PlanetData

Network of Excellence

FP7 – 257641

D2.4 Conceptual model and best practices for high-quality metadata publishing

Coordinator: Max Schmachtenberg

With contributions from: Max Schmachtenberg, Heiko Paulheim, Christian Bizer

1st Quality reviewer: Elena Simperl

2nd Quality reviewer: Andreas Harth

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	M39
Actual delivery date:	M39
Version:	0.1
Total number of pages:	44
Keywords:	Linked Open data, LOD Cloud, data quality, data publishing Best Practices

Abstract

In the past years, the amount of datasets published as Linked Open Data (LOD) has grown considerably. To foster further adoption and ease the development of LOD applications, a number of best practices for Linked Data publishers have been proposed. In an empirical overview of LOD datasets, this deliverable examines the adherence to those best practices in different domains, and identifies both well adopted best practices and open issues.

[End of abstract]

Executive summary

In past years, more and more Linked Open Data (LOD) was published by different parties, such as universities, libraries, government agencies, companies or individual persons, forming the LODCloud. The content released ranges from small Friend-Of-A-Friend (FOAF)-profiles of persons, blog posts, up to large databases on various topics with a large number of resources having rich descriptions.

But real benefit for the Web of Data is only endowed if data publishers put out their data in a way conforming to the idea of Linked Open Data, creating a network of knowledge, accessible and processable by automatic agents. Addressing this issue, Deliverable 2.1 described a set of best practices, giving a guideline to publishers which aspects to consider for the publication of datasets for making them appealing for data consumers [1].

These best practices encourage publishers to interconnect their dataset with others for creating a true web of Linked Data (best practice 1) and to provide provenance and licensing information along with their datasets (best practices 2 and 3). Further, they advise them to use widely-deployed vocabularies (best practice 4) and to make proprietary vocabularies, i.e. those used by just one dataset, dereferencable (best practice 5) and to map them with other vocabularies (best practice 6). Lastly, data publishers should provide dataset-level metadata (best practice 7) which should contain information about alternative access methods, such as SPARQL endpoints and data dumps (best practice 8).

The aim of this deliverable is to evaluate to what extent and in what fashion dataset published in the current LODCloud adhere to these best practices. For this, we crawled and obtained resources from a large number of datasets to get samples from datasets in the LODCloud. This resulted in 918 datasets, which we categorized into nine different categories, depending on their content and origin. Category social web contains data about persons and their relationships, category CMS has datasets with the output of content-management systems, like blogging tools and services. Category lifesciences contains datasets related to biology or medical data and category publications hold dataset for example from libraries. In category cross-domain, datasets containing different kind of information, for example DBpedia, are found whereas in category government, datasets with information from and about governments are included. Datasets about spatial entities are found in category geographic while datasets with information about media, such as the program of BBC, are found in category media. Finally in category user-generated content datasets about content created by users can be found.

These categories differ in their size and appearance. On the one side, the largest category social web, consisting of datasets like FOAF-profiles and data from social online communities, provides 45% of all datasets, but on average has only 19 entities per dataset. Category lifesciences is an example of a category of medium size with nearly 100 datasets, and an average of 1,540 entities per dataset. Lastly, there are rather small categories regarding the number of datasets, for example cross-domain, containing, among others, different local versions of DBpedia, holding only 53 datasets, which have an average of 48,064 entities per dataset or category media having only 33 datasets, but on average 77,384 entities per dataset.

Based on the datasets and their categorization, we evaluate the best practices by creating metrics for each of them, analysing the adherence both for all datasets as well as individual categories.

We found that more than a half of all dataset do not link to other datasets at all. The extent of interlinking differs between categories, for example dataset from category social web data, more often link to datasets and on average link to more datasets than governmental datasets, were on average, a dataset linked to less than one other dataset.

The supplementation of provenance information in the form of used vocabularies which are tailored to supply such information was analysed next. We found out that although many dataset use Dublin Core to express basic provenance information, dedicated vocabularies for provenance information are used rarely. In the area of CMS and social web, a special vocabulary indicating with which agent the dataset was created (e.g. the CMS system) finds some use.

By searching for triples denoting license information, we evaluated to what extent dataset include license information. Although there is some adoption of this best practice with around every sixth dataset providing license terms, the differences between different kind of datasets is strong, ranging from 6% (social web datasets) up to 50% (governmental datasets) of dataset publishing license information in a category.

Taking a look at vocabulary usage, we distinguish between two classes of vocabularies, proprietary and non-proprietary/widely-used vocabularies, based on their usage, meaning that we define a vocabulary as proprietary if it is used only by one dataset. We identified a total of 618 vocabularies, of these 393 being characterised as proprietary. All datasets use at least one non-proprietary vocabulary, while the extent of proprietary vocabulary usage differs for the categories. When taking a look at the dereferencability of proprietary vocabularies, we see that two thirds of these vocabularies are not dereferencable at all, a problem as they are not understandable by automatic agents. Of the dereferencable proprietary vocabularies, 28% connect terms to other vocabularies, most often by specializing existing terms, most prominent those from the SKOS vocabulary.

In order to evaluate to what extent metadata about datasets is provided, we search for the usage of the VoID vocabulary and especially for if a VoID file was supplied, either linked to from within the dataset, or at a standardised location of the domain. Our results show that only a minority of dataset providers use the VoID vocabulary or supply a VoID file. Again, the adoption of this best practice differs between different kind of datasets, with those providing geographical data, lifesciences data or governmental data showing a relative strong uptake, while those providing social web or MCS data exhibiting a very low uptake.

Within these metadata, we searched for the reference to additional access methods, such as a SPARQL endpoint or a downloadable data dump. For most datasets providing a VoID description, we also found reference to such access methods. Again, there are large differences between the adoption for different groups of datasets, with lifesciences and geography datasets more often providing such alternative access methods. One can also see a tendency to supply SPARQL endpoints rather than data dumps.

In summary, we can see that none of the best practices exhibit a universal adherence. To a large extent, it is depending on the category of datasets. For example while datasets of category social web more strongly adhere to the best practice of interlinking, they do generally not adhere to provide dataset level metadata. On the other side do datasets from category government often supply metadata, but are not linked to other dataset. These results show that it has to be continued to promote these practices and publishers have to be supported to adhere to them, for example by supplying tools adapted to different kinds of datasets which ease to follow or automatically implement these best practices.

Document Information

IST Project Number	FP7 - 257641	Acronym	PlanetData
Full Title	PlanetData		
Project URL	http://www.planet-data.eu/		
Document URL			
EU Project Officer	Leonhard Maqua		

Deliverable	Number	D2.4	Title	Conceptual model and best practices for high-quality metadata publishing
Work Package	Number	WP2	Title	Quality assessment and context

Date of Delivery	Contractual	M36	Actual	M39
Status	version 0.1		final □	
Nature	prototype □ report ☒ dissemination □			
Dissemination level	public ☒ consortium □			

Authors (Partner)				
Responsible Author	Name	Max Schmachtenberg	E-mail	max@informatik-uni-mannheim.de
	Partner	UMA	Phone	+49 621 181 3705

Abstract (for dissemination)	In the past years, the amount of datasets published as Linked Open Data (LOD) has grown considerably. To foster further adoption and ease the development of LOD applications, a number of best practices for Linked Data publishers have been proposed. In an empirical overview of LOD datasets, this deliverable examines the adherence to those best practices in different domains, and identifies both well adopted best practices and open issues.
Keywords	Linked Open data, LOD Cloud, data quality, data publishing Best Practices

Version Log			
Issue Date	Rev. No.	Author	Change
13/12/2013	01	Max Schmachtenberg	First Version
28/12/2013	02	Max Schmachtenberg	Adressing review comments
30/12/2013	03	Max Schmachtenberg	Final Version

Table of Contents

Executive summary 3
Document Information 5
Table of Contents 6
List of Figures 7
List of Tables 8
1 Introduction 9
2 Sample Acquisition 10
2.1 Notion of Datasets 11
2.2 Basic Crawl Corpus Properties & Comparison With Other Corpora 12
2.3 Categorization of Datasets 14
3 Best Practices 16
3.1 Providing Links to other Datasets 17
3.1.1 Definition 17
3.1.2 Results 18
3.2 Providing Provenance Data 21
3.2.1 Definition 21
3.2.2 Results 23
3.3 Provide Licensing Data 25
3.3.1 Definition 25
3.3.2 Results 26
3.4 Using terms from widely deployed vocabularies 28
3.4.1 Definition 29
3.4.2 Results 29
3.5 Dereferencability of proprietary vocabulary terms 31
3.5.1 Definition 31
3.5.2 Results 31
3.6 Mapping of proprietary vocabularies to others 34
3.6.1 Definition 34
3.6.2 Results 34
3.7 Provide dataset-level metadata 36
3.7.1 Definition 36
3.7.2 Results 37
3.8 Referring to additional access methods 38
3.8.1 Definition 38
3.8.2 Results 38
4 Related Work 40
5 Conclusion 42
References 44

List of Figures

Figure 1: Entity distribution by datasets 13
Figure 2: Frequency and quota of datasets in different categories 15
Figure 3: Hierarchy for classification of license information. Percentages are relative to the number of datasets on step above in the hierarchy 28
Figure 4: Derreferencability Quota of partially dereferencable vocabularies in descending order 33

List of Tables

Table 1: Overlap between our crawl and other dataset collections....13
Table 2: number of datasets with outlink, average out- and indegree by category....19
Table 3: Top 10 datasets and their categories with highest outdegree/indegree 19
Table 4: Top 10 of predicates that have outlinks from and inlinks to datasets....20
Table 5: Provenance Vocabularies, including namespace and prefix (taken from prefix.cc) 22
Table 6: Usage of different provenance vocabularies....23
Table 7: Usage of provenance vocabularies by dataset category. The percentage values are relative to all datasets of a category....24
Table 8: Generator agents which are used by more than one dataset 25
Table 9: Usage of property terms used to indicate license information by number of datasets. Stars indicate non-dereferencable terms 26
Table 10: Number of datasets using license predicates per category....27
Table 11: Vocabularies used by more than 5% of datasets with prefix, namespace, usage count and quota of usage 30
Table 12: Usage of proprietary vocabularies by datasets in different categories. 30
Table 13: General dereferencability of vocabularies 32
Table 14: Dereferencability of proprietary vocabularies used in different categories 33
Table 15: Number of Vocabularies using connecting Terms 34
Table 16: Number of proprietary vocabularies mapping by category 35
Table 17: Target vocabularies of proprietary vocabulary mappings 35
Table 18: Usage of VoID vocabulary and supply of VoID files 37
Table 19: Datasets providing VoID description by category 37
Table 20: Availability of additional access methods 38
Table 21: Availability of alternative Access Methods by dataset category 39

1 Introduction

The last decade saw the rise of Linked Open Data (LOD), as more and more datasets, offered by different parties, were published to be consumed by intelligent agents. Over time, the cloud of interLinked Datasets grew from a few datasets to a large network of hundreds, housing millions of resources with billions of triples.[1] Apart from the academic community, libraries, like the Library of Congress[2] or the Deutsche Nationalbibliothek[3] and government agencies, such as the British government[4], have also adopted the idea of Linked Data for publishing their data in a machine-processable way.

To unleash the usefulness of Linked Data, it is necessary that automatic agents can browse, discover and consume it easily. For example, a dataset should have links to others, enabling to discover information related to it. Also, adding describing metadata to datasets, for example license information, so that intelligent agents discovering a dataset can assess under which legal terms it is allowed to use it for their purposes, greatly enhances quality and usability of a dataset.

Giving a guideline to practitioners, Deliverable 2.1 of the PlanetData project described a set of best practices on how to publish high quality Linked Data. In general, these best practices advise data publishers to link their data to other datasets, to provide metadata along with their dataset, to use vocabularies widely used and understood, and if a dataset provider defines and uses its own vocabulary, to supply its definition and map it to other vocabularies.

With "The State of the LODCloud", the adherence to these best practices where first surveyed in 2011 [2]. It was based on descriptions dataset publishers provided when adding their data to the lod-cloud group at datahub.io[5], a catalogue of Linked Datasets. More than two years have passed since the last time the adherence of the best practices has been evaluated. The LODCloud has changed, as datasets have gone permanently offline while others have been added and existing ones changed their properties. In this deliverable, we will again evaluate the extent best practices are adhered to. For this, we performed a crawl of Linked Data, gathering a sample from a high number of datasets. Based on this sample, we evaluate if the best practices from Deliverable 2.1 are adopted or overlooked, and if the former is the case, how they are followed.

The remainder of this deliverable is organized as follows. The next section describes the crawling process for gathering sample data from a large number of different datasets. We explain on how we separated datasets that originated from the same pay-level domain (PLD), to also account for different datasets being published under one PLD. To give an impression of the crawl's coverage, we compare it to other crawls and datasets catalogues. After this, we describe how we categorized datasets, based on an existing categorization scheme, for example indicating that a dataset contains governmental data or that its data describes publications. Using this categorization, we are able to show to what extent best practices are adhered to by different categories of datasets.

Section 3 describes the adherence to the eight best practices described in deliverable 2.1. For every one of them, we discuss its motivation and the benefits of implementing it. We then explain how we operationalize its measurement for assessing the extent of its adherence, outlining advantages as well as limitations of our approach. Following this, we report our results, both in general for all datasets, as well as regarding different categories, both regarding the adherence in general as well as the way best practices are implemented.

To put our finding in context to others, we compare our results with the existing literature on quality of Linked Open Data in Section 4. Finally, we draw our conclusion in Section 5. Here, we first summarize our findings and make suggestions on how adaption of different best practices can be improved.

1 See [13] and [2] for an overview on this development

2 http://id.loc.gov/

3 http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

4 http://data.gov.uk

5 http://datahub.io/de/group/lodcloud

2 Sample Acquisition

To analyse if and how publishers of Linked Data adhere to the best practices, information about Linked Data datasets available on the web is required. A way chosen by [2] is to base the analysis on information provided by dataset publishers. They were asked to enter properties about their datasets into the lod-cloud group at datahub.io, a catalogue of Linked Data datasets, on which basis the adherence to the best practices was derived. This approach has some drawbacks. First, one has to either actively or passively contact publishers to ask them for information about their datasets. The first approach can be done by posting on mail lists and other Linked Data-related information sources, which does not require much effort, but has the risk that publishes overlook such an appeal. The latter approach requires more effort, as publishers (and their contact information) have to be identified first, requiring to locate datasets, identifying their publishers and obtaining the publisher's contact information, a laborious and no always fruitful effort, as for some datasets the publisher might not be traceable or can't be contacted. Even if publishers are somehow reached, not every one of them might answer, and the population of answering publishers will be biased towards datasets from publishers who readily answer to such requests.

An alternative approach to evaluate the adherence of the best practices is to analyse the dataset itself, which for example was done by [3]. Such an analysis requires actual content form datasets, which can be acquired from existing dumps and crawls, or by performing crawls to obtain Linked Data. A main advantage of this approach is that one does not depend on the publisher's willingness to provide information about their datasets. On the other side, some disadvantages exist. Some information, which a publisher can easily convey, such as the number of triples in a dataset or the extent of linking to other datasets, may be difficult to assess[6], as the completeness of a crawl cannot be guaranteed. Also, information about a dataset, such as alternative access methods, which the publisher is aware of, but which is not expressed as Linked Data, are not obtainable. Initially appearing as a drawback, this factor can also be seen as a virtue, as the approach to only obtain information about a dataset expressed in Linked Data is more in accordance to the idea that an automatic agent should be able to discover and process such information, a main goal of Linked Open Data.

As the effort to identify and contact publishers is significant, for our survey, we chose to evaluate best practices based on the data itself. Our is to crawl and obtain samples of Linked Data from a large number of data sources and based on these samples to evaluate the extent the best practices are followed. As we take samples of Linked Data from datasets, information about a dataset might be missed if the sampled resources do not have properties found in other ones of the same dataset, for example used vocabularies. This danger can be mitigated by making the sample sufficiently large, making it less likely important properties of the dataset being missed.

To obtain our sample, we first perform crawls starting from various seeds, taken from a crawl and a data catalogue described below:

Billion Triple Challenge Dataset 2012 [4]: The Billion Triple Challenge Dataset from 2012 was crawled during May/June 2012 by Karlsruhe Institute of Technology (KIT). It was created by importing the DBpedia 3.7 dump, as well as performing three crawls from different seeds. The first group were all examples from the lod-cloud group at datahub.io, from which a breadth-first crawl with 4-hop expansion was performed. The second group were all Freebase URIs to which an identity link existed from the DBpedia dump. Here, no links where expanded. The third group were all other URIs that where at the object position of a triple having an owl:sameAs predicate in the DBpedia 3.7 dump. These were used as the seeds of a breadth-first crawl with a two-hop expansion. Lastly, Tim Berners-Lee FOAF profile was used as seed, performing a breadth-first crawl with six hop expansion.

The resulting corpus includes around 1.4 billion triples from 845 different PLDs. From this corpus, we initially take 500,000 URIs from the subject position of triples. In order to ensure that from every PLD a fair share of URIs is available, we tried to randomly sample the same share from each PLD[7], or less if not enough is available. As PLDs often had less URIs, we then take the remaining sample mass and distribute it

[6] Giving the case that no adequate metadata is available

[7] 500,000 dived through 845

proportionally to the size[8] of every PLD or, if the samples to be retrieved from a PLD is greater than the number of available PLDs, the total number of URIs available.

Secondly, we create another seed list, using the URIs of the Semantic Web Documents (SWDs) (i.e. the URLs appearing at the fourth (context) position of the corpus' quad files). This is done because in the first seed set, too many resources turned out to be non-Linked Data resources, thereby lowering the number of actual seeds too much.

LOD-cloud group at Datahub.io: The lod-cloud group at datahub.io is a catalogue of Linked Data datasets. Information about datasets is entered by dataset publishers, describing properties such as the topic, the size, links to other datasets, and metadata such as license, alternative access methods or contact addresses. At the time of the crawl process, the group contained 340 datasets from 180 PLDs.

When entering information about a dataset, publishers are able to indicate example resources of a dataset. We recovered these examples from the lod-cloud group by using the API provided by datahub.io[9]. Using this approach, we were able to recover examples for 283 (83.24%) datasets, while for 57 (16.76%) datasets, no examples were given. For these datasets, we manually searched for an example resource to be included to the seed list.

Although the BTC2012 also used these examples as a basis for its crawl, we nonetheless repeated to perform a crawl using the example as seeds. As between our crawl and the BTC2012 crawl, around one year had passed, one can assume that the list of available data sources has changed. This gives us the opportunity to include new data sources, which possibly are not included in the BTC2012 crawl. Also, we can ensure that the data from these datasets are up-to-date.

Based on these seeds, we perform crawls, aiming to retrieve a sample of up to 10,000 entities for every PLD. For this, we use LDSpider, a web crawler for the Linked Data Web. [5] Due to limitations of our crawling machine, each iteration of the crawl is conducted separately, meaning for each, we perform a crawl with a depth of zero (practically downloading every URL in the seed list). If after an iteration the threshold of 10000 entities has been reached for a PLD, its resources was removed from the seed list of the next iteration and no resources of the PLD was crawled in further iterations. As this check could only be done between iterations, it is possible that more than 10,000 entities are crawled for a PLD. This was repeated till for all PLDs the limit had been reached or no new documents were found.

In order to further enhance the number of datasets in our crawl, we add datasets identified by the LODStats project.[10] This project computes basic statistics for dumps and SPARQL endpoints found on the web. From 2289 datasets they tried to retrieve and parse, 870 datasets, composed of 717 dumps and 156 SPARQL endpoints, were both retrievable and parsable. After performing our crawl, we cross-checked if LODStats listed dumps from PLDs that did not appear in our crawl. If a dump was successfully analysed by them, we downloaded it too, either the whole dump or a random subset of up to 10,000 entities, if the dump was larger. We left out dumps that contained only a vocabulary and also did not retrieve data from SPARQL-Endpoints.

2.1 Notion of Datasets

With the approach mentioned above, we obtained a set of Linked Data documents, which can be grouped into datasets. According to the definition from the void:Dataset term of the VoID vocabulary, a dataset is: "(...) a set of RDF triples that are published, maintained or aggregated by a single provider. (...) the term has a social dimension: we think of a dataset as a *meaningful* collection of triples, that deal with a certain topic, originate from a certain source or process, are hosted on a certain server, or are aggregated by a certain custodian." [6]

This definition shows that different perspectives can be assumed when defining a dataset, for example by defining it with respect to its topic, the source or process of the data, the host location, the publisher or a combination of them.

8 As number of triples

9 http://datahub.io/api, see http://docs.ckan.org/en/latest/api.html for a description of the API

10 http://stats.lod2.eu/

An approach used in the literature, for example by [3], is to define a dataset based on the host location, more specifically the PLD of the documents it is composed of. While this approach is simple, as only the document URLs are required to group documents to a dataset, for our approach, it also means that one assumes that one publisher has control over a PLD and is consequently the publisher of all data hosted there. On the other side, it ignores the possibility that datasets with very different properties, e.g. topics or used vocabularies, are hosted under one PLD. For instance, at "270a.info", datasets such as the European Central Bank dataset (http://ecb.270a.info), a dataset containing with World Bank data (http://worldbank.270a.info) or a dataset with data from the Food and Agriculture Organization (http://fao.270a.info) are published. Looking at the PLD's main website[11], one can see that a distinction between these different datasets is made there, hinting that one is advised to treat these different datasets individually.

An alternative approach is to use the information of what publishers consider as individual datasets. A source of this knowledge is the lod-cloud group catalogue. There, every entry describes one dataset. These descriptions often contain a namespace, which is common for all documents of the dataset. Naturally, multiple datasets may be hosted at one PLD as in the case of "270a.info", being PLD of eight different datasets described in the catalogue. This way of defining datasets should be more accurate, as publishers themselves differentiate datasets. A limitation for us lies in the fact that not all documents can be attributes to a namespace.

Lastly, the project LODStats declares individual dump files as datasets. When dumps are created, one could assume that their content is meaningful composition of data curated by the publisher. Thus, they may be treated as individual datasets.

One can see that there is no clear cut definition of what a dataset is. Thus, we combine these above mentioned notions of a dataset. We use the approach to subsume data published under one PLD to one dataset, enhancing it with knowledge from dataset publishers conveyed at the lod-cloud group, if available. For the crawl data, we group documents by their PLDs as datasets. If the lod-cloud group holds information about multiple datasets being published under one PLD, we split this dataset further up, by matching the namespaces of the individual datasets with the documents URLs. Lastly, the dumps obtained from LODStats are treated as separate datasets.

This approach is limited by the fact that multiple datasets[12] hosted under one PLD might be treated as one, if no information about them is available in the lod-cloud group. On the other side, it should not be the case that we falsely joined to datasets, as 1) to our knowledge no dataset is hosted at different PLDs 2) we only use knowledge of the publisher themselves to split up datasets, who can be considered the authority regarding their datasets. For dumps from LODStats on the other side, there is the possibility that it would be more intuitive to join them to one dataset.

2.2 Basic Crawl Corpus Properties & Comparison With Other Corpora

With the data gathering and assembling procedure described above, we crawled a sample of 83,237,909 triples and 8,399,097 entities in 743,989 documents coming from 1026 different datasets. These datasets also included vocabularies crawled and datasets with malformed and empty documents, which we removed from our further analysis (see details in the next section). This resulted in a total of 918, with Illustration 1 showing the distribution of the entities with a log scale. Notice that most datasets have less than 100 entities. Also note that 102 datasets have more than 10,000 entities. This is due to limitations of the crawler, which is not able to crawl the precise numbers of entities, as only between crawl iterations the number of downloaded documents was checked.

In order to get a sense of the completeness of our crawl regarding datasets, one can compare it to the ones available in other dataset collections. We compared the overlap of datasets regarding the three collections mentioned above, namely the BTC2012 crawl, the lod-cloud group at datahub.io collection and LODStats. We define overlap as the fraction of datasets of a collection that also appears in our dataset. As all three collections have different approaches for collecting data and also use different notions of what a dataset is, the comparison cannot be done ad hoc, but needs some explanation for every case. A general overview over the properties of the datasets and the coverage of our dataset for every case is given at Table 1.

[11] See http://270a.info/

[12] What the publishers consider individual datasets

The BTC2012 crawl, as outlined above, was assembled from multiple crawls with different seeds as well as importing the DBpedia 3.7 dump. In total, the BTC2012 corpus contains data from 845 different PLDs, ranging from the whole DBpedia dump to small FOAF profiles. Comparing the PLDs present, our dataset on the other side has a higher number of 849 PLDs. Taking a look at the overlap of PLDs, our crawl contains 65.7% of the PLDs present in the BTC2012 are also present in our corpus. As outlined above, we performed crawls using seeds from every PLD. This means one can assume that the data hosted at the remaining 44.3% PLDs are offline. This is quite possible since between the time of our crawl and the crawls performed for BTC2012, one year has passed.

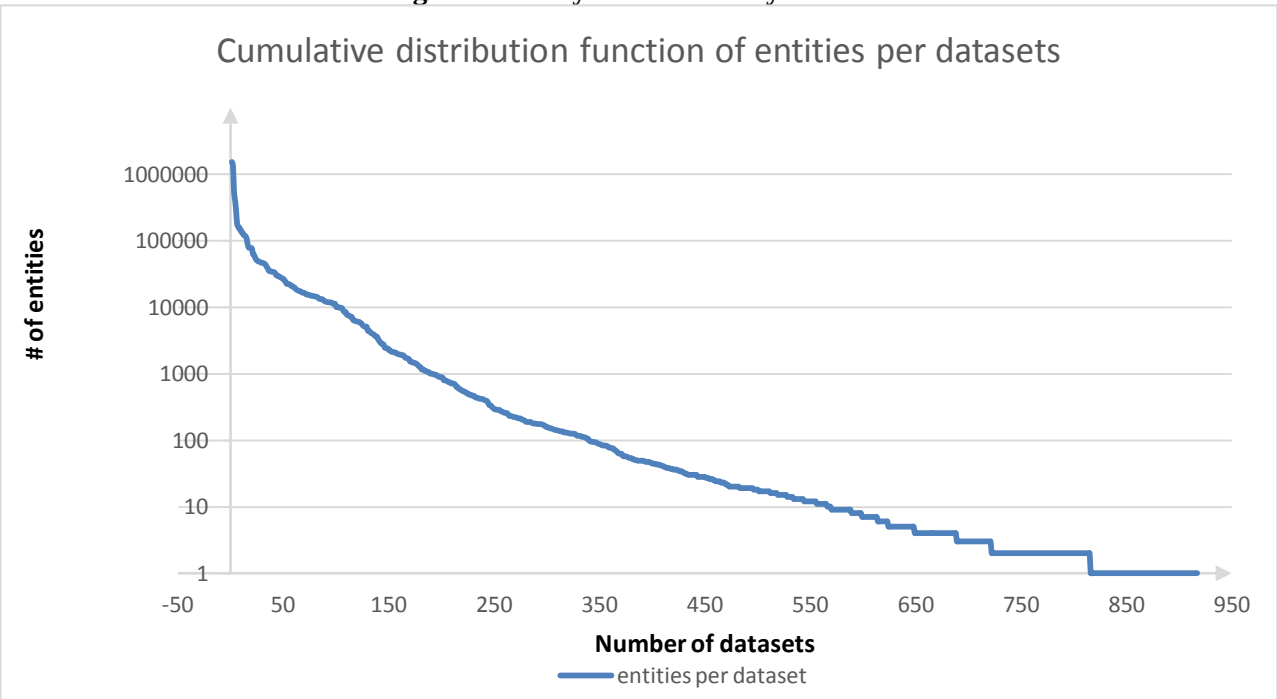
One can also compare our crawl to the data collection from the lod-cloud group at datahub.io. Around the date of our crawl, it housed 340 datasets, coming from 180 different PLDs. The overlap between our crawl and their datasets at that time was 67.8% when considering PLDs. For the lod-cloud group, there are multiple reasons for a dataset not appearing in our crawl particular to the properties of the group. First, datasets, like for the BTC2012 dataset, may have gone offline. This was the case for some datasets we investigated. Also, PLDs of some datasets host a robots.txt that prohibits crawling for all user agents. As we respect robots.txt, these datasets where not crawled. The practice to disallow all agents might not be appropriate for Linked Data, which is predisposed to be accessed by automatic agents.

Lastly, one can compare our crawl to LODStats. Again, we left out SPARQL endpoints. The 717 data dumps come from 229 PLDs. Here, we have coverage of 71.2% Again, like before reasons for this can be the general unavailability of dumps. But during the import process of dumps from PLDs that we did not find in our dataset, we rejected those dumps that contained only vocabularies.

Table 1: Overlap between our crawl and other dataset collections

	BTC2012	**lod-cloud group**	**LODStats**
Size	845 PLDs	337 datasets/180PLDs	717 dumps from 229 PLDs
Date of creation/update frequency	May/June 2012	regularly updated	last updated May 2013
overlap with our crawl	65.7% (PLDs)	67.8% (PLDs)	71.2% (PLDs)

Figure 1: Entity distribution by datasets



2.3 Categorization of Datasets

Published datasets contain data about different topics and are generated by different kind of publishers. In the lod-cloud group, publishers who add their dataset into the catalogue can attach different tags[13] to it, indicating for example that the dataset contains geographic information (category "geographic"), or that it is some form of government data, for example that it was published by a government agency (category "government"). A subset of these tags were also used in [2] for describing different categories. These tags are:

- media
- geographic
- lifesciences
- publications
- government
- user-generated content
- cross-domain

We also included category "social web" from the initial list of tags, used for classifying datasets comprising of social web data, e.g. FOAF profiles or other social web services. Additionally, we introduced category CMS, which includes the output of content-management systems, which example provide blog posts as Linked Data.

In the lod-cloud group, a dataset may contain multiple tags, for example indicating that it is both a governmental dataset and contains data about publications. For instance "http://prov.vic.gov.au", the dataset of the Public Record Office of Victoria, Australia, holds information about the history of Victoria, for example notable places and events. While it is a government agency, it also publishes records on the history of Victoria, which often have also a geographic component.

Although datasets often have multiple topics, we tried to work out the most important tag, as to ease comparison between different groups of datasets by avoiding multiple categorizations, at the prize of losing some classification detail. While doing classification, 22 datasets have more than one category, with 19 having two tags, two having three tags and one having five tags.

The classification of datasets was done manually, by examining the Linked Data itself as well as visiting the website of a datasets if the analysis of the Linked Data deemed no conclusive classification. The result of this classification is shown in Illustration 2. Social web with 40.54% holds by far the most datasets, which comes from the fact that the crawl includes many FOAF-profiles. The second most frequent is category CMS, which houses 133 datasets, or 12.96% of all datasets. The third biggest category, lifesciences, contains a total of 102 (9.9%) datasets, many of them (around 48%) coming from bio2rdf.org. The fourth biggest category with 95 datasets or a fraction of 9.5% is publications.

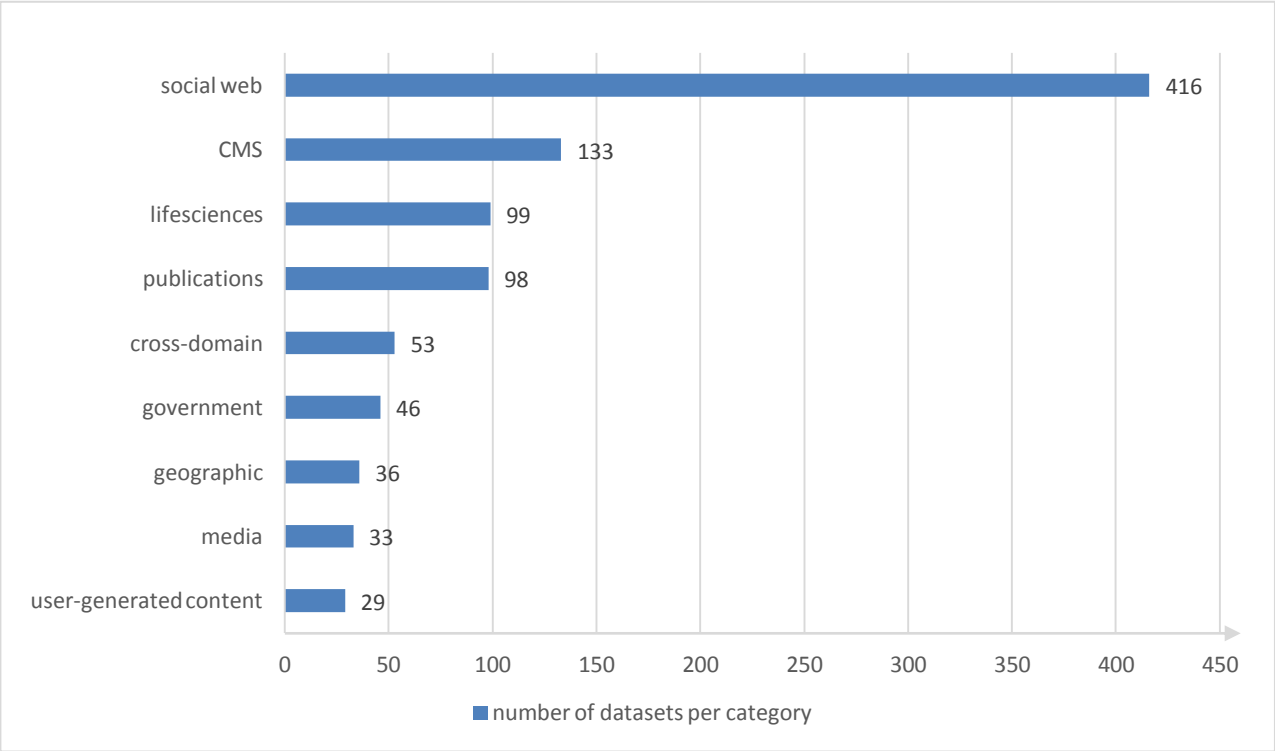
The number of datasets per category is comparatively lower for the other datasets. Cross-domain, which holds for example DBpedia, contains 53 datasets, 5.1% of the whole population. Finally categories geographic, media and user-generated content have 36 (3.5%), 33 (3.2%) and 29 (2.8%) datasets.

Some datasets were unclassifiable, usually if the content of the files downloaded were empty or their contents were malformed. In absolute numbers, these where 21 datasets, a fraction of 2% of all datasets. We excluded these datasets in our further analysis.

During manual categorization, it turned out that the crawl contained a considerable fraction of datasets that constitute vocabularies, meaning that they only contained vocabularies along with meta-information about them. We identified a total of 89 datasets we classified accordingly, which is a fraction of 8.6% of all datasets. We also did not consider these datasets in our further analysis.

This left us with a total of 918 different datasets that we classified into nine different categories. In the following, we will analyse these different datasets regarding their adherence to best practices, putting the results in context to the classes we just defined.

Figure 2: Frequency and quota of datasets in different categories



3 Best Practices

Deliverable 2.1 of the PlanetData project described a set of best practices on how to publish high quality Linked Data. They aim at providing concrete guidelines and recommendations on how to publish data that is as self-descriptive as possible, making it possible to explore the information flexibly by both humans and automatic agents.

These best practices constitute more general rules of how to publish high quality Linked Data. Each of the eight can be expressed as a question directed towards dataset publishers:

1. Does your data provide links to other datasets?
2. Do you provide provenance metadata?
3. Do you provide licensing metadata?
4. Do you use terms from widely deployed vocabularies?
5. Are the URIs of proprietary vocabulary terms dereferencable?
6. Do you map proprietary vocabulary terms to other vocabularies?
7. Do you provide dataset-level metadata?
8. Do you refer to additional access methods?

In this section, we evaluate to what extent they are adhered to with an up-to data sample from datasets in the LODCloud i.e. the set of all Linked Open Data datasets. This sample consists of Linked Data from all 916 datasets crawled and categorized into one of the nine categories.

Each section that describes one of the eight best practices has the same layout. First, we explain the rationale behind a best practice and what benefit its adherence endows to dataset users and the LODCloud in general. Next, we will define how we measure the adherence to the best practice, which is important, as they are aimed for practitioners and tend to be penned more generally, but also because we have to find approaches different from previous ones, given that we evaluate their adherence by analysing data directly. Thus, we not only describe the definition itself, but also discuss its rationale, advantages and limitations.

Lastly, we present the results of our evaluation. After describing the general level of adherence our sample exhibits in its entirety, we go into more detail, describing the level of adherence with respect to the categories of datasets we defined in the last section. If datasets adhere to a best practice, we go further into detail, elaborating how this adherences assumes shape, again also with respect to different categories of datasets.

3.1 Providing Links to other Datasets

This rule corresponds to the fourth rule of the Design Issues on Linked Data [7], which states that publishers should link their data to other Linked Data datasets. Those that adhere to this rule are categorized as a five star dataset, the highest rating according to the rating system for Linked Open Data.

The high rating for interlinked Datasets underlines the importance of this best practice. The benefit of including links to other resources lies in the resulting ability to traverse Linked Data along these connections. By interconnecting a dataset with others, its resources can be put into relation to resources of other datasets, and by following the links, additional information related to a resource can be found. If on the other side a dataset is not connected to others, it becomes an isolated data island, and related information outside of it cannot be discovered automatically, limiting available information to those provided by the publisher.

To create links to other datasets, two steps have to be mastered. First, datasets with information related to the content of the one to be linked have to be identified. Such datasets can for example be found with the aid of catalogues such as the lod-cloud group, were datasets are categorized and described, enabling a publisher to identify suitable candidates for linking. In a second step, individual resources between two datasets have to be connected. Here, tools like Silk [8] can help data publishers to easily connect resources of datasets, lowering the need for manual labour.

3.1.1 Definition

In general, three types of external links can be distinguished: identity links, relationship links and vocabulary links [9]:

- Identity links point to URI aliases of the same real-world object or abstract concept, enabling to find additional information about the same entity. Germany for instance is both described in geonames.org[14] and DBpedia.org[15], with these two resources being connected by an identity link.
- Relationship links on the other side point to resources that are somehow related to the resource the link is outgoing from. A description of a person can for example link to a town in another dataset, indicating that she lives in that town.
- Vocabulary links point to the definition of vocabulary terms used to represent the data. Such vocabulary links make the data self-descriptive and allows integrating data across vocabularies.

When analysing to what extent links to other datasets are provided, we are mainly interested in relationship links and identity links and leave vocabulary links out, as they do not connect two datasets, but a dataset with a vocabulary. This means that according to our definition, external links using the predicate “rdf:type” will be excluded from our analysis.

If a triple exists whose subject and object belong to two different datasets, then this triple constitutes an external link between the two datasets.

External RDF links between two datasets are constituted by a triple were the resources at the subject and object position are each part of the two different datasets. External RDF links that connect to datasets can then be seen as an arc, were the direction depends on to which dataset the linking triple belongs to. This triple can be part of one of three different datasets:

- It can be part of a dataset which namespace is used by the subject. In this case, the triple forms an arc from the dataset whose has the same namespace as the subject to the dataset with the namespace of the object.
- It can be part of the dataset which namespace is used by the object. In this case, the arc runs from the object's dataset to the one of the subject.
- It can be part of a dataset where neither subject nor object uses its namespace. As we aim at taking a look to what extent publishers augment their data with external RDF links, we do not count the third

[14] http://sws.geonames.org/2921044/

[15] http://dbpedia.org/resource/Germany

type of connection as a link between two datasets, as not publishers of the datasets involved, but someone else did the interlinking.

As we crawled a sample of resources for every dataset, we do not always know if the external link a triple constitutes leads to another Linked Data resource, if the targeted resource is not in the sample. It could also point to a non-Linked Data resource, such as a web page, a PDF file or some sort of media file, like a picture. As we want to analyse RDF links, such triples should not be considered to be external RDF links, as they do not pose a connection between Linked Data resources. For being able to only consider links that probably link to another resource, we use two heuristics. First, we only consider links between datasets in our crawl, as we know that they contain Linked Data. Second, we only consider predicates that are used to establish external links between resources that are in the sample, indicating that the triple is used for connecting Linked Data resources. The 634 different predicates complying with this requirement are then manually analysed, to filter out those which are typically not used in our sample to connect Linked Data resources or instance data. Examples for these are DBpedia:externalLink, foaf:homepage or pim:mailbox, but also rdfs:subClassOf, rdfs:subPropertyOf and owl:equivalentClass. In total, we filter out 70 different properties. Additionally, we compare all property names in lower-case, to allow for (mistaken) variations in the capitalization.

To carve out to what extent publishers link their datasets to others, we analyse linking on a dataset level, not individual resources. The rationale behind this is that for the best practice, the fact that the other dataset is found and linked to is of importance, while the extent of linking is less important. Thus, we define the outdegree of a dataset as the number of datasets it links to. The indegree indicates the number of datasets, that link to the dataset in question.

Our approach has naturally some shortcomings. First, we only consider links between datasets we know of. This may lower the outdegree for some datasets which link to datasets not in our sample. Second, we manually filtered out predicates that appeared to us as being mainly used for linking non-Linked Data. While we first may be erroneous in filtering out a certain predicate, others may be used dually, with publishers using a predicate that is usually used to link to non-Linked Data resources for linking their data to Linked Data resources. Both might lower the outdegree of some datasets. On the other side, some links may be included that link to Linked Data datasets, despite the fact that the link leads to non-Linked Data, for example the human-readable main page describing the dataset. In this case, the degree of a dataset will be overrated.

3.1.2 Results

We first take a look at the outdegree of all datasets. In total, 44.10% of all datasets have an outdegree higher than zero with an average outdegree of 3.068. This means that more than half of all dataset do not have any outgoing links at all, showing that generally this best practice is not adhered to universally. Taking a look at the first column at Table 2, one can see the number and fraction of datasets of a category that have outgoing links. Apart from categories CMS and lifesciences, the fraction of datasets with outgoing links is above the global average. This indicates that especially publishers of datasets in these two categories do not adhere to the best practice. Datasets from categories cross-domain, geographic and media on the other side exhibit a higher than average adherence, with up to 60% of the datasets of a category linking to other datasets.

Table 2 shows average outdegree and indegree for the different categories. Taking a look at the discrepancy between indegree and outdegree, one can see three different patterns. First, there are categories were the difference between indegree and outdegree is small. Categories social web, CMS, lifesciences, publications, government and media have a difference between the outdegree and indegree which is less than one. The second group, comprised of categories cross-domain and geographic, have a much higher indegree than outdegree. Datasets from these categories thus more used to reference to, hence the higher indegree and lower outdegree. As can be seen from the top 10 list at Table 3, DBpedia in category cross-domain and geonames.org in category geographic are the datasets with a very high indegree, contributing a large amount to the categories' averages. Lastly, datasets from category user-generated content tend to link to many other datasets, but being referenced by only a small number of datasets. This indicates that the output generated by users is generally not often referenced to by other datasets.

Table 2: number of datasets with outlink, average out- and indegree by category

Category	number of datasets with outdegree > 0 (% of datasets in category)	average outdegree	average indegree
social web	208 (50%)	4.207	3.814
CMS	48 (36.09%)	1.481	1.566
lifesciences	18 (18.18%)	0.976	1.408
publications	47 (47.95%)	2.01	2.598
cross-domain	32 (60.38%)	5.087	7.966
government	22 (52.17%)	0.976	0.604
geographic	21 (58.33%)	0.867	4.7
media	19 (57.57%)	2.034	1.8
user-generated content	16 (55.17%)	6.5	2.143
Total Average	406 (44.22%)	3.068	3.068

Additionally, one can examine a category's average indegree and outdegree relative to the global average. As the in- and outdegree for the categories with similar degrees is relatively similar, one can easily compare them to the global average. Here, one can see that datasets from social web are generally higher interlinked that the average, both regarding in- as well as outdegree. Datasets from category CMS on the other side have on average smaller in- and outdegree compared to the global average. For lifesciences, which has a somewhat stronger difference between the degrees, the outdegree is even smaller compared to the global average. Also, datasets from category publications are not as strongly interlinked regarding in- as well as outdegree, compared to the global average. The lowest connectivity for outdegree as well as indegree is for datasets from category government. Datasets from category media on the other side has a higher general connectivity, albeit it is lower than the global average.

Table 3 shows the ten datasets with the highest outdegree and the ones with the highest indegree. Taking a look at the list for outdegree, we see that bibsonomy.org, which allows users to share bookmarks and literature lists, has with 95 the highest value. This is followed by semanlink.net, a personal information management system, from category user-generated content with an outdegree of 89 and deri.org from category cross-domain with 71. The next five datasets in the list are all from category social web and constitute FOAF-profiles (harth.org) or online communities who often employ an instance of a statusNet, a microblogging server which publishes its content as Linked Data. The last two datasets are from the Open Knowledge Foundation (okfn.org) with an outdegree of 42 and fragdev.com with an outdegree of 35.

Table 3: Top 10 datasets and their categories with highest outdegree/indegree

dataset	Category	outdegree	dataset	category	indegree
bibsonomy.org	social web	95	DBpedia.org	cross-domain	150
semanlink.net	user-generated content	89	geonames.org	geographic	115
deri.org	cross-domain	71	w3.org	cross-domain	82
quitter.se	social web	66	quitter.se	social web	64
harth.org	social web	65	status.net	social web	63
skilledtests.com	social web	57	postblue.info	social web	56
postblue.info	social web	53	skilledtest.com	social web	55
status.net	social web	42	fragdev.com	lifesciences	41
okfn.org	cross-domain	42	russwurm.org	social web	32
fragdev.com	lifesciences	35	morphtown.de	social web	31

Taking a look at the top 10 list of indegree, DBpedia has the highest value with 150. Second is geonames.org, having an indegree of 115. The indegrees for datasets at places three to ten are considerably lower. Dataset w3.org, has an indegree of 82, ldodds.com, and status.net, both have an indegree of 63 and 64. Again, we can see many datasets from category social web on the lower ranks, and fragdev.com from the domain lifesciences at place eight.

Table 4 lists predicates[16] that are most often used for linking by dataset publishers to external resources. The first two columns show how many datasets use foaf:knows to link to others. The last two columns, display the top 10 predicates regarding the number of datasets which have an ingoing link with this predicate.

The top two predicates used for linking datasets are foaf:knows and owl:sameAs. The number of dataset that use them for linking is lower that the number of datasets they link to, indicating that on average, publishers tend to use those predicate to link to multiple datasets.

Predicate foaf:basedNear is also a popular predicate for linking, but it is not used to link to many different datasets. In fact, it is used to link to only seven different datasets, mostly geonames.org (from 87 datasets), DBpedia.org (from 26 datasets) and linkedgeodata.org (from one dataset).

There are some other predicates used by many dataset publishers to point to only a few others. One example is admin:generatorAgent, which is used by 92 datasets, but is linked in most cases to ldodds.com, where an editor for creating FOAF-profiles is maintained. Another one is cc:license, which is used by 25 different datasets, but which only points to one, namely creativecommons.org, to indicate the license.

On the other side predicate rdfs:seeAlso is used by 71 dataset publishers to link to 183 other datasets. Although it is in both twelve case used to link to DBpedia.org and w3.org, the indegree of all other datasets for rdfs:seeAlso is usually low.

Table 4: Top 10 of predicates that have outlinks from and inlinks to datasets

Predicate	#datasets w. outlink	Predicate	#datasets w. inlink
http://xmlns.com/foaf/0.1/knows	179	http://xmlns.com/foaf/0.1/knows	232
http://www.w3.org/2002/07/owl#sameAs	120	http://www.w3.org/2002/07/owl#sameAs	202
http://xmlns.com/foaf/0.1/based_near	113	http://www.w3.org/2000/01/rdf-schema#seealso	183
http://rdfs.org/sioc/ns#follows	88	http://rdfs.org/sioc/ns#follows	94
http://xmlns.com/foaf/0.1/interest	79	http://purl.org/dc/terms/haspart	73
http://www.w3.org/2000/01/rdf-schema#seealso	71	http://xmlns.com/foaf/0.1/account	27
http://creativecommons.org/ns#license	25	http://purl.org/dc/terms/references	27
http://purl.org/dc/terms/license	21	http://xmlns.com/foaf/0.1/currentproject	22
http://xmlns.com/foaf/0.1/currentproject	20	http://rdfs.org/sioc/ns#links_to	21
http://web.resource.org/cc/license	19	http://xmlns.com/foaf/0.1/made	14

In summary, one can see that this best practice is not adhered to sufficiently with less than half of all datasets linking to others. One has to note though that different categories of datasets show different linking patterns, with for example social web having a comparatively high degree of datasets with outgoing links and a high average outdegree, while datasets from category government both having low fraction of linking datasets as well as a low average outdegree. Also, some datasets stand out, having a high outdegree compared to the rest of the datasets, both globally and compared to their category.

[16] Prefixes taken from prefix.cc

3.2 Providing Provenance Data

Best practice number two advises publishers to supply provenance metadata as part of a dataset. Provenance information of a resource "[…] describe entities and processes involved in producing and delivering or otherwise influencing that resource"[17]. This may be as simple as stating who the author of a dataset was and when it was created. It may also include more complex information, for example by describing the origin of the data served and describing steps and agents involved in transforming this initial raw data to the dataset made available by the publisher.

By providing provenance information, applications and users get the opportunity to learn of the origin and the process that led to the data available and the parties involved. Such knowledge enable both users and agents to better understand the data, judge its quality, and assess the level of trust it can into it.

For example, if data consumers want to consume sensor data published as Linked Data, provenance information are of great value to them. When knowing about provenance of a measurement, for example the sensors used for measuring or the algorithms used for transforming raw sensor information to Linked Data, they might be aware of characteristics of the sensor's output and peculiarities of algorithms involved. This can help them to better judge how accurate the data is and to what extent they can rely on it, enabling them to take steps for managing these characteristics.

3.2.1 Definition

For providing provenance metadata, various vocabularies have been developed over time. In 2010, the W3C Provenance Working Group curated a list of provenance vocabularies as the starting point of their activity, describing the state of the art at that point in time[18]. Later, a main result of this work group was the development of the "W3C PROVenance Interchange vocabulary" [19].
Another possibility to identify provenance vocabularies is to search for them in vocabulary catalogues that are available online. One of such catalogues is the Linked Open Vocabularies Project (LOV)[20], which not only describes and reviews a vocabulary, but also categorizes it into different vocabulary categories. One of these, called "Upper & Meta", houses the subcategory "Quality, Provenance and Trust". In this category, one can find vocabularies that are used to describe topics of quality, provenance information and trust issues. From this list, we identified additional vocabularies used for describing provenance information. For identification, we mostly rely on information of the profile page of the vocabulary in LOV, but also examine the vocabulary itself. We only consider vocabularies that are used for provenance information, which are reachable and which are a final version, not a draft or begin in development.

Based on our own experience, we also add the MetaVocab[21] to the list of provenance vocabularies. It defines two properties, admin:generatorAgent and admin:errorReportsTo, which are used to denote the generator tools used to created Linked Data and an URI to reference to a contact in case of errors. It has found some application in CMS systems (see results section for details).

Using these sources, we identified a total of 17 different vocabularies, which are used to express provenance information. A list of all vocabularies, their name spaces and prefixes[22] (if available) are shown in Table 5. Note that Dublin Core takes a special position in this list, as it is a vocabulary which is not exclusively used for providing provenance information. For example, DBpedia uses dct:subject to indicate the category of a resource. In order to see which data publishers use Dublin Core for expressing provenance information, we manually review all predicates of the vocabulary, identifying those which are usable for expressing provenance information.[23] These are: dct:contributor, dct:creator, dct:date, dct:publisher, dct:source,

17 http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

18 http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

19 http://www.w3.org/ns/prov#

20 http://lov.okfn.org

21 http://webns.net/mvcb/

22 Taken from http://prefix.cc

23 If the property was included in dc-elements, it was also used with this namespace

dct:provenance, dct:source, dct:accuralPolicy, dct:accrualPeriodicity, dct:available, dct:conformsTo, dct:coverage, dct:created, dct:isVersionOf and dct:modified.

Table 5: Provenance Vocabularies, including namespace and prefix (taken from prefix.cc)

Vocabulary	namespace (prefix)
Open Provenance Model	http://purl.org/net/opmv/ns (OPMV) http://openprovenance.org/model/opmo (OPMO) http://openprovenance.org/model/opmx (OPMX)
Provenir Ontology	http://knoesis.wright.edu/provenir/provenir.owl# (provenir)
Provenance Vocabulary	http://purl.org/net/provenance/ns# (prv) http://purl.org/net/provenance/types (prvtypes) http://purl.org/net/provenance/files http://purl.org/net/provenance/integrity (prviv)
Proof Markup Language	http://inference-web.org/2006/06/pml-provenance.owl (PML-P) http://inference-web.org/2006/06/pml-justification.owl (PML-J) http://inference-web.org/2006/06/pml-trust.owl (OML-T)
Dublin Core (elements and terms)	http://purl.org/dc/elements/1.1/ (dc/dce) http://purl.org/dc/terms/ (dc/dct) http://dublincore.org/documents/dces/ (dc)
Premis	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl# (premis)
Web of Trust	http://xmlns.com/wot/0.1/# (WOT)
Semantic Web Applications in Neuro-medicine Ontology	http://purl.org/swan/1.2/discourse-elements/ (swande) http://purl.org/swan/1.2/discourse-relationships/ (swandr) http://purl.org/swan/1.2/pav/ (swanpav) http://purl.org/swan/1.2/qualifiers/ (swanqs) http://purl.org/swan/1.2/swan-commons/ (swanco) http://purl.org/swan/1.2/citations/ (swanci) http://purl.org/swan/1.2/agents/ (swanag) http://purl.org/swan/1.2/qualifiers/ (swanq)
Semantic Web Publishing Vocabulary	http://www.w3.org/2004/03/trix/swp-2/ (swp)
Changeset Vocabulary	http://purl.org/vocab/changeset/schema# (cs)
W3C PROVenance Interchange	http://www.w3.org/ns/prov# (prov)
The Cert Ontology	http://www.w3.org/ns/auth/cert# (cert)
The Data Quality Management Vocabulary	http://purl.org/dqm-vocabulary/v1/dqm# (dqm)
Evaluation and Report Language	http://www.w3.org/ns/earl# (earl)
Open Annotation Data Model	http://www.w3.org/ns/oa# (oa)
Provenance, Authoring and Versioning	http://www.w3.org/ns/oa# (oa)
PML2 provenance ontology	http://inference-web.org/2.0/pml-provenance.owl# (pmlp)
Trust Assertion Ontology	http://vocab.deri.ie/tao# (tao)
Vocabulary for Dataset Publication Projects	http://data.lirmm.fr/ontologies/vdpp# (vdpp)
Vocabulary Of Attribution and Governance	http://voag.linkedmodel.org/voag# (voag)
MetaVocab	http://webns.net/mvcb/ (admin)

Thus we define a dataset to adhere to the best practice of providing provenance metadata if it uses one of the vocabularies outlined above, i.e. it has triples with predicates defined in the vocabulary.

It is important to note that different vocabularies are used to describe different aspects of provenance. For example Dublin Core is more tailored to documents, where the author of a document or a piece of information is central. On the other side, "advanced" provenance vocabularies, such as prv or prov[24], are developed to indicate different transformation steps of data and the agent involved, thus enable to deliver more sophisticated and complex descriptions of the origin history of data.

Our approach has naturally some shortcomings. First, our sample might not include provenance data of a dataset as it was included in the crawl. Also, provenance information might have been expressed in a vocabulary not on the list, for example with a vocabulary defined by the publisher. Lastly, provenance vocabularies, as described above for Dublin Core, might have been used to for example to describe the provenance of the resource itself, not the data.

3.2.2 Results

Taking a look at the results for the usage of different provenance vocabularies at Table 6, one can see that of all vocabularies listed by us, only a small fraction is used for expressing provenance metadata. Of the 21 vocabularies that are outlined above, only four are actually used by the datasets in our corpus.

Table 6: Usage of different provenance vocabularies

Prefix	Vocabulary	# datasets	Fraction of all datasets
dc dct	http://purl.org/dc/elements/1.1/ http://purl.org/dc/terms/	332	36.17%
Admin	http://webns.net/mvcb/	154	16.78%
Prv	http://purl.org/net/provenance/ns#	15	1.63%
Prov	http://www.w3.org/ns/prov#	15	1.63%
Any provenance vocabulary		417	45.42%

Of all datasets examined, 417 datasets (45.42%) use at least one provenance vocabulary. Of all vocabularies described above, Dublin Core hast the highest usage fraction. A total of 332 datasets, which is 36.17 percentage points of all datasets, use this vocabulary to express provenance information. The second most used vocabulary is MetaVocab, which is used by 154 datasets or 16.78 percentages of all datasets. Vocabularies that are specialized provenance vocabularies prv and prov, are only used a few times, with both of them by 1.63% of all datasets.

As the sum of the datasets that use any provenance vocabulary in Table 6 is higher than the sum of the individual vocabulary users, some datasets use multiple vocabularies for expressing provenance information. In total, 318 datasets use only one vocabulary, of which 236 use Dublin Core and 76 use MetaVocab. Five use prov and two use prv. Two vocabularies are used by 96 datasets, with nearly always one of the vocabularies being Dublin Core. In 76 cases the second vocabulary is MetaVocab, in ten cases prv and in nine cases prov. One dataset uses prv in combination with MetaVocab. Lastly, two vocabularies use three provenance vocabularies. Both use Dublin Core and prv, while one uses MetaVocab as the third vocabulary, the other one using prov as the third one.

Thus, we see that dc/dct is rather used complementary to other provenance vocabularies. Vocabularies used for supplying more complex information on the other side are used rarely, with only a small fraction using vocabularies such as prov or prv.

Next, we take a look at the usage of different vocabularies within different dataset categories, where some differences on the usage patterns of different categories of datasets. Regarding the usage of provenance

[24] Provenance vocabulary (prv) and W3C PROVenance Interchange (prov), in the following referenced by their prefix

vocabularies in general, datasets from categories publication and government have a high quota. , often use provenance vocabularies. They have a high fraction of datasets using Dublin Core and in the case of governmental datasets, also advanced provenance vocabularies find some usage while for publications, MetaVocab is used relatively often.

In categories CMS, geographic and user-generated content, around 55% of all datasets provide some form of provenance information. For these, user-generated content and CMS both show a similar pattern, with Dublin Core being used in around 45% of all cases while MetaVocab is being used in 20-28% of all cases, while advanced vocabularies are not used very often. The latter is even more true for the category CMS, where only one dataset uses an advanced provenance vocabulary. Datasets from category geographic on the other side more often use Dublin Core, but only in one case MetaVocab, while advanced vocabularies are used relatively often.

Regarding the fraction of datasets providing provenance information, categories cross-domain and media form a third group. Datasets from the category media more often use Dublin Core, while never using advanced provenance vocabularies. MetaVocab is used in both categories to a similar amount.

Category social web has the most datasets which use provenance vocabularies but as to the size of the category, the fraction is comparatively low. Most often, Dublin Core and MetaVocab is used while advanced provenance vocabularies on the other side are used barely.

Finally, datasets from category lifesciences have with 22.22% the lowest quota of provenance vocabulary usage. Like other groups, they mainly use Dublin Core, while MetaVocab and advanced provenance vocabularies are not often used.

Table 7: Usage of provenance vocabularies by dataset category. The percentage values are relative to all datasets of a category

Category	Dataset usage			# using any
	Dublin Core	MetaVocab	PROV/PRV	
social web	97 (23.32%)	86 (20.67%)	5 (1.2%)	153 (36.78%)
CMS	61 (45.86%)	38 (28.57%)	1 (0.75%)	70 (52.63%)
lifesciences	19 (19.19%)	5 (5.05%)	3 (3.03%)	22 (22.22%)
publications	62 (63.27%)	10 (10.20%)	6 (6.12%)	68 (69.39%)
cross-domain	17 (32.08%)	5 (9.43%)	4 (7.55%)	24 (45.28%)
government	30 (65.22%)	0 (0.0%)	7 (15.23%)	30 (65.22%)
geographic	19 (52.78%)	1 (2.78%)	3 (8.33%)	20 (55.56%)
media	14 (42.42%)	3 (9.09%)	0 (0.0%)	14 (42.42%)
user-generated content	13 (44.83%)	6 (20.69%)	1 (3.45%)	15 (51.72%)

As it shows some distribution, we finally take a look at the usage of the MetaVocab, specifically the property generatorAgent, indicating which agents are used to create Linked Data of the dataset. Table 8 shows the generator agent used by more than one dataset. We omitted version variances, which especially occurred for wordpress.org and movabletype.org.

One can distinguish two different kinds of generator agents. The first kind are scripts for creating FOAF- or Description of a Project (DOAP) files, like foaf-a-matic, doap-a-matic, foafgenerator and foaf-O-matic. Especially foaf-a-matic has been used relatively often, being referred to by 79 datasets. On the other side, we see CMS and blogging tools and services like wordpress, movabletype and typepad, which also annotate their content with Linked Data.

In summary, one sees that less than half of all datasets provide provenance information, indicating that this best practice is adhered to a certain extent. While Dublin Core sees a wide use, only a small fraction of provenance vocabularies are used, and of the ones being used, they are used only by a small fraction of all

datasets. MetaVocab on the other side has seen some distribution especially for creating simple FAOF or DOAP files and in CMS systems. Again, we see differences between categories of datasets, with datasets from categories from publications and government showing a high fraction of datasets providing provenance information, while for example datasets from category lifesciences having a relatively small fraction.

Table 8: Generator agents which are used by more than one dataset

generator agent	# datasets declaring generator agent
http://www.ldodds.com/foaf/foaf-a-matic	79 (50.32%)
http://wordpress.org	29 (18.47%)
htpp://movabletype.org	20 (12.74%)
http://doapy.bonjourlesmouettes.org/doap-a-matic	10 (6.37%)
http://www.typepad.com/	3 (1.91%)
http://www.dcs.shef.ac.uk/~mrowe/foafgenerator.html	2 (1.27%)
http://www.okkam.org/foaf-O-matic/	2 (1.27%)

3.3 Provide Licensing Data

The third best practice states that a publisher should provide licensing metadata. A dataset should be as self-descriptive as possible, including legal restrictions that apply to its use. A common way to express such restrictions is to indicate a license, outlining what a user is legally allowed to do with the data, under which terms he is allowed to use the data and which restrictions apply.

Ideally, such license information is available in a machine-processable form, allowing an automatic agent to automatically decide if it can use a dataset in accordance to the restrictions it is programmed with. For example, an intelligent agent used by a company should not use datasets which prohibit commercial use of the data, as the agent's output should be commercially exploitable. Another example would be the automatic management of attributions, where an automatic agent can dynamically add attributions to its output, depending on the datasets used and the attribution terms attached to it.

3.3.1 Definition

In order to provide licensing information, one can use predicates from different vocabularies to link a resource to license information. Such properties exist in various vocabularies, for example dct:license and dc:rights from Dublin Core or cc:license from the Creative Commons vocabulary[25]. Also, the indication of waiver statements is possible, indicating that the publisher abandons certain rights he has on the data. For this, the waiver vocabulary[26] might be used.

To capture all possible properties used for indicating the license of a dataset, we follow the approach from [3]. In a corpus of datasets, they searched for predicates containing the string "licen". Following this approach, we obtain a list of 37 predicates from our corpus.

Then, they filtered out obviously irrelevant properties. Such properties are not used to indicate the license of the data it is used in, but license of the resource it describes, like a piece of software. We filter such predicates out, by manually assessing if a property was used to indicate the license of the dataset or rather the license of the resource the data is about. Using this strategy, we filter out 28 property terms. Three terms were not dereferencable, which will be marked in our further analysis[27]. This leaves us with nine predicates on our list.

[25] http://creativecommons.org/ns#, with prefix “cc”

[26] http://vocab.org/waiver/terms/, with prefix “wv”

[27] Here, one can argue that on the one side, an intelligent agent would not be able to understand the predicate while on the other side, the data publisher showed an effort to indicate a license

Like [3], we also add the two predicates dc:rights and dct:rights, as advised by [2] to the list and we ourselves also included predicates from the waiver vocabulary. This leaves us with a total of twelve properties, which are used to indicate a license. The predicates identified for indicating a licensed are listed in the first column of Table 9.

When providing licensing information, two principle ways for publishing them as Linked Data are possible. First, license information can be published individually for every resource. This means that the description of a resource includes triples indicating rights and waiver statements. As an alternative, one can publish licensing information for a whole dataset. In this case, licensing information is published at one central point, for example in a VoID file. In our analysis, we both checked for these triples in the dataset itself as well as in VoID files, if they are available[28].

We thus evaluate license usage by searching for the number of datasets providing licensing information through the predicates described above, defining a dataset to provide licensing information if it uses one of those predicates.

The approach we chose may not find license statements of datasets, due to the sample nature of the datasets. This is not necessarily a drawback, as an intelligent agent should easily find license information without searching the whole dataset. Secondly, predicates that indicate a license might either indicate the license of the resource talked about (for example a piece of software) or the license of the data itself. While we excluded predicates based on the usage we observed, no clear distinction is made between these two cases with regard to the predicates. If license information are attached to the resource, which is why we might have wrongly included or excluded license information if the usage of predicates was unusual. This is a generally shortcoming which we cannot solve, but which makes it necessary to indicate more clearly to what the license refers to, through special predicates solely used for dataset licenses or through some other mechanism.

3.3.2 Results

In total, 155 datasets, which is 16.88% of all datasets in our corpus, provide some form of license information. Taking a look at Table 9, one can see that 7.30% of all datasets use cc:license or its redirect.[29]

Table 9: Usage of property terms used to indicate license information by number of datasets. Stars indicate non-dereferencable terms

Term	# of datasets (fraction of all datasets)
http://creativecommons.org/ns#license http://web.resource.org/cc/license	67 (7.30%)
http://purl.org/dc/terms/license	47 (5.12%)
http://purl.org/dc/terms/rights	43 (4.68%)
http://purl.org/dc/elements/1.1/rights	29 (3.16%)
http://purl.org/dc/elements/1.1/license	8 (0.87%)
http://www.w3.org/1999/xhtml/vocab#license	2 (0.22%)
http://vocab.org/waiver/terms/norms	1 (0.11%)
http://purl.org/dc/terms/licence*	1 (0.11%)
http://creativecommons.org/schema.rdflicense*	1 (0.11%)
http://data.semanticweb.org/ns/misc#licenseDoc*	1 (0.11%)
http://vocab.org/waiver/terms/waiver	0 (0%)
Any license predicate	155 (16.88%)

[28] See section 8 for details

[29] http://web.resource.org/cc/license redirects to the original cc:license definition

Predicate dct:license is the second most often used predicate, being used by 5.12% of all datasets. Property dct:rights and dc:rights are used by 4.68% and 3.16% of all datasets. All other terms are used by less than one percent of all datasets. Non-dereferencable predicates are marked by a star. As one can also see, these terms are used by only one dataset each, showing that their usage is not widespread.
The indication of license information for different categories is described in Table 10. As one can see, half of all datasets from category government adhere to the best practice. Followed is this category by lifesciences datasets, for which 30.3% provide license information.

Categories publications and user-generated content both a similar fraction of datasets providing license information. For both of them, around one quarter of all datasets provide some form of licensing information. The quota is a bit lower for categories cross-domain and geographic, where 22 to 23 percentages of all datasets provide license information.

Within the last three dataset categories, CMS, media and social web, less than ten percentages of all datasets provide license information. One should note that social web and CMS are the two largest categories, yet they exhibit very low quotas. This is especially true for social web, which is the largest category, but has the smallest quota.

Table 10: Number of datasets using license predicates per category

Category	# of datasets having license statements (fraction)
social web	35 (8.39%)
CMS	13 (9.77%)
lifesciences	30 (30.30%)
publications	24 (24.49%)
cross-domain	12 (22.64%)
government	23 (50.00%)
geographic	8 (22.22%)
media	3 (9.09%)
user-generated content	7 (24.14%)
Total	155 (16.88%)

After we investigated the general provision of licensing information, we also take a look at the license information itself, i.e. the information at the object position of a triple that has one of the predicates listed above. For this, we devised a scheme for arranging different kinds of information, which can be seen in Of all datasets that link to an URI, 63.97% link to one of the licenses outlined above, making their license terms understandable for automatic agents. 47.05 percentages of all datasets refer to an URI which is not a known license. Of datasets that use a known license, 58.62% refer to a license that is OKFN-conformant, while 41.37% refer to a creative commons license that is not OKFN conformant, because they include one of the rights that disqualify them as an open license, for example a non-commercial clause.

Figure 3. It was created by partitioning licensing information in an iterative process. Starting from all licensing information provided, we separated the information into groups until we had a classification tree constituting a meaningful division.

A first partition can be achieved if we separate the license information based on whether they are an URI or a literal. From all datasets providing licensing information, 87.74% provide licensing information in form of an URI, while 19.35% provide only a literal text.

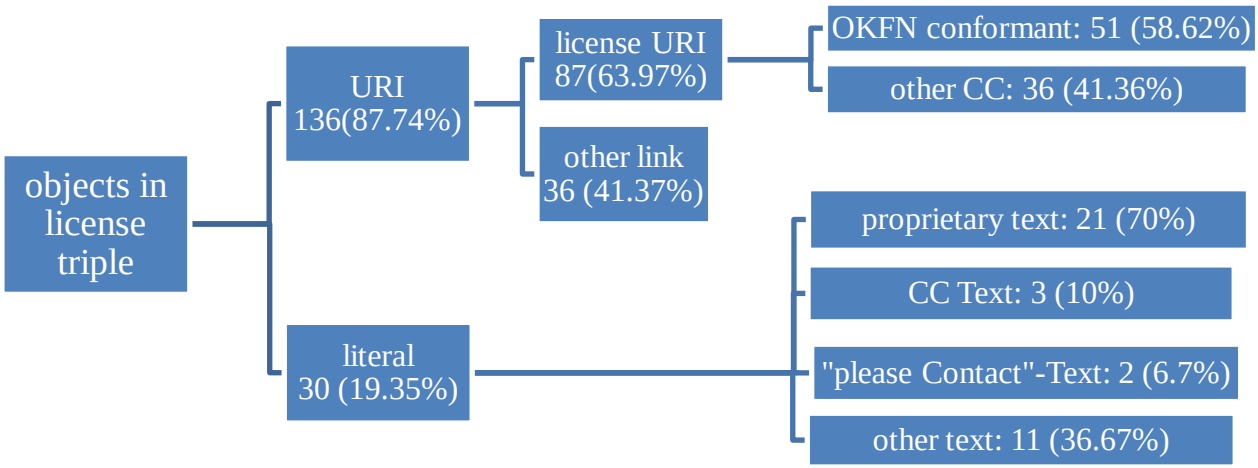
For the datasets which provide URIs, we are mainly interested if their URI references a Linked Data resource, which would make their licenses understandable for an automatic agent. For this, we look if some forms of creative commons license or one of the licenses that are considered by the Open Knowledge-Foundation to be “open” [30] are stated. These licenses are creative commons licenses without non-commercial

[30] http://opendefinition.org/licenses/

or non-derivative works clauses, Open Data Commons licenses as well as some other licenses, such as the open government licenses from the UK and Canada.

Of all datasets that link to an URI, 63.97% link to one of the licenses outlined above, making their license terms understandable for automatic agents. 47.05 percentages of all datasets refer to an URI which is not a known license. Of datasets that use a known license, 58.62% refer to a license that is OKFN-conformant, while 41.37% refer to a creative commons license that is not OKFN conformant, because they include one of the rights that disqualify them as an open license, for example a non-commercial clause.

Figure 3: Hierarchy for classification of license information. Percentages are relative to the number of datasets on step above in the hierarchy



On the other side, 30 datasets provide a literal as license information. Further analysing these, we search if the text contains “copyright”, “rights reserved” or some variation of it[31], or if contained the symbol “©”. Based on our experiences, such keywords indicate a copyright notice by the publishing party. Of all datasets that publish literals for license information, 70% included those strings. Secondly, we analysed if a creative commons license was specified in the text by searching for the string “creative commons”. This was true for 10% of datasets providing textual license information. Lastly, we searched for strings were one was asked to contact a party regarding licensing information by searching for the strings “information” and “contact” or some variation of it. With this, we found that two datasets or 6.7% of those who provide textual descriptions provide such information. Lastly, there was a group of “other text” which were those not categorisable above, including strings without any comprehensible content.

In summary, a relative small fraction of datasets provide license information, leading to the conclusion that this best practice is only partially followed. Datasets from category government have the highest fraction of datasets with attached license information, but this fraction is still only 50% of all datasets. Other categories have a much lower fraction, and especially category social web, media and CMS have fractions of less than 10%. The information provided is often an URI, the majority of them being creative commons or OKFN conformant licenses.

3.4 Using terms from widely deployed vocabularies

The best practice to use terms from widely deployed vocabularies advises publishers to use vocabularies that have seen a widespread adoption in the LOD-Cloud. Using such vocabularies helps automatic agents to understand the data more easily. Vocabulary re-use also makes the development of agents and data integration easier, as resources from different datasets are more often described with the same terms. This eases ontology alignment between the vocabularies used by two different datasets, or makes it unnecessary altogether.

[31] We accounted for spelling mistakes and some variations, for example “right reserved”

3.4.1 Definition

For assessing whether a vocabulary is used in a dataset, we need to clarify what we consider usage. For this, we distinguish two cases. First, a triple can define a resource to be of a certain type using the rdf:type predicate, in which case a term of a vocabulary is at the object position, constituting the use of the vocabulary. Second, the use of a predicate in any triple of a dataset (including rdf:type) constitutes the use of the vocabulary it is part of by the dataset. A datasets thus uses a vocabulary if one of its terms appears: 1) at the predicate position of a dataset's triple or 2) at the object position of a dataset's triple defining the class of the subject. An important point to note is that we do not count the usage of RDF when serialising Linked Data as RDF/XML as the usage of the RDF vocabulary. First, this has a practical reason, as the crawler used converts crawled document into quads, meaning that the original serialization gets lost. Second, in such cases, RDF is not used to represent knowledge about a resource, but for serializing this knowledge.

Additionally, the evaluation of this best practice requires drawing a line between proprietary and widely used (or non-proprietary) vocabularies, with proprietary meaning that a vocabulary "belongs" to a dataset, a distinction which will also be relevant for the two best practices to follow. For attaching a vocabulary to a distinct dataset, thereby defining it proprietary, the aspects of hosting or of usage can be considered as a criterion. In the first case, the vocabulary is proprietary if it shares the PLD with the dataset using it while in the second case, the vocabulary is proprietary if it is used by only one dataset. Both views have their advantages and shortcomings for characterizing the one to one relation of vocabularies and datasets.

For both definitions, the case where a vocabulary and a dataset share the same PLD can be ambiguous. The dataset can be a showcase for the vocabulary, without the intent of the vocabulary creators to create the vocabulary for the dataset, but making it proprietary for both definitions. On the other side, the vocabulary might have actually created for use in a specific dataset, in which case the definition based on the hosting aspect would be adequate.

On the other side, a vocabulary created for a certain dataset might over time be adopted by other dataset publishers. With more datasets using it, it would become non-proprietary, regardless of the original intent behind the vocabulary creation. A definition based on the hosting site of the vocabulary on the other side would always define such a vocabulary proprietary, regardless of its distribution in the LODCloud.

Taking these aspects into account, we decide to base the notion of a proprietary vocabulary on its usage, defining a vocabulary as proprietary it if used by only one dataset. The reason for this lies in the fact that for Linked Open Data, the actual use of vocabularies is of great importance. A vocabulary used by many other datasets, regardless of where it is hosted, should be considered to be non-proprietary, and thus its use should be advocated. Because the actual use of a vocabulary, not who defines it, incites application builders to tailor their application towards it and makes it easier to integrate a dataset with others.

3.4.2 Results

Following [2], a first point to look at is the extent "standard vocabularies", namely RDF, RDFS and OWL are used. We found a total of 853 datasets, or 93% of all dataset, use one of these vocabularies. The lowest quota of their usage is by datasets of category lifesciences, where only 53% of all datasets use one of these vocabularies. Media, cross-domain and government have higher quotas with 69, 75 and 76 percentage points. They are followed by category geographic with 80 percentage points and category user-generated content with 89%. The highest usage of standard vocabularies is in categories CMS (93%), publications (94%) and social web, also with 94%. As noted above, we did not count the usage of RDF for encoding RDF/XML, in which case the rate of use would be much higher. Even so, the result shows that nearly all datasets use standard vocabularies in their datasets.

Next, we take a look at the usage of the other, non-standard vocabularies. Vocabularies that have no namespace in w3.org, purl.org or xmlns.org were reduced to the PLD of the namespace in order to ease the issue of distributed or modular vocabularies, were it is hard to assess if vocabulary fragments belong together. For the top list, we substituted the PLD for the whole name space if we knew only one vocabulary was hosted, in order to make it clearer which vocabulary was meant.

The most used vocabularies with more of 5% usage by datasets can be seen in Table 11. Friend-of-a-friend shows the highest usage with nearly 70 percent. Following this, the quota of usage is much lower, with dce used in 34% of the datasets and dcterms used in 30% of the cases. Following this, we see 26% of use of the

wgs84/pos vocabulary, used for geographical coordinates and both around 17% for SIOC and the MetaVocab vocabulary.

Taking a broader look at the issue of widely used vocabularies, we analyse to what extent different categories make use of proprietary vocabularies. As pointed out, proprietary vocabularies are those which are used by only one dataset. Of all 622 vocabularies we encountered, 393 (63.18%) vocabularies can be considered proprietary according to our definition, while 229 (36.82%) are non-proprietary. The latter one also includes RDF, RDFS and OWL.

Table 11: Vocabularies used by more than 5% of datasets with prefix, namespace, usage count and quota of usage

Prefix	Vocabulary	# of datasets	Fraction of all datasets
foaf	http://xmlns.com/foaf/0.1/	638	69.57%
dce	http://purl.org/dc/elements/1.1/	315	34.35%
dcterm	http://purl.org/dc/terms/	281	30.64%
pos	http://www.w3.org/2003/01/geo/wgs84_pos#	239	26.06%
sioc	http://rdfs.org/sioc/ns#	159	17.34%
admin	http://webns.net/mvcb/	157	17.12%
bio	http://purl.org/vocab/bio/0.1/	124	13.52%
skos	http://www.w3.org/2004/02/skos/core#	108	11.78%
rss	http://purl.org/rss/1.0/	100	10.91%
w3con	http://www.w3.org/2000/10/swap/pim/contact#	73	7.96%
doap	http://usefulinc.com/ns/doap#	61	6.65%
void	http://rdfs.org/ns/void#	58	6.32%
bibo	http://purl.org/ontology/bibo/	52	5.67%
cert	http://www.w3.org/ns/auth/cert#	51	5.56%
bio2rdf	http://bio2rdf.org/	48	5.23%

In total, 256 (27.89%) datasets use proprietary vocabularies, while 911 (99%) datasets use non-proprietary vocabularies. This means that a very small fraction of datasets (less than one percentage) only use proprietary vocabularies. On the other side, 655 datasets (71.3%) use only terms from non-proprietary vocabularies.

As every proprietary vocabulary is used by only one dataset, which in turn has a categorization, we can attach categories to the vocabularies themselves, describing in which categories proprietary vocabularies are mostly used. This gives us a notion in which areas proprietary vocabularies are more used and where non-proprietary vocabularies find more adoption.

Table 12 displays how strongly proprietary vocabularies are used by datasets from different categories. In the first column, one can see the total number of proprietary vocabularies that are used by a dataset from a category. Note that a dataset can use more than one proprietary vocabulary, making this number higher than the actual number of datasets that use proprietary vocabularies in the dataset. The attached percentage number refers to the quota of all vocabularies used in the category that are proprietary.

Table 12: Usage of proprietary vocabularies by datasets in different categories.

Category	Number of proprietary vocabularies used by datasets	Number of datasets using a proprietary vocabulary
social web	82 (34.45%)	63 (15.14%)
CMS	68 (45.3%)	56 (42.11%)
Lifesciences	29 (31.18%)	21 (21.21%)
Publications	69 (32.7%)	39 (39.8%)

cross-domain	57 (35.18%)	13 (24.53%)
Government	27 (27.84%)	19 (41.3%)
Geographic	26 (28.57%)	16 (44.45%)
Media	17 (21.78%)	15 (45.45%)
user-generated content	18 (25.35%)	14 (48.28%)
Total	393 (62.18%)	256 (27.89%)

Many vocabularies are used in datasets from category social web, followed by the categories publications, CMS and cross-domain. Proprietary vocabularies are used less in category government, geographic, user-generated content and media. Especially in category CMS, many of the vocabularies used are proprietary.

The last column indicates the number of datasets that use proprietary vocabularies in each category. The percentages are relative to the number of datasets per category. Category social web has with 15% a low fraction of datasets which use proprietary vocabularies. Categories cross-domain and lifesciences both have a quota of around on quarter. All other categories have a fraction of around 40 to 48 percentage points.

In summary, nearly all datasets use at least one non-proprietary vocabulary, which is why one can say that the best practice is followed. While nearly all datasets use standard vocabularies such as RDF, OWL and RDFS, widely-distributed vocabularies include FAOF, Dublin Core, WGS84 vocabulary and SIOC. The usage of proprietary vocabularies differs between categories. While category social web has a few datasets using proprietary vocabularies, other categories such as CMS, government or user-generated content have a relative high quota.

3.5 Dereferencability of proprietary vocabulary terms

The dereferencability of proprietary vocabulary terms plays a crucial role in understanding a dataset that is using such a vocabulary. An agent with no specific configuration can only understand a dataset if the definition of its proprietary terms is available. This is both true for datasets that are widely used, for example FOAF, but especially important if the vocabulary is not used broadly, but only by one single dataset.

For example, if an automatic agent encounters a term it does not know, it should be able to dereference it, obtaining its definition. From it, the agent may learn what properties with which domains the term has, if it is related to other terms, for example being a subclass of another, more widely used term, or it can retrieve a human-readable description of the term. With such knowledge, an automatic agent can process a resource using the term more easily, having learned about the term's definition.

If on the other side, the definition of a term from a proprietary vocabulary is not known, then the meaning of the term remains unknown to an agent or even a human, making it impossible to use of the knowledge encoded with this vocabulary without the involvement of human labour.

3.5.1 Definition

To measure the dereferencability of a proprietary vocabulary, we attempt to retrieve a definition for every term of a proprietary vocabulary. Only successful responses delivered valid Linked Data are counted as a dereferencable term. This means that for example a http-get request for the term, which returned a file with malformed content, is not counted as a dereferencable, as such content would be of no use for an automatic agent.

For every vocabulary, we then define the quota of terms that are dereferencable. This leads to three vocabulary categories. The first includes vocabularies where all terms are dereferencable, i.e. which have a dereferencability quota of one. The second group consist of vocabularies that are partially dereferencable, meaning not all terms using the namespace of the vocabulary are dereferencable. Finally, non dereferencable vocabularies, those where no term used by datasets can be dereferenced, form the third group.

3.5.2 Results

The results for the dereferencability of proprietary vocabularies are shown in Table 13. Of all proprietary vocabularies, nearly 280, or more than two third of all vocabularies, are not dereferencable at all. Given the importance of this rule, this is a very high value. Of the remaining vocabularies, 66 (16.78%) are fully

dereferencable, meaning that for every term, a definition in form of Linked Data was returned, while 47 vocabularies, or 11.95% of all proprietary vocabularies, are partially dereferencable.

Table 13: General dereferencability of vocabularies

Vocabulary which are fully dereferencable	66 (16.78%)
Vocabularies which are partially dereferencable	47 (11.95%)
Vocabularies which are not dereferencable	280 (71.24%)

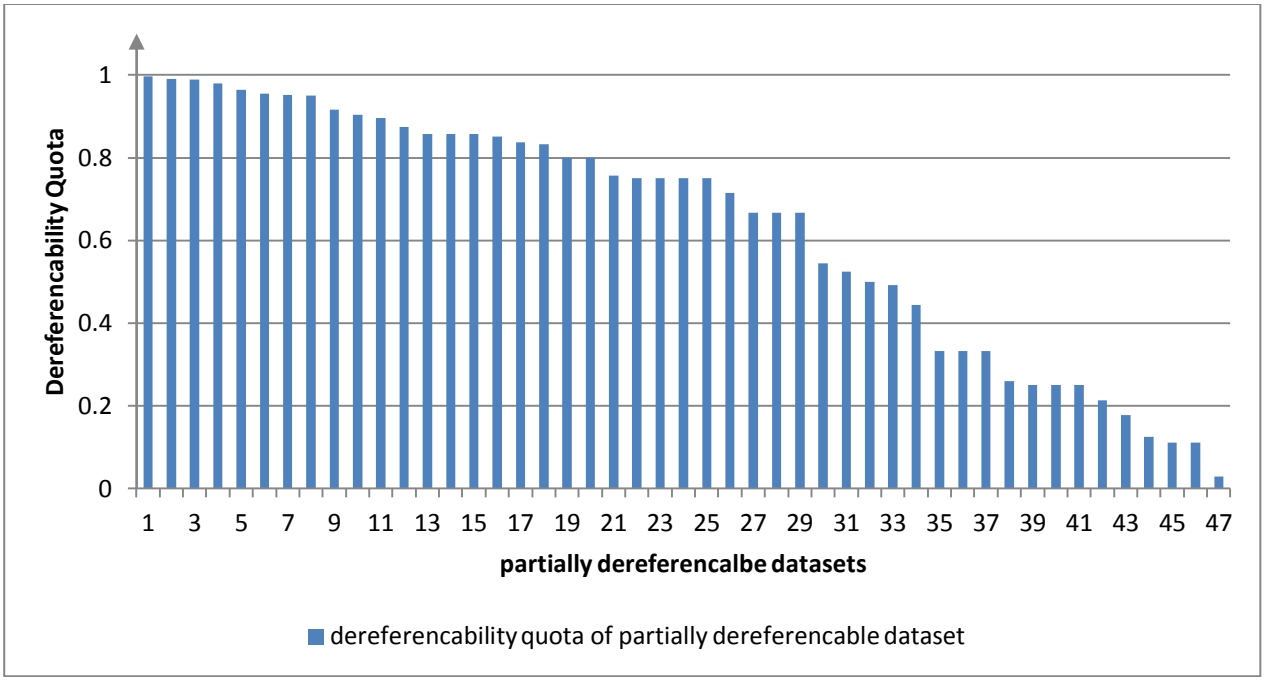
Figure 4 plots the dereferencability quotas in descending order for the 47 partially dereferencable vocabularies. Ten vocabularies have dereferencability quotas higher than 0.9, 15 vocabularies have a dereferencability quota between 0.75 and 0.90. Seven vocabularies have a dereferencability quota between 0.75 and 0.5 percentage points. Finally, 15 vocabularies have a dereferencability quota that is lower than 0.5.

When thinking of the reasons for a partial dereferencability of a vocabulary, some possibilities arise. First, the user of a vocabulary might have used vocabulary terms that are not defined in the vocabulary, or might have made spelling errors using the terms, effectively creating terms in the namespace of the vocabulary. The case that the user of the vocabulary does not own a namespace, but nonetheless (consciously or consciously) mints terms in it without the permission of the owner of the domain, is called namespace squatting.[32] One can assume that such squatting especially occurs for vocabularies with a high dereferencability quota, as it is unlikely that a high quota of non-dereferencable terms is squatted.

In the alternative case, a term that was properly defined in the namespace and has been used for creating a dataset, but after this, its definition has gone offline. As we analysed some vocabularies on a PLD level, it might be possible that a whole vocabulary or the module of a distributed vocabulary has gone offline. This for example seems to be the case for kit.edu, where terms from the namespaces “http://www.aifb.kit.edu/id“, “http://km.aifb.kit.edu/projects/numbers/number#” and “http://people.aifb.kit.edu/nlpdeppos#” have been used in datasets. Terms from the latter were not dereferencable, lowering the fraction of all vocabularies in this PLD negatively.
Similarly, some terms from a vocabulary can have used in a dataset, but afterwards the vocabulary was changed, removing the term that formerly was included in the vocabulary. In this case, the vocabulary owner did not properly handle the maintenance of the vocabulary, for example by keeping vocabulary term definitions online but marking them as dereferencable, to still supply users with the term definition, but also indicating that it is a term that should not be used.

[32] http://www.w3.org/wiki/NamespaceSquatting

Figure 4: Derreferencability Quota of partially dereferencable vocabularies in descending order



Again, we take a look at the difference level of adherence to this best practice by datasets in different categories. Table 14 describes for every category how many proprietary vocabularies are fully, partially or not dereferencable, while the percentages in brackets indicate the quota with respect to all proprietary vocabularies.

Datasets from category publications have the highest number of full dereferencable or partially dereferencable vocabularies. Relative to the number of proprietary vocabularies, the quota is highest for category geographic. The category with the highest number of non-dereferencable vocabularies, both absolute and relative, is category social web. This is also the category with most proprietary vocabularies. In general, no clear connection between the number of proprietary vocabularies and the tendency to make them dereferencable can be found, as for example category media and user-generated content both have a quota of higher than 70% for non-dereferencable vocabularies. In generally, no connection between the number of proprietary vocabularies used in a dataset and the quota of dereferencability can be found.

Table 14: Dereferencability of proprietary vocabularies used in different categories

category	# prop. vocabs	dereferencable	partially dereferencable	non-dereferencable
social web	82	10 (12.2%)	8 (9.76%)	64 (78.05%)
CMS	68	8 (22.81%)	7 (12.28%)	53 (64.91%)
lifesciences	29	6 (20.69%)	2 (6.9%)	21(72.41%)
publications	69	13 (18.84%)	8 (11.59%)	48(69.57%)
cross-domain	57	13 (16.67%)	7 (11.9%)	37 (71.43%)
government	27	3 (11.11%)	7 (25.93%)	17(62.96%)
geographic	26	9 (34.62%)	3 (11.54%)	14(53.84%)
media	17	2 (11.76%)	2 (11.76%)	13(76.47%)
user-generated content	18	2 (11.11%)	3 (16.67%)	13(72.2%)
Total	393	66 (16.79%)	47 (11.96%)	280 (71.25%)

Summarizing the results, a high quota of proprietary vocabularies is not dereferencable at all. As the dereferencability of terms is important for automatic agents, one can conclude that this best practice is not being followed sufficiently. Again, the quota of dereferencable vocabularies varies in different categories.

3.6 Mapping of proprietary vocabularies to others

If a proprietary vocabulary is used, one should map those vocabularies to others. This puts its terms into relation to other, and an automatic agent who encounters a term with a connection to another can put the term to relation.

If for example the agent search for foaf:person resources and encounters a resource of class proprietary:director, he might learn from the term's definition that it is connected through rdfs:subClassOf to foaf:person. Then, the agent is able to treat the resource just like a foaf:person resource, enabling him to make use of it, which would have not been possible in an automatic way if the connection between these terms did not exist.

3.6.1 Definition

To operationalise the mapping of proprietary vocabularies, we can use the data obtained for evaluating the last best practice. For interlinking of vocabularies, multiple possible properties are usable. Bizer et al. listed in [2] terms that should be used for mapping vocabularies:

- owl:equivalentClass
- owl:equivalentProperty
- rdfs:subClassOf
- rdfs:subPropertyOf
- skos:broadMatch
- skos:narrowMatch

Of course, the interlinking of properties can only be evaluated for those proprietary vocabularies which are dereferencable. Special cases are links that define an explicit link to the term owl: Thing. As each user-defined class is implicitly a subclass of owl: Thing, such a connection is redundant and will not be counted explicitly.

We define a proprietary vocabulary to link to other vocabularies if it uses one of the properties mentioned above to link to one or more other vocabularies.

3.6.2 Results

First, we take a look to what extent vocabulary terms link to other vocabularies. Of all 133 proprietary vocabularies, that are at least partially dereferencable, 37 or 27.82% also have at least one term with a connection to another. When considering all proprietary vocabularies, this quota becomes 8.16%, indicating that many vocabularies are not interconnected with others.

Table 15: Number of Vocabularies using connecting Terms

Connection Term	# prop. vocabs using term (% dereferencable)
http://www.w3.org/2000/01/rdf-schema#subClassOf	32 (28.31%)
http://www.w3.org/2000/01/rdf-schema#subPropertyOf	26 (23.01%)
http://www.w3.org/2002/07/owl#equivalentClass	5 (4.42%)
http://www.w3.org/2002/07/owl#equivalentProperty	3 (2.6%)

Table 15 shows how many vocabularies us terms usable for connecting different vocabularies. The percentages show the quota relative to dereferencable proprietary vocabularies. Interestingly, skos:broadMatch and skos:narrowMatch are never used by the proprietary vocabularies. One can see also a clear difference between the usage of rdfs:subClassOf and rdfs:subProperty on the one side and

owl:equivalentClass and owl:equivalentProperty on the other side. The former two are used by a rather large number of vocabularies, while the latter ones are barley used. One might conclude that mapped vocabularies are rather a specialization of existing vocabularies, extending already existing terms.

Proprietary vocabularies used by different categories of datasets might map the vocabularies to a different extent. Table 16 displays this in detail. The one category with a high number of mapped proprietary vocabularies is publications, where 14 proprietary vocabularies, 20.29% of all in this category, use have mappings to other vocabularies. A similar quota is only reached by user-generated content, but only because the general number of proprietary vocabularies is much lower. Proprietary vocabularies used by datasets in category lifesciences on the other side do not map to other vocabularies at all.

Table 16: Number of proprietary vocabularies mapping by category

category	# prop. vocabularies with mappings (% of prop. vocabs)
social web	3 (3.66%)
CMS	5 (7.35%)
lifesciences	0 (0%)
publications	14 (20.29%)
cross-domain	4 (7.02%)
government	3 (11.11%)
geographic	3 (11.54%)
media	1 (5.88%)
user-generated content	4 (22.23%)
Total	37 (9.41%)

An interesting question is to which vocabularies mapped ones link to. In total, we found 43 different vocabularies that are being referenced. Table 17 displays those that are used by at least 5% of all proprietary vocabularies that map to others. A large fraction, nearly one third of all vocabularies, link to SKOS. Other vocabularies often linked to are Dublin core, FOAF, RDFS and OWL. In general, vocabularies mapped to are those which are wide distributed, which is positive, as these are also well understood and supported by automatic agents.

Table 17: Target vocabularies of proprietary vocabulary mappings

Target Vocabulary	#vocabularies mapped to (% of mapping vocabs)
http://www.w3.org/2004/02/skos/core#	12 (32.43%)
http://purl.org/dc/terms/	8 (21.62%)
http://xmlns.com/foaf/0.1/	8 (21.62%)
http://www.w3.org/2000/01/rdf-schema#	8 (21.62%)
http://purl.org/dc/elements/1.1/	7 (18.91%)
http://www.w3.org/2002/07/owl#	5 (13.51%)
http://purl.org/NET/c4dm/event.owl#	4 (10.81%)
http://purl.org/NET/scovo#	2 (5.41%)
http://rdfs.org/sioc/ns#	2 (5.41%)
http://purl.org/vocab/aiiso/schema#	2 (5.41%)
http://www.w3.org/2003/01/geo/wgs84_pos#	2 (5.41%)
http://www.geonames.org/ontology#	2 (5.41%)

http://purl.org/ontology/bibo/	2 (5.41%)

In summary, only a small fraction of proprietary vocabularies map to others. As the extent of linking is limited, one can conclude that the adherence to this best practice is limited. Mappings to other vocabularies are often done by extending existing terms from well-known vocabularies, especially SKOS. Of the different categories, only category publications has a significant number of vocabularies which map to others.

3.7 Provide dataset-level metadata

Providing dataset-level metadata is essential for describing the basic properties of a dataset, for example who created the dataset, if and how the dataset can be used and if there are alternative access methods (which will be analysed in detail in the next part).

Although there are different possibilities to publish metadata about a dataset, providing a VoID description is a generally accepted form, according to the WC3 Semantic Web Interest Group [6]. Such a VoID file can serve for an automatic agent as a point of reference it can visit to inform itself about basic properties of the dataset. If for example a crawler that searches for data to be included to a data catalogue encounters a dataset, it may search for a VoID file, were it find information such as the creator and its contact address, the size of the dataset, additional access methods and so on. According to W3C interest group, the discovery of VoID descriptions can either happen by providing links from other documents of the dataset or by discovering the VoID document using well-known URIs. In the first case, one should use a triple with the predicate void:inDataset to link to a VoID file representing the whole document. In the second option, a VoID document should be at the "well-known location", a path prefix for HTTP(S) URIs standardized by the Internet Engineering Task Force (IETF) [10]. For instance at the PLD example.com, a VoID file would be located at "http://example.com/.well-known/void", giving a universal place for an intelligent crawler to search for dataset-level metadata first.

3.7.1 Definition

As we crawled a sample from different datasets, our approach to evaluate if a dataset provides metadata is to assess whether it uses the VoID vocabulary, be it "in the wild" or in a separate VoID file. As outlined before, a separate VoID file can be both referred to via a backlink within the data itself, or it can be served on the well-known location URI. For the first possibility, we simply look for any dataset if it uses any VoID predicates. Even if the crawler failed to crawl the VoID document itself, an existing back link to the document using the void document would exist, making it possible to find the VoID document itself. Thus we manually inspected if a back link exists and if this is the case, make sure that the VoID description had been crawled.

For the latter possibility, we additionally assess for every dataset the path of the well-known location, "./well-known/void", in order to determine if a VoID description is available or not. Paulheim et al. indicated that VoID files may not always be found at the well-known location directly at the host, but further down the path, for example because the dataset publisher might not have access to the host server [11]. To account for this, we also searched at well-known locations further up the path of a document. For every dataset, we randomly sampled 10 document URLs and appended the well-known prefix at every part of the path. Based on the return code, we then downloaded the files and inspected them if they contained Linked Data using Raptor, a RDF Syntax Library.[33]

As such, our approach will not find if dataset level metadata is available at other locations. Publishers might publish metadata at different locations not covered by our search or might not use the backlink mechanism for linking to them. Also, metadata will not be included if it is not Linked Data. In both cases, an automatic agent would not be able to find it, which is the ultimate goal of the best practice.

Lastly, it is possible to express metadata in other ways, especially by creating a proprietary vocabulary for expressing this kind of information. But as using VoID or positioning a VoID file is the generally accepted way of expressing metadata, we only search for metadata published in the way advised in by [9].

[33] http://librdf.org/raptor/

3.7.2 Results

From the data retrieved by the crawl, a total of 58 datasets used the VoID vocabulary. This constitutes around 6% of all available data sources. Interestingly, not all users of the VoID vocabulary also link to VoID files, but use the vocabulary "in the wild". In six cases, it was used to directly link to a SPARQL-Endpoint within the data. For the remaining vocabulary users which linked to VoID files, we had to additionally download the VoID files in 22 cases, as they were not part of the crawl.

When looking at the propagated default location for void-files at "/.well-known/void", we found three cases where a VoID-File was provided at the location advised by the Internet Task force, which we call the strict version of the well-known location. In one case though, the file only contained an RDF tag and a name space definition for RDFS. In all three cases, the datasets did not use the backlink mechanism to refer to their VoID file in their dataset. Trough the approach propsed by Paulheim et al., we were able to retrieve additional 11 VoID descriptions from well-known locations. In seven of these cases, the data was also accessible through the backlink meachanism.

Table 18: Usage of VoID vocabulary and supply of VoID files

Usage of VoID vocabulary	Number of datasets (% relative to all datasets)
Linking to VoID-Files with VoID description	52 (5.88%)
Using VoID Predicates	6 (0.54%)
providing VoID-File at ".well-known/void" (strict)	3 (0.22%)
providing VoID-File at ".well-known/void" (relaxed)	11 (1.19%)
Total	65 (7.08%)

When taking a look at the usage of VoID files by dataset category at Table 19, we see that many datasets from the category lifesciences provide VoID descriptions, but in relative terms, category geographic has the highest fraction. Category government has also a high fraction of 17.39%, but every other category has fractions of less than ten percentage points, with the lowest value being found for category social web.

In summary, the best practice of providing data-set level metadata, which we understood as using the VoID vocabulary and, similarly, the publishing of a VoID file is not often adhered to. When taking a look at how the meta-data can we found, we find that the reference via back-links from Linked Data is mostly used, while providing a void-file at the ".well-known/void" location is also not used very often. This means that the few meta-information published about Linked Data can best be discovered from within Linked Data and not by a general request to the servers hosting them, especially because often, both ways lead to a VoID description.

Table 19: Datasets providing VoID description by category

Category	Number of datasets (% relative to category size)
social web	10 (2.4%)
CMS	3 (2.26%)
lifesciences	21 (21.21%)
publications	9 (9.18%)
cross-domain	3 (5.66%)
government	8 (17.39%)
geographic	8 (22.23%)
media	1 (3.03%)
user-generated content	2 (6.9%)

3.8 Referring to additional access methods

Referring to additional access methods allows potential data consumers to access Linked Data in a way that fits them best. When requiring the whole data, a dump gives the possibility to download the dataset in a few zipped files, saving the need to crawl the data. Publishers on the other side can store such a dump on another server, thus reducing the workload for the server responsible for dereferencing URIs.

If a SPARQL endpoint is provided for a dataset, data consumers with special interests in certain aspects of the data can send SPARQL requests to that endpoint. Again, this saves data consumers the need to manually crawl the data, searching for resources that fit their demands or download all data and then process the dump to filter out the information they need. For data publishers, such an endpoint can help to limit bandwidth, as only those information need to be sent which are relevant to the data consumers.

In summary, providing alternative means to access the data, namely data dumps and SPARQL endpoints can both help data consumers in offering a more flexible and easy to use dataset, but also the data provider, lowering both workload and bandwidth.

3.8.1 Definition

As the central vocabulary for describing a dataset, the VoID vocabulary provides some opportunities to refer to additional access methods. Using the predicate "void:sparqlEndpoint", data providers can point to a SPARQL-Endpoint serving their dataset. RDF dumps can be announced via the predicate "void:dataDump".

From the previous section, we already examined how well data-set level metadata in the form of VoID descriptions is provided by dataset publishers. Based on this findings (the content of the VoID file itself or the triples using predicates from the VoID vocabulary), we evaluate how many of those descriptions link to alternative access methods, namely to a SPARQL-Endpoint or data dumps. When evaluating if additional access methods are referred, we can only analyse the subset of datasets which employ a VoID description.

This means that we will not be able to find SPARQL endpoints or dumps of a dataset which are not properly linked to from within the dataset or the VoID file. Which seems to be a limitation of our approach is actually more in accord with the idea of Linked Data as only in the case where such access methods are referred to in the Linked Data itself, do they become discoverable by automatic clients without the help of manually curated catalogues.

For datasets having some form of VoID description, we analyse whether one of the described properties for referring to other access methods is available. A dataset is providing a SPARLQL endpoint if it uses the predicate void:SPARQLEndpoint and a data dump if it uses void:datadump.

3.8.2 Results

From all datasets, 65 use the VoID vocabulary, 52 (80 %) provide some form of additional access methods, while 13 (20%) do not provide any additional access methods at all.

Table 20: Availability of additional access methods

Additional access method	# of datasets (% of VoID description providers/% all datasets)
SPARQL	40 (76.9%/ 4.36%)
dump	35 (67.3%/ 3.81%)
Both	23 (35.38%/ 4.42%)
Any	45 (80%/ 4.9%)

From the 52 datasets providing additional access methods, 40 (76.92%) provide a link to a SPARQL endpoint, while 35 (67.31%) provide links to a data dump. Of all datasets providing a SPARQL endpoint, 17 (42.5%) only provide a SPARQL endpoint and no dump while for the datasets providing a dump download, 12 (34.29%) do not provide a SPARQL endpoint. That means that from all 65 datasets providing a VoID description, 35.38% provide both access methods, 44.61% provide one access method and 20% provide no

access method. To put these numbers in perspective to all datasets, this means 2.51% of all datasets provide two access methods, 3.16% one access method and 94.34% none.

In Table 21, an overview of the availability of access methods in general as well different kinds of access methods is given. Similar to the results for data-set level metadata in general, datasets from category lifesciences provide the most alternative access methods and also tend to both offer SPARQL-Endpoints and dumps. This category has also the most datasets with alternative access methods relative to its size. Also regarding to its size, category geographic has many datasets with alternative access methods, with nearly every dataset provides a SPARQL endpoint. From the categories publications, social web, and geographic, five to eight datasets provide alternative access methods. This usually tends to be a SPARQL endpoint rather than a data dump. For the categories government, cross-domain, user-generated content, CMS and media only have few to no alternative access methods. Like before, SPARQL-Endpoints seem to be more often used than data dumps.

Table 21: Availability of alternative Access Methods by dataset category

Category	alt. access methods	Only Dump	Only SPARQL	Both
social web	7 (1.68%)	1 (0.24%)	5 (1.2%)	1 (0.24%)
CMS	3 (2.26%)	1 (0.75%)	1 (0.75%)	1 (0.75%)
lifesciences	21 (21.21%)	2 (2.02%)	4 (4.04%)	15 (15.1%)
publications	8 (8.16%)	2 (2.04%)	4 (4.08%)	2 (2.04%)
cross-domain	3 (5.66%)	1 (1.89%)	1 (1.89%)	1 (1.89%)
government	2 (4.34%)	0 (0%)	2 (4.34%)	0 (0%)
geographic	7 (19.44%)	1 (2.78%)	3 (8.33%)	3 (8.33%)
media	1 (3.03%)	1 (3.03%)	0 (0%)	0(0%)
user-generated content	0 (0%)	0 (0%)	0 (0%)	0 (0%)

In summary, only very few datasets refer to alternative access methods. While not for every dataset, alternative access methods are necessary and sensible, the quota of datasets providing and referring to such methods is still low.

4 Related Work

There has been some research on quality of the LODCloud where it was tried to assess the quality of datasets based on samples or information obtained through publishers.

An analysis bearing the strongest resemblance to our work has been done by Bizer et al. [2]. In this analysis, the authors evaluated the same best practices we investigated. However, the datasets analysed as well as the way adherence to the best practice were evaluated is different to ours.

First, they based their analysis to datasets catalogued at the lod-cloud group at datahub.io, only including datasets from this catalogue in their analysis if they had have at least 50 outlinks or at least one dataset with at least 50 links pointing to the dataset. Consequently, many datasets we have in our corpus are not included, as the dataset is not catalogued or not interlinked enough to be included. For instance FOAF-profiles, which are included in our corpus, are usually not part of the lod-cloud group. Their approach resulted in them having 295 in their corpus for analysis. As our approach does not pose such restrictions on the datasets included, our corpus includes with 918 a larger number of datasets. Consequently, the distribution of dataset categories is differs, also because their analysis considered multiple categories per dataset. A large number of our datasets are from category social web, which is not found in their analysis. Additionally, we created category CMS for output of content management systems, a category which is not part of their categories.

Second, as the basis for the evaluation of the best practices, they the authors information publishers provided to the catalogue. A central issue here is the possibility that information provided and those derived from the dataset itself might differ, either because changes in the dataset are not be reflected in the catalogue information or because information given in the description are not derivable from the dataset itself. This is especially true for metadata, like VoID files and additional access methods. While in the analysis of Bizer et al., the reference to a VoID file was sufficient, our analysis further requires that the dataset references to the VoID file as it is not retrievable for us otherwise.

Looking at the results for the best practice on interlinking, some differences arise. Our fraction of datasets with no outgoing links is much higher compared to their results. This might be due to their policy of only including datasets that are interlinked to their analysis. Also, there are more datasets with a higher outdegree compared to ours. An explanation could be that in [2], the authors asked publishers to state the number of links to other datasets within the lod-cloud group, without having the possibility to state links to datasets not part of this group. Their top 10 list of datasets and their outdegree regarding datasets has not much resemblance to ours, as it is dominated by datasets from the PLD rkbexplorer.com. This can be explained by the fact that datasets from this PLD cannot be crawled due to their restrictive robots.txt settings. The data set with the highest number of Linked Datasets has with 35 a much lower outdegree that ours, which has an outdegree higher than 100. This dataset, bibsonomy.org, does not seem to be included into their catalogue, showing that some strongly interlinked datasets are not part of this group, thus distorting the results.

Taking a look at best practices related to metadata, some differences are apparent. For provenance metadata, with 45%, we found a higher quota of datasets providing provenance metadata, compared to their quota of 36.63%. An explanation could be that the additional datasets included in our analysis have a stronger tendency to publish such metadata. Also, old datasets may have been updated without these changes being reflected in the dataset's description at the lod-cloud group.

Regarding licensing metadata, the quotas of both investigations are comparatively similar with 15% of all data sets publishing in our analysis, while their quota is with 17.84% only a little bit higher. Thus even for datasets not included in the lod-cloud group, the fraction of datasets with license information remains roughly similar.

For dataset-level metadata, the Bizer et al. analysis included the adoption of semantic sitemaps which we left out, as they are no longer endorsed by the W3C group [6]. Regarding VoID descriptions, we found only 6% of all datasets using the VoID vocabulary and thus giving some dataset-level metadata, according to our definition, compared to the 32.2% the analysis from Bizer et al. found. As publishers just had to provide the VoID file itself, it was not evaluated if it is reachable via backlink or hosted at a well-known location. In absolute terms, 95 dataset provide a VoID description. Assuming that all datasets using VoID vocabulary we found are part of these 95 datasets, this would mean that 21% of the descriptions are not reachable by automatic agents. This number would even be higher if there are datasets adhering to this best practice, but not part of the lod-cloud group.

Comparing the results for alternative access methods, the quota of publishers providing a SPARQL endpoint or a dump file is with 68.14% and 39.66% much higher, compared to fractions of 4.36% and 3.81% in our results. Again, publishers were asked directly for such access methods and it was not required that they were identifiable through the dataset-level metadata. Looking at the absolute numbers, we see that in our evaluation, 52 datasets with alternative access methods are found compared to 201 in their analysis. This might indicate that the problem is more the discoverability of access methods through Linked Data itself rather than the existence of alternative access methods.

Lastly, they also analysed best practices that occupy themselves with vocabulary usage. Differences in the list of most used vocabularies are the more prominent status of FOAF, along with the inclusion of the MetaVocab vocabulary and RSS having a more prominent position. Looking at the usage of proprietary vocabularies, our results show a higher fraction of datasets not using any proprietary vocabularies. This might be either a result of our different definition of "proprietary" (them defining at as a vocabulary hosted at the same PLD like the data) or from our inclusion of datasets not present in the lod-cloud group, were possibly more standard vocabularies are used. This is also hinted by the more prominent positions of the FOAF and RSS vocabularies, which are used for FOAF profiles and by CMS systems, both being present in categories social web and CMS, both categories not found in the analysis done by Bizer et al.

For the dereferencabiliy of datasets, we have a much lower quota of dereferencable vocabularies, with nearly two thirds of them not being dereferencable at all. Again, this might be a result of the different definition of proprietary, but it might also be possible that the fraction of non-dereferencable proprietary vocabularies used in datasets not included in the lod-cloud group is higher. Finally, regarding vocabulary mappings, the numbers between our and their analysis resemble strongly. While in their analysis 7.89% vocabularies provide mappings, the quota is similar in our analysis with 8.16%, showing that providing vocabulary mappings for proprietary vocabularies is an issue to be addressed.

Another analysis similar with aspects similar to ours was done by Hogan et al. [3], who analyse the conformance of Linked Data to a set of guidelines compiled from a tutorial by Bizer et al. [12]. Similar to our approach, they based their evaluation on a crawled dataset, namely the BTC2010 corpus, a predecessor of the BTC2012 corpus we used. They restricted themselves to PLDs with more than 1000 quadruples, lowering their sample to 188 PLDs. Also, they did consider a PLD to be an atomic dataset, while we distinguished between different datasets hosted under one PLD, if according information was available.

Similar to us, they analysed the extent of a dataset's linking to others by counting the number of datasets a given dataset connects to, like us only considering PLDs they know of having RDF. But unlike our analysis, they did not try to assure that only those links are counted which point to Linked Data resources.

Looking at their results, the number of datasets that do not link at all to others is with five out of 188 lower than our quota. Also, the average degree is with 20.4 much higher, compared to ours of 3.084. Comparing their top five linking datasets with ours, we see that their number two, status.net, also appeared in our top five links in their analysis to 191 datasets, while it links to only 42 in our analysis. This shows that different approaches to the extent of linking can lead to different results.

Next, they also evaluate the guideline to re-use terms from well-known vocabularies, which is similar to our fourth best practice. For their approach, they counted the usage of individual terms by a dataset and weight it with the use of the terms by all datasets. They conclude that there exists a non-trivial level of re-use of terms. But they also note that most class and property terms are minted in local namespaces. While their results are not comparable directly to our insights due to different measures of vocabulary use and definition of proprietary vocabularies, they are nonetheless similar to ours.

Lastly, the authors addressed topic of dataset metadata, focussing on the deliverable of metadata in general and the attachment licensing information to resources. For evaluating if metadata about a resource are publishes, they use an approach different to ours. They evaluate to what extent triples appear in a dataset, where the subject of the triple is the URI of an information resource. We on the other side explicitly looked for dataset-level metadata, which incorporates the use of the VoID vocabulary, especially within a separate VoID file. For licensing information, we adopted their approach for identifying license information by searching for predicates containing the string "licen". For filtering though, we imposed a stricter filter regime, for example filtering out doap:license, as it appeared to us that it was mainly used to denote the license of software, not of an information resource. The general results though are much in line with our findings. They found license information for 14.4% of all datasets, a value similar to ours.

5 Conclusion

In this deliverable, we analysed to what extent dataset publishers adhere to a set of best practices described in Deliverable 2.1 [1]. For this examination, we gathered samples of Linked Data from a large number of datasets by mainly crawling them and classifying them into different categories. For evaluation, we developed metrics for assessing the adherence of best practices.

- Best practice one advises to link a dataset to others for creating a web of interconnected datasets, enabling them to discover new relevant information by browsing the interconnected datasets. Our results show that more than half of all datasets do not adhere to this best practice, while datasets of category government more often do not adhere to this best practice, while datasets of category social web and lifesciences tend to adhere to this best practice more often. To promote linking to other datasets, we think that both the visibility of linking opportunities as well as the task of linking should be eased. The former can include offering more ways to discover linkable datasets, for example by fostering dataset catalogues or search engines for Linked Data. For establishing links between datasets on the other side, tools like Silk need to be enhanced, easing this task. Another possibility would be to apply tools like DBpedia spotlight, which identifies references to resources in written text, with non-Linked Data parts of the Linked Data, for example blog posts, bridging the gap between unstructured and structured data. This would especially be interesting for CMS data, which often uses Linked Data only to structure its content.
- Regarding the best practice of adding provenance metadata, we see a considerable adoption, at least regarding basic provenance metadata. The usage of vocabularies for expressing more complex provenance information on the other side is comparatively low. A way to heighten the quota of datasets providing provenance information could be to integrate their generation in the data generation process itself, for example to automatically generate such data by tools used for generating the data itself. While this might be easy for datasets like social web or content management systems, such integration is difficult to achieve for more complex and custom data generation processes, as they may involve different agents, which all need to support the generation of provenance information. For example for the generation of provenance information from linked sensor data, this would require that both the sensors itself as well as tools transforming these data add their provenance information to the data. This complexity might be the reason for the low uptake of advance provenance vocabularies like prov.
- Licensing metadata on the other side has a relative low uptake, being supplied by around every sixth dataset. Again, the automatic integration into the data production process might help to heighten the number of datasets adhering to this best practice. On the other side, there is a considerable number of datasets with textual descriptions of their license, which might indicate that there is a need to express license information as Linked Data different to the possibilities currently existing, for example the possibility to express a copyright notice as Linked Data.
- When looking at vocabulary usage, we see that on the one side, well-known vocabularies are widely used, with most datasets using at least one vocabulary used by others and a top list of vocabularies that have high fractions of datasets using them. On the other side, a considerable number of proprietary vocabularies exist, being used by only one dataset. A way to mitigate the use of proprietary vocabularies would be to make it easier for publishers to find vocabularies that are well known and suitable for their need. This can be achieved for example with easy-to use vocabulary catalogues and tools that make recommendations for vocabularies to be used, for example for the case where existing database tables are to be converted to Linked Data.
- The majority of proprietary vocabularies are not dereferencable at all, making them essentially not usable to an automatic agent not manually adapted to them. Apart from helping publishers to find widely used vocabularies for his dataset, making it unnecessary to define a proprietary vocabulary, the only way to enhance the dereferencability of terms is to urge publishers to publisher their vocabularies and to explain its use to them.
- Additionally, only a limited number of dereferencable proprietary vocabularies have mappings to other vocabularies. If proprietary vocabularies are connected to others, the mapping puts the terms in a specializing relation to existing terms. Again, mapping between vocabularies might be enhanced

by giving publishers powerful and easy to use mapping tools, making requiring less effort by them to adhere to this best practice.

- The last two best practices, the supply of dataset-level metadata in the form of VoID files and the addition of alternative access method, have a general low adherence, while the result of the second is a direct result of the first, as without the usage of the VoID vocabulary, we were not able to identify alternative access methods. While it seems admittedly disproportionate to expect from someone hosting his FOAF profile to additionally host a VoID file about his profile, the adherence to this best practice is still too low. Again, software used to create, host and administer Linked Data could be used to automatically created and publish such files, linking to alternative access methods.

Taking a look at the different categories, we see that the adherence to the best practices is generally different for the datasets. For example datasets in category social web have a relatively high tendency to adhere to best practice one, and also tend to use widely-used vocabularies, but fail to provide metadata, alternative access methods and properly dereferencable and mapped vocabularies. Categories government, publications and geographic on the other side do not adhere to the first best practice, but often provide metadata like provenance, license and dataset-level metadata. As different categories of datasets have different levels of adherence to best practices, the focus on what to improve should be different for categories, for example by focussing on the providence of metadata for social web datasets and providing means to enhance linking for governmental datasets.

In summary, some best practices are better adhered to than others, while the degree of adherence always depending on the category of datasets examined. In order to stimulate the adherence of such rules, we think that data publishers should be supported, for example with powerful and easy to use tools or with software that decreases the labour necessary for adhering to the best practices. By helping publishers creating high quality datasets, the LODCloud can become not only a large net of data, but also a large net of data comprised of datasets with high quality.

References

[1] C. Bizer, P. N. Mende, Z. Miklos, J.-P. Calbimonte, A. Moraru and G. Flouris, "D2.1 Conceptual model and best practices for high-quality metadata publishing," PlanetData, 2012.

[2] C. Bizer, A. Jentzsch and R. Cyganiak, "State of the LOD Cloud," March 2011. [Online]. Available: https://lod-cloud.net/state/.

[3] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres and S. Decker, "An empirical survey of Linked Data conformance.," *Journal on Web Semantics*, vol. 14, pp. 14-44, 2012.

[4] A. Harth, *Billion Triples Challenge data set*, http://km.aifb.kit.edu/projects/btc-2012/, 2012.

[5] R. Isele, J. Umbrich, C. Bizer and A. Harth, "LDSpider: An open-source crawling framework for the Web of Linked Data," in *ISWC Posters&Demos 658*, 2010.

[6] K. Alexander, R. Cyganiak, M. Hausenblas and J. Zhao, "Describing linked datasets with the void vocabulary," *W3C Interest Group Note, W3C*, 2011.

[7] T. Berners-Lee, "Linked data-design issues (2006)," 2011. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html.

[8] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, "Silk-A Link Discovery Framework for the Web of Data.," in *Linked Data on the Web*, 2009.

[9] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1-136, 2011.

[10] M. Nottingham and E. Hammer-Lahav, "Defining Well-Known Uniform Resource Identifiers (URIs)," 2010. [Online]. Available: http://tools.ietf.org/html/rfc5785?chocaid=397.

[11] H. Paulheim and S. Hertling, "Discoverability of SPARQL Endpoints in Linked Open Data," in *Proceedings of the ISWC*, 2013.

[12] T. Heath, C. Bizer and R. Cyganiak, "How to publish linked data on the web," in *Tutorial in the 7th International Semantic Web Conference*, Karlsruhe, Germany, 2007.

[13] C. Bizer, "The emerging web of linked data," *Intelligent Systems, IEEE*, vol. 24, no. 5, pp. 87-92, 2009.