

PlanetData

Network of Excellence

FP7 – 257641

D4.1 PlanetData data sets, vocabularies and provisioning tools catalogue and access portal

**Coordinator: Pablo N. Mendes, Steffen.Stadt Müller, Christian
Bizer**

**1st Quality reviewer: Elena Simperl
2nd Quality reviewer: Zoltán Miklós**

Deliverable nature:	R
Dissemination level: (Confidentiality)	PU
Contractual delivery date:	September 30 th , 2011
Actual delivery date:	September 20 th , 2011
Version:	0.4
Total number of pages:	33
Keywords:	

Executive summary

In this report we describe a first release of the catalogue and access portal for data sets and vocabularies published in a self-descriptive manner. We have chosen CKAN (<http://ckan.org>) as the supporting software for the PlanetData catalogue, since it is freely available as open source, includes all of the necessary features for our initiative and has been validated in a number of deployments.

PlanetData has reused OKFNs TheDataHub.org portal and previous results from cataloguing activities in LATC. We have extended the metadata schema used in LATC, developed a set of guidelines¹ and extended validation scripts to match those guidelines. We individually provided through the PlanetDataEditor² user on TheDataHub.org a total of 49 new packages and over 540 edits to existing entries. Moreover, we supported the community in providing metadata about their own datasets, following this process with an in-house quality assurance step. After the quality assurance process, 276 data sets were added to the lodcloud group³, totalling 31.5 billion triples and almost 500 million links between data sets. The metadata about these datasets is available for humans and machines from TheDataHub.org.

Moreover, we made the results available via an access portal summarizing important characteristics of the datasets, including their topic, size, level of interconnection and compliance with best-practice recommendations. The textual content of the portal is included in this report, for reference, in section 2. The interactive access portal is available at: http://www4.wiwiss.fu-berlin.de/lodcloud/state/2011_09_index.html Users can use shortcuts from the access portal directly to the package listings at TheDataHub.org.

In order to ease for data producers the process of finding suitable vocabulary terms for describing their data sets, we have also provided an access portal for searching vocabulary terms, available from <http://vocab.cc>.

1 <http://www4.wiwiss.fu-berlin.de/lodcloud/ckan/validator/levels.html>

2 <http://ckan.net/user/PlanetDataEditor>

3 <http://thedatahub.org/package?q=lod&groups=lodcloud>

Table of Contents

Executive summary	2
Document Information.....	3
Table of Contents.....	4
Abbreviations.....	5
Definitions.....	6
1. Data Sets Catalogue.....	7
1.1 The data sets catalogue portal software.....	8
1.2 The data sets catalogue metadata schema.....	9
1.2.1 Level 1 metadata.....	9
1.2.2 Level 2 metadata.....	10
1.2.3 Level 3 metadata.....	11
1.3 The data sets cataloguing process.....	13
1.4 An overview of the resulting data sets catalogue.....	15
1.5 The data sets catalogue access portal.....	17
2. State of the LOD Cloud.....	19
2.1.1 Provide dereferencable URIs.....	19
2.1.2 Set RDF links pointing at other data sources.....	19
2.1.3 Use terms from widely deployed vocabularies.....	21
2.1.4 Make proprietary vocabulary terms dereferencable.....	22
2.1.5 Map proprietary vocabulary terms to other vocabularies.....	23
2.1.6 Provide provenance metadata.....	23
2.1.7 Provide licensing metadata.....	24
2.1.8 Provide dataset-level Metadata.....	24
2.1.9 Refer to additional access methods.....	25
3. Vocabulary Catalogue.....	26
3.1 Analysis of existing Linked Data vocabularies.....	27
3.1.1 The Billion Triple Challenge Dataset.....	27
3.1.2 Extraction of information.....	27
3.2 vocab.cc.....	28
3.2.1 URI lookup.....	28
3.2.2 URI search.....	28
3.2.3 Web portal.....	28
3.2.4 Linked Vocab Services.....	29
3.3 Further work.....	31
4. Conclusions.....	32
5. References.....	33

Abbreviations

API – Application Programming Interface

CMS – Content Management System

LOD – Linking Open Data

RDF – Resource Description Framework

Definitions

Crowdsourcing: “is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a 'crowd'), through an open call.”

(Source: en.wikipedia.org/wiki/Crowdsourcing)

Package: is a slightly ambiguous concept but it is effectively a group of files and services that are useful to consider together and any metadata that relates to them.

(Source: http://wiki.ckan.net/FAQ#What_is_a_.27package.27.3F)

1. Data Sets Catalogue

The PlanetData Project, through its Task 4.1 will gather, document and maintain a catalogue of large-scale open-license data sets. The catalogue will contain data sets ranging from distributed Web data over linguistic corpora to real-time sensor data and climatic data. The catalogue will lay the foundation for the experiments conducted in the research work packages and will be offered to the participants of the PlanetData Programs. In this report we describe a first release of the catalogue and access portal for data sets and vocabularies that publish data in a self-descriptive manner. The catalogue of data provisioning tools has been moved to WP5, and its first release has been included in D5.1.

We have chosen CKAN (<http://ckan.org>) as the supporting software for the PlanetData catalogue, since it is freely available as open source, includes all of the necessary features for our initiative and has been validated in a number of deployments. More details on this choice are provided in section 1.1. CKAN was originally developed to support TheDataHub.org for cataloguing, publishing, sharing and finding data. TheDataHub.org is maintained by members of the Open Knowledge Foundation, the Linking Open Data community as well as other communities interested in open data. It currently contains meta-data about 2125 data sets. Due to high activity from the open data community at TheDataHub.org, and due to its collaborative nature, we have decided to contribute to this effort instead of starting a new parallel cataloguing effort.

Although TheDataHub.org also includes some vocabularies in the form of data sets - a collection of all terms in a vocabulary can be packaged and shared as a data set – we have also extended the vocabulary catalogue with statistics of their usage “in the wild”. We have crawled linked data from the Web and shared the statistics collected through a search portal that can help users to choose terms to use when describing their data sets.

In collaboration with LATC, we have designed a meta-data schema for describing data sets – as presented in section 1.2. As part of this deliverable, we have provided new entries to TheDataHub.org, and extended existing entries to comply with the metadata schema we developed. The cataloguing process is described in section 1.3. As a positive effect of our choice to contribute to an unified catalogue, many data publishers have followed our metadata schema and updated their catalogue entries accordingly.

We analysed the cataloguing results, providing an overview of the types of data it contains, publishing techniques used, level of interconnection and reuse, etc. We present these results in section 1.4. The resulting catalogue is available on the Web in its raw form from TheDataHub.org, both for humans and machines. We provided, in addition, a State of the LOD Cloud statistics page, where we summarize the information in the catalog, providing clickable links that take the users directly to the package listings in TheDataHub.org. One can, for example, find the list of datasets using a given vocabulary, or only those published by third parties. More details are presented in section 1.5.

1.1 The data sets catalogue portal software

CKAN is an open-source data portal software. The objective of CKAN is to make it easy to publish, share and find data. It provides a powerful database for cataloguing and storing datasets, with an intuitive web front-end and API. CKAN was developed by the non-profit Open Knowledge Foundation in 2007 to run the Data Hub, a public registry of open datasets. It is managed by a full-time CKAN development team, with professional-level support. The key features of CKAN are:

- Complete catalog system with easy-to-use web interface and powerful API
- Open source software, written in Python
- Strong integration with third-party CMS's like Drupal and WordPress
- Tagging, rating, arbitrary metadata, dataset relationships
- Workflow support including moderated editing, full change history
- Fine-grained access control
- Linked data support
- Powerful API with clients in Python, PHP, Perl, Javascript etc
- Extension and Plugin framework
- Integrated storage for storing data
- Federated structure: easily set up new instances with common search

Since its initial version, CKAN has grown in maturity and usage, having been used to deploy more than 29 regional data catalogues⁴, such as data.gov.uk and PublicData.eu, as well as the Dutch and Norwegian governments.

4 <http://wiki.ckan.net/Instances>

1.2 The data sets catalogue metadata schema

We have developed a data set description schema that extends the general information about datasets as collected on TheDataHub.org. Illustration 1 shows an overview of the information collected.

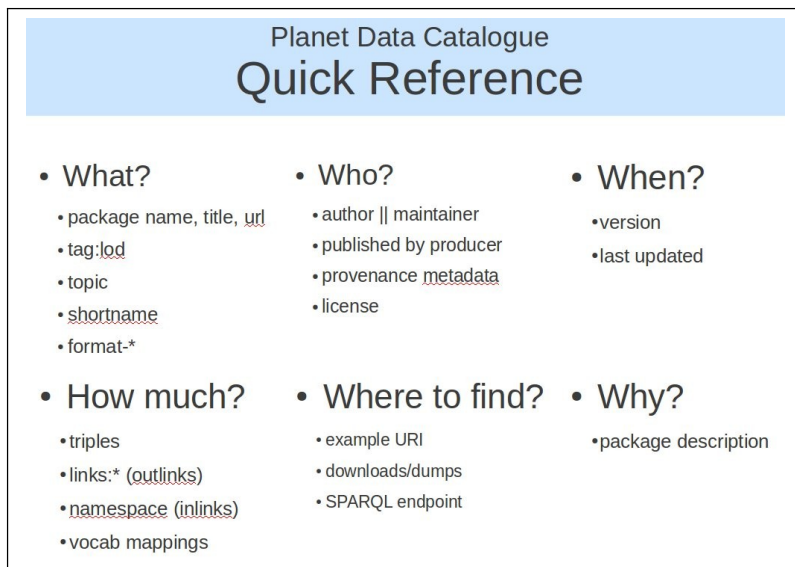


Illustration 1: Planet Data catalogue quick reference guide

Our metadata schema benefits from the extensibility of CKAN in order to collect custom information. A data set is called “package” on CKAN. A package description page contains three main sections: Basic Information, Resources and Extras. We have developed a set of guidelines⁵ on how to use these sections to provide the information collected for D4.1. In order to allow an incremental cataloguing process that caters to different types of datasets, we have divided the metadata into four levels, with basic, minimal, complete and finalized information, respectively. We describe these levels in the remainder of this section and present the cataloguing process in more details in section 1.3.

1.2.1 Level 1 metadata

Metadata in level 1 aims at collecting basic information about a data set, including its name, contact information as well as if this dataset provides Linked Data.

CKAN / Basic Information

Field name	Description	Format/Examples
Name	Unique ID for the data set on CKAN	[a-z0-9-]+ "my-dataset"
Title	Full name of the data set	"My Dataset"
URL	Link to data set homepage	http://example.com/my-ds
Author	Name of publishing org and/or person	"FUB (Pablo Mendes)"
Author email Maintainer email	Contact email	contact@pablomendes.com

⁵ <http://www4.wiwiw.fu-berlin.de/lodcloud/ckan/validator/levels.html>

CKAN / Basic Information / Tags

Tag	Purpose
lod	Identifies the data set as Linked Data

1.2.2 Level 2 metadata

The level 2 metadata aims at collecting complementary information that we consider minimal for our cataloguing effort: the topic of the data, the data formats used to encode data, ways to access the data and the dimensions of the set in triples and links to other datasets.

CKAN / Basic Information / Tags

Tag	Purpose
<topic>	One of: media geographic lifesciences publications (including library and museum data) government ecommerce socialweb (people and their activities) usergeneratedcontent (blog posts, discussions, pictures, ...) schemata (structural resources, including vocabularies, ontologies, classifications, thesauri) crossdomain

CKAN / Resources

What	Format	Description
RDF example link	Any of: <ul style="list-style-type: none"> • example/rdf+xml • example/turtle • example/ntriples • example/x-quads • example/rdfa • example/x-trig 	Link to an example data item within the data set in the corresponding format (e.g. RDF/XML)

CKAN / Resources

What	Format	Description
SPARQL endpoint	api/sparql	SPARQL endpoint
Direct link to each RDF download file (preferred)	Any of: <ul style="list-style-type: none"> • application/rdf+xml • text/turtle • application/x-ntriples • application/x-nquads • application/x-trig 	Download
Download page with list of downloads (accepted)	-	Download (for multiple files)

CKAN / Extras - via "Add more information (Groups, authors etc)"

New key	With value	Format/Examples
triples	Approximate size of the data set in RDF triples	100000, 62345123
links:xxx	Number of RDF links pointing at data set xxx. Please provide separate links xxx statements for each data set linked to	20000

New key	With value	Format/Examples
sparql_graph_name	Named graph in SPARQL store (if used by the SPARQL endpoint)	http://species.geospecies.org

1.2.3 Level 3 metadata

The metadata collected in level 3 completes the schema with additional information used in our catalogue. Those include versioning information, licensing, provenance information, etc.

CKAN / Basic Information

Field name	Description	Format/Examples
Version	Last modification date or version of the data set	"2010-04 (3.5)", "2006", "beta"
Notes	Description of the data set	some free text
License	Standard license drop-down	OSI approved::MIT license

CKAN / Extras - via "Add more information (Groups, authors etc)"

New key	With value	Format/examples
shortname	Short name for LOD bubble	"NY Times"
license_link	Custom license link	http://example.com/so-sue-me
namespace	Instance namespace	http://dbpedia.org/resource/

CKAN / Resources

Purpose	Format	Description
void file	meta/void	void description
XML Sitemap	meta/sitemap	XML Sitemap
RDF Schema	meta/rdf-schema	Download link to RDF/OWL Schema used by the data set (in addition to having dereferenceable vocabulary URIs)
Vocabulary Mappings, e.g., OWL, RDFS, RIF, R2R	mapping/<format>	If the data set provides vocabulary mappings to other vocabularies (owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, and/or rdfs:subPropertyOf links), provide a link to the mapping file in the <i>Downloads & Resources</i> section, using the following format: <i>mapping/<format></i> . Replace <format> with the mapping/rule language used, like R2R or RIF.

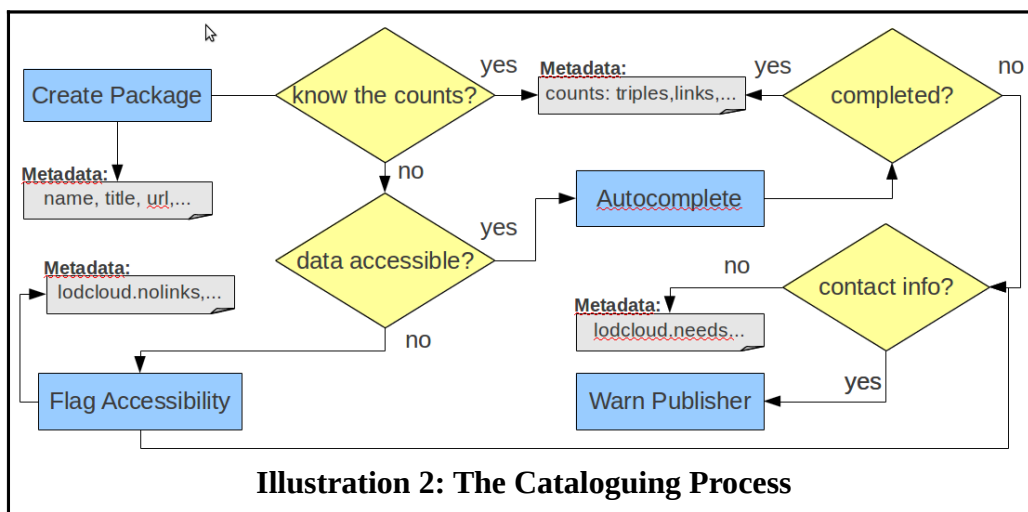
CKAN / Basic Information / Tags

Tag	Purpose
One of: <ul style="list-style-type: none"> no-proprietary-vocab deref-vocab no-deref-vocab 	The tag no-proprietary-vocab indicates that your data set does not use a proprietary vocabulary (defined within your top-level domain). The other two tags indicate that your dataset uses proprietary vocabulary terms (the ones that are defined within your top-level domain) and they are (deref-vocab) or are not (no-deref-vocab) dereferenceable according to the best practices for Publishing RDF Vocabularies
vocab-mappings no-vocab-mappings	Indicates whether mappings for proprietary vocabulary terms are provided (by setting <code>owl:equivalentClass</code> , <code>owl:equivalentProperty</code> , <code>rdfs:subClassOf</code> , and/or <code>rdfs:subPropertyOf</code> links, or publish mapping expressed as RIF rules or using the R2R Mapping Language).
provenance-metadata no-provenance-metadata	Indicates whether the data set provides provenance meta-information (creator of the data set, creation date, maybe creation method) as document meta-information or via a <code>void</code> description. For instance, using the <code>dc:creator</code> or <code>dc:date</code> properties.
license-metadata no-license-metadata	Indicates whether the data set provides licensing meta-information as document meta-information or via a <code>void</code> description. For instance, using the <code>dc:rights</code> property.
published-by-producer published-by-third-party	Indicates whether the data set is published by the original data producer or a third party.
limited-sparql-endpoint	Indicates whether the SPARQL endpoint is not serving the whole data set.
format- <i><prefix></i>	A vocabulary used by the data set, e.g., <code>format-skos</code> , <code>format-dc</code> , <code>format-foaf</code>
lodcloud.nolinks	Data set has no external RDF links to other datasets.
lodcloud.unconnected	Data set has no external RDF links to or from other datasets.
lodcloud.needsinfo	The data provider or data set homepage do not provide minimum information (and information can't be determined from SPARQL endpoint or downloads).
lodcloud.needsfixing	The dataset is currently broken. Provide details in the Notes.

1.3 The data sets cataloguing process

The metadata collection process of the Planet Data catalogue involved three main activities: manual entries, crowdsourcing, and quality assurance. After the metadata schema for the catalogue was developed, we started producing new entries and updating existing ones at TheDataHub.org. Concurrently, we invited the Linking Open Data community to join the cataloguing effort through an e-mail⁶ message sent to the public-lod mailing list hosted by the W3C. During the crowdsourcing period we provided guidance to data publishers through e-mail exchanges, and after the end of the crowdsourcing period we conducted a review of all entries for quality assurance.

The cataloguing process was open for participation by any interested parties. Participants were asked to register a user name with TheDataHub.org before editing or adding packages, and to confirm if data sets do not already exist on the catalogue before adding a new data set. They were then requested to provide as much additional information as possible as described on the metadata guide (section 1.2). The flowchart depicted on Illustration 2 summarizes the cataloguing process. First users had to create a package and provide basic information. For participants that knew basic statistics for a data set, they were asked to provide this information. When this information was not available, we attempted to obtain an “autocomplete” for the metadata by running automated scripts. Since there is the possibility of encountering parsing errors or unavailable downloads, we flagged the packages as “needs fixing” when the automated scripts did not complete. Whenever possible, given available contact information, we attempted to warn publishers of problems with their data.



The incentives for providing high quality information for data publishers are clear, since all the information on TheDataHub.org is made available via the CKAN API and can be used by search engines or data consumers to find new datasets to which to link. As additional incentive for incremental improvement, we devised a validation page with a badge scheme⁷. The validation page checks for required, recommended and optional information in each catalogue entry. The ultimate aim of the validation page is to enhance the quality of the metadata provision by highlighting information that is considered key for users searching for data to consume. Data sets that provide minimal information for each level are promoted to the next level. Meanwhile, data sets that provide complete descriptions are awarded a medal, which is displayed alongside the data set name on the validation page. Data publishers involved in the crowdsourcing process were given access to

6 <http://lists.w3.org/Archives/Public/public-lod/2011Jul/0059.html>

7 <http://www4.wiwiw.fu-berlin.de/lodcloud/ckan/validator/>

the validation script, which was also used internally after the crowdsourcing process as a guide for our manual review of entries. An example validation page is displayed on Illustration 3.

CKAN LOD Validator Freie Universität Berlin

[LOD Datasets on CKAN](#) | [Validate](#) | [Help](#)

Search CKAN package:

DBpedia-Live

- Package name: dbpedia-live
- [Package on CKAN](#)

LOD Cloud Diagram Compliance Level

Level: 1 (basic) ✓

Level: 2 (minimal) ✓

Level: 3 (complete) ✓

Level: 4 (reviewed and added to lodcloud group) ✗

Warning

Unconnected. An editor has indicated that the dataset has no external RDF links to or from other datasets and thus can not be added to the LOD cloud.

No external links. An editor has indicated that the dataset has no external RDF links to other datasets.

Enhanced Information

Please provide the following additional information about the data set. This information helps the community to know more about the development state of the Web of Linked Data. Moreover, following best practices for dataset description will make easier for users (and search engines) to consume this dataset. Find more help [here](#).

- Missing download(s).** Please provide a link to the download file(s) if available in the *Downloads & Resources* section, using one of the following formats: *application/rdf+xml*, *text/turtle*, *application/x-ntriples*, *application/x-nquads*, *application/x-trig*, *text/n3*.
- Missing size.** No data set size found. Please provide the approximate size of the data set in RDF triples using the custom CKAN field *triples*.
- Missing link count information.** Please provide the custom CKAN field *links:target_data_set* with the number of RDF links pointing at data set *target_data_set*. Please provide separate *links:target_data_set* statements for each data set to which *dbpedia-live* links.
- Missing instance namespace.** Please provide the namespace used for instances of the dataset. For example, the namespace for DBpedia instances is *http://dbpedia.org/resource/*. This will be used to detect who links to your dataset.
- Missing void or Semantic Sitemap.** Please provide a link to a void description or XML Sitemap if available in the *Downloads & Resources* section, using one of the following formats: *meta/void*, *meta/sitemap*.
- Missing mapping(s).** If the data set provides vocabulary mappings to other vocabularies, provide a link to the mapping file in the *Resources* section, using the following format: *mapping/format*. Replace *format* with the mapping/rule language used, like R2R or RIF.
- Missing schema.** If the data set uses a proprietary vocabulary, provide a download link to the RDF/OWL Schema used by the data set (in addition to having dereferenceable vocabulary URIs) in the *Resources* section, using the following format: *meta/rdf-schema*.

Preliminary assessment:

For each resource within the accepted download and example formats, we attempt to connect to your URL, to check if the link is broken. Results are below.

Example resource: http://dbpedia.org/resource/The_Lord_of_the_Rings

- Availability ✓: Link seems currently OK.
- Interpretability ✓: Format informed is OK.

[\[more\]](#)
SPARQL endpoint availability: 99% last month [2]

[Edit this package on CKAN.](#)

Illustration 3: Validator Page for the data set DBpedia-Live

As a final incentive for participation, data sets that pass the level 3 of the validation script were manually reviewed and added to the lodcloud group. Data sets in this group are promoted to level 4 and added to the LOD cloud diagram, which is one of the most used images in publications and presentations in the Semantic Web and Linked Data communities. The latest LOD cloud diagram depicts an overview of the interconnected sets of linked data on the Web as of September 2011 and includes data generated through the cataloguing process of D4.1.

Life sciences	42	3,039,978,308	9.65 %	191,825,949	38.44 %
User-generated content	14	115,072,057	0.37 %	3,431,983	0.69 %
	276	31,506,115,868		499,051,305	

Linked Data technologies are being used to share data covering a wide range of different topical domains. Table 1 gives an overview of the amount of triples as well as the amount of RDF links per domain. The number of RDF links refers to out-going links that are set from data sources within a domain to other data sources.

It can be observed from Illustration 5 and Illustration 6 that although the Government domain is the most prolific data publisher, it is the Life Sciences domain that provides the most interconnected data sets.

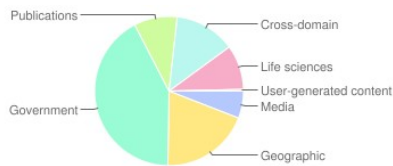


Illustration 6: Triples by domain

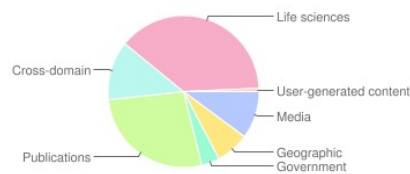


Illustration 5: Links by domain

Initially, the Linked Data best practices were adopted mainly by research projects and Web enthusiasts. These third-parties took existing data sets, converted them into RDF and served them on the Web. Alternatively, they implemented Linked Data wrappers around existing Web APIs. Today, Linked Data technologies are increasingly adopted by the primary data producers themselves and are used by them to provide access to their data sets. As of August 2011, out of the 276 datasets in the LOD cloud 101 (36.59 %) are published by the data producers themselves, while 175 (63.41 %) are published by third-parties.

Most data sets (64.49 %) reuse terms from other non-proprietary vocabularies. Vocabulary reuse is an important indicator of interpretability, as it is harder to understand data if each provider uses a different vocabulary. However, in some cases there are no available vocabularies for a given purpose, and data providers are forced to introduce new vocabularies. In those cases, providers are advised to return definitions of the vocabulary terms from the Web according to Linked Data principles. A total of 85.14% of the data sources that use proprietary vocabulary terms follow this recommendation.

In general, many sources provide additional ways to access their data. Altogether, 67.75 % of the linked data sets provide a SPARQL endpoint, while 38.41 % out of the 276 data sets provide RDF dumps.

1.5 The data sets catalogue access portal

The data sets catalogue is available in its raw form from TheDataHub.org. We provide an additional view of the catalogue as a portal entitled “State of the LOD Cloud⁹”. The portal provides statistics about the structure and content of the LOD cloud. It also analyses the extend to which LOD data sources implement nine best practices that are either recommended W3C or have emerged within the LOD community. All statistics within this document are based on the data set catalogue that is maintained on TheDataHub.org.

The format of the portal was chosen with two uses in mind. First, it serves as a quick-reference document containing best practices and their level of adoption. It can be used as a guide for data publishers, or as a quick reference for researchers describing the linked data ecosystem. Second, it serves as a shortcut into the catalogue to facilitate its exploration. The portal provides access to the catalogue through the use of the CKAN API. When describing each feature of the data sets, we link directly to lists of packages in TheDataHub.org containing those features, providing a direct link to explore the catalogued data sets by their most prominent characteristics. It is possible, for example, to quickly jump from a chart displaying the size of datasets by domain (Illustration 6: Triples by domain) into a listing of all datasets in the media domain¹⁰.

The catalogue is also available in machine readable format via <http://semantic.ckan.net>. It is described using the vocabularies VoID¹¹, DCAT¹², MOAT¹³, among others. A snippet of a dataset description for DBpedia is shown in Text 1.

The machine readable descriptions can be queried via SPARQL (<http://semantic.ckan.net/snorql/>), allowing the exploitation of the interconnections between datasets among other features stored in the catalogue in order to find datasets of interest.

9 http://www4.wiwiwiss.fu-berlin.de/lodcloud/state/2011_09_index.html

10 <http://ckan.net/package/search?q=tags:media+AND+groups:lodcloud+AND+-tags:lodcloud.unconnected+AND+-tags:lodcloud.needsfixing>

11 <http://rdfs.org/ns/void#>

12 <http://www.w3.org/ns/dcat#>

13 <http://moat-project.org/ns#>

```

<http://ckan.net/package/dbpedia>
  moat:taggedWithTag
    <http://ckan.net/tag/crossdomain>,
    <http://ckan.net/tag/deref-vocab>,
    <http://ckan.net/tag/format-foaf>,
    ...
    <http://ckan.net/tag/wikipedia> ;
  dc:contributor [
    foaf:mbox <mailto:dbpedia-discussion@lists.sourceforge.net> ;
    foaf:name "DBpedia Team - http://wiki.dbpedia.org/Imprint"
  ] ;
  dc:created "2007-04-10T21:19:38Z"^^xsd:dateTime ;
  dc:description ""<h3>Description</h3> ... ""
  dc:modified "2011-09-01T16:40:08Z"^^xsd:dateTime ;
  dc:rights lic:cc-by-sa ;
  dc:title "DBpedia" ;
  rev:rating "2.50"^^xsd:float ;
  void:exampleResource
    <http://dbpedia.org/page/DBpedia>,
    <http://dbpedia.org/resource/Berlin> ;
  void:sparqlEndpoint <http://dbpedia.org/sparql> ;
  void:subset [
    void:target <http://ckan.net/package/dbpedia>,
    <http://ckan.net/package/freebase> ;
    void:triples 3400000 ;
    a void:Linkset ];

```

Snippet from: <http://semantic.ckan.net/record/dcc6715c-bf94-4a89-bbf3-35933da795a5.ttl>

Text 1: Machine readable (Turtle) description of the DBpedia dataset

2. State of the LOD Cloud

The document at http://www4.wiwiss.fu-berlin.de/lodcloud/state/2011_09_index.html provides statistics about the structure and content of the LOD cloud. It also analyses the extend to which LOD data sources implement nine best practices that are either recommended W3C or have emerged within the LOD community. All statistics within this document are based on the LOD data set catalogue that is maintained on TheDataHub.org.

The promise of the Web of Linked Data is to enable applications to discover and integrate data from an global Web of interconnected data sources. In order to make it as easy as possible for applications to access and process Linked Data, data providers should publish data according to a set of best practices. These best practices recommend to make data accessible using the [Web's standard access mechanism](#) (HTTP) and represent data using [standardized Web formats](#) (i.e. RDF/XML, RDFa, XML with GRDDL). On the other hand, the best practices aim at making data as [self-descriptive](#) as possible. The best practices are either recommended directly by W3C or have emerged within the [LOD community](#).

This section analyses to which extend data sources in the LOD cloud implement these best practices.

2.1.1 Provide dereferencable URIs

The basic idea of Linked Data is to make data accessible using the Web's standard retrieval algorithm. This means that every entity of interest, for instance a person, place, company or abstract concept, should be identified with its own http URI ([Linked Data Principle 1](#)). On the other hand, these URIs should be made dereferencable into an RDF description of the entity ([Linked Data Principle 2 and 3](#)). The W3C Interest Group Note [Cool URIs for the Semantic Web](#) describes the different technical options for realizing such lookup operations (hash vs. slash URIs).

100% of the data sources in the LOD cloud fulfill this best practice as it is a pre-condition for being included into the LOD cloud.

There are two validators available for checking whether URIs fulfill this best practice:

- [RDF:Alerts](#)
- [Vapour Linked Data Validator](#)

2.1.2 Set RDF links pointing at other data sources

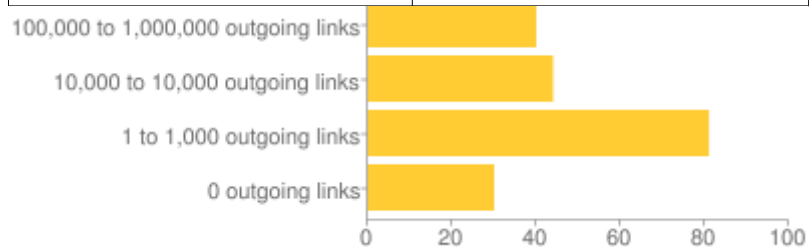
The [4th Linked Data principle](#) is to set RDF links pointing into other data sources. These RDF links connect data from different sources into a single global RDF graph and enable Linked Data browsers and crawlers to navigate between data sources.

The absolute number of RDF links in the LOD cloud is given in [Section 1.2](#). Table 2 categorizes the datasets in the LOD cloud by the absolute number of outgoing RDF links.

Table 2: Data sets categorized by absolute number of outgoing links

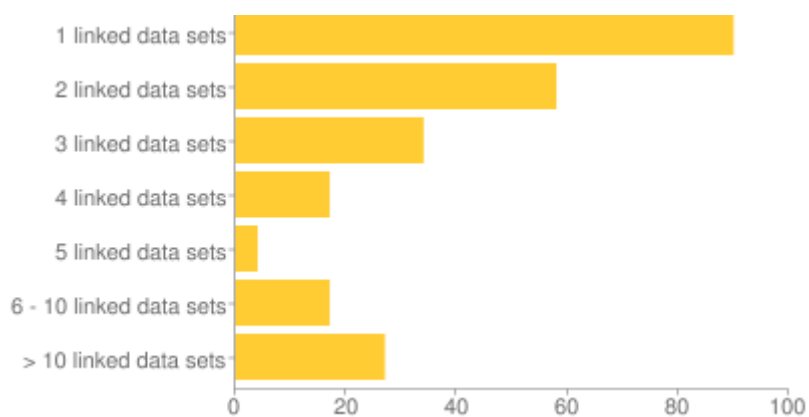
(Out-)Links	Number of datasets
up to 1,000	30 (10.87 %)
up to 1,000	81 (29.35 %)

10,000 to 10,000	53 (19.20 %)
10,000 to 100,000	44 (15.94 %)
100,000 to 1,000,000	40 (14.49 %)
more than 1,000,000	28 (10.14 %)
	223



The next table categorizes the LOD data sets by the number of other data sources that are target of outgoing RDF links.

Number of linked datasets	Number of datasets
more than 10	27 (9.78 %)
6 to 10	17 (6.16 %)
5	4 (1.45 %)
4	17 (6.16 %)
3	34 (12.32 %)
2	58 (21.01 %)
1	90 (32.61 %)
	247



The table below lists the 10 LOD data sets with the highest number of linked other LOD data sets.

Number of linked datasets	Dataset
35	rkb-explorer-dblp
31	rkb-explorer-southampton
31	rkb-explorer-eprints

31	rkb-explorer-acm
29	rkb-explorer-wiki
28	dbpedia
27	rkb-explorer-rae2001
27	rkb-explorer-citeseer
25	rkb-explorer-newcastle
25	rkb-explorer-kisti

2.1.3 Use terms from widely deployed vocabularies

In order to make it easier for applications to understand Linked Data, data providers should use terms from widely deployed vocabularies to represent data wherever possible.

Nearly all data sources in the LOD cloud use terms from the W3C base-vocabularies RDF, RDF Schema, and OWL. In addition 178 (64.49 %) of the 276 data sources in the LOD cloud use terms from other non-proprietary vocabularies. When calculating this number, we consider a vocabulary to be proprietary if it is defined in the same top-level domain that is also used to serve the instance data.

The table below lists the most widely used vocabularies and provides links to the data sources that use a specific vocabulary.

Vocabulary prefix	Vocabulary link	Number of usages in data sets	Data sets that use the vocabulary
dc	http://purl.org/dc/elements/1.1/	88 (31.88 %)	Data sets that use dc
foaf	http://xmlns.com/foaf/0.1/	79 (28.62 %)	Data sets that use foaf
skos	http://www.w3.org/2004/02/skos/core#	54 (19.57 %)	Data sets that use skos
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	25 (9.06 %)	Data sets that use geo
xhtml	http://www.w3.org/1999/xhtml/vocab#	18 (6.52 %)	Data sets that use xhtml
akt	http://www.aktors.org/ontology/portal#	17 (6.16 %)	Data sets that use akt
mo	http://purl.org/ontology/mo/	13 (4.71 %)	Data sets that use mo
bibo	http://purl.org/ontology/bibo/	13 (4.71 %)	Data sets that use bibo
sioc	http://rdfs.org/sioc/ns#	9 (3.26 %)	Data sets that use sioc
cc	http://creativecommons.org/ns#	8 (2.90 %)	Data sets that use cc
vcard	http://www.w3.org/2006/vcard/ns#	7 (2.54 %)	Data sets that use vcard
frbr	http://purl.org/vocab/frbr/core#	6 (2.17 %)	Data sets that use frbr
geonames	http://www.geonames.org/ontology#	6 (2.17 %)	Data sets that use geonames
time	http://www.w3.org/2006/time#	5 (1.81 %)	Data sets that use time
xsd	http://www.w3.org/2001/XMLSchema#	5 (1.81 %)	Data sets that use xsd
event	http://purl.org/NET/c4dm/event.owl#	5 (1.81 %)	Data sets that use event
dbpedia	http://dbpedia.org/resource/	5 (1.81 %)	Data sets that use dbpedia
dbo	http://dbpedia.org/ontology/	4 (1.45 %)	Data sets that use dbo
ore	http://www.openarchives.org/ore/terms/	4 (1.45 %)	Data sets that use ore
bio	http://purl.org/vocab/bio/0.1/	4 (1.45 %)	Data sets that use bio
tag	http://www.holygoat.co.uk/owl/redwood/0.1/tags/	3 (1.09 %)	Data sets that use tag

uniprot	http://purl.uniprot.org/core/	3 (1.09 %)	Data sets that use uniprot
umbel	http://umbel.org/umbel#	3 (1.09 %)	Data sets that use umbel
http	http://www.w3.org/2006/http#	3 (1.09 %)	Data sets that use http
rev	http://purl.org/stuff/rev#	3 (1.09 %)	Data sets that use rev
void	http://rdfs.org/ns/void#	3 (1.09 %)	Data sets that use void
dbp	http://dbpedia.org/property/	3 (1.09 %)	Data sets that use dbp
scovo	http://purl.org/NET/scovo#	3 (1.09 %)	Data sets that use scovo
mpeg7		2 (0.72 %)	Data sets that use mpeg7
doap	http://usefulinc.com/ns/doap#	2 (0.72 %)	Data sets that use doap
metalex		2 (0.72 %)	Data sets that use metalex
geospecies	http://rdf.geospecies.org/ont/geospecies#	2 (0.72 %)	Data sets that use geospecies
vu-wordnet		2 (0.72 %)	Data sets that use vu-wordnet
wot	http://xmlns.com/wot/0.1/	2 (0.72 %)	Data sets that use wot
api		2 (0.72 %)	Data sets that use api
admingeo	http://data.ordnancesurvey.co.uk/ontology/admingeo/	2 (0.72 %)	Data sets that use admingeo
wdrs	http://www.w3.org/2007/05/powder-s#	2 (0.72 %)	Data sets that use wdrs
sawsdl	http://www.w3.org/ns/sawsdl#	2 (0.72 %)	Data sets that use sawsdl
txn	http://lod.taxonconcept.org/ontology/txn.owl#	2 (0.72 %)	Data sets that use txn
bib	http://zeitkunst.org/bibtex/0.1/bibtex.owl#	2 (0.72 %)	Data sets that use bib
gr	http://purl.org/goodrelations/v1#	2 (0.72 %)	Data sets that use gr
compass	http://purl.org/net/compass#	2 (0.72 %)	Data sets that use compass
rdfg	http://www.w3.org/2004/03/trix/rdfg-1/	2 (0.72 %)	Data sets that use rdfg
tl	http://purl.org/NET/c4dm/timeline.owl#	2 (0.72 %)	Data sets that use tl
dcam	http://purl.org/dc/dcam/	2 (0.72 %)	Data sets that use dcam
swrc	http://swrc.ontoware.org/ontology#	2 (0.72 %)	Data sets that use swrc

An alternative view on vocabulary usage “in the wild” - i.e. in data crawled from the Web in contrast to data catalogued on TheDataHub.org – is given in section 3.

2.1.4 Make proprietary vocabulary terms dereferencable

Widely deployed vocabularies often do not provide all terms that are needed to publish the complete content of a data set on the Web. Thus, data providers often define proprietary terms that are used in addition to terms from widely deployed vocabularies.

Currently:

- 175 (63.41 %) out of the 276 data sources use proprietary vocabulary terms.
- 100 (36.23 %) out of the 276 data sources do not use proprietary vocabulary terms.

In order to enable applications to automatically retrieve the definition of vocabulary terms from the Web, URIs identifying vocabulary terms should be made dereferencable. Guidelines for doing this are given in the W3C Note [Best Practice Recipes for Publishing RDF Vocabularies](#).

Currently:

- [149 \(85.14 %\)](#) out of these 175 data sources make proprietary term URIs deferencable.
- [26 \(14.86 %\)](#) out of these 175 data sources do not make proprietary term URIs deferencable.

Split by topical domain, the numbers look as follows:

Domain	Proprietary vocabulary terms	Dereferencable proprietary term URIs	Not dereferencable proprietary term URIs	No proprietary vocabulary terms
Media	17/27 (62.96 %)	13/27 (48.15 %)	4/27 (14.81 %)	10/27 (37.04 %)
Geographic	21/26 (80.77 %)	14/26 (53.85 %)	7/26 (26.92 %)	5/26 (19.23 %)
Government	34/44 (77.27 %)	30/44 (68.18 %)	4/44 (9.09 %)	10/44 (22.73 %)
Publications	58/86 (67.44 %)	52/86 (60.47 %)	6/86 (6.98 %)	28/86 (32.56 %)
Cross-domain	26/36 (72.22 %)	23/36 (63.89 %)	3/36 (8.33 %)	9/36 (25.00 %)
Life sciences	13/42 (30.95 %)	11/42 (26.19 %)	2/42 (4.76 %)	29/42 (69.05 %)
User-generated content	5/14 (35.71 %)	5/14 (35.71 %)	0/14 (0.00 %)	9/14 (64.29 %)

2.1.5 Map proprietary vocabulary terms to other vocabularies

Proprietary vocabulary terms should be related to corresponding terms within other (widely used) vocabularies in order to enable applications to understand as much data as possible and to translate data into their target schemata (see [RDF and the Self-Describing Semantic Web](#)). The W3C recommendations define the following terms for representing such correspondences

(mappings): [owl:equivalentClass](#), [owl:equivalentProperty](#), or if a looser mapping is desired: [rdfs:subClassOf](#), [rdfs:subPropertyOf](#), and [skos:broadMatch](#), [skos:narrowMatch](#).

Currently [15 \(8.57 %\)](#) out of the 175 data sources that use proprietary terms provide mappings to other vocabularies for their terms.

- [List of data sources that provide mappings for proprietary terms.](#)
- [List of data sources that do not provide mappings for proprietary terms.](#)

2.1.6 Provide provenance metadata

In order to enable applications to be sure about the origin of data as well as to enable them to assess the quality of data, data source should publish provenance meta data together with the primary data. A common means for providing provenance information is to represent it as document level metadata as described in the [How to publish Linked Data](#) tutorial. A widely deployed vocabulary for representing provenance information is [Dublin Core](#) ([dc:creator](#), [dc:publisher](#), [dc:date](#)). Alternative vocabularies which provide for representing more details about the data creation process include the [Open Provenance Model](#) as well as the vocabularies examined by the [W3C Provenance Incubator Group](#).

Currently:

- 83 (35.17 %) out of the 276 data sources provide provenance information.
- 153 (64.83 %) out of the 276 data sources do not provide provenance information.

Split by topical domain, the figures look as follows:

Domain	Provenance information
Media	5/27 (18.52 %)
Geographic	12/26 (46.15 %)
Government	9/44 (20.45 %)
Publications	41/86 (47.67 %)
Cross-domain	7/36 (19.44 %)
Life sciences	2/42 (4.76 %)
User-generated content	6/14 (42.86 %)

2.1.7 Provide licensing metadata

Web data should be self-descriptive concerning any restrictions that apply to its usage. A common way to express such restrictions is to attach a [data license](#) to published data. Doing so is essential to enable applications to use Web data on a secure legal basis. A common means to attach licenses to Linked Data is to use `dc:rights` links pointing at the license as document-level metadata. An example of this is given in the [How to publish Linked Data](#) tutorial.

Currently:

- 40 (16.95 %) out of the 276 data sources provide licensing information.
- 196 (83.05 %) out of the 276 data sources do not provide licensing information.

Split by topical domain, the figures look as follows:

Domain	Licensing information
Media	4/27 (14.81 %)
Geographic	9/26 (34.62 %)
Government	6/44 (13.64 %)
Publications	8/86 (9.30 %)
Cross-domain	7/36 (19.44 %)
Life sciences	1/42 (2.38 %)
User-generated content	4/14 (28.57 %)

2.1.8 Provide dataset-level Metadata

In addition to making instance data self-descriptive, it is also desirable that data publishers provide metadata describing characteristic of complete data sets, for instance the topic of a dataset and more detailed statistics. A vocabulary for representing such metadata is the [voID vocabulary](#). A second means for representing dataset-level metadata are [Semantic Sitemaps](#).

Currently,

- 89 (32.25 %) out of the 276 data sources provide a void description.
- 50 (18.12 %) out of the 276 data sources provide a Semantic Sitemap.
- 102 (36.96 %) out of the 276 data sources provide either a void description or Semantic Sitemap.
- 174 (63.04 %) out of the 276 data sources do not provide either a void description or Semantic Sitemap.

Split by topical domain, the figures look as follows:

Domain	void	Semantic Sitemap	void or Semantic Sitemap
Media	6/27 (22.22 %)	0/27 (0.00 %)	6/27 (22.22 %)
Geographic	11/26 (42.31 %)	3/26 (11.54 %)	12/26 (46.15 %)
Government	18/44 (40.91 %)	1/44 (2.27 %)	19/44 (43.18 %)
Publications	39/86 (45.35 %)	35/86 (40.70 %)	43/86 (50.00 %)
Cross-domain	6/36 (16.67 %)	5/36 (13.89 %)	9/36 (25.00 %)
Life sciences	3/42 (7.14 %)	6/42 (14.29 %)	7/42 (16.67 %)
User-generated content	5/14 (35.71 %)	0/14 (0.00 %)	5/14 (35.71 %)

2.1.9 Refer to additional access methods

The primary way to publish Linked Data on the Web is to make the URIs that identity data items dereferencable into RDF descriptions. In addition, various LOD data providers have chosen to provide two alternative means of access to their data:

1. SPARQL endpoints which allow expressive queries to be asked against the datasets.
2. They provide RDF dumps of the complete dataset for download from a separate URL.

Altogether 187 (67.75 %) out of the 276 data sources provide a SPARQL endpoint. 106 (38.41 %) out of the 276 data sources provide RDF dumps.

The [void vocabulary](#) provides terms for pointing applications from the description of a single entity to these alternative means of access. For this void recommends to use a link of the type `dcterms:isPartOf` to point from the entity description to a void description of the complete dataset. This dataset description in turn may contain `void:sparqlEndpoint` and `void:dataDump` links pointing at SPARQL endpoint and the download URI of RDF dataset dumps.

Currently,

- 43 (42.16 %) out of the 102 void descriptions and Semantic Sitemaps point to the data source's SPARQL endpoint.
- 38 (37.25 %) out of the 102 void descriptions and Semantic Sitemaps point to the data source's RDF dump(s).

3. Vocabulary Catalogue

The usefulness of publicly exposed data can be significantly improved by using metadata vocabularies that can be used to describe and enrich data sets in a machine-understandable manner. This principle was especially leveraged in the establishment of Linked Data (LD) on the Web, which has ushered in a remarkable new era of widespread semantics-based interoperation. In particular, the use of RDF and SPARQL allows for a high degree of generic tool support, lightweight composition of data sources and enables a new level of automation in terms of discoverability and processing of the data.

Thousands of vocabularies exist on the Web that can be used to provide information about published data. However, which domains are covered by these existing vocabularies and their relevance for industry and Web technology providers is so far not finally determined. This leads to a situation in which data providers and RDF developers often face the difficult question, if they have to devise and publish their own new vocabulary to describe the topic of their data, or if they can make use of already existing vocabularies.

The reuse of such existing vocabularies improves the interlinkage of datasets in the Web and should therefore, if possible and appropriate, be preferred over the creation of new metadata. This gives rise to a need of data publishers to swiftly acquire information about already accessible vocabularies, which can potentially serve to describe the topics the publishers are concerned with. Furthermore, when such candidates for vocabularies are found, developers need the means to decide on their relevance.

3.1 Analysis of existing Linked Data vocabularies

To provide a first insight in the existing vocabularies on the Web of Data and to ease the task of identifying the relevant ones for Data publishers and developers, we analysed a crawled LD dataset and extracted information from it. This information can be used as an indicator to answer the question if vocabularies for certain domains exist and how often they are used in the Web of Data. The latter can be seen as a pointer to the relevance of a specific dataset.

3.1.1 The Billion Triple Challenge Dataset

As basis for our analysis we used the Billion Triple Challenge Dataset 2011¹⁴ (BTCD), which was crawled from the Web in May/June 2011 using a random sample of URIs from the BTC 2010 dataset as seed URIs. The BTCD contains over 2.1 billion statements in N-Quad¹⁵ format.

This format extends the N-Triple syntax for RDF with a fourth context node. The additional node is used in the BTCD to provide provenance information about a found LD triple (i.e. the original dataset a triple was found in during the crawl). Furthermore, N-Quads inherit the practical advantages of N-Triples: simple parsing; succinctness compared to alternatives such as reification or multi-document archives; effective streaming and processing with line-based tools.

3.1.2 Extraction of information

Considering the size of the BTCD (~20 GB), we used Apache Hadoop¹⁶ software library for the analysis of the data. Hadoop is an open source implementation of Google MapReduce and allows for the parallel and distributed processing of large data sets across clusters of computers.

For the necessary infrastructure we made use of the KIT OpenCirrus¹⁷ Hadoop Cluster. OpenCirrus1 is a collaboration of several organizations to provide an open cloud-computing research test bed designed to support research into the design, provisioning and management of services at a global scale. For our analysis we used 54 work nodes, each with a 2.27 GHz 4-Core CPU and 100GB RAM.

We extracted from the BTCD all URIs, that were used as predicates (i.e. properties) and all URIs that represent a class. The latter were identified by being in object position in a triple with “*rdf:type*” as predicate. Since vocabularies provide their meta-information essentially by defining properties and classes, we gain a first insight in the existing vocabularies.

For each identified class and property, we counted how often it was used in the BTCD overall. Beyond that, by regarding the given provenance information, we ascertained for every identified URI how many of the original datasets made use of it. This established two measurements for the commonness of the URI (each for classes and properties) and therefore for the underlying vocabulary, which in turn is an indicator for relevance. Eventually it allows creating two rankings for classes and properties respectively: One with respect to the overall appearance and one normalised over the datasets they appear in.

Finally we extracted all labels the identified URIs were tagged with (i.e., all strings in Object position of a triple with “*rdf:label*” as predicate and one of the relevant URIs as subject), to gain information about the domains that apply to these URIs. To account for the fact, that many URIs are intended to be human-readable, we additionally extracted the last segment of every identified URI and added it to the set of labels for this property or class. Since the BTCD is collected from the web and as such of varying quality, we had to perform several optimizations to exclude noisy and useless data from the labels.

¹⁴<http://km.aifb.kit.edu/projects/btc-2011/>

¹⁵<http://sw.deri.org/2008/07/n-quads/>

¹⁶<http://hadoop.apache.org/>

¹⁷<https://opencirrus.org/>

3.2 vocab.cc

To make this preliminary information easily accessible for RDF developers, we devised and published a Web portal called *vocab.cc*¹⁸ that allows to lookup information about URIs and search for them via a clean interface thus realising a highly intuitive user interaction. Additionally all information is offered as RDF itself via Linked Services. The source code of the services is licensed under LGPLv3 the content under CCBYSA.

3.2.1 URI lookup

Vocab.cc allows users to specify a URI. If this URI was used in the BTCD as Property or class, it is identified as such and the information about it are returned. This information includes the number of overall appearances in the BTCD as well as the number of datasets it appeared in. Additionally the positions in the rankings of classes or properties based on these numbers are returned.

This functionality allows developers, who consider the use of a certain classes and properties to provide information about their data, to gain a first indicator about these classes and properties. This can help them to decide whether a vocabulary is well-established and therefore potentially useful and of high relevance for him, or if it is rather exotic and not well adopted.

3.2.2 URI search

If a data publisher does not know which vocabulary to use to describe his data or even if one exists, he can make use of vocab.cc to find and compare them by defining an arbitrary query (i.e., a string of words) for his domain of interest. The search functionality matches the words in this query with the labels found for the URIs and returns all classes and properties, which labels contain all of the specified words. To increase the number of potential result sets, words in the query that do not appear in any label are ignored.

This allows to specify an object or a relation that appears in a dataset, and to receive a list of URIs from vocabularies, that can be used to classify the object or identify the relation respectively. To ease the decision for a URI (and therefore a vocabulary), the number overall appearances in the BTCD is provided directly by the search functionality. For a more detailed comparison, the lookup functionality can be used as described.

3.2.3 Web portal

The Web portal provides a clean interface with a central input field (Illustration 7), where users can specify a URI or type their query. The portal distinguishes between these possibilities automatically. This is intended to realise a highly intuitive user interaction. To increase this further, URIs can be specified with their common namespace rather than their fully qualified name. The namespaces are resolved automatically, by making use of data from prefix.cc¹⁹, which also inspired name and layout from the vocab.cc Web portal.

¹⁸<http://www.vocab.cc/>

¹⁹<http://prefix.cc/>

type a specific uri or an arbitrary query

Illustration 7: vocab.cc main page

The List of URIs returned from a search query is directly linked to the lookup functionality for the individual URIs. This allows for an integrated use of both functionalities and an easy comparison between the results (Illustration 8 and Illustration 9). Links to the underlying vocabularies, that are generically provided by the URIs are indicated with orange arrows.

Search Results

maybe these URIs represent what you are looking for:

URI	Occured Overall	Type
http://openresearch.org/wiki/Special:URIResolver/Property-3AHas_program_chair →	1589	Property
http://semanticweb.org/id/Property-3AHas_program_chair →	448	Property
http://semanticweb.org/id/Property-3AHas_area_program_chair →	45	Property
http://semanticweb.org/id/Property-3AHas_program_committee_chair →	4	Property

Illustration 8: vocab.cc search query results

Property

http://semanticweb.org/id/Property-3AHas_program_chair →

Occured overall 448 times
and in 233 datasets.

Is in Position 6 558 in the overall ranking
and in Position 6 741 of the dataset ranking.

Illustration 9: vocab.cc URI lookup result

Finally vocab.cc provides listings of the top 100 URIs in all the devised rankings.

3.2.4 Linked Vocab Services

Beyond the human readable way to access vocab.cc via the Web portal, the functionalities of vocab.cc can also be accessed as Linked Service. This allows for an easy integration of the functionalities in other applications, fostering the Linked Data principles. By combining LD technologies with RESTful services 1., Linked Services (LS) 2. offer Web service functionalities as

RDF prosumers and aim to establish client-service functionality against the background of the Web of Data.

In case of vocab.cc two such LS resources are offered so far: for the URI lookup²⁰ and the URI search²¹. In the payload of an HTTP POST operation RDF data can be submitted to these resources. Depending on the addressed service, this submitted data can either specify URIs for the lookup functionality or a query string for the search functionality. The payload of the HTTP response contains RDF data as well, detailing the URIs found by the search service or usage information from the lookup service as described in 3.2.1.

The structure of this RDF data can be inferred via graph patterns in the service descriptions provided in the syntax of the SPARQL query language. The use of SPARQL graph patterns provide the advantage of familiarity to Linked Data producers and consumers, but also of a more thorough description of what should be communicated and the possibility for increased tool support (e.g., for service discovery 4.).

Following the motivation to provide an easy interlinkage between datasets 3., the output RDF can also be accessed directly via content negotiation: The service resource URIs extended with key/value pairs, which specify the user input (i.e., a concrete URI to lookup or a query string to search for) can be addressed with an HTTP GET method. If this HTTP request specifies in its header to accept the content-type *text/html*, the html documents, described in 3.2.3 are delivered. If instead *application/rdf+xml* or *text/N3* is accepted by an agent, the response contains RDF data.

²⁰<http://vocab.cc/lookup>

²¹<http://vocab.cc/search>

3.3 Further work

Even though the information provided by the vocab.cc portal and services are already very useful for data publishers to find information about existing vocabularies, they only represent a first insight into the available data.

We intend to develop vocab.cc further and to provide additional information and functionalities: By accounting for subclass and subproperty hierarchies more vocabularies in the Web of Data could be identified. Furthermore, by including information about the structure of the BTCD into the analysis (e.g., the size, number and links between the crawled original datasets), other valuable information can arise.

Finally it should be noted, that vocab.cc is not meant to compete with other approaches (like Linked Open Vocabularies²²) to analyse and structure the vocabulary space in the Web of Data , but rather to complement them and to provide additional functionalities, thus augmenting their results. In order to achieve this, we intend to make an effort to link existing information about LD vocabularies to the functionalities of vocab.cc.

²²<http://labs.mondeca.com/dataset/lov/index.html>

4. Conclusions

This report has presented an overview of the current data sources that expose linked data on the Web, a Web-accessible catalogue of these entries, as well as the process with which the catalogue was built. The catalogue includes results of a collaboration with LATC²³, a Specific Support Action in the context of the FP7 ICT Challenge 4, as well as the Open Knowledge Foundation (OKFN) a not-for-profit organization that promotes open knowledge, including open content and open data.

PlanetData has reused OKFNs TheDataHub.org portal and previous results from cataloguing activities in LATC. We have extended the metadata schema used in LATC, developed a set of guidelines²⁴ and extended validation scripts to match those guidelines. We individually provided through the PlanetDataEditor²⁵ user on TheDataHub.org a total of 49 new packages and over 540 edits to existing entries. Moreover, we supported the community in providing metadata about their own datasets, following this process with an in-house quality assurance step. After the quality assurance process, 276 data sets were promoted to Level 4 and added to the LOD cloud, totalling 31.5 billion triples and almost 500 million links between data sets.

To provide a first insight into the existing vocabularies on the Web of Data and to ease the task of identifying the relevant ones for Data publishers and developers, we analysed a crawled LD dataset and extracted information from it. This information can be used as an indicator to answer the question if vocabularies for certain domains exist and how often they are used in the Web of Data. The latter can be seen as a pointer to the relevance of a specific dataset. We have also provided a vocabulary search portal (<http://vocab.cc>) specially targeted at data producers in search for descriptors for their data sets. We hope that this will help data providers in increasing the understandability of their data sets.

The cataloguing guidelines and the validation script produced will remain available to catalyse the organic expansion of the catalogue. With the ever-increasing number of data producers on the Web, we expect many more data producers to become aware of the LOD cloud diagram and The State of the LOD cloud page, and actively use TheDataHub.org to catalogue their data sets. A new release of the catalogue is planned for M30, which will have the potential to include many more quality indicators that are being developed and will be described in D2.1 (M12).

23 <http://latc-project.eu/>

24 <http://www4.wiwiss.fu-berlin.de/lodcloud/ckan/validator/levels.html>

25 <http://ckan.net/user/PlanetDataEditor>

5. References

1. Roy Thomas Fielding. Architectural Styles and the Design of Network-based Software Architectures. PhD thesis, University of California, Irvine, 2000.
2. Barry Norton, Reto Krummenacher, Adrian Marte, and Dieter Fensel. Dynamic linked data via linked open services. Workshop on Linked Data in the Future Internet at the Future Internet Assembly, pages 1-10, 2010.
3. Sebastian Speiser and Andreas Harth. Integrating linked data and services with linked data <http://www.ics.forth.gr/isl/PlanetData/services>. In Proceedings of 8th Extended Semantic Web Conference, ESWC 2011, pages 170-184, 2011.
4. Barry Norton and Steffen Stadtmüller. Scalable discovery of linked services. In Proceedings of the Fourth International Workshop on REsource Discovery, volume 737, Heraklion, Greece, Mai 2011. RED Workshop, CEUR-WS.