Chapter 9
Technological Aspects of (Linked) Open Data

Stanislav Ronzhin, Erwin Folmer and Rob Lemmens

Contents

Abstract Choices of technologies to be used for publishing open data influence data reusability. In general, these technologies should be based on open standards and be aligned with the technologies adopted within the communities of data users. However,

S. Ronzhin (✉) · R. Lemmens
Faculty of Geo-Information Science and Earth Observation (ITC), Department of Geo-information Processing (ITC-GIP), University of Twente, Enschede, The Netherlands
e-mail: s.ronzhin@utwente.nl

R. Lemmens
e-mail: r.l.g.lemmens@utwente.nl

E. Folmer
Faculty of Behavioural Management and Social Sciences, Department of Industrial Engineering and Business Information Systems, University of Twente, Enschede, The Netherlands
e-mail: e.j.a.folmer@utwente.nl

B. van Loenen et al. (eds.), *Open Data Exposed*, Information Technology and Law Series 30, https://doi.org/10.1007/978-94-6265-261-3_9

even though open standards are used in many industries ensuring interoperability within a specific domain, there is a lack of interoperability across sectors. This is due to a semantic heterogeneity of cross-domain information. Linked Data is an approach, which aims to achieve interoperability at the widest scale (the World Wide Web) by using the proven architecture of the World Wide Web, based on fully open standards. This chapter provides an overview of the five star model for open data and introduces the need for publishing open data along Linked Data design rules. Examples of Linked (Open) Data use at the Dutch Kadaster as well as at the University of Twente, is used to illustrate the main aspects of the technology. Analysis of current trends and future challenges in Linked Open Data are provided at the end of the chapter.

Keywords Linked Data · API · Five Star Model · OGC · LOD cloud · Open Data

9.1 Introduction

For decades, computer systems are used for data storage. A plethora of technological solutions were developed to support efficient data maintenance, retrieval and dissemination. However, often, the intellectual property of these technologies belong to individuals or organizations. By protecting rights to intellectual property, a vendor sets constraints on the use of the technology. Existence of such constraints creates barriers for interoperability between computer systems, thus, hampering access to the data, its consumption and exchange on technical level.

For example, consider a typical situation—a public company issues an annual report under an open license. The report is free and can be downloaded from the company's data portal as a RAR[1] archive file. Even though it can be very convenient for the company since they use RAR archives as a de facto standard, RAR is a proprietary file format that can be open only by WinRAR[2] software. As a result, potential data (re)use is complicated by the need of acquiring this specific software. For some (re)users who have the software, it might not be a problem, but for others it could be an obstacle.

In the abovementioned example, data is legally open but access to the data is not straightforward. The latter can be solved differently depending on technical capabilities of data consumers. Therefore, for data to be open, it needs to be open both on legal and technical levels. Technical openness, however, goes beyond the mere requirement of using interoperable non-proprietary formats. It also takes into account attainability of the data by technical means of particular users, including the technology that supports the meaningful connection between data sets that are made open at heterogeneous sources.

[1] Roshal Archive Compressed file.

[2] WinRAR is an archive manager. See https://www.rarlab.com/.

In this chapter, we provide an overview of the technological aspects that impact on the level of open data reusability. Section 9.2 introduces and discusses a generic framework for open data infrastructures. Section 9.3 elaborates on the five star model for open data and explains data structures and their impact on searchability and discoverability of the data. After that, several open formats and serializations supporting the discussed data structures are introduced. An overview of access channels are given in Sect. 9.4 illustrated by real cases drawn from the work of the Netherlands' Cadastre, Land Registry and Mapping Agency—in short Kadaster. Section 9.5 explains the concept of Linked (Open) Data. The latter is an innovative technique that will be key in the upcoming years. The current trends are discussed at the end of this section. Remaining challenges are given in Sect. 9.6 followed by conclusions drawn in Sect. 9.7.

9.2 Technology: Bridging People and Data

A framework[3] to describe relationships between people, data and technological components is presented in Fig. 9.1. This model consists of two categories. People and data comprise the first category. The second category contains the access network, policy and standards—the main technological components. Due to the pace of technological developments, the nature of the second category is very dynamic.

In practice, the availability of access channels defines the ways to acquire the data. Different user communities have different de facto standards and technologies. For example, GIS communities[4] are used to deal with spatially enabled technologies, such as the Open Geospatial Consortium[5] (OGC)[6] Web Feature Service[7] (WFS)

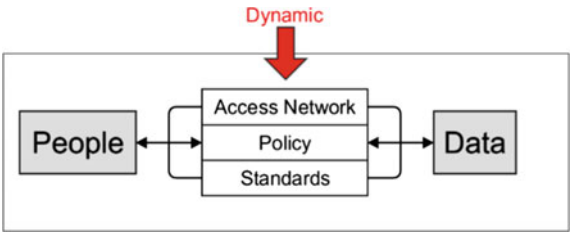

Fig. 9.1 Nature and relations between SDI components [*Source* Rajabifard et al. 2002]

[3] Rajabifard et al. 2002.

[4] GIS stands for Geographic Information System.

[5] http://www.opengeospatial.org/. Accessed May 2018.

[6] The OGC (Open Geospatial Consortium) is an international not for profit organization committed to making quality open standards for the global geospatial community. These standards are made through a consensus process and are freely available for anyone to use to improve sharing of the world's geospatial data.

[7] http://www.opengeospatial.org/standards/wfs. Accessed May 2018.

and the XML-based serialization of the Simple Feature Access model. In contrast, communities with a strong IT background would be more likely to build their services around RESTfull (Representational state transfer) APIs (Application Programming Interface) using JSON-like (JavaScript Object Notation) formats for data exchange. Therefore, the technical capabilities of the user community together with the way the data is provided from a technical perspective, is crucial for the potential (re)usability of the data.

However, prior to reusing the data, first data have to be discovered. To allow data discovery, metadata about the data need to be published on the Web. Collections of structured descriptions about the data content, provenance information and data formats are exposed on the web through catalogue services. Web portals publish such catalogues, which allows discovery and exploration of services and datasets based on their metadata. Search engines (e.g., Google) can significantly increase discoverability of the data. However, for data to be indexed by a search engine there is a need to publish data as HTML pages enriched with semantic annotations using the Schema.org vocabulary.[8]

9.3 Five Stars of Open Data

This section introduces the five star model for open data[9] as a generic framework for data publishing. This model is often used to classify the technical level of advancement of the dataset offering (see Fig. 9.2). Many data registries have implemented the model into their metadata model.

In the model, the first star represents a dataset that is published on the Web under an open license (OL), but without requirements on the data format; e.g., a handwritten document stored in PDF with an open license is one star open data.

Fig. 9.2 Five star model for open data [*Source* http://5stardata.info/en/]

[8] https://developers.google.com/search/docs/guides/intro-structured-data. Accessed May 2018.

[9] http://5stardata.info/en/. Accessed May 2018.

In order to be findable, the data must be prepared and organized in a way approachable by computers. This implies that the data are published using data formats that can be read and processed by a computer. Data formats are used to formally specify and implement data structures for processing by computers. The second star adds to the open license, the ability that the data is machine-readable, e.g., Microsoft Excel format. When a proprietary format is replaced with an open format (OF), such as Comma-separated values (CSV) or Extensible Markup Language (XML), the dataset receives three stars.

There is a clear distinction and a big implementation gap between the first three and the last two stars. Most datasets reside at the first three stars, and for many datasets three stars is the endpoint. For this reason, we discuss the technologies needed to bring the data to the third star in the following section. Several real-life use cases from different domains (e.g., academia and government) are provided to highlight the pros and cons of different approaches to data offering.

The fourth and the fifth star require the use of the Linked Data rules, which is discussed in detail in Sect. 9.5.

9.3.1 From Unstructured to Structured Data

Any data can be classified into two major categories depending on the degree of organization—structured and unstructured data. A simple example to illustrate the difference is to compare a spreadsheet and a free text. A spreadsheet arranges the data into columns and rows, thus, providing a uniform structure. Computers and algorithms rely upon this structure to access, sort and search the data values. Therefore, structured data is machine-readable and processable.

The structures in data can be a linear, hierarchical and graph structure.[10] Spreadsheets as well as relational databases are examples of linear structures. The table rows, if written one after another, form a linear sequence of values. In general, hierarchical and graph structures are very close to each other because they both organize the data in nodes and links. Nodes contain values and can be compared with cells of a spread sheet. Links represent relations between nodes. The difference is that graphs contain circular structures of nodes whereas hierarchies do not. This difference can be illustrated by comparing a tree and a fish net. Every branch of a tree has only one "parent" branch. If there is a need to go from one branch to another, the route will always go up in the hierarchy of branches until a common ancestor branch, then, down to the needed branch. In a fish net, there is more than one (often many) way to traverse the net between any two points. Computations on graph structures heavily utilize concepts from the fields of graph theory and theory of sets. These are well developed fields of mathematics, which ensure efficiency and scalability of computations on graph structures.

[10] Frakes and Baeza-Yates 1992.

To allow computers to work with structures of human language, they need to be preprocessed to discover and to reveal hierarchies and dependencies in the data. Techniques such as data mining and Natural Language Processing (NLP) provide methods for finding and interpreting patterns in human language. These findings are then formalized by tagging to separate semantic elements and enforce hierarchies of records. Formalized structures of information can be used to 'ground' the elements in a dataset to one or more data structures that formally present concepts and relations between them. Such formal systems are called *vocabularies or ontologies*. They can range from simple descriptions of concepts to complex networks of formalized concepts and their constructs such as *sub-super class* relationships, *same-as* and *part-of* relationships.

Unstructured information being enriched with semantic tags from formal ontologies is called semi-structured data. The inferred structure allows machine reasoning with associative and hierarchical relationships. For example, a building has floors, a house is a building, and consequently, a house has floors. As a result, it becomes possible to search in data on a lower level of granularity.

Let us consider an example. For one of its courses, the Faculty of Geo-Information Science and Earth Observation (ITC),[11] University of Twente, uses a textbook on GIS and Remote Sensing. The content of the book is published on the web as a wiki. The wiki software supports changes over time with input of teachers and students, using different types of media and facilitating peer discussions.

However, the original book is in plain text (unstructured format). The text in the book can only be searched with a keyword search, which does not allow meaningful exploration of the content of the textbook. To improve this, the text has to be enriched with semantic tags to allow access to the content within the text blocks. As a result, a different type of textbook was created, The Living Textbook.[12]

The Living Textbook is based on an ontology of the concepts from the domains of GIS and Remote Sensing. The concepts within the text are linked to the ontology, thus, allowing to browse the content with finer granularity than would be possible with plain text. In addition, the ontology is visualized as a network of concepts and relations (a *concept map*) to allow interactive navigation in the content. The concept map allows students to see and to explore visually the relationships between concepts in a non-pre-described order of reading and learning. Figure 9.3 shows the improved interface of the Wiki with the concept map. If a user clicks on a concept, information tagged with the related concept is shown in the Wiki interface. For instance, in the figure a node representing the Relational Data model is selected (on the right). The relations to other concepts (nodes in the visualization) are highlighted. The view of the concept map is linked to the Wiki, a click of a selected concept retrieves descriptions related to it (in the wiki text on the left).

The example of the Living Textbook illustrates the difference in usability of unstructured data and structured semantically annotated data. When a data publisher

[11] https://www.itc.nl/. Accessed May 2018.

[12] https://itc-giscience.utwente.nl/. Accessed May 2018.

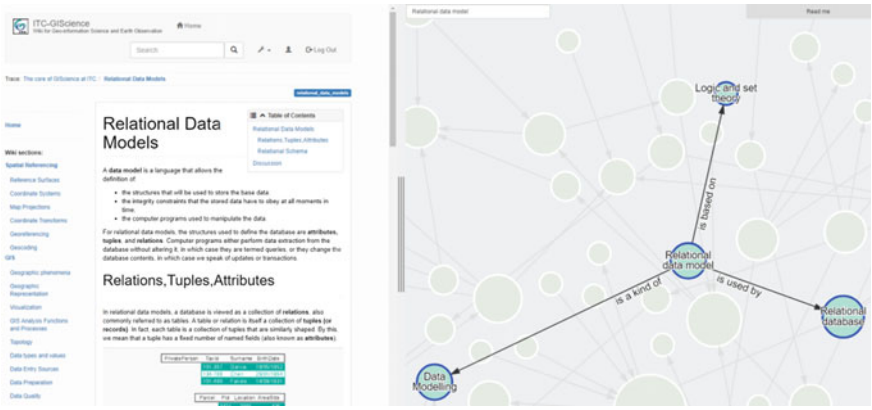

Fig. 9.3 The Living Textbook interface with the concept map and a Wiki page [*Source* The authors]

is interested in improving the (re)usability of data, it is preferable to make the data structured and enriched with semantic annotations.

9.3.2 Formats and Serializations

The data structures discussed above are abstract concepts about data organization based on rigorous mathematical concepts. To enable computers to store and exchange the data, the data should be represented as a sequence of bits in such a manner that a computer can read this sequence and recreate a clone of the original data. The process of translating a data structure into a standardized format is called serialization. This is similar to a translation of a common concept into different human languages. For example, the notion of bread has different names in different cultures, but all of them refer to the same idea of a cooked dough. Data formats can be seen as languages used by computers to store and communicate information. Similar to spoken language, data formats have elements such as grammar and syntax.

There is a great variety of formats that are used nowadays. Some of them have an open specification and are software independent (e.g., Scalable Vector Graphics (SVG) for vector graphics) when others have commercial licenses and can be processed by software products of a particular vendor only (e.g., Adobe Photoshop file format). In the realm of open data, obviously, open formats are used and we name some of them below.

Comma-separated values (CSV) is a widespread open format for spreadsheets and simple databases. It uses plain text to store tabular data. Each record consists of one or more fields (columns), separated by commas. This format is not standardized, but the idea to separate values using an agreed upon delimiter is very simple. There are many variants including tab-separated values and space-separated values.

Extensible Markup Language (XML)[13] encodes data in way that is both human-readable and machine-readable. The World Wide Web Consortium (W3C)[14] developed a family of XML-based specifications to ensure simplicity, usability and interoperability of data exchange across the Internet. It is a textual data format, in which structure is captured by the XML schema definition (XSD). XSD is attached to XML; and defines a set of rules and constrains to describe formally the elements in an XML data source.

Regardless of the format, there are always elements that create redundancy in the data. Data compression identifies and eliminates such repetitions to increase efficiency of data exchange. Most prominent open formats supporting data compression are ZIP[15] and 7z.[16]

9.4 Access Channels

Data formats, structures and serializations discussed in the previous section have an impact on data storage and processing. By using formats with open specification, publishers ensure interoperability of the data between computer systems. However, before processing the data, data have to be acquired in some way.

The following section elaborates on access channels. We use experience of the Dutch Kadaster to illustrate different possibilities for data offering.

9.4.1 Downloadable Data

The easiest way to offer a dataset is via a full download of data file. The advantage is that the user has a full copy of the original data as it is in the source database. There are, however, several major drawbacks to full download. First of all, it is not efficient: if hundred government agencies use a full download option and load that data in their own databases, then hundred databases, including licenses are needed. This creates an unnecessary growth of the number of redundant tasks for data management within government agencies.

Another significant limitation of the full download is the need of another full download every time the data is updated. Consequently, copies of the same dataset mutate within various organizations in the period between updates. If a dataset is not used for high impact decision making, several months between data updates might be not that important. However, in many situations, datasets are not stable, and changes

[13] https://www.w3.org/XML/. Accessed May 2018.

[14] https://www.w3.org/. Accessed May 2018.

[15] https://www.iso.org/standard/60101.html. Accessed May 2018.

[16] http://www.7-zip.org/. Accessed May 2018.

are so rapid and significant that when a download of the dataset is finished, the downloaded dataset is already outdated.

In contrast, when data are used for high-value purposes, such as issuing permits, the data should be up to date. In order to avoid risks related to transformation of the original sources (e.g., semantic loss), the data should be consumed directly from the source. For such cases, service requests are needed instead of downloads. With services it is possible to make specific calls to the dataset and receive the answer in real-time.

9.4.2 Data Services

Services provide a layer on top of the data for the purpose of controlling and easing the use of data sources. Therefore, services differentiate according to the intended use of the data and provide a so-called interface to the user, which may be a human user but can also be a machine application.

It is important that service offerings are well understood by their users. This can be achieved by standardizing the interface and is part of the mission of many standardization organizations. The Open Geospatial Consortium has been active in this domain for two decades and has produced geodata and map service standards, which support the interoperability of independent computer systems and the access to the data sources they hold. This is key to the essence of open data. Software manufacturers that want to play a role in data infrastructures are inclined to implement these standards through their end-user interface or through an Application Programming Interface (API) for third party applications developers.

9.4.3 Choosing Between Download and Services

Choices between file download and services depend on the characteristics of a dataset: a stable code table can be published as a downloadable file, while a million records dataset with daily updates is better published as a service.

It is getting more complicated simply because the users of data are heterogeneous[17] (just as the data itself). The technical expertise of the users differs (from professionals to lay users), the roles differ (from development to analysts), and more important, the context of their potential application differs enormously. In practice, this implies that there are different needs for the technical features of a dataset. No one size fits all, but different offerings for different users. One (or a limited) technical option for data usage will limit the (re-)usability in practice.

[17] See Chap. 3 of this book.

This can be illustrated by the constellation of data services provided by the Dutch Kadaster.[18] Kadaster collects and registers administrative and spatial data on property and the rights involved. The organization publishes many large authoritative datasets including several key registers of the Dutch Government (e.g., Key Register Topography (BRT),[19] and Key Register Addresses and Buildings (BAG)).[20]

As an example, BAG data can be accessed as several independent data products accessible via different access channels. Table 9.1 gives details of BAG data services providing descriptions, access channels, intended users, use cases and related fees. For instance, with the BAG Extract service an organization can download its own copy of the BAG database. The BAG *Digilevering* (Digi-delivery) service allows subscribers to receive changes that are pushed within an hour after an update. BAG Compact provides only a subset of the dataset (address and related objects) and can be delivered once or based on a monthly subscription.

BAG *Bevragen* (querying) is the web service product by which individual objects can be requested. There is also a BAG viewer on the web to access the BAG data based on a user interface. More BAG products are being offered in traditional geographic formats and delivered (free of charge to the end user) via Publieke Dienstverlening Op de Kaart (PDOK),[21] including Web Map Service (WMS) and Web Feature Service (WFS).

Apart from legal issues related to the use of the service, the Service Level Agreement (SLA) sets conditions about the availability of the service. This is important because this sets the requirements for technical implementation. SLA clarifies technical conditions of the service implementation for users who plan to develop applications building on a specific service.

Another aspect is related to the actuality of the content of the data. Again, data with a dynamic nature (such as weather forecast) needs to be updated. Many datasets in practice are published once, without updates. The user, however, requires up-to-date data, and guarantees of the update frequency of a dataset.

9.4.4 Lessons Learned

Many technical aspects of a dataset offering are independent of the openness of a dataset. Fee-based proprietary datasets are often offered more professionally than open datasets. However, organizations that offer open data can adopt technical standards and SLAs of the same level as proprietary datasets to improve usability of the data. When designing an access channel frequency of the data updates should be taken into account together with technical skills of users and intended use cases.

[18] http://www.kadaster.nl. Accessed May 2018.

[19] https://brt.basisregistraties.overheid.nl/. Accessed May 2018.

[20] https://bag.basisregistraties.overheid.nl/. Accessed May 2018.

[21] https://www.pdok.nl/; (English: Public Services on the Map). Accessed May 2018.

Table 9.1 Constellation of services to disseminate the BAG data [*Source* The authors]

Data offering	Description	Access channel	Use case and users	Fee
BAG Extract[a]	XML files with data from the BAG	File-based full download, subscription or mutation subscription (daily or monthly updates)	For all the users who needs BAG as a whole	Payable
BAG Digilevering[b]	Service based data mutation delivery	Service providing updates of the BAG within an hour	For governments to synchronize versions of BAG	Payable
BAG Compact[c]	XML files with addresses and address-related elements only	File-based full download or subscription	Lightweight distribution of addresses	Payable
BAG Bevragen[d]	Web service	Service for querying any data from the BAG via a Web interface. One object at a time. No full download	To build web applications using BAG data	Payable
BAG Viewer[e]	A map interface to browse the BAG data	A web-based application that visualize BAG data providing search functionality	For the public and professionals to browse the data	Free
BAG Web Map Service[f]	Web Map Service (WMS, WMTS)	A service to render and disseminate tiled raster maps with the BAG objects	For (re)use in a GIS or a web map	Free
BAG Web Feature Service[g]	Web Feature Service (WFS)	A service to access vector BAG data	For (re)use in a GIS or a web map	Free

[a]https://www.kadaster.nl/-/bag-extract. Accessed May 2018
[b]https://www.kadaster.nl/-/bag-digilevering. Accessed May 2018
[c]https://www.kadaster.nl/-/bag-compact. Accessed May 2018
[d]https://www.kadaster.nl/-/bag-bevragen. Accessed May 2018
[e]https://bagviewer.kadaster.nl. Accessed May 2018
[f]https://www.pdok.nl/nl/producten/pdok-services/overzicht-urls/b. Accessed May 2018
[g]https://www.pdok.nl/nl/producten/pdok-services/overzicht-urls/a. Accessed May 2018

The case of BAG data has shown that there is a large and growing amount of different ways a dataset can be offered. This requires good data management to be able to maintain the services efficiently and to avoid a non-maintainable mess of service offerings. One particular challenge is that every service should return the same data (on the same object), which is especially relevant for the authoritative datasets since they are used for issuing permits.[22]

9.5 Linked (Open) Data

The five star model of open data assigns the fourth star and the fifth star to the data published using the Linked Data rules. In this section, these rules are explained and the difference between the fourth and the fifth star are showcased.

9.5.1 Four Rules of Linked Data

Linked Data is a method of publishing structured data so that they can be linked and queried. It is built on the Semantic Web technology, and is driven by open standards set by the World Wide Web Consortium[23] (W3C). The term was introduced by Sir Tim Berners-Lee[24] who also set the four design rules,[25] which are often used to define linked data.[26] They are as follows:

1. Use URIs (Uniform Resource Identifiers) as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF and Simple Protocol and RDF Query Language (SPARQL)).[27]
4. Include links to other URIs so that people can discover more things.

Although these design rules look simple, it is far more complicated when diving into the world of linked data. The first two rules touch upon the use of HTTP URIs to name things. In the concept of the Semantic Web HTTP URIs are used as names for real-world objects and abstract concepts rather than as addresses for web documents. The content of a dataset is structured using a simple graph-based data model—the Resource Description Framework (RDF).[28] In RDF, a resource is described as a set of

[22] Eckartz and Folmer 2015.

[23] https://www.w3.org/. Accessed May 2018.

[24] Berners-Lee et al. 2001.

[25] Berners-Lee 2006.

[26] https://www.w3.org/DesignIssues/LinkedData.html. Accessed May 2018.

[27] https://www.w3.org/TR/rdf-sparql-query/. Accessed May 2018.

[28] https://www.w3.org/RDF/. Accessed May 2018.

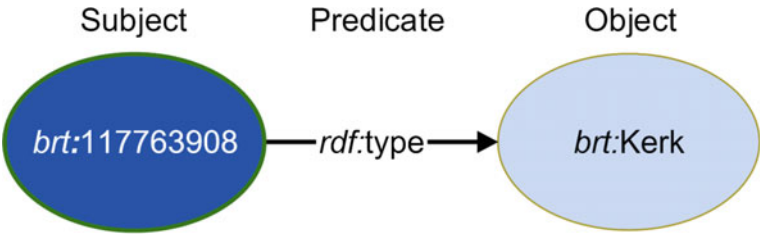

Fig. 9.4 Triple is a basic building block to structure on the Semantic Web consisting of a subject expressed as a URI, predicate (also URI) and an object that can be either a liter or a URI [*Source* The authors]

statements called *triples or facts*. A triple represents the basic structure of a sentence consisting of three parts, namely (1) a subject, (2) a predicate and (3) an object.

These three parts can be expressed as URIs, but objects can also be a literal value (e.g., an integer number). In general, the subject defines the described resource, the predicate shows what kind of relation exists between subject and object, and object is another resource that has a relation with the subject. There are several serializations of RDF, but XML-based is the most common format.

Figure 9.4 shows an example triple representing an object registered in the Dutch Key Register Topography (BRT) with the number 117763908. This is a church (*kerk* in Dutch). The first part of a triple, the subject (brt:117763908), is represented by a URI coined to hold the unique identification of the building. The URI resides in the BRT namespace[29] and is shortened in Fig. 9.4 to "*brt*". Any registered attribute of this particular building will be linked to the subject via a predicate. In the figure, the predicate (*RDF* namespace)[30] defines a type of the building, for instance, it is a church (kerk). A URI to define churches represents an object of a triple. This URI originates from the BRT ontology (the same *BRT* namespace), a collection of formal concept definitions and relations.

The example given in Fig. 9.4 demonstrates the use of the first three Linked Data design rules. As shown, the result of implementation of those three rules corresponds to the four star data in the five star model. Moreover, implementation of the fourth Linked Data design rule prepares for the next step to raise the level of the data up to the fifth star. The following example (see Fig. 9.5) walks through the transformation steps necessary to improve usability of an open dataset by changing its structure from a plain text to the RDF structure.

Let us imagine that the Dutch Kadaster registered a building of the Saint Catharine church erected in 1900 with a certain registration ID. In Fig. 9.5, this is shown as free text. To make it machine-readable the data has to be structured, for example, as a table. In this way, the free text is decomposed into smaller bits of information, which are written in a formal structure. The columns of the table contain semantically similar

[29] http://brt.basisregistraties.overheid.nl/def/top10nl#. Accessed May 2018.

[30] http://www.w3.org/1999/02/22-rdf-syntax-ns#. Accessed May 2018.

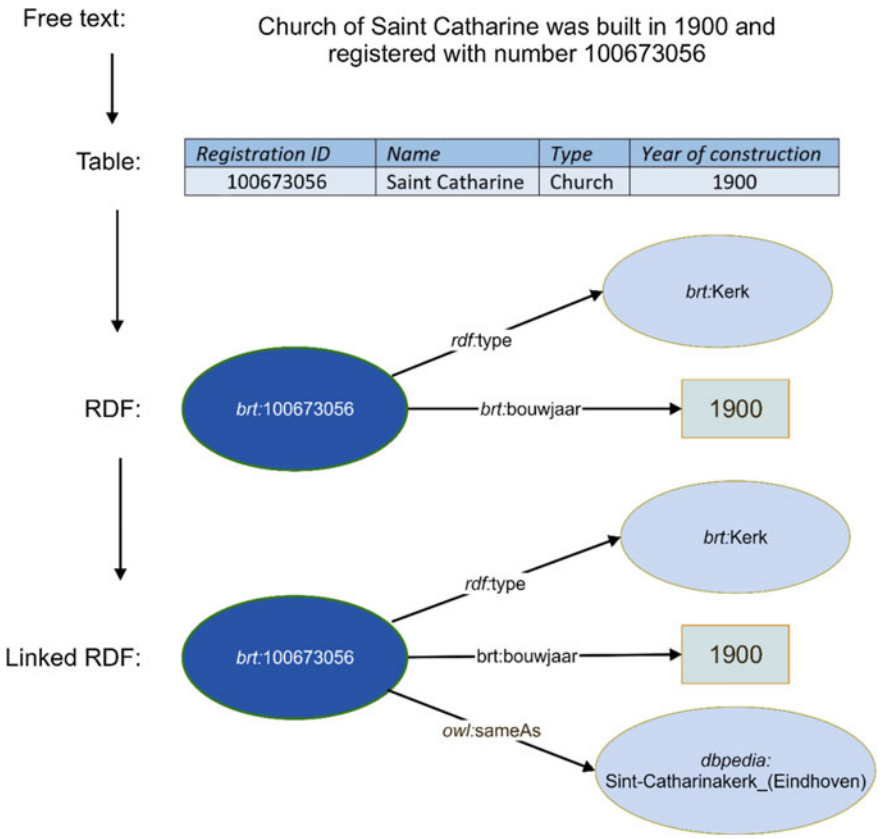

Fig. 9.5 Data transformation from an unstructured free text into structured table and RDF [*Source* The authors]

information. For example, it is assumed that the column with the name Registration ID hold only records with such ID, the opposite would be considered an error.

However, even though the computers can read the tables, they cannot make sense of such structures. To enable reasoning, a formal representation of the concepts and relations used in the data should be decoupled from the data structure.[31] The RDF data model together with ontologies provides means for capturing formal semantics independently from the data structure. As can be seen from the Fig. 9.5, the table resulted in two triples, one related to the building being a church (*kerk*) and the other triple related to the year of construction (*bouwjaar*) being 1900. The former is an example of a triple where all three parts are represented as URIs when in the latter, the object is a literal value ("*1900*" is an integer number). The URIs used in the triple, explicitly define unique identifications for types and relations between data

[31] Janowicz et al. 2014.

items. The Linked RDF, in the figure, has an additional third triple that states that the registered building is the same building as described in DBpedia[32] (owl:sameAs is used). This makes the Linked Data *linked*.

Many data suppliers,[33] especially those that publish official government data are diving into the world of linked data as they see potential for their data. The reasons are diverse. One motivation is to reduce the number of copies of datasets, as these organizations are worried that decision-making takes place based on outdated data (from an outdated copy). Keep the data at the source, and use it at the source, is the adagio of these government data publishers. Another worry is incidental misuse of the data, for which linked data is the potential solution. Wrong interpretation of data can be avoided using metadata (including provenance) as part of the data, practice with linked data. Other organizations put more emphasize on the linkability of data by having unique resolvable identifiers on the web, indexability by search engines, and the possibility of querying the data on the web.

9.5.2 The Linked Open Data (LOD) Cloud

An attempt to provide an overview of Linked Open Data occurrences is made by the publication of the Linked Open Data cloud diagram (see Fig. 9.6). It portrays 1000+ datasets that are connected via Linked Open Data, although it does not mean that all datasets are open from a legal perspective. The metadata of each dataset is accessible via DataHub.[34]

To be included in the diagram the dataset needs to meet the following criteria: it must contain at least 1000 triples, it must connect to another dataset in the Cloud (the size of the circles corresponds to the number edges connected to each dataset), and must be accessible via an endpoint with resolvable web links (URIs). The LOD Cloud diagram provides a good entry point to a large number of datasets. The main prominent resources are DBpedia, a linked data extract of Wikipedia, and LinkedGeoData, an LD version of Openstreetmap.[35] However, it does not necessarily enable the immediate meaningful connection between datasets. This needs a more careful evaluation of the semantics of the triples that are subject to such connection, possibly resulting in the semantic enrichment of them.

[32] Bizer et al. 2009a, b.

[33] Folmer and Beek 2017.

[34] https://datahub.io. Accessed May 2018.

[35] Auer et al. 2009.

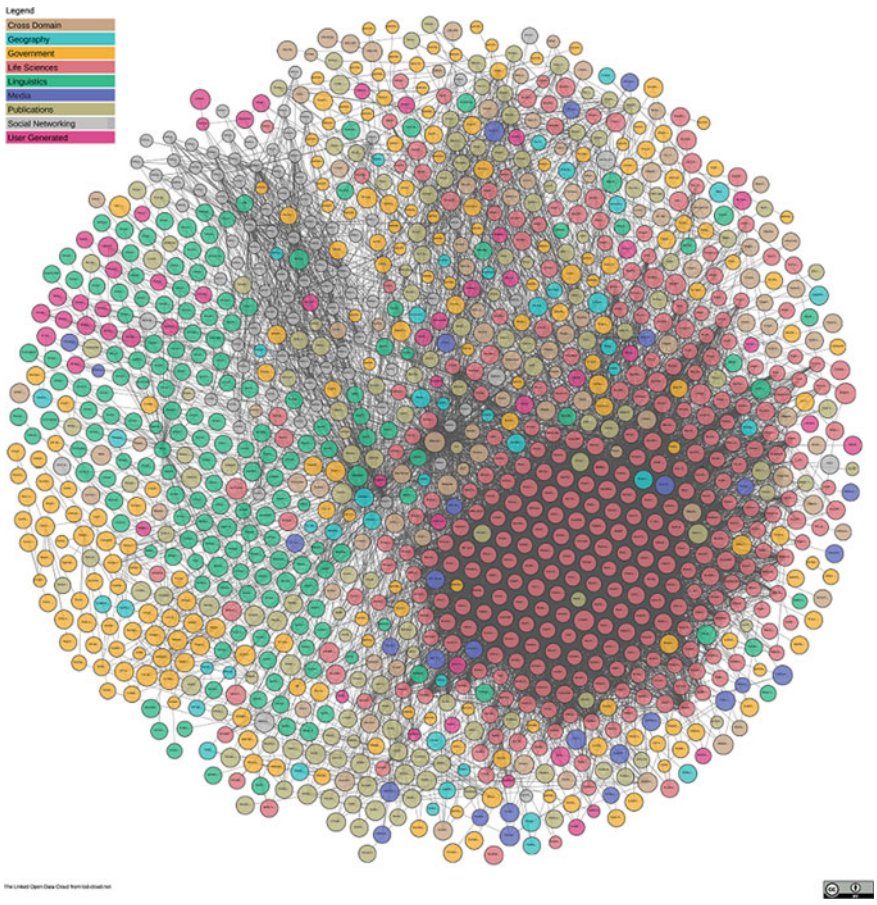

Fig. 9.6 Linked Open Data cloud diagram [*Source* Abele et al. 2017]

9.5.3 Current Trends

The technical investment of bringing the data to the fifth star are relatively high and the benefits of publishing a simple and static dataset as five star data are limited. Therefore, a general policy in creating all datasets as five star open data is inefficient. As a solution, a user driven differentiation of data offering is needed. An example approach is developed within the Platform Linked Data Netherlands[36] to bridge the gap between the third and fourth star in practice and to have a better alignment with the needs of developers. For this reason, a six star[37] model was introduced. In the model, the first three stars stay

[36] http://platformlinkeddata.nl. Accessed May 2018.

[37] http://www.pilod.nl/w/images/3/32/Linked_Data_in_beeld_2014.pdf. Accessed May 2018.

intact, the fourth and fifth stars are given if the dataset is provided through a developer friendly JSON API (4th star) and semantically rich API serving JSON-LD (5th star). The sixth star means the data is available as an open SPARQL endpoint.

The Linked Data technology is widely implemented, but in general, it is hard to get a good understanding of the adoption stage of the technology. When browsing through data registries, the amount of linked data is limited, but the LOD cloud provides another impression. Although, LOD cloud contains a great number of datasets, the degree of interconnectedness of them is not clear. Preliminary research[38] shows that resources interlinked unevenly across and within the LOD sets.

Phil Archer, a W3C veteran who worked on standardization of the data on the Web for many years, suggested linked data has reached many enterprises, but many implementations are not visible for the public. The technology is used within the organizations for data management and especially for creating various information products.[39] Information products are customized to a specific user need, and might involve data elements from different datasets. This is a user centric approach instead of a data push approach. Linked Data helps with this shift from data push to user centric by lowering the cost of data integration and simplifying the creation of information products. Exposing the results through a developer-friendly JSON API

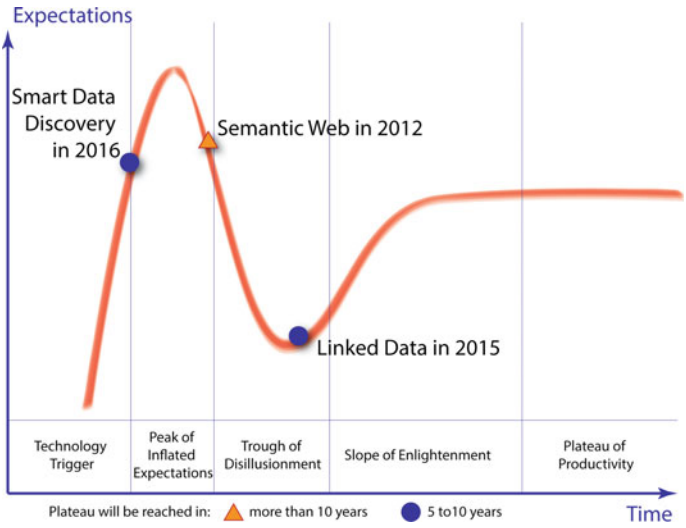

Fig. 9.7 A Gartner diagram summarizing the trends in Advanced Analytics and Data Science from 2012 and 2015 combined with the latest diagram for the emerging technology of 2016 [*Source* The authors]

[38] Adlan forthcoming.

[39] https://www.w3.org/blog/2017/06/possible-future-directions-for-data-on-the-web/. Accessed May 2018.

with high availability lowers technological barriers between communities thus increasing the chance of data reuse.

One of the major advantages of publishing the key (formal) register as linked data is the availability of persistent URIs that are maintained officially. These URIs form a backbone to support linking between datasets coming from other governmental and public organizations. This is one of the main motivations[40] to deploy linked data by a national mapping authority.

The maturity of the Linked Data technology can be judged by comparing two Gartner's Hype Cycle diagrams for Advanced Analytics and Data Science from 2012[41] and 2015[42] (see Fig. 9.7). If in 2012, Semantic web was to reach the Plateau of Productivity in more than 10 years, the diagram of 2015 placed Linked Data in the middle of the Trough of disillusionment with expectation that it reaches the Plateau in 5–10 years. However, in 2016,[43] emerging technologies that were fundamentally based on linked data (e.g., Smart Data Discovery) started climbing towards the Peak. This means that the community has adopted the Linked Data and Semantic Web technologies and a new generation of derivative technologies (e.g., context aware APIs serving JSON-LD) has emerged.

9.6 Future Challenges

In this section, we highlight remaining challenges in relation to technical aspects of linked open data that may set the research agenda for future work.

The Semantic Web is based on its own technology stack,[44] which is different from the OGC standards. This creates a situation where both technologies and their infrastructures co-exist separately. The challenge is to create a semantic layer on top of existing data services that would allow using Semantic Web reasoners over their metadata and data. One of the approaches is to provide semantic descriptions (markup) for Web services.[45] Several works propose the development of Linked Data proxies, software components that allow users of both infrastructures to share data and services.[46] Linked Data proxy provides functionality to publish OGC services using Linked Data rules and to create and maintain URIs for data items and metadata descriptions.[47]

---

[40] Goodwin et al. 2008.

[41] https://ablvienna.wordpress.com/tag/gartner/. Accessed May 2018.

[42] http://www.datasciencecentral.com/profiles/blogs/big-data-falls-off-the-hype-cycle. Accessed May 2018.

[43] http://www.gartner.com/newsroom/id/3412017. Accessed May 2018.

[44] http://www.w3c.it/talks/2005/openCulture/slide7-0.html. Accessed May 2018.

[45] Lemmens et al. 2006.

[46] Janowicz et al. 2012.

[47] https://geo4web-testbed.github.io/topic4/. Accessed May 2018.

As was discussed in the previous sections, users have different questions in mind that they would like to answer by using data. However, to judge if data are fit for answering their questions, users need to compare the context of their tasks to the context of the data provider. In other words, users from different domains interpret semantics of the data differently. Therefore, the meaning of the data should be communicated to the target community using the domain language of that particular community. This can be achieved by translating the data between different semantic representations in a peer-to-peer fashion with respect to a provider's context.[48] This implies that the current practice of publishing data together with static semantic annotations (e.g., using a domain ontology) may rather decrease reusability and that dynamic typing (ontology transformation to fit user context) approaches are required instead.

With the growing amount of datasets published as linked data, it becomes increasingly difficult to view all the datasets as a whole. For example, consider a national mapping authority publishing a variety of Linked Datasets. The goal is to provide access to all of them in order to allow querying datasets. The challenge is twofold: on one hand, a publisher needs to interlink features presented in the sets and on the other hand, this cannot be done without a comprehensive and coherent ontology that encompasses all the notions in the sets.

Finally, due to the vast heterogeneity of geo-data, the creation of knowledge graphs spanning across multitude of datasets will require novel interaction strategies and interfaces that support users in finding relevant data. Users will not know what is inside these voluminous and highly interconnected data spaces. They will need assistance in browsing and navigating these data. As a result, data providers will need to augment their infrastructures with components enabling exploratory search.[49]

9.7 Conclusion

The technical choices a data supplier makes in the offering of a dataset, impacts data re-usability. However, on the downside, the same choices have also an impact on the costs involved in publishing the data. In all cases, a business case will help in making the right choice.

The five star model of open data promotes publishing data in machine-readable structured non-proprietary formats. It is always essential to have a good understanding of the potential user of the data. Different user communities have different technical skills and what is more important they have different questions in mind to be answered with the data. Therefore, the data should be offered through a variety of access channels to broaden the range of potential (re)use cases. Examples can vary from a

[48] Scheider et al. 2012.

[49] Marchionini 2006.

lay-user oriented simple file-based download to a queryable powerful API serving the data in many formats suitable for building web applications of high availability.

Application developers rely heavily on such data services. Therefore, SLAs help them to clarify technical aspects of the service (e.g., availability) as well as to define responsibilities of the data providers. The discussed five star model provides an approachable framework for assessing the technical level of advancement of the dataset offering. Although publishing the data as five star data requires significant efforts and investments, it also lowers the costs of further integration by allowing semantic interoperability between datasets. The latter is of importance because it enables serendipitous scientific discoveries by combining datasets that were difficult or impossible to integrate due to semantic heterogeneity. Linked Data mitigates this by providing means for explicit representation of the data semantics.

In this sense, the value of Linked Open Data increases together with the number of published datasets. This is a so-called network effect: the graph of data becomes bigger, and the potential of querying the graph becomes much more valuable. Therefore, it makes sense to stimulate publishing open data as Linked Data and specially to create links between datasets. Once a dataset is converted into Linked Data, it becomes a part of the LOD cloud, an unbounded data space where standardized methods for data access and retrieval can be used.

The new generation of data applications is dependent on the technologies allowing seamless integration of semantically heterogeneous datasets. Linked Data provides the needed mechanisms and conventions. A fundamental shift is to stop thinking in datasets, although this is very logical from a data supplier point of view. Thinking and acting in linked data instead is an emerging trend in the open data initiative.

References

Abele A, McCrae JP, Buitelaar P, Jentzsch A, Cyganiak R (2017) The Linked Open Data Cloud. http://lod-cloud.net/. Accessed May 2018

Adlan C (forthcoming) State of Linked Open GeoData Cloud. MSc Thesis, University of Twente

Auer S, Lehmann J, Hellmann S (2009) Linkedgeodata: Adding a spatial dimension to the web of data. The Semantic Web-ISWC 2009, pp. 731–746

Berners-Lee T (2006) Linked Data - Design Issues. Retrieved 1 October 2014. http://www.w3.org/DesignIssues/LinkedData.html. Accessed May 2018

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific American 284(5):2837

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

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009b) DBpedia-A Crystallization Point for the Web of Data. Web Semantics, Science, Services and Agents on the World Wide Web 7(3):154–165

Eckartz SM, Folmer EJA (2015) BOMOD: Management and development model for open data. TNO

Folmer E, Beek W (2017) Kadaster Data Platform - Overview Architecture. Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings: Vol. 17, Article 23. Available at: http://scholarworks.umass.edu/foss4g/vol17/iss1/23. Accessed May 2018

Frakes WB, Baeza-Yates R (1992) Information retrieval: Data structures and algorithms

Goodwin J, Dolbear C, Hart G (2008) Geographical linked data: The administrative geography of Great Britain on the semantic web. Transactions in GIS 12(s1):19–30

Janowicz K, Scheider S, Pehle T, Hart G (2012) Geospatial semantics and linked spatiotemporal data—Past, present, and future. Semantic Web 3(4):321–332

Janowicz K, Van Harmelen F, Hendler JA, Hitzler P (2014) Why the data train needs semantic rails. AI Magazine

Lemmens R, Wytzisk A, By R, Granell C, Gould M, Van Oosterom P (2006) Integrating semantic and syntactic descriptions to chain geographic services. IEEE Internet Computing 10(5):42–52

Marchionini G (2006) Exploratory search: From finding to understanding. Communications of the ACM 49(4):41–46

Rajabifard A, Feeney MEF, Williamson IP (2002) Future directions for SDI development. International Journal of Applied Earth Observation and Geoinformation 4(1):11–22

Scheider S, Janowicz K, Adams B (2012) The observational roots of reference of the semantic web. arXiv preprint arXiv:1206.6347