end in itself. And as we write this new editorial, the discussion regarding the question what aspects of linked datasets are actually important for quality, is still ongoing.

Table 1
Linked datasets analysis

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
A Curated and Evolving Linguistic Linked Dataset [6]	2012	ASit Vocabulary	Geonames Ontology		4	
Amsterdam Museum Linked Open Data [14]	2012	Extends EDM, OAI-ORE model	EDM, OAI-ORE, Thesaurus to AATNed, DBPedia, ULAN,Thesaurus to GeoNames, RDA		4	
BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF [41]	2012	BioPortal Metadata Ontology extends OMV	SKOS, OMV		4	BioPortal Metadata Ontology extends OMV – Ontology Metadata Vocabulary – to track a set of metadata related to each ontology in the system
datos.bne.es: a Library Linked Data Dataset [49]	2012	Modelled using well established vocabularies: RDA, Dublin Core Metadata Element Set, BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF & IFLA approved ontologies: FRBR, FRAD, FRSAD, ISBD	VIAF, GND18, DBpedia, Libris, SUDOC, Lexvo		4	
Europeana Linked Open Data – data.europeana.eu [27]	2012	EDM	Aligned to: Dublin Core, SKOS, OAI-ORE, CIDOC-CRM, GEMET Thesaurus, Semium ontology	The Amsterdam Museum Prototype uses EDM	5	Europeana Semantic Elements (ESE) XML Schema to EDM is done manually
Facebook Linked Data via the Graph API [51]	2012	Extends RDFS, OWL	RDFS, OWL		4	
Fiction Literature as Linked Open Data - the BookSampo Dataset [33]	2012	Aligned with the ontologies of CultureSampo, KOKO a lightweight ontology comprising 14 domain ontologies joined under Finnish National Upper Ontology YSO	KOKO lightweight ontology comprising 14 domain ontologies joined under Finnish National Upper Ontology YSO, LEXVO language ontology		4	Metadata was transformed to RDF format from legacy libraries and then manually enriched by dozens of Finnish Librarians using SAHA a metadata editor

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
Linked Brazilian Amazon Rainforest Data [30]	2012	OLA, TISC as well as other well-established vocabularies	TISC Vocabulary		4	
Linked data for Potential Algal Biomass Production [45]	2012	WGS84 ontology, spatial relations ontology, the Geonames ontology and the NeoGeo ontology, extension of Neo Geo Geometry Ontology, QUDT Ontology, NUTS Vocabulary	All vocabularies used are established ones. The LEAPS datasets are interlinked with each other through their respective vocabularies		4	
Linked Nomenclature of Territorial Units for Statistics [12]	2012	Extends OS (Ordinance Survey) Ontology, OWL Time Ontology	OS Ontology, OWL Time Ontology		4	
The AGROVOC Linked Dataset [9]	2012	AGROVOC (SKOS-XL)	Linked to 13 Vocabularies, Thesauri & Ontologies: The Library of Congress Subject Headings (LCSH), RAMEAU Répertoire d'autoritématière encyclopedique et alphabetique unifie, EUROVOC, DBpedia, experimental Linked Data version of the Dewey Decimal Classification, NAL Thesaurus for agriculture, GEMETfor environment, STW for Economics, TheSoz for social science, both GeoNames and FAO Geopolitical Ontology cover countries and political regions, ASFA17 covers aquatic science & Biotechnology glossary covers biotechnology	Aligned to the following Vocabularies: ASFA, Biotechnology Glossary (FAO), Chinese Agriculture Thesaurus (CAT), DBPedia, Dewey Decimal Classification (DDC), EUROVOC, GEMET, GeoNames, Geopolitical Ontology, Library of Congress Subject Headings (LCSH), NAL Thesaurus, RAMEAU Répertoire d'autoritématière encyclopedique et alphabetique unifie, STW – Thesaurus for Economics, TheSoz – Thesaurus for Social Sciences, SWD (Schlagwortnormdatei), EARTh	5	SKOS-XL (SKOS Extension for Labels) has been used for the EUROVOC Thesaurus

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences [53]	2012	SKOS, SKOS-XL	Creative Commons, Provenance	VoiD	5	SKOS-XL (SKOS Extension for Labels) has been used for the EUROVOC Thesaurus
TourMISLOD: a Tourism Linked Data Set [40]	2012	TOURMIS ontology – Schema based on dbpedia.com, xsd & text	DBPedia,GeoNames		4	Closed license because of the heterogeneous nature of its contributors
Transforming Meteorological Data into Linked Data [3]	2012	AEMET ontology (an extension of W3C SSN – Semantic Sensor Network ontology) comprises: 1. Measurement Ontology – Mainly reuse from SSN. 2. Location Ontology – Mainly reuse from wgs84_pos ontology; also reuse from GeoBuddies ontology. 3. Time Ontology – Reuse from OWL Time ontology. 4. Sensor Ontology – Extension of SSN ontology	GeoLinkedDataset, DBPedia		4	
Translational research combining orthologous genes and human diseases with the OGOLOD dataset [35]	2012	OGO Ontology	GO, ECO, NCBI taxonomy, RO, HPO		4	
Converting neXtProt into Linked Data and nanopublications [11]	2013	PTM nanopublication	BioPAX ontology, Semantic science Integrated Ontology (SIO), Weighted Interests Vocabulary		4	Many of the reference links provided do not work
eagle-i: biomedical research resource datasets [46]	2013	ERO (Eagle-I Resource Ontology)	Gene Ontology, DBPedia through OWL sameAs	Has been generalized and re-used in Reagent Ontology (ReO) and the Agent, Resource & Grant ontology	5	

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
EARTh: an Environmental Application Reference Thesaurus in the Linked Open Data Cloud [1]	2013	SKOS – extends GEMET	Interlinks to: AGROVOC, EUROVOC, DBPEDIA,UMTHES Thesauri, each of them having their own vocabulary	NatureSDIplus & eENVplus employs EARTh as a backbone thesaurus	5	
Geospatial Dataset Curation through a Location-based Game [10]	2013	Urbanopoly Ontology, Human Computation Ontology. Also linked to ontologies: OpenStreetMap, LinkedGeoData	Human Computation Ontology extends the PROV-O ontology		4	Any Human Computation effort can be modelled using the Human Computation ontology
Increasing the Financial Transparency of European Commission Project Funding [34]	2013	FTS Ontology	Sub-Property links to DBPedia (close connection to the DBPedia Vocabulary)		4	
Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud [15]	2013	Lexvo Ontology	GEneral Multilingual Environmental Thesaurus (GEMET), the United Nations FAO AGROVOC thesaurus, the US National Agricultural Library Thesaurus, EuroVoc, RAMEAU subject headings. Links to upper ontologies such as OpenCyc are provided as well		4	
Linked European Television Heritage [44]	2013	EBU Core ontology – RDF representation of the EBU Class Conceptual Data Model (CCDM), MetaData Model: EU Screen Harvesting Schema	W3C Media Annotation Ontology (W3C MAWG), EDM for provenance related information		4	
Linked SDMX Data [8]	2013	RDF Data Cube vocabulary, SKOS, XKOS, PROV-O	RDF Data Cube vocabulary, SKOS, XKOS, PROV-O		4	

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
Migrating Bibliographic Datasets to the Semantic Web: the AGRIS case [2]	2013	BIBO, FOAF, Dublin Core	AGROVOC, RDF/SKOS – XL Concept Scheme aligned to 16 other vocabularies:ASFA, FAO Biotechnology Glossary, Chinese Agriculture Thesaurus (CAT), EARTh, Eurovoc, GEMET, Library of Congress Subject Headings (LCSH), NAL, RAMEAU, STW – Thesaurus for Economics, TheSoz – Thesaurus for the Social Sciences, Geopolical Ontology, Dewey Decimal Classification (DDC), DBPedia, SWD (Schlagwortnormdatei), GeoNames		4	
Public spending as LOD: the case of Greece [47]	2013	PS Ontology	DBPedia, Geonames		4	
Publishing and Interlinking the Global Health Observatory Dataset [54]	2013	RDF Data Cube			3	Vocabulary has meta data but is not linked to any other vocabulary
Semantic Quran a Multilingual Resource for Natural-Language Processing [43]	2013	The Quran Schema Vocabulary (qvoc)	Extends GOLD linguistic ontology & OLIA Arabic linguistic ontology		4	
Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with lemon [23]	2014	LEMON,SIMPLE-OWL ontology			5	Vocabulary has meta data but is not linked to any other vocabulary. Base Vocabulary used makes it a 5-star dataset
Countering language attrition with PanLex and the Web of Data [52]	2014	PanLex RDF vocabulary	LexVo, Dbpedia, GOLD, LEMON		4	

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF [42]	2014	Dbnary Ontology	Based on the LEMON ontology		4	Classes & properties are added on to LEMON when required
EventMedia: a LOD Dataset of Events Illustrated with Media [31]	2014	LODE ontology, Media Resources ontology	W3C Ontologies for Media Resources, SIOC,VCard		4	
lemonUby – a large, interlinked, syntactically-rich lexical resource for ontologies [19]	2014	UBY-LMF is mapped to Lemon called lemonUby model, Ontology lexica, OWL/DL ontology, UBYCat (to link to ISOCat)	UbyCat linked to Ontologies of Linguistic Annotations (Olia)		4	Through Olia, it is possible to interlink lemonUby with other LLOD linked to either Olia or any of the terminology repositories like GOLD & ISOCat
LinkedSpending: OpenSpending becomes Linked Open Data [25]	2014	RDF Data Cube Vocabulary, SDMX, XSD, DBPedia, LinkedGeoData	DBPedia, Linked GeoData SDMX		4	The cube models underlying both Linked GeoData SDMX as well as RDF Data Cube vocabularies are compatible
PAROLE/SIMPLE 'Lemon' ontology and lexicons [50]	2014	PAROLE/SIMPLE Model	WordNet, lemonUby, DBpedia		4	Mapped onto LEMON Model
CEDAR: The Dutch Historical Censuses as Linked Open Data [36]	2015	CEDAR	RDF Data Cube, Open Annotation, SDMX, PROV vocabularies		4	
Facilitating Scientometrics in Learning Analytics and Educational Data Mining – the LAK Dataset [16]	2015	BIBO, FOAF, SWRC, Schema.org, SWC	DBPedia is used as base vocabulary		4	
Publishing DisGeNET as Nanopublications [37]	2015	DisGeNET nanopublications	SIO, NcIt, PROV-O, Provenance, Authoring and Versioning (PAV) vocabulary, Provenance Vocabulary Core Ontology Specification (PRV), Weighted Interests vocabulary (WI), Evidence Codes Ontology (ECO)		4	

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
The Open University Linked Data – data.open.ac.uk [13]	2015	FOAF, SKOS, SIOC, OWL, Dublin Core, Bibo, XCRI	FOAF, SKOS, SIOC, OWL, Dublin Core, Bibo, XCRI		4	Reuses vocabularies to the most extent possible (57 vocabularies)
A Linked Data Wrapper for CrunchBase [21]	2016	CrunchBase Schema	Schema description in OWL & vocabulary description in VoiD		4	Many of the reference links provided do not work
DM2E: A Linked Data Source of Digitised Manuscripts for the Digital Humanities [4]	2016	DM2E model based on OAI-ORE, Dublin Core and SKOS	Reused vocabularies: BIBFRAME, BIBO, CIDOC-CRM, FABIO, PRO, rdaGr2, VIVO		4	
JRC-Names: Multilingual Entity Name variants and titles as Linked Data [20]	2016	JRC-Model	Reuses: Lemon, Lexinfo, OLiA, JRC links to vocabulary of NYT - New York Times (but loosely)		4	
Linked Web APIs Dataset: Web APIs meet Linked Data [18]	2016	Linked Web APIs Ontology, LWAPIS (RDF/OWL based)	Integrates: FOAF, Dublin Core, WSMO-Lite, MSM, PROV Ontology		4	
Meta-Data for a lot of LOD [38]	2016	LOD Laundromat based on: Bio2RDF, PROV, Dce, Dct, VANN, VOID, VOID-ext	PROV-O, HTTP Vocabulary, Error Ontology		4	
Migration of a library catalogue into RDA linked open data [7]	2016	RDA based on models: FRBR, FRAD, FRSAD	OWL Time ontology, FOAF, Dublin Core, Library of Congress MetaData Authority Description Schema, DBPedia-OWL ontology		4	
The ACORN-SAT Linked Climate Dataset [32]	2016	W3C RDF Data Cube Vocabulary, W3C Semantic Sensor Network ontology, acorn-sat observation ontology, acorn-series time series ontology, climate ontology, raindist rainfall district ontology	Vocabulary of Interlinked Datasets: RDF Data Cube Vocabulary, Simple Knowledge Organization System, Semantic Sensor Network ontology, GeoNames ontology (version 3.1), Basic Geo (WGS84 lat/long) vocabulary, Time Ontology in OWL, DOLCE+DnS Ultralite	External datasets already SSN Ontology based link to it	5	

Table 1

(Continued)

Linked dataset	Published date	Base vocabulary	Linked to vocabulary	Linked from vocabulary	Star rating	Comments
The Apertium Bilingual Dictionaries on the Web of Data [22]	2016	Lemon, Lexinfo, Lemon translation module	Lemon, lexinfo	Lemon along with lemon translation module serves as the basis for the lemon-ontolex model	5	
The debates of the European Parliament as Linked Open Data [48]	2016	Vocabulary for LinkedEP	Links to 4 external sources (2 on politicians' backgrounds, 1 geographical database and 1 on topic taxonomy): 1) LinkedEP to DBPedia - RDF, YAGO Ontology 2) LinkedEP MEP to the Italian Parliament - dati.camera.it (dublin core, foaf, BIO vocabularies) 3) LinkedEP Countries to GeoNames - GeoNames Ontology 4) LinkedEP Agenda Items to Eurovoc Thesaurus		4	
The Rijksmuseum Collection as Linked Data [17]	2016	EDM & RDA (Resource Description and Access Vocabulary), AAT (Arts	Architecture Thesaurus), IconClass, STCN (Short Title catalogue Netherlands) are used to extend it		4	Though the base model is EDM, others access it only through APIs
Wikidata through the Eyes of DBpedia [28]	2016	Own data model to better capture provenance information (not RDF based but using this as base, different RDF serializations are possible). DBPedia ontology and RDF reification	OWL punning used to define owl:equivalentClass to links between DBPedia classes and related Wikidata items		3	Part of the DBPedia publishing work-flow

Likewise, the point of publishing our analysis herein lies in stimulating further discussion.

Based on our analysis, a case may be made for interchanging the orders of the third and fourth stars of the 5-star model. The first 3 stars would then be a reflection on the completeness of information on the vocabulary itself, while the next 2 stars would be a reflection on the vocabulary's interactions with other vocabularies. Furthermore, this would avoid the occurrence of a dataset getting a fourth star without getting the third star, in some cases. Of course, if the metadata of the vocabulary were not defined, then we could still see this happening.

Our assessment shows that all linked datasets with corresponding papers in SWJ have between 3 and 5 stars for their vocabulary, with the overwhelming majority having 4 stars. If we accept the categories as a quality measure, then the datasets score high, but most do not score the top 5 stars. But we can at this stage only speculate as to the reasons for this. Due to the journal's depublishing strategy, we cannot easily track whether rejected dataset description papers were less than 4 stars on average.

We note that obtaining the fifth star is in fact not quite easy, because it actually requires third parties to acknowledge the independent value of the used vocabulary by establishing links. This means that the authors of a dataset cannot always actively improve their dataset in order to obtain the fifth star. The fifth star thus constitutes an indirect quality measure, by understanding a reuse by third parties as a type of *endorsement* of the vocabulary. Authors of a dataset can actively only obtain the fifth star by basing their dataset on a vocabulary that has already been established and is already linked to by others. But of course such a vocabulary may not exist yet in the topic area of the linked dataset being constructed. Note, that a 4 star dataset may become a 5 star dataset as time progresses and the utilized vocabulary is gaining impact.

Regarding the overall good quality of the vocabularies with respect to our rating, we of course acknowledge that the corresponding papers (and thus the underlying datasets) have undergone a rigorous review for the journal, and although a rating regarding our 5-star categories was not part of the review process, some of the categories – stars one through four – are arguably rather natural and would often be adhered to by authors concerned about a certain minimum quality of the artifact they create. And correspondingly, reviewers could be expected to look into such aspects.

This of course raises the open question where other datasets, i.e., datasets which do not have corresponding dataset description papers in the Semantic Web journal, fall with respect to the 5-star rating. If the profiles look very similar to the ratings described herein, i.e., they would mostly be four stars, then perhaps this could be taken as an argument that the 5-star rating is actually not very helpful in the sense that it is an already established practice. If the profiles look different, in particular if they include a significant body of ratings with fewer stars, then the hypothesis arises that the categories may indeed correlate with aspects of quality, although research will have to advance on the Linked Data quality front before this hypothesis can be investigated effectively.

More importantly, however, the fact that accepted dataset description papers typically clock in at 4 stars means that vocabularies and their quality are indeed considered important aspects of proper Linked Data publishing and that it may indeed be best practice to select vocabularies wisely and following certain criteria. The star rating proposed back in 2014 reminds us that in earlier years many Linked Data enthusiasts questioned the need for shared (and formal) vocabularies altogether and that this sentiment seems to be changing. We believe that proper vocabularies are a key driver for dataset discovery and reuse. Interestingly, the reuse of vocabularies themselves and best strategies for doing so (e.g., direct usage versus alignment) remains an open issue.

References

[1] R. Albertoni, M.D. Martino, S.D. Franco, V.D. Santis and P. Plini, EARTh: An environmental application reference thesaurus in the Linked Open Data cloud, *Semantic Web* **5**(2) (2014), 165–171. doi:10.3233/SW-130122.

[2] S. Anibaldi, Y. Jaques, F. Celli, A. Stellato and J. Keizer, Migrating bibliographic datasets to the Semantic Web: The AGRIS case, *Semantic Web* **6**(2) (2015), 113–120. doi:10.3233/SW-130128.

[3] G.A. Atemezing, Ò. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, D. Vila-Suero and B. Villazón-Terrazas, Transforming meteorological data into linked data, *Semantic Web* **4**(3) (2013), 285–290. doi:10.3233/SW-120089.

[4] K. Baierer, E. Dröge, K. Eckert, D. Goldfarb, J. Iwanowa, C. Morbidoni and D. Ritze, DM2E: A linked data source of digitised manuscripts for the digital humanities, *Semantic Web* **8**(5) (2017), 733–745. doi:10.3233/SW-160234.

[5] T. Berners-Lee, Linked data, 2006, https://www.w3.org/DesignIssues/LinkedData.html.

[6] E.D. Buccio, G.M.D. Nunzio and G. Silvello, A curated and evolving linguistic linked dataset, *Semantic Web* **4**(3) (2013), 265–270. doi:10.3233/SW-2012-0083.

[7] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, Migration of a library catalogue into RDA linked open data, *Semantic Web* (2018), To appear.

[8] S. Capadisli, S. Auer and A.-C.N. Ngomo, Linked SDMX data, *Semantic Web* **6**(2) (2015), 105–112. doi:10.3233/SW-130123.

[9] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques and J. Keizer, The AGROVOC linked dataset, *Semantic Web* **4**(3) (2013), 341–348. doi:10.3233/SW-130106.

[10] I. Celino, Geospatial dataset curation through a location-based game, *Semantic Web* **6**(2) (2015), 121–130. doi:10.3233/SW-130129.

[11] C. Chichester, O. Karch, P. Gaudet, L. Lane, B. Mons and A. Bairoch, Converting neXtProt into linked data and nanopublications, *Semantic Web* **6**(2) (2015), 147–153. doi:10.3233/SW-140149.

[12] G. Correndo and N. Shadbolt, Linked data representation of the nomenclature of territorial units for statistics, *Semantic Web* **4**(3) (2013), 251–256.

[13] E. Daga, M. d'Aquin, A. Adamou and S. Brown, The Open University linked data – data.open.ac.uk, *Semantic Web* **7**(2) (2016), 183–191. doi:10.3233/SW-150182.

[14] V. de Boer, J. Wielemaker, J. van Gent, M. Oosterbroek, M. Hildebrand, A. Isaac, J. van Ossenbruggen and G. Schreiber, Amsterdam Museum linked open data, *Semantic Web* **4**(3) (2013), 237–243. doi:10.3233/SW-2012-0074.

[15] G. de Melo, Lexvo.org: Language-related information for the linguistic linked data cloud, *Semantic Web* **6**(4) (2015), 393–400. doi:10.3233/SW-150171.

[16] S. Dietze, D. Taibi and M. d'Aquin, Facilitating scientometrics in learning analytics and educational data mining – The LAK dataset, *Semantic Web* **8**(3) (2017), 395–403. doi:10.3233/SW-150201.

[17] C. Dijkshoorn, L. Jongma, L. Aroyo, J. van Ossenbruggen, G. Schreiber, W. ter Weele and J. Wielemaker, The Rijksmuseum collection as linked data, *Semantic Web* **9**(2) (2018), In this issue.

[18] M. Dojchinovski and T. Vitvar, Linked web APIs dataset, *Semantic Web* (2018), To appear.

[19] J. Eckle-Kohler, J.P. McCrae and C. Chiarcos, LemonUby – A large, interlinked, syntactically-rich lexical resource for ontologies, *Semantic Web* **6**(4) (2015), 371–378. doi:10.3233/SW-140159.

[20] M. Ehrmann, G. Jacquet and R. Steinberger, JRC-Names: Multilingual entity name variants and titles as linked data, *Semantic Web* **8**(2) (2017), 283–295. doi:10.3233/SW-160228.

[21] M. Färber, C. Menne and A. Harth, A linked data wrapper for CrunchBase, *Semantic Web* (2018), To appear.

[22] J. Gracia, M. Villegas, A. Gómez-Pérez and N. Bel, The apertium bilingual dictionaries on the web of data, *Semantic Web* **9**(2) (2018), In this issue.

[23] R.D. Gratta, F. Frontini, F. Khan and M. Monachini, Converting the PAROLE SIMPLE CLIPS lexicon into RDF with lemon, *Semantic Web* **6**(4) (2015), 387–392. doi:10.3233/SW-140168.

[24] P. Hitzler and K. Janowicz, Linked data, big data, and the 4th paradigm, *Semantic Web* **4**(3) (2013), 233–235. doi:10.3233/SW-130117.

[25] K. Höffner, M. Martin and J. Lehmann, LinkedSpending: OpenSpending becomes linked open data, *Semantic Web* **7**(1) (2016), 95–104. doi:10.3233/SW-150172.

[26] A. Hogan, P. Hitzler and K. Janowicz, Linked dataset description papers at the semantic web journal: A critical assessment, *Semantic Web* **7**(2) (2016), 105–116. doi:10.3233/SW-160216.

[27] A. Isaac and B. Haslhofer, Europeana linked open data – data.europeana.eu, *Semantic Web* **4**(3) (2013), 291–297. doi:10.3233/SW-120092.

[28] A. Ismayilov, D. Kontokostas, S. Auer, J. Lehmann and S. Hellmann, Wikidata through the eyes of DBpedia, *Semantic Web* (2018), To appear.

[29] K. Janowicz, P. Hitzler, B. Adams, D. Kolas and C. Vardeman, Five stars of linked data vocabulary use, *Semantic Web* **5**(3) (2014), 173–176. doi:10.3233/SW-140135.

[30] T. Kauppinen, G.M. de Espindola, J. Jones, A. Sànchez, B. Gräler and T. Bartoschek, Linked Brazilian Amazon rainforest data, *Semantic Web* **5**(2) (2014), 151–155. doi:10.3233/SW-130113.

[31] H. Khrouf and R. Troncy, EventMedia: A LOD dataset of events illustrated with media, *Semantic Web* **7**(2) (2016), 193–199. doi:10.3233/SW-150184.

[32] L. Lefort, A. Haller, K. Taylor, G. Squire, P. Taylor, D. Percival and A. Woolf, The ACORN-SAT linked climate dataset, *Semantic Web* **8**(6) (2017), 959–967. doi:10.3233/SW-160241.

[33] E. Mäkelä, K. Hypén and E. Hyvönen, Fiction literature as linked open data – The BookSampo dataset, *Semantic Web* **4**(3) (2013), 299–306. doi:10.3233/SW-120093.

[34] M. Martin, C. Stadler, P. Frischmuth and J. Lehmann, Increasing the financial transparency of European Commission project funding, *Semantic Web* **5**(2) (2014), 157–164. doi:10.3233/SW-130116.

[35] J.A.M. narro Giménez, M.E. na Aranguren, B. Villazón-Terrazas and J.T. Fernández-Breis, Translational research combining orthologous genes and human diseases with the OGOLOD dataset, *Semantic Web* **5**(2) (2014), 145–149. doi:10.3233/SW-130109.

[36] A.M. no Peñuela, A. Ashkpour, C. Guéret and S. Schlobach, CEDAR: The Dutch historical censuses as linked open data, *Semantic Web* **8**(2) (2017), 297–310. doi:10.3233/SW-160233.

[37] N. Queralt-Rosinach, T. Kuhn, C. Chichester, M. Dumontier, F. Sanz and L.I. Furlong, Publishing DisGeNET as nanopublications, *Semantic Web* **7**(5) (2016), 519–528. doi:10.3233/SW-150189.

[38] L. Rietveld, W. Beek, R. Hoekstra and S. Schlobach, Metadata for a lot of LOD, *Semantic Web* **8**(6) (2017), 1067–1080. doi:10.3233/SW-170256.

[39] V. Rodriguez-Doncel, A.A. Krisnadhi, P. Hitzler, M. Cheatham, N. Karima and R. Amini, Pattern-based linked data publication: The linked chess dataset case, in: *Proceedings of the 6th International Workshop on Consuming Linked Data Co-Located with 14th International Semantic Web Conference (ISWC 2105)*, Bethlehem, Pennsylvania, US, October 12th, 2015, O. Hartig, J. Sequeda and A. Hogan, eds, CEUR Workshop Proceedings, Vol. 1426, 2015.

[40] M. Sabou, I. Arsal and A.M.P. Braşoveanu, TourMISLOD: A tourism linked data set, *Semantic Web* **4**(3) (2013), 271–276. doi:10.3233/SW-2012-0087.

[41] M. Salvadores, P.R. Alexander, M.A. Musen and N.F. Noy, Bioportal as a dataset of linked biomedical ontologies and terminologies in RDF, *Semantic Web* **4**(3) (2013), 277–284. doi:10.3233/SW-2012-0086.

[42] G. Sérasset, DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF, *Semantic Web* **6**(4) (2015), 355–361. doi:10.3233/SW-140147.

[43] M.A. Sherif and A.-C.N. Ngomo, Semantic Quran, *Semantic Web* **6**(4) (2015), 339–345. doi:10.3233/SW-140137.

[44] N. Simou, J.-P. Evain, N. Drosopoulos and V. Tzouvaras, Linked European television heritage, *Semantic Web* **4**(3) (2013), 323–329. doi:10.3233/SW-130104.

[45] M. Solanki, J. Skarka and C. Chapman, Linked data for potential algal biomass production, *Semantic Web* **4**(3) (2013), 331–340. doi:10.3233/SW-130105.

[46] C. Torniai, D. Bourges-Waldegg and S. Hoffmann, eagle-i: Biomedical research resource datasets, *Semantic Web* **6**(2) (2015), 139–146. doi:10.3233/SW-130133.

[47] M. Vafopoulos, M. Meimaris, I. Anagnostopoulos, A. Papantoniou, I. Xidias, G. Alexiou, G. Vafeiadis, M. Klonaras and V. Loumos, Public spending as LOD: The case of Greece, *Semantic Web* **6**(2) (2015), 155–164. doi:10.3233/SW-140155.

[48] A. van Aggelen, L. Hollink, M. Kemman, M. Kleppe and H. Beunders, The debates of the European Parliament as linked open data, *Semantic Web* **8**(2) (2017), 271–281. doi:10.3233/SW-160227.

[49] D. Vila-Suero, B. Villazón-Terrazas and A. Gómez-Pérez, datos.bne.es: A library linked data dataset, *Semantic Web* **4**(3) (2013), 307–313. doi:10.3233/SW-120094.

[50] M. Villegas and N. Bel, PAROLE/SIMPLE 'Lemon' ontology and lexicons, *Semantic Web* **6**(4) (2015), 363–369. doi:10.3233/SW-140148.

[51] J. Weaver and P. Tarjan, Facebook linked data via the graph API, *Semantic Web* **4**(3) (2013), 245–250. doi:10.3233/SW-2012-0078.

[52] P. Westphal, C. Stadler and J. Pool, Countering language attrition with PanLex and the web of data, *Semantic Web* **6**(4) (2015), 347–353. doi:10.3233/SW-140138.

[53] B. Zapilko, J. Schaible, P. Mayr and B. Mathiak, TheSoz: A SKOS representation of the thesaurus for the social sciences, *Semantic Web* **4**(3) (2013), 257–263. doi:10.3233/SW-2012-0081.

[54] A. Zaveri, J. Lehmann, S. Auer, M.M. Hassan, M.A. Sherif and M. Martin, Publishing and interlinking the Global Health Observatory dataset, *Semantic Web* **4**(3) (2013), 315–322. doi:10.3233/SW-130102.

[55] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for linked data: A survey, *Semantic Web* **7**(1) (2016), 63–93. doi:10.3233/SW-150175.