CHAPTER 3

Linked Open Data

In contracts to the isolated data silos of the conventional Web, the Semantic Web interconnects open data, so that all datasets contribute to a global data integration, connecting data from diverse domains, such as people, companies, books, scientific publications, films, music, reviews, television and radio programs, medicine, statistics, online communities, and scientific data. The union of the structured datasets forms the Linked Open Data Cloud, the decentralized core of the Semantic Web, where software agents can automatically find relationships between entities and make new discoveries. Linked Data browsers allow users to browse a data source, and by using special (typed) links, navigate along links into other related data sources. Linked Data search engines crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data. To support data processing for new types of applications, Linked Open Data (LOD) is used by search engines, governments, social media, publishing agencies, media portals, researchers, and individuals.

Linked Data Principles

Conventional web pages are hypertext documents connected with hyperlinks (or simply links). These hyperlinks point to other documents or a part of another document; however, they do not hold information about the type of relationship between the source and the destination resources. While the link relation can be annotated using the `rel` attribute on the `link`, `a`, and `area` markup elements, they are suitable for annotating external CSS files, script files, or the favicon. As mentioned before, some microformats such as `rel="tag"` and XFN also declare link relations. Other specific relation types can be defined in the Atom syndication format and XLink. On the Semantic Web, links can be typed using `rdf:type` or its equivalent in other serializations, such as the `datatype` attribute in RDFa, to provide a machine-interpretable definition for an arbitrary relationship between the source and destination resources. Those structured datasets derived from different resources that are published with such *typed links* between them are called *Linked Data* (also known as *Linking Data*) [1].

Berners-Lee outlined four *Linked Data principles* for publishing and interlinking data on the Web in a human- and machine-readable way, using Semantic Web technologies, so that all published data becomes part of a single global data space [2].

1. Use URIs as names for the "things" of the Web of Data (real-world objects and people). In other words, a *dereferenceable* Uniform Resource Identifier (URI), such as a web address, is assigned to each resource rather than an application-specific identifier, such as a database key or incremental numbers, making every data entity individually identifiable.

■ **Note** The dereferenceable URIs must be *HTTPRange-14*-compliant. For years, HTTPRange-14 was a design issue of the Semantic Web, because when HTTP was extended from referring only to documents to referring to "things" (real-world objects and persons), the domain of HTTP GET became undefined, leading to ambiguous interpretations of Semantic Web resources. The resolution is to check the web server's answer to the `GET` request, and if an HTTP resource responds with a 2xx response, the resource identified by that URI is an information resource; if it is a 303 (See Other) response, the resource identified by that URI could be any resource. A 4xx (error) response means that the nature of the resource is unknown [3].

2. Use HTTP URIs, so that people can look up the resource names. In other words, provide the URIs over the HTTP protocol into RDF representations for dereferencing.
3. When someone looks up a URI, provide useful information using Semantic Web standards, such as RDF. By providing metadata about the published data, clients can assess the quality of published data and choose between different means of access.
4. Include links to other URIs, so that users can discover related information. When RDF links are set to other data resources, users can navigate the Web of Data as a whole by following RDF links.

The benefits of Linked Data are recognized by more and more organizations, businesses, and individuals. Some industrial giants that already have LOD implementations are Amazon.com, BBC, Facebook, Flickr, Google, Thomson Reuters, The New York Times Company, and Yahoo!, just to name a few.

The Five-Star Deployment Scheme for Linked Data

Publishing Linked Data (following the Linked Data principles) does not guarantee data quality. For example, the documents the URIs in LOD datasets point to might be documents that are difficult to reuse. Pointing to a fully machine-interpretable RDF file is not the same as pointing to a PDF file containing a table as a scanned image. A five-star rating system is used for expressing the quality of Linked Data which are not open, and Linked Open Data (open data and Linked Data at the same time) [4]. The five-star rating system is cumulative, meaning that on each level, the data has to meet additional criteria beyond the criteria of the underlying level(s) [5]:

★	Data is available on the Web in any format, which is human-readable but not machine-interpretable, due to a vendor-specific file format or lack of structure. All following stars are intended to make the data easier to discover, use, and understand. For example, a scanned image of tabular data in a PDF file is one-star data. Data reusability is limited.
★★	Data is available as machine-readable structured data. For example, tabular data saved in an Excel file is two-star data.
★★★	Data is available in a nonproprietary (vendor-independent) format. For example, tabular data saved as a CSV file is three-star data.

(*continued*)

★★★★	Published using open standards from the W3C (RDF and SPARQL). For example, tabular data in HTML with RDFa annotation using URIs is four-star data.
★★★★★	All of the above plus links to other, related data to provide context. For example, tabular data in HTML with RDFa annotation using URIs and semantic properties is five-star data. Maximum reusability and machine-interpretability.

The expression of rights provided by licensing makes free data reuse possible. Linked Data without an explicit open license[1] (e.g., public domain license) cannot be reused freely, but the quality of Linked Data is independent from licensing. When the specified criteria are met, all five ratings can be used both for Linked Data (for Linked Data without explicit open license) and Linked Open Data (Linked Data with an explicit open license). As a consequence, the five-star rating system can be depicted in a way that the criteria can be read with or without the open license. For example, the Linked Open Data mug can be read with both green labels for five-star Linked Open Data, or neither label for five-star Linked Data, as shown in Figure 3-1. For example, Linked Data available as machine-readable structured data is two-star Linked Data, while the same with an open license is two-star Linked Open Data.

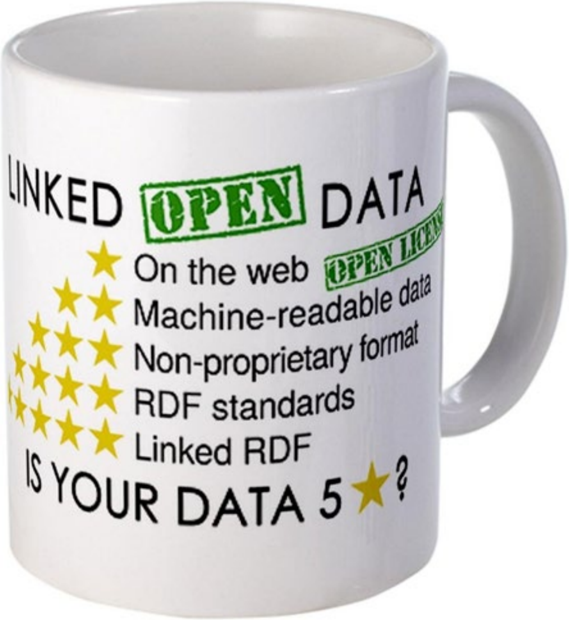


***Figure 3-1.** The requirements of 5 ★ Linked Data and 5 ★ Linked Open Data*

Because converting a CSV file to a set of RDF triples and linking them to another set of triples does not necessarily make the data more (re)usable to humans or machines, even four-star and five-star Linked Open Data have many challenges. One of the challenges is the lack of provenance information, which can now be provided about Linked (Open) Data using standards such as the PROV-O ontology [6]. Another challenge

[1] This licensing concept is used on the conventional Web too, in which the term *Open Data* refers to the free license.

is querying Linked Data that do not use machine-readable definitions from a vocabulary, which is difficult and almost impossible to interpret with software agents. Furthermore, the quality of the definitions retrieved from vocabularies and ontologies varies greatly, and the used vocabularies might not restrict the potential interpretations of the used classes and roles towards their intended meaning.

LOD Datasets

A meaningful collection of RDF triples covering a field of interest according to the Linked Open Data principles is called an *LOD dataset*. LOD datasets collect descriptions of entities within the field of interest, and these descriptions often share a common URI prefix (as, for example, `http://dbpedia.org/resource/`). The authors of the largest datasets provide advanced features that enable easy access to their structured data, such as downloadable compressed files of the datasets or an infrastructure for efficient querying.

RDF Crawling

Similar to the web crawlers that systematically browse conventional web sites for indexing, Semantic Web crawlers browse semantic contents to extract structured data and automatically find relationships between seemingly unrelated entities. LOD datasets should be published in a way so that they are available through *RDF crawling*.

RDF Dumps

The most popular LOD datasets are regularly published as a downloadable compressed file (usually Gzip or bzip2), called an *RDF dump*, which is the latest version of the dataset. RDF dumps should be valid RDF files. The reason why the RDF dump files are compressed is that the datasets containing millions of RDF triples are quite large. The size of Gzip-compressed RDF dumps is approximately 100MB per every 10 million triples, but it also depends on the RDF serialization of the dataset. Table 3-1 summarizes the most popular RDF dumps.

Table 3-1. *Popular RDF Dumps*

Dataset	RDF Dump
DBpedia	`http://wiki.dbpedia.org/Downloads2014`
WikiData	`http://dumps.wikimedia.org/wikidatawiki/`
GeoNames	`http://download.geonames.org/all-geonames-rdf.zip`
LinkedGeoData	`http://downloads.linkedgeodata.org/releases/`
Open Directory	`http://rdf.dmoz.org/`
MusicBrainz	`ftp://ftp.musicbrainz.org/pub/musicbrainz/data/`

SPARQL Endpoints

Similar to relational database queries in MySQL, the data of semantic datasets can also be retrieved through powerful queries. The query language designed specifically for RDF datasets is called *SPARQL* (pronounced "sparkle," it stands for *SPARQL Protocol and RDF Query Language*), which will be discussed in detail in Chapter 7. Some datasets provide a *SPARQL endpoint*, which is an address from which you can directly run SPARQL queries (powered by a back-end database engine and an HTTP/SPARQL server).

Frequently Used Linked Datasets

LOD datasets are published in a variety of fields. Interdisciplinary datasets such as DBpedia (http://dbpedia.org) and WikiData (http://www.wikidata.org) are general-purpose datasets and are, hence, among the most frequently used ones. Geographical applications can benefit from datasets such as GeoNames (http://www.geonames.org) and LinkedGeoData (http://linkedgeodata.org). More and more universities provide information about staff members, departments, facilities, courses, grants, and publications as Linked Data and RDF dump, such as the University of Florida (http://vivo.ufl.edu) and the Ghent University (http://data.mmlab.be/mmlab). Libraries such as the Princeton University Library (http://findingaids.princeton.edu) publish bibliographic information as Linked Data. Part of the National Digital Data Archive of Hungary is available as Linked Data at http://lod.sztaki.hu. Even Project Gutenberg is available as Linked Data (http://wifo5-03.informatik.uni-mannheim.de/gutendata/). Museums such as the British Museum publish some of their records as Linked Data (http://collection.britishmuseum.org). News and media giants publish subject headings as Linked Data, as for example the New York Times at http://data.nytimes.com. MusicBrainz (http://dbtune.org/musicbrainz/) provides data about music artists and their albums, served as Linked Data and via available through a SPARQL endpoint. Data about musicians, music album releases, and reviews are published as Linked Data by BBC Music at www.bbc.co.uk/music, which is largely based upon MusicBrainz and the Music Ontology. The Linked Movie DataBase (LinkedMDB) at http://www.linkedmdb.org is an LOD dataset dedicated to movies, with high quality and quantity of interlinks to other LOD data sources and movie-related web sites. More and more government portals publish publicly available government data as Linked Data, as, for example, the US government's http://data.gov or the UK government's http://data.gov.uk. Some of the most popular LOD datasets will be discussed in the following sections.

DBpedia

The hundreds of concept definitions on schema.org are suitable to annotate common knowledge domains, such as persons, events, books, and movies, but complex machine-readable statements require far more.

DBpedia, hosted at http://dbpedia.org, extracts structured factual data from Wikipedia articles, such as titles, infoboxes, categories, and links. Because Wikipedia contains nearly 5 million articles in English, DBpedia is suitable to describe virtually anything in a machine-readable manner. DBpedia contains approximately 3.4 million concepts described by 1 billion triples.

■ **Note** Wikipedia infoboxes are the most straightforward for DBpedia extraction, because they contain attribute-value pairs of the corresponding Wikipedia page to be displayed on the right-hand side of the article as a summary of the most important facts in a tabular form. However, structured data extraction is challenging, because the template system changed over time on Wikipedia, resulting in the lack of uniformity, whereby the same attributes have different names, such as `placeofbirth` and `birthplace`.

The unique resource identifiers of DBpedia are written as URI references of the form http://dbpedia.org/resource/*Name*, where *Name* is derived from the URL of the Wikipedia article of the form http://en.wikipedia.org/wiki/*Name*. As a result, each resource is a direct mapping of a Wikipedia article. The DBpedia URI references of the form http://dbpedia.org/resource/*Resource:Name* are set up (through content negotiation, where the same content is served in a different format, depending on the query of the client) to return the machine-readable description in RDF when accessed by Semantic Web agents, and the same information in XHTML, when accessed by traditional web browsers (see Figure 3-2).

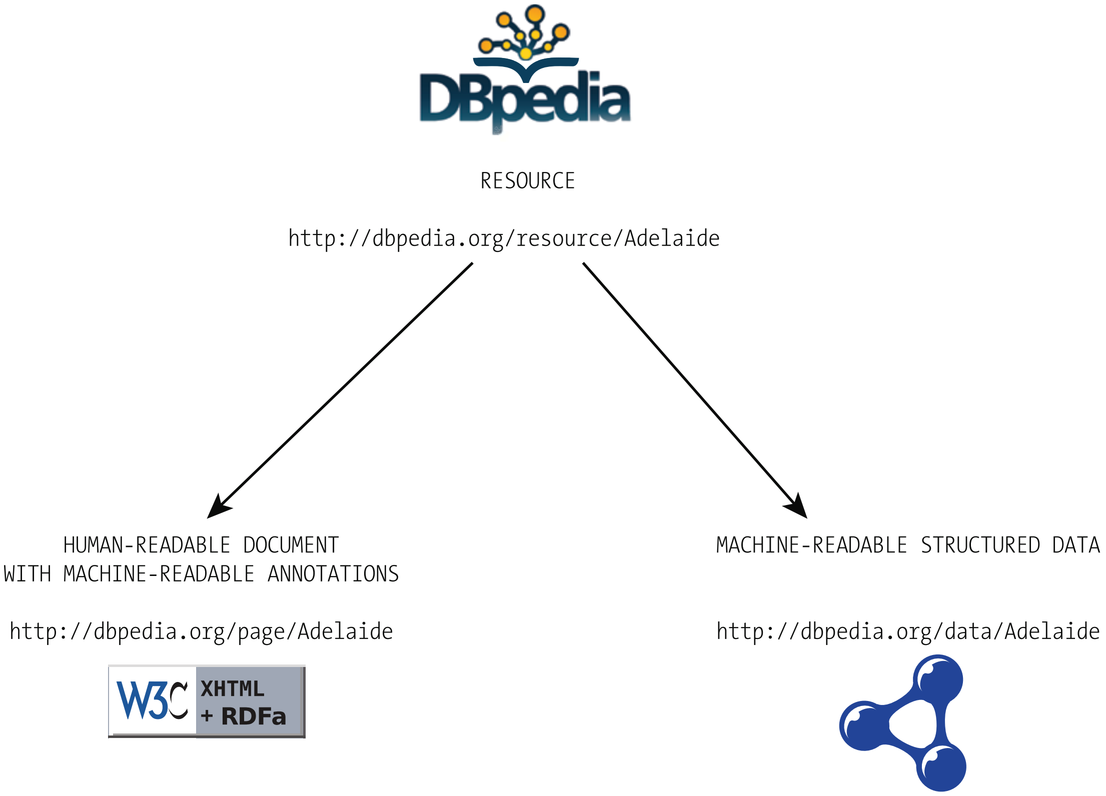

DBpedia
RESOURCE
http://dbpedia.org/resource/Adelaide
HUMAN-READABLE DOCUMENT
WITH MACHINE-READABLE ANNOTATIONS
http://dbpedia.org/page/Adelaide
W3C XHTML + RDFa
MACHINE-READABLE STRUCTURED DATA
http://dbpedia.org/data/Adelaide

Figure 3-2. *DBpedia resources return XHTML or RDF through content negotiation*

Assume we want to describe a Semantic Web researcher in RDF who lives in Adelaide, is interested in Web standards, and is a member of the W3C. To do this, we need the corresponding DBpedia URIs that identify the non-information resources (in the form http://dbpedia.org/resource/*Resource:name*) declared as the attribute value of rdf:resource (see Listing 3-1).

Listing 3-1. Linking to DBpedia Resources

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF ↩
xmlns:foaf="http://xmlns.com/foaf/0.1/" ↩
 xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#" ↩
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <foaf:person rdf:about="http://www.lesliesikos.com/datasets/sikos.rdf#sikos">
    <foaf:name>Leslie Sikos</foaf:name>
    <foaf:based_near rdf:resource="http://dbpedia.org/resource/Adelaide" />
    <foaf:topic_interest rdf:resource="http://dbpedia.org/resource/Web_standards" />
    <contact:nearestAirport rdf:resource="http://dbpedia.org/resource/Adelaide_Airport" />
  </foaf:person>
  <rdf:Description rdf:about="http://dbpedia.org/resource/W3C">
    <foaf:member rdf:resource="http://www.lesliesikos.com/datasets/sikos.rdf#sikos" />
  </rdf:Description>
</rdf:RDF>
```

The SPARQL endpoint of DBpedia is http://dbpedia.org/sparql, where you can run queries on DBpedia resources, say, the list of people born in Budapest before the 20th century (see Listing 3-2). Querying with SPARQL will be described later, in Chapter 7.

Listing 3-2. A SPARQL Query on DBpedia

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?name ?birth ?death ?person WHERE {
     ?person dbo:birthPlace :Budapest .
     ?person dbo:birthDate ?birth .
     ?person foaf:name ?name .
     ?person dbo:deathDate ?death .
     FILTER (?birth < "1901-01-01"^^xsd:date) .
}
ORDER BY ?name
```

Wikidata

Wikidata is one of the largest LOD databases that features both human-readable and machine-readable contents, at http://www.wikidata.org. Wikidata contains structured data from Wikimedia projects, such as Wikimedia Commons, Wikipedia, Wikivoyage, and Wikisource, as well as from the once popular directly editable Freebase dataset, resulting in approximately 13 million data items.

In contrast to many other LOD datasets, Wikidata is collaborative—anyone can create new items and modify existing ones. Like Wikipedia, Wikidata is multilingual. The Wikidata repository is a central storage of structured data, whereby data can be accessed not only directly but also through client Wikis. Data is added to items that feature a label, which is a descriptive alias, connected by site links. Each item is characterized by statements that consist of a property and property value. Wikidata supports the Lua Scribunto parser extension to allow embedding scripting languages in MediaWiki and access the structured data stored in Wikidata through client Wikis. Data can also be retrieved using the Wikidata API.

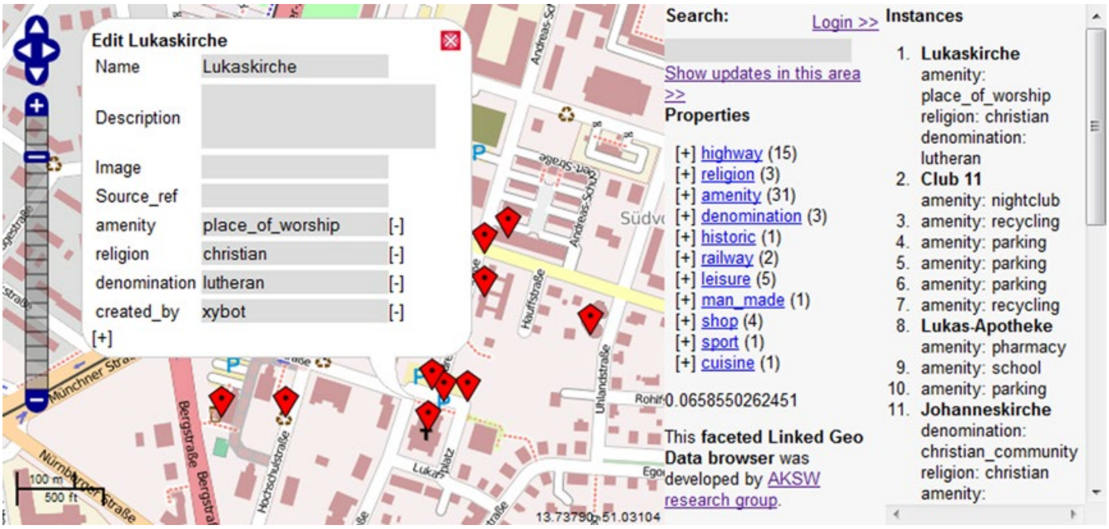
GeoNames

GeoNames is a geographical database at http://www.geonames.org that provides RDF descriptions for more than 7,500,000 geographical features worldwide, corresponding to more than 10 million geographical names. All features are categorized as one of the nine feature classes, and subcategorized into one of the 645 feature codes. Place-names are stored in the database in multiple languages. GeoNames also contains data such as latitude and longitude, elevation, population, administrative subdivision, and postal codes of cities. The coordinates are expressed in the World Geodetic System 1984 (WGS84) standard used in cartography, geodesy, and navigation.

The GeoNames resources use 303 (See Other) redirection to distinguish a concept (thing as is) from the document describing the resource. For example, the city of Adelaide has two addresses on GeoNames: http://sws.geonames.org/2078025/ and http://sws.geonames.org/2078025/about.rdf. The first represents the city (in a form used in Linked Data references); the second is a document with information about Adelaide.

LinkedGeoData

The *LinkedGeoData* dataset at http://linkedgeodata.org uses the information collected by OpenStreetMap data (a free editable world map), makes it available as an LOD dataset, and interlinks this data with other LOD datasets. The authors of the dataset provide their own semantic browser, called *LGD Browser and Editor*, at http://browser.linkedgeodata.org (see Figure 3-3).



Edit Lukaskirche

Name	Lukaskirche	
Description		
Image		
Source_ref		
amenity	place_of_worship	[-]
religion	christian	[-]
denomination	lutheran	[-]
created_by	xybot	[-]

[+]

Search:

Login >>

Show updates in this area >>

Properties

- [+] highway (15)
- [+] religion (3)
- [+] amenity (31)
- [+] denomination (3)
- [+] historic (1)
- [+] railway (2)
- [+] leisure (5)
- [+] man_made (1)
- [+] shop (4)
- [+] sport (1)
- [+] cuisine (1)

0.0658550262451

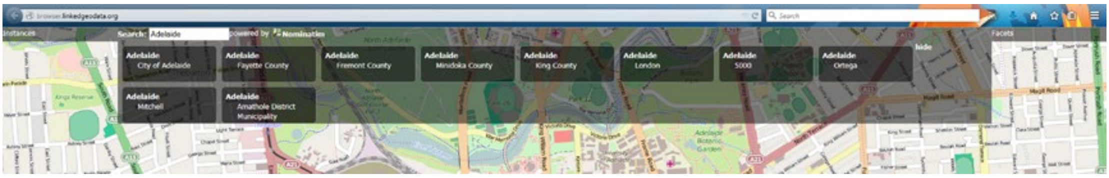
This faceted Linked Geo Data browser was developed by AKSW research group.

Instances

1. Lukaskirche
   amenity: place_of_worship
   religion: christian
   denomination: lutheran
2. Club 11
   amenity: nightclub
3. amenity: recycling
4. amenity: parking
5. amenity: parking
6. amenity: parking
7. amenity: recycling
8. Lukas-Apotheke
   amenity: pharmacy
9. amenity: school
10. amenity: parking
11. Johanneskirche
   denomination: christian_community
   religion: christian
   amenity:

Figure 3-3. *LinkedGeoData in the LGD Browser and Editor*

A good example for the unambiguity on the Semantic Web is searching for "Adelaide" in the LGD Browser. Because there is a city with this name in South Australia, another one in South Africa, and three in the United States (one in Colorado, one in Idaho, and one in Washington), the software will ask for clarification and provide the city map and details according to your choice (see Figure 3-4).



browser.linkedgeodata.org

Instances

Search: Adelaide powered by Nominatim

- Adelaide City of Adelaide
- Adelaide Fayette County
- Adelaide Fremont County
- Adelaide Minidoka County
- Adelaide King County
- Adelaide London
- Adelaide 5000
- Adelaide Ortega
- Adelaide Mitchell
- Adelaide Amathole District Municipality

Figure 3-4. *Linked Data is unambiguous*

YAGO

YAGO (Yet Another Great Ontology) is a dataset containing more than 10 million entities and 120 million facts about them, which are automatically extracted from Wikipedia categories, redirects, and infoboxes; synsets and hyponymy from the lexical database WordNet; and GeoNames.

In contrast to other datasets automatically extracting data from LOD datasets, YAGO is more accurate, as a large share of facts is manually evaluated. YAGO entities and facts are often linked to the DBpedia ontology. The SPARQL endpoint of YAGO is `http://lod2.openlinksw.com/sparql`, but queries can also be executed through a web interface at `https://gate.d5.mpi-inf.mpg.de/webyagospotlx/WebInterface`.

LOD Dataset Collections

LOD datasets can be registered and managed using *Datahub* at `http://datahub.io`, an open data registry. Datahub is used by governments, research institutions, and other organizations. Powered by structured data, datahub.io provides efficient search and faceting, browsing user data, previewing data using maps, graphs, and tables. As you will see, a datahub.io registry is a prerequisite for merging new datasets of the LOD Cloud Diagram with existing ones.

Ontobee, available at `http://www.ontobee.org`, is a SPARQL-based linked ontology data server and browser that has been utilized for more than 100 ontologies containing over 2 million ontology terms.

The LOD Cloud Diagram

The *LOD Cloud Diagram* represents datasets with at least 1,000 RDF triples and the links between them (Figure 3-5) [7]. The size of the bubbles corresponds to the data amount stored in each dataset. In the middle of the cloud, you can see the largest datasets, DBpedia and GeoNames, followed by FOAF-Profiles, Freebase, and the W3C.

If you have a big enough dataset that contains at least 1,000 triples and fulfills the requirements of Linked Open Data, you can make a request to add it to the LOD Cloud Diagram. The resources of the dataset must have resolvable `http://` or `https://` URIs that resolve, with or without content negotiation, to RDF data as RDFa, RDF/XML, Turtle, or N-Triples. The dataset must be connected through at least 50 RDF links to arbitrary datasets of the diagram. The dataset must be accessible via *RDF crawling*, via an *RDF dump*, or via a *SPARQL endpoint*. The dataset must be registered on Datahub, and you have to e-mail the authors of the LOD Cloud Diagram (`richard@cyganiak.de` and `mail@anjajentzsch.de`).

An alternate visualization of the LOD Cloud Diagram is created by Stanford University's Protovis, using the CKAN API, and published at `http://inkdroid.org/lod-graph/` (see Figure 3-6).

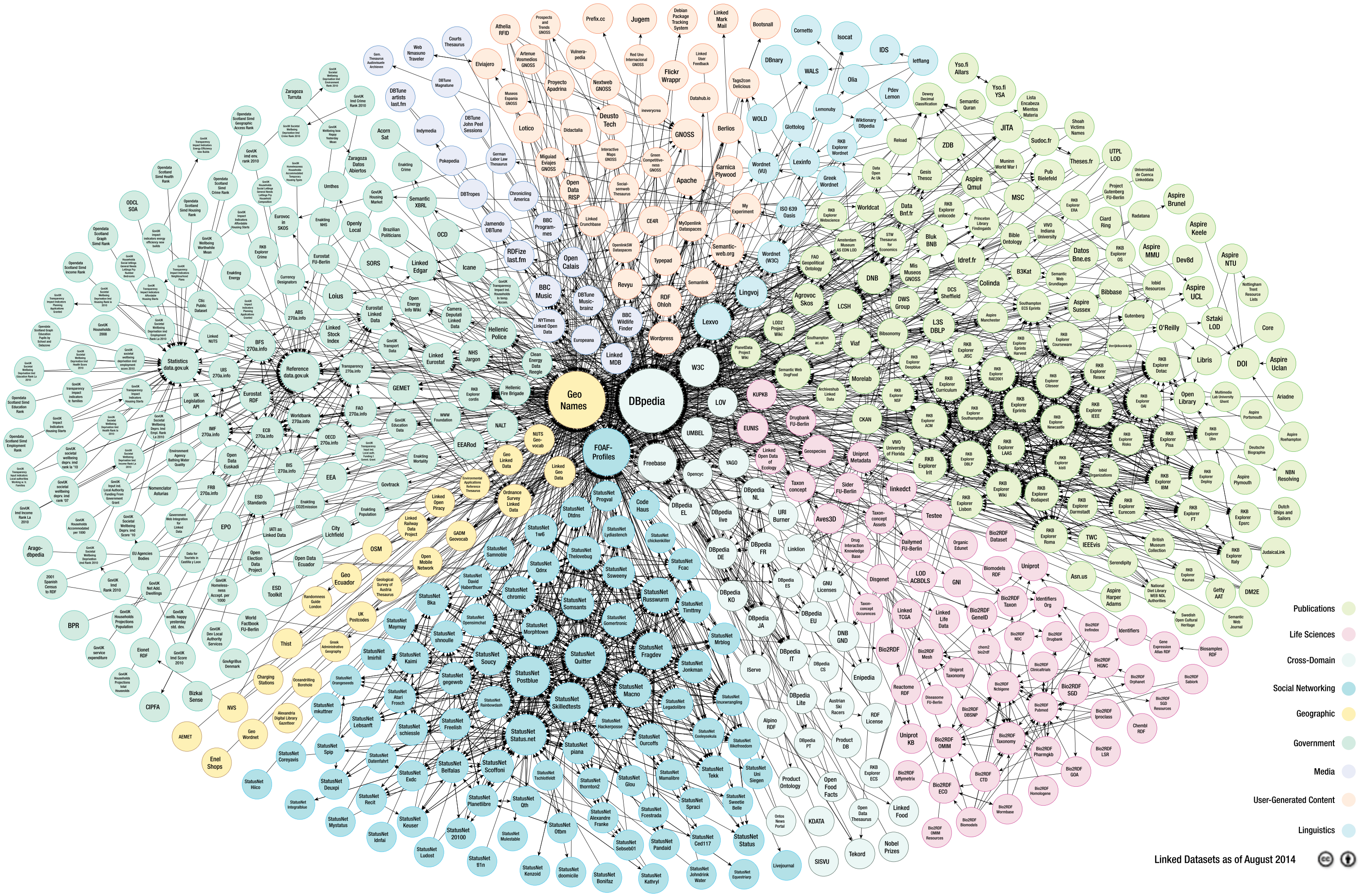


Figure 3-5. *The LOD Cloud Diagram (courtesy of Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak)*

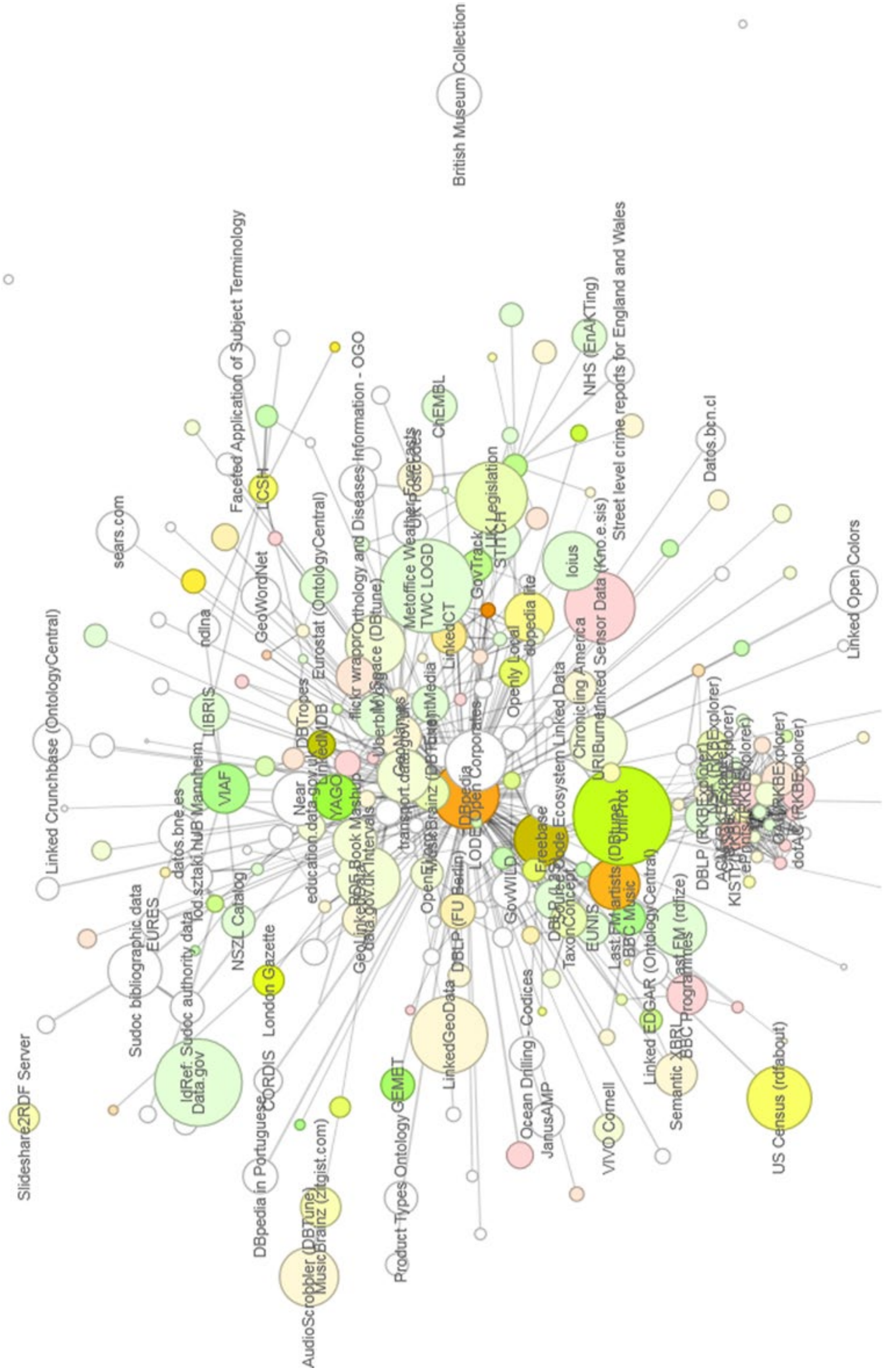


Figure 3-6. The LOD Graph generated by Protovis

The CKAN ratings are represented by colors, by which datasets with high average ratings are shown in green, and the ones with low average ratings in red. The intensity of the color signifies the number of received ratings, where white means no rating, and the darker the color the higher the rating.

Creating LOD Datasets

While large datasets are generated with software tools, in the following sections, you will see how to create LOD datasets manually.

RDF Structure

Let's create a dataset file in RDF/XML! The first step is to create a UTF-8 encoded text file with an `.rdf` extension and to add the XML prolog (see Listing 3-3).

Listing 3-3. XML Prolog

```
<?xml version="1.0" encoding="UTF-8"?>
```

The document content will be between `<rdf:RDF>` and `</rdf:RDF>`. The list of namespaces is declared as `xmlns` attributes on `rdf:RDF`. For example, if you want to use any definitions from the FOAF vocabulary, you have to declare its namespace to abbreviate it throughout the document (see Listing 3-4).

Listing 3-4. The Main Container with One Namespace Declaration

```
<rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/">

</rdf:RDF>
```

From now on, you can use the `foaf` prefix to abbreviate the Friend of a Friend (FOAF) namespace, as shown in Listing 3-5.

Listing 3-5. Using the `foaf` Prefix to Abbreviate the FOAF Namespace

```
<foaf:Person rdf:about="http://www.lesliesikos.com/metadata/sikos.rdf#sikos">
  <foaf:firstname rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Leslie ➥
   </foaf:firstname>
  <foaf:surname rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Sikos</foaf:surname>
</foaf:Person>
```

The list of namespaces is typically extended by the namespace declarations of RDF, RDFS, OWL, and so on, along with Dublin Core, schema.org, etc., depending on the vocabulary terms you use in the dataset (see Listing 3-6). During development, the list will be extended constantly.

Listing 3-6. Multiple Namespace Declarations

```
<rdf:RDF ➥
   xmlns:dc="http://purl.org/dc/elements/1.1/" ➥
   xmlns:dcterms="http://purl.org/dc/terms/" ➥
   xmlns:foaf="http://xmlns.com/foaf/0.1/" ➥
   xmlns:owl="http://www.w3.org/2002/07/owl#" ➥
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" ➥
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" ➥
   xmlns:schema="http://schema.org/">
```

After the namespace list, one can provide dataset information, including licensing, followed by the actual data (RDF statements) of the dataset that will be linked to classes and entities of other LOD datasets with typed links.

Licensing

Linked Open Data without an explicit license is just Linked Data. To make our LOD datasets truly "open," we have to declare the license explicitly, which prevents potential legal liability issues and makes it clear to users what usage conditions apply.

The licensing information of a dataset can be provided in the dataset file or in an external metadata file, such as a VoID (Vocabulary of Interlinked Datasets) file. The license under which the dataset has been published can be declared using the `dcterms:license` property. The most frequently used license URIs for Linked Open Data are the following:

- `http://opendatacommons.org/licenses/pddl/`
  Public Domain Dedication and License (PDDL)—"Public Domain for data/databases"
- `http://opendatacommons.org/licenses/by/`
  Open Data Commons Attribution (ODC-By)—"Attribution for data/databases"
- `http://opendatacommons.org/licenses/odbl/`
  Open Database License (ODC-ODbL)—"Attribution Share-Alike for data/databases"
- `https://creativecommons.org/publicdomain/zero/1.0/`
  CC0 1.0 Universal—"Creative Commons public domain waiver"
- `https://creativecommons.org/licenses/by-sa/4.0/`
  Creative Commons Attribution-ShareAlike (CC-BY-SA)
- `http://gnu.org/copyleft/fdl.html`
  GNU Free Documentation License (GFDL)

The first four licenses are specifically designed for data, so their use for LOD dataset licensing is highly recommended. Licensing of datasets is a complex issue, because datasets are collections of facts rather than creative works, so different laws apply. Creative Commons and GPL are quite common on the Web; however, they are based on copyright and are designed for creative works, not datasets, so they might not have the desired legal result when applied to datasets.

Community norms (nonbinding conditions of use) can be expressed using the `waiver:norms` property (`http://vocab.org/waiver/terms/norms`). A common community norm is ODC Attribution ShareAlike (`www.opendatacommons.org/norms/odc-by-sa/`), which permits data use from the dataset, but the changes and updates are supposed to be public too, along with the credit given, the source of the data linked, open formats used, and no DRM applied. For example, if we have an `ExampleLOD` dataset published under the terms of the Open Data Commons Public Domain Dedication and License, and users are encouraged, but not legally bound, to follow the community norms mentioned above, the licensing of the dataset would be as shown in Listing 3-7.

Listing 3-7. LOD Dataset Licensing Example

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF ↵
   xmlns:dcterms="http://purl.org/dc/terms/" ↵
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" ↵
   xmlns:wv="http://vocab.org/waive/terms/">
```

```
<rdf:Description rdf:about="http://www.examplelod.com/loddataset.rdf#examplelod">
  <dcterms:license rdf:resource="http://www.opendatacommons.org/odc-public-domain-↵
   dedication-and-licence" />
  <wv:norms rdf:resource="http://www.opendatacommons.org/norms/odc-by-sa/" />
  <wv:waiver rdf:datatype="http://www.w3.org/2001/XMLSchema#string">To the extent ↵
   possible under law, Example Ltd. has waived all copyright and related or neighboring ↵
   rights to this dataset.</wv:waiver>
</rdf:Description>
```

■ **Note** The Datahub registration requires an open license to be selected from a drop-down list as a field value, and the license URI as a separate field set as the dataset properties, not just the licensing information provided in the dataset or VoID file.

RDF Statements

The most generic objects of datasets are collected in rdf:description containers. Those objects that are representations of real-world objects already defined in a machine-readable vocabulary are usually collected under the corresponding object class (persons in schema:person, books in schema:book, and so on). Because a basic requirement of Linked Open Data is to identify everything with a dereferenceable web address, make sure that the addresses and fragment identifiers are correct. Whenever possible, use typing to differentiate string literals, numbers, dates, and so on.

Making a statement about another statement is called *reification*. It allows triples to be used in multiple contexts but can affect the formal semantics of your dataset.

Interlinking

Government agencies, large enterprises,[2] media institutes, social media portals, and researchers work with large amounts of data that can be represented as structured data and published as Linked Data. Describing your government data, university research department, colleagues, books, or any other knowledge domain in RDF results in an isolated dataset file, which is not part of the Semantic Web until it is linked to other datasets.

Creating links between the structured datasets of the Semantic Web is called *interlinking*, which makes isolated datasets part of the LOD Cloud, in which all resources are linked to one another. These links enable semantic agents to navigate between data sources and discover additional resources. Interlinking typically happens with owl:sameAs, rdfs:seeAlso, foaf:holdsOnlineAccount, sioc:user, and similar predicates. In contrast to conventional hyperlinks of (X)HTML documents, LOD links are *typed links* between two resources. The URIs of the subject and the object of the link identify the interlinked resources. The URI of the predicate defines the type of the link. For example, an RDF link can state that a person is employed by a company, while another RDF link can state that the person knows other people. *Dereferencing* the URI of the link destination yields a description of the linked resource, usually containing additional RDF links that point to other, related URIs, which, in turn, can also be dereferenced, and so on.

Consider the machine-readable description of the book *Web Standards: Mastering HTML5, CSS3, and XML*, at http://www.masteringhtml5css3.com/metadata/webstandardsbook.rdf#book, which declares the title of the book using the title property from the Dublin Core vocabulary, and, among many other properties, declares a machine-readable resource describing the author, using schema:author (Figure 3-7). Further resources related to the book could be declared using rdfs:seeAlso.

[2] If the Linked Data is behind a corporate firewall, it is called Linking Enterprise Data.

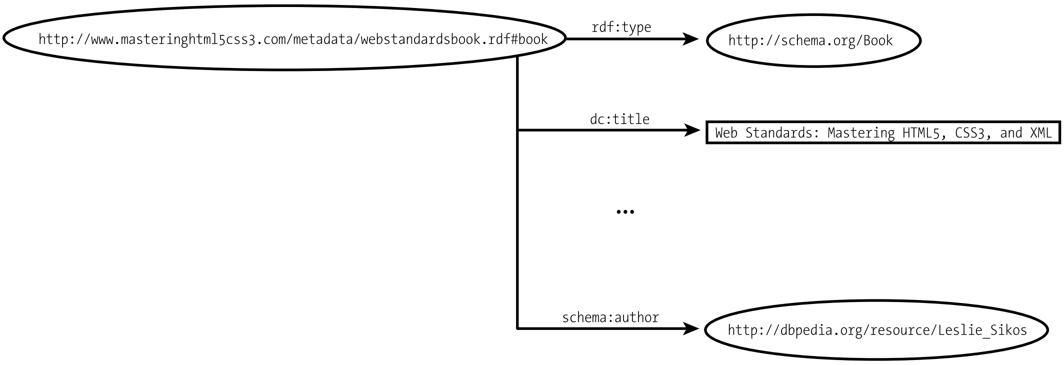

http://www.masteringhtml5css3.com/metadata/webstandardsbook.rdf#book
rdf:type
http://schema.org/Book
dc:title
Web Standards: Mastering HTML5, CSS3, and XML
...
schema:author
http://dbpedia.org/resource/Leslie_Sikos

Figure 3-7. *Linking a dataset to a related entity of another dataset*

The DBpedia resource of the author reveals properties of the author, such as his home page address, defined using `foaf:homepage`, and, among many other classifications, links the author to Australian writers, with YAGO (Figure 3-8).

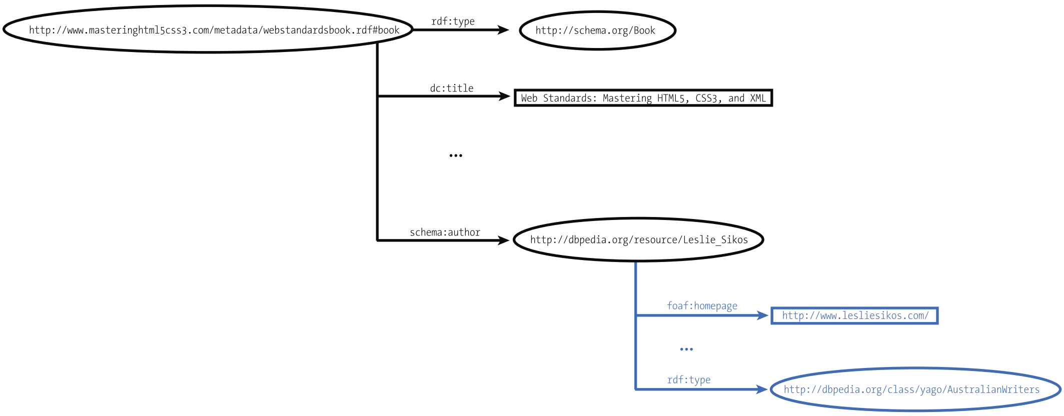

http://www.masteringhtml5css3.com/metadata/webstandardsbook.rdf#book
rdf:type
http://schema.org/Book
dc:title
Web Standards: Mastering HTML5, CSS3, and XML
...
schema:author
http://dbpedia.org/resource/Leslie_Sikos
foaf:homepage
http://www.lesliesikos.com/
...
rdf:type
http://dbpedia.org/class/yago/AustralianWriters

Figure 3-8. *The DBpedia resource links the dataset to another dataset*

Based on `yago:AustralianWriters`, semantic agents will find other authors in the same category (Figure 3-9). By linking to datasets already on the LOD Cloud Diagram (such as pointing to a definition on DBpedia), your dataset will become part of the LOD Cloud.

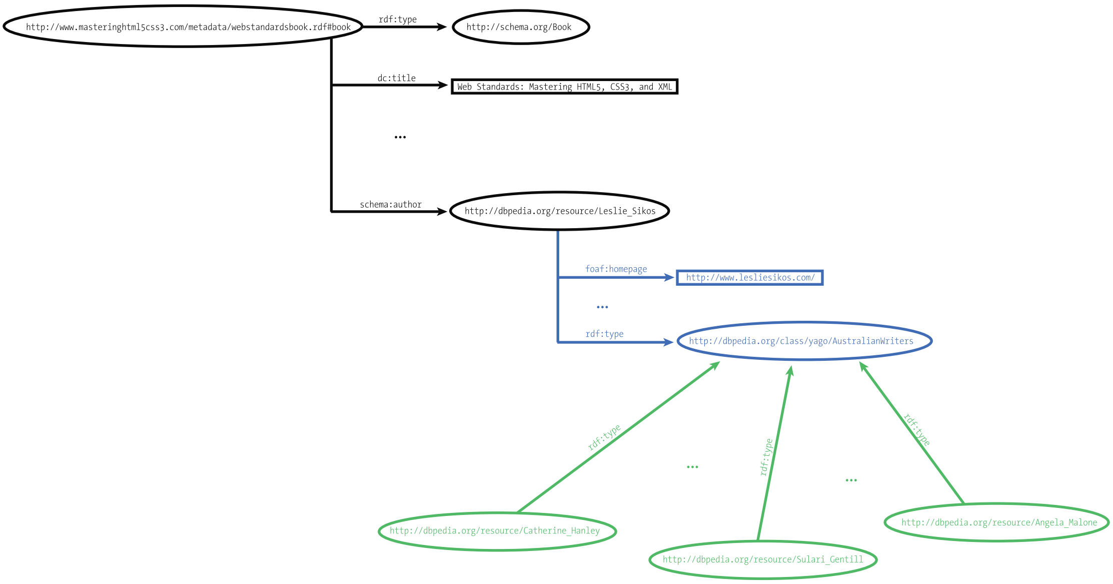

http://www.masteringhtml5css3.com/metadata/webstandardsbook.rdf#book
rdf:type
http://schema.org/Book
dc:title
Web Standards: Mastering HTML5, CSS3, and XML
...
schema:author
http://dbpedia.org/resource/Leslie_Sikos
foaf:homepage
http://www.lesliesikos.com/
...
rdf:type
http://dbpedia.org/class/yago/AustralianWriters
rdf:type
...
rdf:type
...
rdf:type
http://dbpedia.org/resource/Catherine_Hanley
http://dbpedia.org/resource/Sulari_Gentill
http://dbpedia.org/resource/Angela_Malone

Figure 3-9. *Two RDF graphs sharing the same URI merge*

The *Giant Global Graph*, the other name for the Web of Data coined by Tim Berners-Lee, is the supergraph of all the automatically merged LOD graphs [8].

Registering Your Dataset

To be considered for inclusion in the LOD Cloud Diagram, your dataset must be registered on `http://datahub.io`. To be able to register, you need an affiliation with a company or research institute already on Datahub, or if it is not yet registered, you have to request a new company registration. Candidate datasets of the LOD Cloud Diagram are validated through four compliance levels.

Level 1 (basic compliance) requires basic metadata about your dataset, including name, title, URL, author, and contact e-mail, as well as the `lod` tag added to your dataset on Datahub. Level 2 (minimal compliance) requires a topic tag for your dataset, which can be one of the following: `media`, `geographic`, `lifesciences`, `publications`, `government`, `ecommerce`, `socialweb`, `usergeneratedcontent`, `schemata`, or `crossdomain`. You have to provide an example link URI in the `Data and Resources` section with the `example/serialization_format` (`example/rdf+xml`, `example/turtle`, `example/ntriples`, `example/x-quads`, `example/rdfa`, or `example/x-trig`), to help people get a feel for your data before they decide to use it. You also have to provide links to the dump file or the SPARQL endpoint of the dataset. Level 3 (complete compliance) requires additional information, such as the last modification date or version of the dataset (as the value of the field `version`), a dataset description (`notes`), the open license of your dataset (selected from the drop-down menu), a short name for LOD bubble (`shortname`), a link to the license of the dataset (`license_link`), and the instance namespaces (`namespace`). Beyond these custom fields, Level 3 compliance also requires metadata such as VoID file (`meta/void` format), XML sitemap (`meta/sitemap` format), RDF Schema (`meta/rdf-schema` format), and vocabulary mappings (`mapping/format`). Depending on whether you use proprietary vocabularies defined within your top-level domain or not, you have to add the `no-proprietary-vocab` tag (you don't use proprietary vocabularies), or either the `deref-vocab` tag (use dereferenceable propriatory vocabularies) or the `no-deref-vocab` tag (use proprietary vocabularies that are not dereferenceable). Once you are ready with the dataset registration, you can validate it using `http://validator.lod-cloud.net/`.

■ **Note** Since the introduction of the LOD Cloud Diagram, some of the CKAN/Datahub fields used by the Datahub LOD Validator have been changed or discontinued. As a consequence, the Validator might give errors, even on correctly registered datasets. If this happens to your dataset, you have to contact the authors of the LOD Cloud Diagram, via e-mail, for manual approval.

The last step is to e-mail the authors (`richard@cyganiak.de` and `mail@anjajentzsch.de`). Further tags are used by the authors of the LOD Cloud Diagram to annotate whether your dataset has any issues or when it is ready to be added to the next update of the LOD Cloud Diagram. Level 4 compliance means that your dataset has been reviewed and added to `lodcloud` group by the authors who use this group to generate the LOD Cloud Diagram.

Linked Data Visualization

Linked Data visualization tools make the analysis and manipulation of Linked Data easier. The list of Linked Data visualization techniques includes, but is not limited to, comparison of values, analysis of relationships and hierarchies, analysis of temporal or geographical events, text-based visualizations as tag cloud or network of phrases, and representing multidimensional data.

LOD Visualization (`http://lodvisualization.appspot.com`) can produce visual hierarchies using treemaps and trees from live access to SPARQL endpoints. *LodLive* (`http://en.lodlive.it`) provides a graph visualization of Linked Data resources. Clicking the nodes can expand the graph structure. LodLive can be used for live access to SPARQL endpoints (see Figure 3-10).

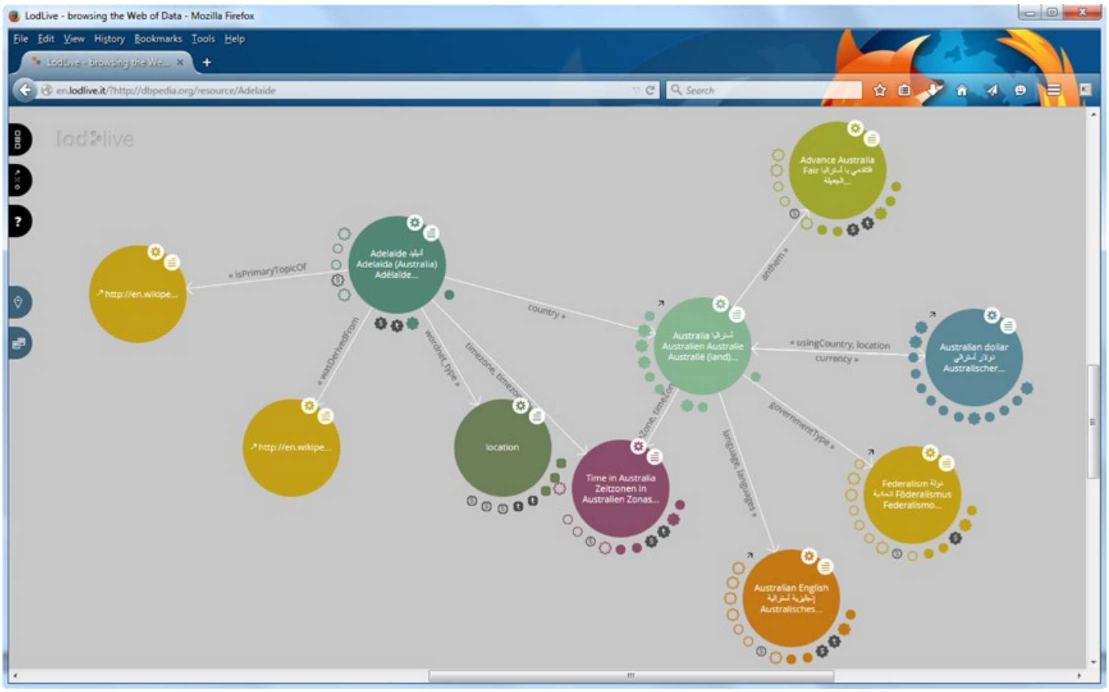

LodLive - browsing the Web of Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
LodLive - browsing the We...
en.lodlive.it/?http://dbpedia.org/resource/Adelaide
Search
lod live
http://en.wikipe...
« isPrimaryTopicOf
Adelaide أديلايد Adelaida (Australia) Adélaïde...
« wasDerivedFrom
http://en.wikipe...
wordnet_type »
timezone, timezone
location
country »
Time in Australia Zeitzonen in Australien Zonas...
Advance Australia Fair تقدمي يا أستراليا الجميلة...
anthem »
Australia أستراليا Australien Australie Australië (land)...
« usingCountry, location currency »
Australian dollar دولار أسترالي Australischer...
Zone, timeZone
language, languages »
governmentType »
Federalism دولة اتحادية Föderalismus Federalismo...
Australian English إنجليزية أسترالية Australisches...

Figure 3-10. *Browsing the Web of Data with LodLive*

The open source graph visualization and manipulation software package Gephi, downloadable from https://gephi.github.io/, is ideal for Linked Data visualization. The program helps explore and understand graphs, modify the representation, and manipulate the structures, shapes, and colors to reveal hidden properties (see Figure 3-11).

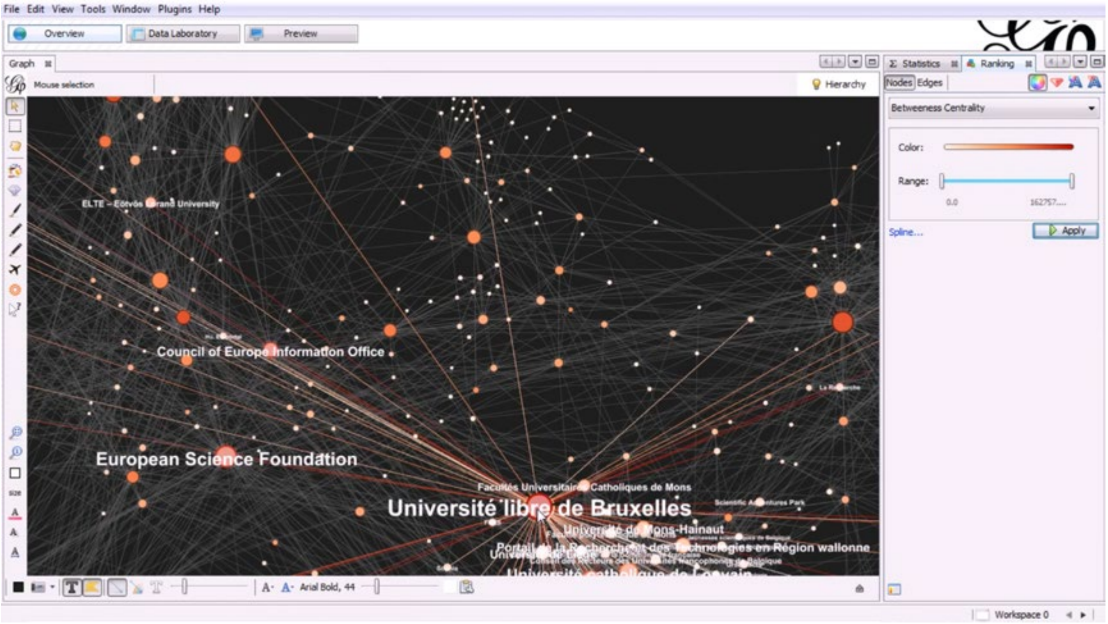

File Edit View Tools Window Plugins Help
Overview
Data Laboratory
Preview
Graph
Mouse selection
Hierarchy
Statistics
Ranking
Nodes Edges
Betweeness Centrality
Color:
Range:
0.0
Spline...
Apply
ELTE – Eötvös Loránd University
Council of Europe Information Office
European Science Foundation
Facultés Universitaires Catholiques de Mons
Université libre de Bruxelles
Université de Mons-Hainaut
Portail de la Recherche et des Technologies en Région wallonne
Université catholique de Louvain
Arial Bold, 44
Workspace 0

Figure 3-11. *Advanced graph visualization with Gephi*

Gephi is powered by sophisticated layout algorithms, to focus on quality (force-based algorithms) or speed (multilevel refinements through *graph coarsening*). The asynchronous OpenGL exploration engine supports flat rendering and 3D rendering with customizable levels of detail and multiple threads. The other engine of the software tool, called the mapping engine, supports refinements, vector rendering, and SVG output, which is perfect for publishing and drawing infographics. The highlight selection makes it easy to work with large graphs. The graphs can be modified by selecting, moving, annotating, resizing, connecting, and grouping nodes. The very powerful real-time graph visualization supports networks with a maximum of 50,000 nodes and 1 million edges and iteration, through visualization using dynamic filtering.

Summary

In this chapter, you learned the concept and requirements of Linked Open Data and about the industries that link thousands of LOD datasets to one another. You understand now how semantic agents can make new discoveries, based on the machine-readable definition of objects and subjects, and the typed links between them. You learned the structure, licensing, and interlinking of LOD datasets.

The next chapter will introduce you to semantic development tools, including ontology editors, reasoners, semantic annotators, extractors, software libraries, frameworks, and APIs.

References

1. Bizer, C., Heath, T., Berners-Lee, T. Linked data—The story so far. Semantic Web and Information Systems 2009, 5(3):1–22.

2. Berners-Lee, T. (2006) Linked Data—Design Issues. www.w3.org/DesignIssues/LinkedData.html. Accessed 25 March 2014.

3. Fielding, R. T. (2005) httpRange-14 Resolved. http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html. Accessed 28 March 2015.

4. The Open Definition—Defining Open in Open Data, Open Content and Open Knowledge. http://opendefinition.org. Accessed 18 January 2015.

5. Hausenblas, M. (2012) 5 ★ Open Data. http://5stardata.info. Accessed 18 January 2015.

6. Lebo, T., Sahoo, S., McGuinness, D. (eds.) (2013) PROV-O: The PROV Ontology. www.w3.org/TR/prov-o/. Accessed 27 March 2015.

7. Schmachtenberg, M., Bizer, C., Jentzsch, A., Cyganiak, R. (2014) The LOD cloud diagram. http://lod-cloud.net. Accessed 18 January 2015.

8. Berners-Lee, T. (2007) Giant Global Graph | Decentralized Information Group (DIG) Breadcrumbs. http://dig.csail.mit.edu/breadcrumbs/node/215. Accessed 28 March 2015.