Adoption of the Linked Data Best Practices in Different Topical Domains

Max Schmachtenberg, Christian Bizer, and Heiko Paulheim

University of Mannheim
Research Group Data and Web Science, Germany
{max,chris,heiko}@informatik.uni-mannheim.de

Abstract. The central idea of Linked Data is that data publishers support applications in discovering and integrating data by complying to a set of best practices in the areas of linking, vocabulary usage, and metadata provision. In 2011, the *State of the LOD Cloud* report analyzed the adoption of these best practices by linked datasets within different topical domains. The report was based on information that was provided by the dataset publishers themselves via the *datahub.io* Linked Data catalog. In this paper, we revisit and update the findings of the 2011 *State of the LOD Cloud* report based on a crawl of the Web of Linked Data conducted in April 2014. We analyze how the adoption of the different best practices has changed and present an overview of the linkage relationships between datasets in the form of an updated LOD cloud diagram, this time not based on information from dataset providers, but on data that can actually be retrieved by a Linked Data crawler. Among others, we find that the number of linked datasets has approximately doubled between 2011 and 2014, that there is increased agreement on common vocabularies for describing certain types of entities, and that provenance and license metadata is still rarely provided by the data sources.

Keywords: Linked Open Data, Web of Linked Data, Best Practices.

1 Introduction

The Web of Linked Data [3,7] has grown from a dozen datasets in 2007 into a large data space containing hundreds of datasets today. In order to enable Linked Data applications to discover datasets as well as to ease the integration of data from multiple sources, Linked Data publishers should comply with a set of best practices [4]. These best practices can be grouped into three areas:

Linking: By setting RDF links, data providers connect their datasets into a single global data graph which can be navigated by applications and enables the discovery of additional data by following RDF links.

Vocabulary Usage: The best practices advise publishers to use terms from widely-used vocabularies in order to ease the interpretation of their data. If data providers use their own vocabularies, the terms of such proprietary vocabularies

P. Mika et al. (Eds.) ISWC 2014, Part I, LNCS 8796, pp. 245–260, 2014.

should be *dereferencable* into their RDF schema or OWL definitions. The definitions of proprietary vocabulary terms should contain RDF links pointing at terms from widely-used vocabularies in order to ease their interpretation.

Metadata Provision: Linked Data should be as self-descriptive as possible, and thus include metadata. An important form of metadata is *provenance* metadata describing the origin of datasets and enabling applications to assess their quality. The best practices also advise to provide *licensing* metadata and *dataset-level metadata*, e.g., in the form of a VoID file[1]. If datasets are accessible via additional access methods, such as a SPARQL endpoint or data dumps, then the VoID file should contain information about these access methods.

The adoption of the Linked Data best practices by datasets belonging to different topical domains was analyzed in the *State of the LOD Cloud* report [7] in 2011. The report is based on information provided by the data publishers themselves via the `datahub.io` Linked Data catalog[2]. In this paper, we revisit and update the findings of the *State of the LOD Cloud* report from 2011 based on a crawl of the Web of Linked Data conducted in April 2014. The paper is structured as follows: Section 2 describes the crawling strategy that was used to gather the data that forms the basis of our analysis. Section 3 explains the categorization of the data by topical domain. Sections 4, 5, and 6 discuss the adoption of best practices in the areas of linking, vocabulary usage, and metadata provision. Section 7 gives an overview of related work. The paper closes with a wrap-up of our findings.

2 Crawl of the Linked Data Web

To evaluate the conformance to the best practices, we have crawled a snapshot of the Linked Data Web. For this, we used *LDSpider*, a framework for crawling Linked Data [6]. We seeded LDSpider with 560 thousand seed URIs originating from three sources: 1. We included all URIs of example resources from datasets contained in the *lod-cloud* group in the `datahub.io` dataset catalog as well as example URIs from other datasets in the catalog marked with Linked Data related tags; 2. We included a sample of the URIs contained in the Billion Triple Challenge 2012 dataset[3]; 3. We collected URIs from datasets advertised on the `public-lod@w3.org` mailing list since 2011. With those seeds, we performed crawls during April 2014 to retrieve entities from every dataset using a breadth-first crawling strategy. Altogether, we crawled 900,129 documents describing 8,038,396 resources. The crawled data is provided for download on the website accompanying this paper[4] so that all results presented in the following can be verified.

For grouping the retrieved resources into datasets, we generally assume that all data originating from one pay-level domain (PLD) belongs to a single dataset.

[1] `http://www.w3.org/TR/void/`
[2] `http://datahub.io/group/lodcloud`
[3] `http://km.aifb.kit.edu/projects/btc-2012/`
[4] `http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/`

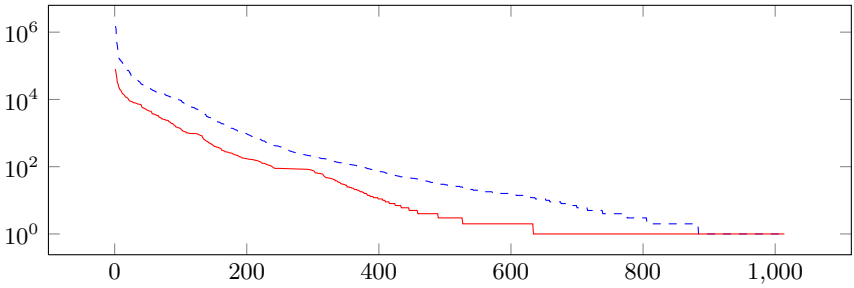

Fig. 1. Distribution of the number of resources (- - -) and documents (——) per dataset contained in the crawl (log scale)

If the *datahub.io* catalog lists multiple datasets for a single PLD, we apply an exception to the general rule and use the dataset definitions from the catalog. Altogether, the crawled data belongs to 1014 different datasets. Figure 1 shows the distribution of the number of resources and documents per dataset contained in the crawl.

Our crawler did respect crawling restrictions expressed by the data sources via `robots.txt` files. Altogether, we discovered 77 linked datasets which do not allow crawling and did not retrieve data from these sources.

3 Categorization by Topical Domain

Since the adoption of the Linked Data best practices might vary depending on the topical domain of the datasets, we classify the datasets into the following topical categories: *media*, *government*, *publications*, *life sciences*, *geographic*, *cross-domain*, *user-generated content*, and *social networking*. This categorization schema is the same as the one used by the 2011 *State of the LOD Cloud* report with the only difference that we added the category *social networking* as we discovered a large number of datasets providing data about people and their social ties. For datasets that are contained in the *datahub.io* dataset catalog, we adopt the topical categorization form the catalog. We manually assigned categories to the newly discovered datasets after inspecting them. In the following, we define the categories and refer to some prominent datasets from each category. Afterwards, we compare the overall number of datasets per category with the findings of the 2011 *State of the LOD Cloud* report.

The *media* category contains datasets providing information about films, music, TV and radio programmes, as well as print media. Prominent datasets within this category are the *dbtune.org* music datasets, the *New York Times* dataset, and the *BBC radio and television program* datasets. The *government* category contains Linked Data published by federal or local governments, including a lot of statistical datasets. Prominent examples include the `data.gov.uk` and `opendatacommunities.org` datasets. The category *publications* holds library datasets, information about scientific publications and conferences, reading

Table 1. Number of datasets in each category and growth compared to 2011

Category	Datasets 2014	Percentage	Datasets 2011	Growth
Media	24 (-2)	2%	25	-4%
Government	199 (-16)	18%	49	306%
Publications	138 (-42)	13%	87	59%
Geographic	27 (-6)	2%	31	-13%
Life Sciences	85 (-2)	8%	41	107%
Cross-domain	47 (-6)	4%	41	15%
User-generated Content	51 (-3)	5%	20	155%
Social Networking	520 (-0)	48%	-	-
Total	1091 (-77)		294	271%

lists from universities, and citation databases. Well known datasets include the German National Library dataset, the L3S DBLP dataset, and the Open Library dataset. The category *geographic* contains datasets like `geonames.org` and `linkedgeodata.org` comprising information about geographic entities, geopolitical divisions, and points of interest. The *life sciences* category comprises biological and biochemical information, drug-related data, and information about species and their habitats. The *cross-domain* category includes general knowledge bases such as DBpedia or UMBEL, linguistic resources such as WordNet or Lexvo, as well as product data. The the category *user-generated content* contains data from portals that collect content generated by larger user communities. Examples include metadata about blogposts published as Linked Data by `wordpress.com`, data about open source software projects published by `apache.org`, scientific workflows published by `myexperiment.org`, and reviews published by `goodreads.com` or `revyu.com`. The category *social networking* contains people profiles as well as data describing the social ties amongst people. We include into this category individual FOAF profiles, as well as data about the interconnections amongst users of the distributed microblogging platform StatusNet. The distinction between the categories *user-generated content* and *social networking* is that the datasets in the former category focus on the actual content while datasets in the later focus on user profiles and social ties. The 2011 *State of the LOD Cloud* report did not contain *social networking* as a separate category since the report did not count individual FOAF profiles as separate datasets and since StatusNet servers did not export Linked Data in 2011.

Table 1 gives an overview of the number of datasets in each category as well as the growth per category compared to the 2011 report. A list with the exact assignments of each dataset to a category is found on the accompanying website. The numbers in brackets in the second column refer to the number of datasets that do not allow crawling. The by far largest category is *social networking* with 520 datasets (48% of all datasets). The second largest category is *government* with 199 datasets (18%), followed by *publications* with 138 datasets (13%). Compared to the 2011 *State of the LOD Cloud* report, we observe a larger number of datasets in all categories except *geographic* and *media* data. The category

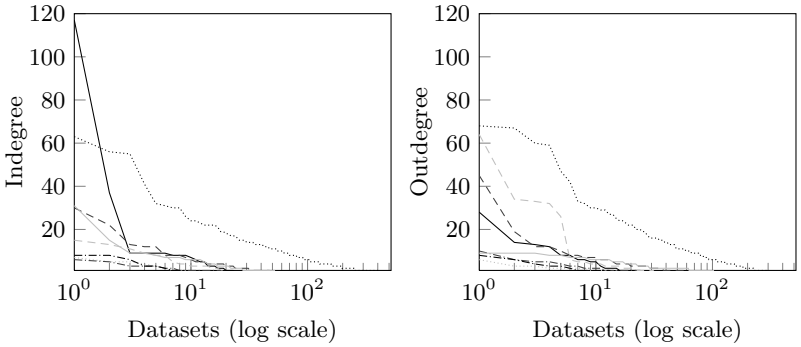

Fig. 2. Degree distributions for datesets belonging to the categories crossdomain(——), user-generated content(- - - -), social networking(·····), publications(- - - -), media(·····), life sciences(-·-·-·), government(——) and geographic(-·-·-·).

government shows the largest growth, followed by the categories *user-generated content* and *life sciences*. Excluding the new category *social networking*, the overall number of Linked Datasets has approximately doubled from 2011 (294 datasets) to 2014 (571 datasets). Including the new category, we observe an overall growth of 271% (from 294 to 1091 datasets).

4 Adoption of the Linking Best Practices

The linking best practice encourages publishers to set RDF links between datasets in order to enable the discovery of additional data and to support the integration of data from multiple sources. For analyzing the linkage between datasets, we aggregate all RDF links by dataset, meaning that we consider two datasets to be linked if there exists at least one RDF link between resources belonging to the datasets.

4.1 Degree Distributions

In total, 56% of all datasets in our crawl set RDF links pointing to at least one other dataset. The remaining 44% are either only the target of RDF links from other datasets or are isolated. Figure 2 shows the distribution of the in- and outdegrees for each category. We see that the in- and outdegrees vary widely with a small number of datasets in each category being highly linked, while the majority of the datasets is only sparsely linked. Overall, datasets from the category *social networking* show the highest degree values. The categories *cross-domain*, *user-generated content*, and *geographic* show an imbalance between in- and outdegrees, with *user-generated content* having larger out- than indegrees, and *cross-domain* and to a lesser extent *geographic* having larger in- than outdegrees (measured as area under the curve).

Table 2. Datasets with the highest in- and outdegrees

Dataset	Category	In	Dataset	Category	Out
dbpedia.org	cross-domain	207	bibsonomy.org	publications	91
geonames.org	geographic	141	semanlink.net	user-gen. cnt.	88
w3.org	cross-domain	117	deri.org	social netw.	70
quitter.se	social netw.	64	harth.org	social netw.	68
status.net	social netw.	63	quitter.se	social netw.	67
postblue.info	social netw.	56	semanticweb.org	user-gen. cnt.	64
skilledtest.com	social netw.	55	skilledtests.com	social netw.	60
reference.data.gov.uk	government	45	postblue.info	social netw.	59
data.semanticweb.org	publications	44	status.net	social netw.	47
fragdev.com	social netw.	41	w3.org	crossdomain	45
lexvo.org	cross-domain	37	data.semanticweb.org	publications	45

Looking at the top ten datasets by in- and outdegree in Table 2, we see that datasets from categories *social networking*, *user-generated content*, and *publications* are among the top ten with respect to outdegree. While datasets with a high indegree are `dbpedia.org` (cross-domain), `geonames.org` (geographic), `w3.org` (cross-domain), `reference.data.gov.uk` (government) as well as several datasets from the category *social networking*.

4.2 Overall Graph Structure

Analyzing the overall graph structure, we find one large weakly connected component which consists of 71.99% of all datasets. In addition, there are three small components, one consisting of three and two consisting of two datasets. Within the large weakly connected component, there exists one large strongly connected component consisting of 36.29% of all datasets.

Figure 3 shows the overall graph structure using the same *LOD cloud* visualization as the 2011 report. The size of the circles reflects the indegree of the corresponding dataset. A zoom-able version of Figure 3 is available on the accompanying website. Note that we have aggregated all individual FOAF profiles into a single circle. Compared to the *LOD cloud* visualization from the 2011 report which was centered around `dbpedia.org` as central linking point, Figure 3 shows a much more decentralized graph with multiple high-degree nodes: The `geonames.org` and `dbpedia.org` datasets are linked by a large number of datasets belonging to different topical categories. In addition, the `statistics.data.gov.uk` and `reference.data.gov.uk` are highly linked from within the *government* category. In the category *publications*, the `Library of Congress Subject Headings (LCSH)` and the `German National Library (DNB)` datasets are highly linked. We can also see in Figure 3 that the category *social networking* is the most densely interlinked.

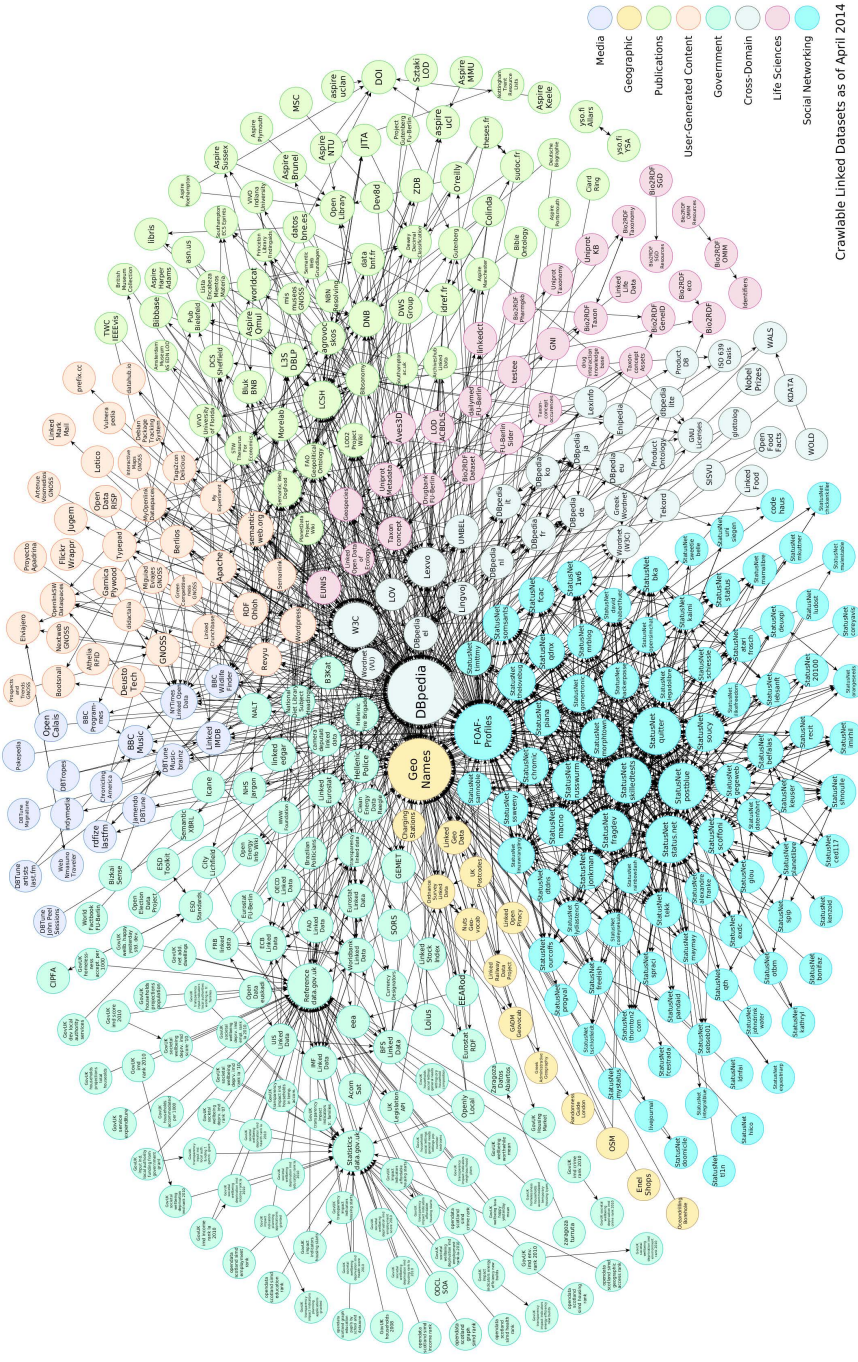

Fig. 3. Overall graph structure and categorization of the datasets by topical domain. The size of the circles reflects their indegree. A zoom-able version of the diagram is available on the accompanying website.

Table 3. Top three linking predicates per category. The percentages are relative to number of datasets within the category which set outgoing links.

Category	Predicate	Usage	Category	Predicate	Usage
social networking	foaf:knows	60.27%	life sciences	owl:sameAs	52.17%
social networking	foaf:based_near	35.69%	life sciences	rdfs:seeAlso	43.48%
social networking	sioc:follows	34.34%	life sciences	dct:creator	21.74%
publications	owl:sameAs	32.20%	government	dct:publisher	47.57%
publications	dct:language	25.42%	government	dct:spatial	30.10%
publications	rdfs:seeAlso	23.73%	government	owl:sameAs	24.27%
user-generated content	owl:sameAs	53.13%	geographic	owl:sameAs	64.29%
user-generated content	rdfs:seeAlso	21.88%	geographic	skos:exactMatch	21.43%
user-generated content	dct:source	18.75%	geographic	skos:closeMatch	21.43%
media	owl:sameAs	81.25%	crossdomain	owl:sameAs	80.00%
media	rdfs:seeAlso	18.75%	crossdomain	rdfs:seeAlso	52.00%
media	foaf:based_near	18.75%	crossdomain	dct:creator	20.00%

4.3 Predicates Used for Linking

Table 3 displays the top three predicates that are used by RDF links within each topical domain. A first observation is that `owl:sameAs` is an important linking predicate within most categories, followed by `rdfs:seeAlso`. The most notable deviance is observed for the category *social networking*, where `foaf:knows` is the most widely used linking predicate.

Due to the outstanding role of `owl:sameAs` as the most widely used linking predicate, we take a closer look at the datasets connected by `owl:sameAs` links. Searching for weakly connected components in the `owl:sameAs` graph, we find one large weakly connected component containing 297 (29.3%) of all datasets. Apart from that, there are only eight further components, out of which three consist of three datasets and the remaining five consist of two datasets. Looking at strongly connected components, we find one large component consisting of 74 datasets (7.3%), one with four and six with two datasets.

Table 4 shows the top ten datasets regarding in- and outdegree, this time considering only `owl:sameAs` links. Compared to Table 2, we observe a much smaller number of datasets from the category *social networking* as this category is dominated by `foaf:knows` links.

5 Adoption of the Vocabulary Best Practices

The vocabularies used to represent data and their interpretability are a key ingredient to make Linked Data *semantic* data. We consider a vocabulary to be *used* by a dataset if a term from the vocabulary appears in the predicate position of a triple from the dataset or at the object position of a `rdf:type` triple from the dataset.

Table 4. Top 10 datasets regarding in- and outdegree for `owl:sameAs` links by category

Dataset	Category	In	Dataset	Category	Out
dbpedia.org	crossdomain	89	bibsonomy.org	publications	91
geonames.org	geographic	29	data.semanticweb.org	publications	31
data.semanticweb.org	publications	24	myopenlink.net	user-gen. cnt.	25
l3s.de	publications	24	dbpedia.org	crossdomain	23
semanticweb.org	user-gen. cnt.	18	semanticweb.org	user-gen. cnt.	18
nytimes.com	media	11	revyu.com	user-gen. cnt.	16
dbtune.org	social networking	11	advogato.org	social networking	16
kit.edu	social networking	9	el.dbpedia.org	crossdomain	13
revyu.com	user-gen. cnt.	8	nl.dbpedia.org	crossdomain	11
w3.org	crossdomain	8	harth.org	social networking	11
it.dbpedia.org	crossdomain	8			

Table 5. Vocabularies used by more than 5% of all datasets

Prefix	Occurrence	Quota	Prefix	Occurrence	Quota
rdf	996	98.22%	void	137	13.51%
rdfs	736	72.58%	bio	125	12.32%
foaf	701	69.13%	cube	114	11.24%
dcterm	568	56.01%	rss	99	9.76%
owl	370	36.49%	odc	86	8.48%
wgs84	254	25.05%	w3con	77	7.60%
sioc	179	17.65%	doap	65	6.41%
admin	157	15.48%	bibo	62	6.11%
skos	143	14.11%	dcat	59	5.82%

5.1 Usage of Well-Known Vocabularies

Table 5 lists the vocabularies that are used by more than five percent of all datasets[5]. The vocabularies *RDF*, *FOAF*, *RDFS*, *DCTerms*, and *OWL* are the top vocabularies used by many datasets from across all topical categories. Compared to the 2011 report, we can state that there is a trend towards the adoption of well-known vocabularies by more datasets. For instance, while the *FOAF* vocabulary was used by 27.46% of all datasets in 2011, it is used by 69.1% of all datasets in 2014. The same is true for the Dublin Core vocabulary which is used today by 56.01% of the datasets and was used by only 31.19% in 2011.

The extent to which well-known vocabularies are used within the different topical categories reveals some differences. In the category *social networking*, there is a high quota of datasets using *FOAF* (85.96%), followed by the Dublin Core and the *WGS84* vocabulary used by 40% and 37% of all datasets. The *admin* vocabulary, which is used by some FOAF generators, finds comparatively wide adoption. In the category *publications*, *DCTerms* is widely used at a quota of 83%. Furthermore, the *bibo* ontology is used by 41.67% of the datasets

[5] Prefixes are taken from `http://prefix.cc`

belonging to this category. The vocabularies *SKOS*, *resourcelist*, which is used to create reading lists, and *SIOC* also find some adoption. In the category *cross-domain*, several vocabularies are used by 10-40% of all datasets: The *dbpedia.org*, *georss.org*, *opengis.net*, *bibo*, the *prov* vocabulary, the *skos* vocabulary, and *void*, showing that a wide variety of topics is covered in this category. In the category *government*, vocabularies for representing statistical data (*cube* with 61.75% and *sdmx* with 26.22%) are found frequently. Vocabularies for expressing metadata, like the *void* vocabulary, the sparql-service-description vocabulary, *prov* and *prv* are also find some use. Within the category *geographic*, 66.67% of all datasets use the *WGS84* vocabulary for encoding geographic coordinates. Other well adopted vocabularies are *skos* or the geonames ontology. In the category *user-generated content*, many datasets use the *FOAF* vocabulary together with the *SIOC* vocabulary (50%) as well as the *RSS* and the *admin* vocabulary (both around 17%). The *DOAP* vocabulary is used by 12.5% of the datasets.

Please note that the *schema.org* vocabulary promoted by Google, Yahoo and Microsoft is not listed in Table 5 as we found this vocabulary to be hardly used in the Linked Data context[6]. In contrast, the vocabulary is very widely used together with the Microdata syntax for annotating HTML pages [2].

5.2 Usage of Proprietary Vocabularies

Widely-used vocabularies often do not provide all terms that are needed to publish the complete content of a dataset on the Web. Thus, data providers often define proprietary terms that are used in addition to terms from widely deployed vocabularies. We have also analyzed to which extent datasets from different categories make use of proprietary vocabularies. We consider a vocabulary to be proprietary if it is used only by a single dataset. Out of the 638 different vocabularies that we encountered in our crawl, 375 vocabularies (58.77%) are proprietary according to our definition, while 263 (41.22%) are non-proprietary. In total, 234 datasets (23.08%) use proprietary vocabularies, while nearly all datasets also use non-proprietary vocabularies. These numbers show that the adoption of the best practice to use common vocabularies is improving compared to the *State of the LOD Cloud* report from 2011 which found 64.41% of all datasets to use proprietary terms. Table 6 further details the usage of proprietary vocabularies by topical category. The second column of the table shows the number of proprietary vocabularies used by datasets from each category. The third column contains the number of datasets in each category that use proprietary vocabularies.

5.3 Dereferencability of Proprietary Vocabulary Terms

In order to enable applications to retrieve the definition of vocabulary terms, the URIs identifying vocabulary terms should be made dereferencable. To assess

[6] One data source that uses the *schema.org* type system in addition to its own type system in order to increase interoperability is *dbpedia.org*.

Table 6. Proprietary vocabularies with dereferencability per category and quota of vocabularies linking to others

Category	Different prop. vocabs. used (% of all prop. vocab.)	# of datasets using prop. vocab. (% of all datasets)	Dereferencability			#of vocabs linking (quota)
			full	partial	none	
Social networking	126 (33.60%)	81 (15.57%)	19.47%	8.8%	77.78%	20 (15.87%)
Publications	59 (15.73%)	33 (34.38%)	22.03%	8.47%	69.49%	15 (25.42%)
Government	47 (12.53%)	34 (18.58%)	21.28%	12.77%	65.96%	16 (34.04%)
Cross-domain	56 (14.93%)	17 (41.46%)	26.79%	10.71%	62.50%	14 (25.00%)
Geographic	13 (3.47%)	8 (38.10%)	15.38%	7.69%	76.92%	2 (15.38%)
Life sciences	36 (9.60%)	27 (32.53%)	27.78%	5.56%	66.67%	4 (11.11%)
Media	12 (3.20%)	12 (54.55%)	0.00%	16.67%	83.33%	2 (16.67%)
User-gen. cnt.	26 (6.93%)	22 (45.83%)	11.54%	11.54%	76.92%	6 (23.08%)
Total	375 (58.77%)	234 (23.08%)	19.47%	8.80%	71.73%	79 (21.07%)

Table 7. Predicates used to link terms between different vocabularies

Term	% of vocabularies	Term	% of vocabularies
rdfs:range	9.87%	rdfs:seeAlso	1.60%
rdfs:subClassOf	8.80%	owl:equivalentClass	1.60%
rdfs:subPropertyOf	6.93%	owl:inverseOf	1.33%
rdfs:domain	5.60%	swivt:type	1.07%
rdfs:isDefinedBy	3.73%	owl:equivalentProperty	0.80%

whether a vocabulary is dereferencable, we requested the definitions of all used terms from the vocabulary via HTTP GET requests. The resulting corpus of vocabulary definitions is provided for download on the accompanying website. We define the dereferencability quota of a vocabulary as the number of dereferencable terms divided by the number of all terms of the vocabulary. In total, 19.47% of all proprietary vocabularies are fully dereferencable (i.e., their quota is 1.0). On the other hand, 71.73% of all proprietary vocabularies are not dereferencable at all. The remaining 8.8% of all proprietary vocabularies are partially dereferencable, meaning that for some terms, but not for all, a definition could be retrieved. Possible causes for partial dereferencability are *namespace squatting*, i.e. accidentally or incidentally using terms not defined in a vocabulary, and vocabularies having changed without proper marking of old terms as deprecated. Columns 4, 5 and 6 in Table 6 show the percentage of fully, partially and not dereferencable proprietary vocabularies per topical category.

5.4 RDF Links to Terms from Other Vocabularies

Vocabulary terms should be related to corresponding terms within other vocabularies in order to enable applications to understand as much data as possible. Table 7 contains the different predicates that are used to link terms between

Table 8. Provenance vocabulary usage and license vocabulary usage by category

Category	Any prov vocab		Dublin Core	Admin	Prv/Prov	Any license vocab
social networking	169	(32.5%)	57.39%	57.39%	1.18%	5.38%
publications	39	(40.63%)	94.87%	5.13%	2.56%	4.17%
government	76	(41.54%)	100.00%	0.00%	1.32%	30.05%
life sciences	20	(24.10%)	100.00%	0.00%	0.5%	3.61%
cross-domain	7	(17.07%)	100.00%	14.29%	0.00%	9.76%
geographic	3	(14.29%)	100.00%	0.00%	33.34%	0.00%
user-gen. content	9	(18.75%)	88.89%	66.67%	0.00%	10.42%
media	4	(18.18%)	100%	0.00%	0.00%	5.41%
Total	372	(36.69%)	29.09%	11.05%	0.79%	9.96%

vocabularies together with the percentage of all vocabularies using each predicate for linking. We see that 9.87% of all vocabularies use the `rdfs:range` predicate to link to other vocabularies (for instance defining the range of a term to be `foaf:Person`). The table also shows that only a very small fraction of the vocabularies provides equivalence links to terms from other vocabularies.

6 Adoption of the Metadata Best Practices

The Linked Data best practices propose that every dataset should provide provenance and licensing information, dataset-level metadata, and information about additional access methods.

6.1 Providing Provenance Information

For our evaluation, we have collected a list of vocabularies that are designed for the representation of provenance information. Information about such vocabularies came from the W3C Provenance Working Group, the LOV vocabulary catalog, as well as our own experience, adding up to a total of 26 vocabularies. Using those vocabularies, we searched for provenance information in our corpus. We followed the approach sugested in [5] and searched for triples using predicates from those vocabularies and containing a document URI as subject.

As shown in Table 8, 36.69% of all datasets use some provenance vocabulary, which is a slight decrease compared to the *State of the LOD Cloud* report from 2011, which reports 36.63% of all datasets to provide provenance information. 29.09% of all datasets use Dublin Core Terms, 11.05% use MetaVocab, while W3C *PRV* and *PROV* are used by only 0.79% of the datasets. The provision of provenance information is widely adopted in the *publications* and *government* domains, while *media* and *geographic* datasets show less adoption. For *government* data, there is also a remarkable increase compared to the *State of the LOD Cloud* document from 2011, which reports only 20.41% for this topical domain.

6.2 Providing Licensing Information

With the help of machine-readable licensing information, Linked Data applications can assess whether they may use data for their purpose at hand. To evaluate whether a dataset provides license information, we again followed the approach proposed in [5] and searched for triples which have the document as their subject and a predicate containing the string *'licen'*. To this list, we added all predicates containing the string *'rights'* as well as the waiver vocabulary, which leads to a total of 47 terms.

In total, 9.96% of all datasets provide licensing information in RDF. This number is lower than the 17.84% reported in the *State of the LOD Cloud* report from 2011, but still higher than the 3.4% reported in [5]. The most important predicates for indicating the license are `dc/dct:license` (7.39%), `cc:license` (2.07%) and `dc/dct:rights` (1.68%). As shown in the last column of Table 8, the provision of licensing information varies widely across topical domains. More than a third of all *government* datasets provide licensing information, while none of the *geographic* datasets provides licensing information. A main cause for the low overall number is the category *social networking* which contains 48% of all datasets and in which only 5.38% of the dataset offer licensing information.

6.3 Providing Dataset Level Metadata

Dataset-level metadata can be provided using the VoID vocabulary, either as inline statements in the dataset or in a separate VoID file. In the latter case, that file has to be linked from the data via backlinks or be provided at a well-known location which is created by appending `/.well-known/void` to the host part of a URI. As reported in [8], the latter condition is often too strict for data providers due to missing root-level access to the servers. Thus, we follow the approach proposed in [8] of relaxing the search for VoID files at well-known locations, appending `/.well-known/void` to any portion of the URI.

In general, dataset-level metadata is still rarely provided by datasets within all topical domains. Some trends towards emerging best practices and de facto standards can be observed: Dataset-level metadata is rather linked to than provided at well-known locations and the Dublin Core vocabulary is becoming the de-facto standard for providing dataset-level provenance information. In total, 149 datasets (14.69%) use the VoID vocabulary. Out of these datasets, 42 (4.14%) use a backlinking mechanism. Columns 2 to 5 of Table 9 show the VoID adoption by topical category.

Compared to the 2011 report, the overall percentage of datasets publishing dataset-level metadata using VoiD has decreased from 32.20% to 14.69%, with the categories *government*, *geographic*, and *life sciences* being exceptions in which the adoption has slightly grown. Again, the category *social networking* is a main cause for the low overall number.

Table 9. Percentage of datasets using the VoID vocabulary and percentage of datasets offering alternative access methods

Category	VoID	Link	Well-known	Inline	Alt. access	SPARQL	Dump
social networking	5 (0.96%)	0.19%	0.77%	0.00%	4 (0.77%)	0.77%	0.19%
publications	13 (13.54%)	6.25%	3.13%	7.29%	13 (13.54%)	12.50%	4.17%
life sciences	30 (36.14%)	28.92%	2.41%	4.82%	20 (24.10%)	24.10%	15.66%
government	72 (42.08%)	2.73%	2.73%	36.61%	63 (34.43%)	31.15%	31.15%
user-gen. content	6 (11.76%)	11.76%	0.00%	0.00%	3 (6.25%)	6.25%	2.08%
geographic	6 (38.10%)	14.29%	9.52%	14.29%	5 (23.81%)	14.29%	19.05%
cross-domain	5 (12.20%)	7.32%	2.44%	4.88%	4 (9.76%)	4.88%	4.88%
media	2 (9.09%)	0.00%	0.00%	9.09%	1 (4.55%)	0.00%	4.55%
Total	149 (14.69%)	4.14%	1.28%	9.27%	113 (11.14%)	9.96%	8.19%

6.4 Providing Alternative Access Methods

According to the 2011 *State of the LOD Cloud* report, many datasets provide additional access methods, such as SPARQL endpoints (68.14%) and dumps (39.66%). In our analysis, the numbers are much lower as shown in columns 6 to 8 of Table 9. Apart from the *government*, *life sciences* and *geographic* domains, almost no information on alternative access methods are found. The deviation can be explained by the fact that we only look at those alternative access methods that can be discovered via VoID descriptions linked from the datasets or provided at well-known URLs. As reported in [8], the actual number of existing SPARQL endpoints may be higher, as many endpoints cannot be discovered from the data. This is a severe problem for automatic agents navigating the Linked Data graph, as they are not capable of discovering alternative access methods. While the numbers for alternative access methods are low, one has to keep in mind that such methods do not always make sense. For example, the large number of small FOAF files in the *social networking* category are mostly datasets contained in exactly one file. In these cases, it does not make sense to provide a data dump, because the file itself *is* a data dump. Likewise, the use of a SPARQL endpoint for a dataset consisting of only a few dozen triples would not justify the provision effort.

7 Related Work

An effort that is closely related to the work presented in this paper is the *LODStats* project[7] which has retrieved and analyzed Linked Data from the Web until February 2014 [1]. The *LODStats* website provides statistics about the overall number of discovered linked datasets as well as the adoption of different vocabularies. What distinguishes *LODStats* from the work presented in this paper is that they do not categorize datasets by topical domain and do not analyze the

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

overall graph structure, as well as the conformance with the best practices in the areas of vocabulary dereferencability and metadata provision. Their results concerning the overall number of accessible datasets (they found 928 datasets) and the adoption of well-known vocabularies are inline with the findings of this paper.

A comprehensive empirical survey of Linked Data conformance is presented by Hogan et al. [5]. Their survey is based on a large-scale Linked Data crawl from May 2010 as well as a series of smaller snapshots taken between March and November 2010. The work presented in this paper can be seen as an update of the results presented by Hogan et al. as we use a crawl from March 2014. Another major difference is that Hogan et al. do not categorize datasets by topical domain and thus can not analyze the differences in the adoption of the best practices in different domains. The article by Hogan et al. contains a detailed and comprehensive discussion of earlier work on analyzing the adoption of the Linked Data practices as well as work in the wider area of characterizing the Semantic Web/Linked Data, its link structure as well as the semantics of its content. The discussion covers related work from the time span of 2005 to 2012. For space reasons, we can not repeat this excellent review of related work here. The general difference between the works discussed by Hogan et al. and our work is that our analysis is more up-to-date and that we distinguish the datasets by topical domain.

8 Conclusion

This paper revisited and updated the finding of the *State of the LOD Cloud* report [7] from 2011 based on a Linked Data crawl gathered in April 2014. Our analysis shows that the overall number of Linked Datasets on the Web has grown significantly since 2011. Looking only at the topical categories covered in the original report, the number of datasets has approximately doubled since 2011. Also taking the category *social networking* into account, the number of datasets has grown by 271%.

Concerning the linkage of the datasets, our analysis shows that there is still a relatively small number of datasets that set RDF links pointing at many other datasets, while many datasets only links to a few other datasets. Compared to the 2011 *LOD cloud*, which was centered around `dbpedia.org` as central linking hub, we have discovered a more decentralized graph structure with `geonames.org` and `dbpedia.org` being linked from many datasets besides of the existence of further category-specific linking hubs. Concerning the types of RDF links that connect datasets, we have found the predicates `owl:sameAs`, `rdfs:seeAlso` and `foaf:knows` to be most widely used.

We have observed a trend towards the adoption of well-known vocabularies by more datasets, the most prominent one being FOAF, which is used by more than two thirds of all linked datasets, independent of their respective topical domain. In parallel, the usage of proprietary vocabularies has decreased from 64.41% in 2011 to 23.08% of all datasets in 2014.

While provenance information is provided for roughly a third of all datasets, only 10% of all datasets provide machine-readable licensing information. A positive exception concerning licensing information is the *government* domain in which licensing information is provided by 30% of all dataset. Compared to the 2011 report, the percentage of datasets providing provenance metadata is approximately the same, while the percentage of datasets providing machine-readable licensing information has dropped from 17% to 10%. The similar negative trend is also found for the percentage of datasets publishing dataset-level metadata using VoID. In 2011, 32.20% of all datasets published VoID while in 2014 only 14.69% provide such metadata. The categories *government*, *geographic*, and *life sciences* are exceptions to this trend and the adoption has slightly grown in these domains.

Acknowledgements. The work presented in the paper was supported by the research project *PlanetData* (Ref.No. 257641), funded by the European Communitys Seventh Framework Programme.

References

1. Auer, S., Demter, J., Martin, M., Lehmann, J.: LODStats – an extensible framework for high-performance dataset analytics. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 353–362. Springer, Heidelberg (2012)
2. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rDFa, microdata, and microformats on the web – A quantitative analysis. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) ISWC 2013, Part II. LNCS, vol. 8219, pp. 17–32. Springer, Heidelberg (2013)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
4. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology 1(1), 1–136 (2011)
5. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. J. Web Sem. 14, 14–44 (2012)
6. Isele, R., Umbrich, J., Bizer, C., Harth, A.: LDSpider: An open-source crawling framework for the web of linked data. In: Proceedings of the ISWC 2010 Posters and Demonstrations Track (2010)
7. Jentzsch, A., Cyganiak, R., Bizer, C.: State of the lod cloud (September 2011), `http://lod-cloud.net/state/`
8. Paulheim, H., Hertling, S.: Discoverability of SPARQL endpoints in linked open data. In: Proceedings of the Posters and Demos Track of ISWC 2013 (2013)