Chapter 13
Linked Data, Towards Realizing the Web of Data: An Overview

Leila Zemmouchi-Ghomari
University of Sciences and Technology Houari Boumediene (USTHB), Algeria

ABSTRACT

Data play a central role in the effectiveness and efficiency of web applications, such as the Semantic Web. However, data are distributed across a very large number of online sources, due to which a significant effort is needed to integrate this data for its proper utilization. A promising solution to this issue is the linked data initiative, which is based on four principles related to publishing web data and facilitating interlinked and structured online data rather than the existing web of documents. The basic ideas, techniques, and applications of the linked data initiative are surveyed in this paper. The authors discuss some Linked Data open issues and potential tracks to address these pending questions.

INTRODUCTION

At present, the web is based on several notions of information sharing (Allemang and Hendler, 2011) that lead to considerable misunderstanding, such as the Anyone can say Anything about Any topic (AAA) slogan, the open world assumption, and the non-unique naming assumption. Thus, there is a pressing need to move from the current framework to one guided by consistent (homogeneous) principles that allow information sharing, cooperation, and collaboration.

Web data often consists of isolated silos (Herman, 2010) that cannot exchange content with other systems on the web. This lack of communication and sharing of data is due to incompatibility among the various online data formats. A contextual interpretation of these datasets is time-consuming and expensive, and it requires developers' intervention. Moreover, data is embedded into web pages and is available only for human consumption. In contrast, the Semantic Web aims to enable machines to understand and process data contained in web pages and online documents. It targets the realization of a structured 'Web of Data' to complement the existing Web of Documents, which is loosely structured at best.

DOI: 10.4018/978-1-5225-5191-1.ch013

Many researchers (Bizer & Heath, 2009; Wood et al., 2014; Hogan, 2014) are convinced that the Linked Data initiative[1] is a promising approach for publishing and connecting structured data on the Internet by means of a set of standards and tools. Linked Data technologies render explicit and transparent the conceptual models underlying the visible data. In particular, they support data integration in dynamic and distributed environments, such as large enterprises, intergovernmental organizations, and the World Wide Web (Mendez & Greenberg, 2012).

Data exchange is possible if everyone uses a shared data format, such as the Resource Description Framework (RDF), as well as common ways of accessing it, such as Simple Protocol and RDF Query Language (SPARQL).

In practice, Linked Data facilitates the creation of web pages using information from multiple web pages. As Wood notes, 'Linked Data enables cooperation without coordination' (Wood et al., 2014).

However, simply publishing Linked Data in the Linked Data cloud does not allow reuse. Publishing requires provenance, quality, credit, attribution, and the implementation of methods to provide reproducibility for the validation of results (Bechhofer et al., 2013). Thus, several challenges remain unaddressed for researchers in this field.

Survey research can be used for exploration, description, or explanation purposes (Pinsonneault & Kraemer, 1993). Survey research in exploration aims to make readers more familiar with a topic and explain its basic concepts. The objective of this paper is to highlight the growing movement of the Semantic Web community towards the realization of the Web of Data using Linked Data principles for publishing and consuming data on the web. We explore the best practices and recommendations of the World Wide Web Consortium (W3C) in terms of web data modelling and querying, interlinking techniques, vocabulary reuse, and key areas for future research.

The remainder of this paper is organized as follows. Section 2 explains the main Web of Data principles with regard to the Web of Documents and describes how this can be realized through different approaches. Section 3 defines Linked Data and states its four principles (Section 3). Section 4 describes RDF, the Linked Data model. Section 5 provides an overview of SPARQL, the Linked Data query language. Sections 6 and 7 discuss existing approaches and tools for the publication and consumption of Linked Data from heterogeneous data sources and formats. Section 8 presents some of the most prominent success stories of the application of Linked Data technologies. Section 9 discusses the challenges faced by Linked Data along with the probable causes and possible solutions. Finally, Section 10 concludes the paper by highlighting the importance of Linked Data as a promising research field.

WEB OF DATA DESCRIPTION AND APPROACHES

The Web of Data (Bizer & Heath, 2011) can be considered as another layer that is linked with the classic document Web, and it has the following features:

The Web of Data is generic and can contain any type of data.

Entities are connected by links, creating a unique giant global graph that extends data sources and allows new data sources to be found.

The Web of Data is open, meaning that applications do not have to be implemented against a fixed set of data sources, and anyone can publish data to the Web.

In order to facilitate a better understanding of the Web of Data, a comparison between the Web of Data and the Web of Documents is presented in Table 1.

Table 1. Comparison between web of documents and web of data

Web of Documents	Web of Data
Primary objects: documents	Primary objects: things or concepts (or description of things)
Links between documents (or parts of them)	Links between things
Degree of structure in data: fairly low	Degree of structure: high (based on RDF data model)
Implicit semantics of contents	Explicit semantics of contents and links
Designed for human consumption	Designed for both machines and humans

The origins of the Web of Data lie in the efforts of the Semantic Web research community, especially in the activities of theW3C Linking Open Data (LOD) project[2] founded in January 2007.

A piece of content or data is open if anyone is free to use, reuse, and redistribute it; indeed, a sustainable and consequent strategy of publishing and linking data on the web requires the data to be open. Most web sites are created using HTML language, which structures textual documents rather than data. As data is embedded into the text, applications cannot extract structured data from these pages.

The following approaches have been proposed to tackle this issue (Bizer & Heath, 2011):

1. Microformats
2. Web APIs
3. Linked Data

Microformats can be used to publish structured data that describes entity types by specifying how to implant data, and thus, applications can extract data from web pages. However, microformats are restricted to represent data about a small set of different types of entities; they only provide a small set of attributes describing these entities. In addition, it is often not possible to express relationships between entities. Therefore, microformats are not suitable for extracting random data on the web.

Web APIs allow the availability of structured data on the web. They provide interrogation facilities of structured data via the HTTP protocol; therefore, these small applications associate data from different sources, each of which is read through an API specific to the data provider.

It is clear that these two approaches are not generic and do not allow extracting and linking arbitrary data on the web. In fact, linking data across the web requires a way for specifying the existence and the semantics of connections between things described in such data. This mechanism is provided by RDF[3]. Compared with HTML documents and web APIs, RDF provides more semantics and more generality for the following reasons:

- RDF links things or concepts, not just documents: RDF deals with entities extracted from web documents
- RDF links are typed: the relationships are explicitly specified

The use of HTTP as a standardized data access mechanism and RDF as a standardized data model simplifies data access and linking. As an example, Python can be used with RDF lib and HTML5 to extract RDF data from datasets and display it as HTML.

LINKED DATA DEFINITION AND PRINCIPLES

Linked Data is described as follows: 'The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web' (Bizer & Heath, 2009). Data on the web that is in RDF and is linked to other RDF data is Linked Data, which forms a giant global database that can be queried using SPARQL[4].

Linked Data allows discovery and consumption (standards-based data sharing), and reduces redundancy. It also provides the following:

1. A unifying data model. Linked Data relies on RDF as a data model
2. A standardized data access mechanism (the HTTP protocol)
3. Hyperlink data discovery. Because URIs are worldwide identifiers for entities, Linked Data are able to connect entities in different data sources via hyperlinks
4. Self-descriptive data. Linked Data eases the integration of data from different sources from shared vocabularies; this makes it easier for data consumers to discover, access, and integrate data

Thus, Linked Data depends on two technologies that are fundamental to the web: Uniform Resource Identifiers (URIs) and the Hypertext Transfer Protocol (HTTP).

The RDF model (Section 4) encodes data in the form of subject, predicate, and object triples expressed by means of URIs. The predicate specifies how the subject and object are related.

The Linked Data life cycle (Figure 1) unfolds in several phases according to Auer (Auer et al., 2012):

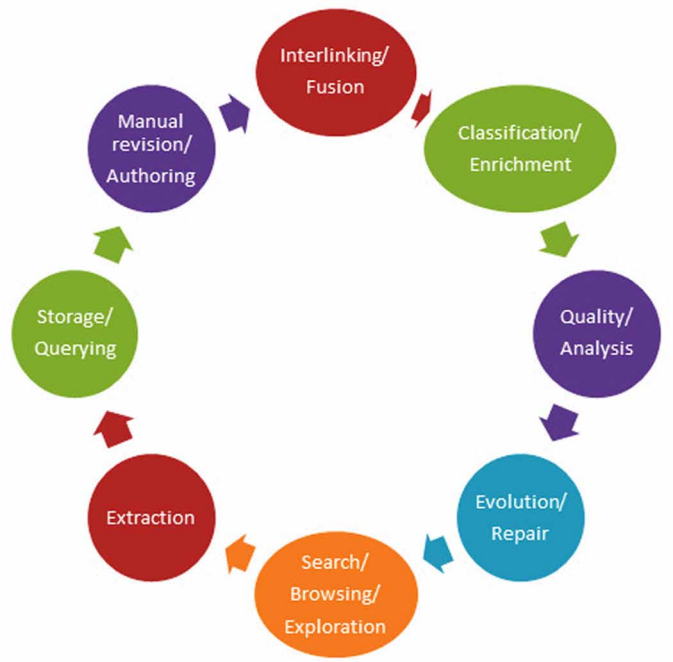
- Storage/querying that involves RDF data management
- Authoring of knowledge bases
- Interlinking between datasets to facilitate data integration
- Classification and integration with upper level ontologies
- Quality analysis according to several dimensions: provenance, context, coverage, and structure
- Evolution/repair in a dynamic environment
- Search/browsing/exploration of explicit web data for end users

As mentioned previously, the term Linked Data denotes a set of best practices for publishing and connecting structured data on the web. These best practices were presented as the Linked Data principles by Tim Berners-Lee in his web architecture notes (Berners-Lee, 2009). These principles are as follows:

1. Use URIs as designations for things: URI references to detect Web documents, concrete objects (e.g. person, book) and abstract concepts (e.g. emotions)
2. Use HTTP URIs so that they can be dereferenced over the HTTP protocol into a description of the identified entity

In this context, there are two different means to make URIs that identify entities dereferenceable, namely, 303 URIs and hash URIs.

Figure 1. Linked Data Life Cycle (Auer et al., 2012)



- 303 URIS: The server does not send the entity over the network; in fact, it answers to the client with the HTTP response code 303 or redirection. In this case, there is a content negotiation with the server: Does the client ask for the HTML or the RDF+XML document?
- The hash URI strategy builds on the characteristic that URIs may contain a special part that is separated from the base part of the URI by a hash symbol (#). This fragment is called the fragment identifier. Thus, a URI that comprises a hash does not automatically identify a Web document.

3. A URI has to provide beneficial information, using a data model for publishing structured data on the Web, i.e. RDF. RDF data on the web can be serialized in different formats. The two RDF serialization formats most commonly used to publish Linked Data on the Web are RDF/XML (Beckett, 2004) and RDFa (Adiba & Birbeck, 2008). RDFa allows RDF to be embedded into HTML code; consequently, the content negotiation mentioned above is unnecessary.
4. Include links to other URIs, so that they can discover more things: Hyperlinks that connect things in a Linked Data context have types that describe the relationship between the things. These links enable applications to access the data.

Concretely, these links can be created using *OWL: Same As* or *RDFS: See Also* to refer to an external equivalent resource. However, these powerful properties may introduce errors because Linked Data includes entailments in addition to the implicit import of properties (Halpin et al., 2010), because *OWL: Same As* is symmetric and transitive.

The example presented in (Wood et al., 2014) is very illustrative. The New York Times used *OWL: Same As* to show the equivalence between three URIs describing the population of Rhode Island. Un-

fortunately, the numbers reported in these resources were different because of the difference in terms of census year. This situation leads to three answers to the query '*what is the population of Rhode Island*?', when a single answer is expected.

The Web of Data is open to arbitrary vocabularies being used in parallel. Despite this general openness, it is considered a good practice to reuse terms from well-known RDF vocabularies, such as Friend Of A Friend (FOAF)[5], Simple Knowledge Organization System (SKOS)[6], Dublin Core[7], and Semantically Interlinked Online Communities (SIOC)[8], wherever possible in order to make it easier for client applications to process Linked Data. Descriptions of some common vocabularies are provided in Table 2.

Most of these vocabularies are specific to a particular domain, except schema.org, which supports a wide range of entities belonging to several domains, such as article, blog, recipe, photography, review, movie, and map.

The LODstats project[9] generates statistics for the most commonly used vocabularies.

LINKED DATA MODEL: RDF

The Semantic Web aims to represent web data with a model that yields a dynamic web of information in a systematic way (Allemang & Hendler, 2011). Linked Data is structured data based on the RDF data model. RDF is a simple and highly flexible data model for semantically describing resources on the web (Domingue et al., 2011), recommended by the W3C. RDF addresses the issue of managing distributed data; it relies on the infrastructure of the web and some of its most common features (Allemang & Hendler, 2011), i.e. URIs.

An RDF triple is of the form subject-property-object. RDF annotates web resources in terms of named properties. The values of the named properties can be URIrefs of web resources or literals, representa-

Table 2. Some common vocabularies

Vocabulary	Description
FOAF (Friend Of A Friend)	Allows description of people and their interests, activities, and relationships with other people
Schema.org	Provides common schema for structured data markup on web pages
VoID (Vocabulary of Interlinked Data)	Describes datasets and sitemaps that describe websites
DOAP (Description Of A Project)	Describes software projects
DC (Dublin Core)	Describes web pages and all types of publications
GoodRelations	Describes e-commerce domains
Geonames	Specifies geographic location
Bibo	Describes citations and bibliographic references

tions of data values such as integers and strings. RDF subjects must be URIs. A set of RDF statements is called an RDF graph (Staab & Studer, 2010).

RDF formats (RDF/XML, Turtle, RDFa, JSON-LD) used in Linked Data are compatible because they share a common data model. Every RDF format can be selected for a user's preferences without negatively influencing data interoperability. RDF/XML was the first RDF format, and it is still used widely in enterprises. Turtle is the simplest and most human-readable format. RDFa is the preferred way to embed RDF data into web pages. JSON-LD is a relatively new format intended for web developers having structured data in JSON. Using RDFa in the development of web pages means that every item related to a vocabulary in such a page is associated with its semantics, as shown in Figure 2.

Thus, RDFa enables search engines to retrieve more relevant results and helps to publish the content as Linked Data on the web. Search engines prefer using RDFa Lite because they care about scalability; however, RDFa Lite has limited expressivity because it consists of only five attributes, namely, vocab, typeOf, property, resource, and prefix. The RDFa elements are used in conjunction with HTML tags. An example of Tim Berners-Lee's information (image, name, job title, and home page) is shown in Figure 3.

LINKED DATA QUERY LANGUAGE: SPARQL

SPARQL aims to be the language for querying data published on the web, but actually, it is a query language for RDF either stored natively as RDF or viewed as RDF via middleware (Domingue et al., 2011). It is a recommendation of the W3C.

SPARQL is to RDF as SQL is to relational databases. The use of triple patterns in the WHERE clause is one of the syntactic differences between SQL and SPARQL; another is the use of prefixes in SPARQL (used to abbreviate long URIs). RDF data can be queried locally (using tools such as TWINKLE) or online (using SPARQL endpoints such as Virtuoso SPARQL editor).

SPARQL endpoints are web query services that return results in different formats such as text, HTML, XML, and JSON. They are not intended to be used only by humans; they accept SPARQL queries in the parameters of HTTP GET or POST requests.

A SPARQL query consists of five parts (Domingue et al., 2011): zero or more prefix declarations (introduce shortcuts for long URIs), a query result clause, zero or more FROM or FROM NAMED clauses, a WHERE clause, and zero or more query modifiers. A SPARQL query can take four forms: SELECT, ASK, CONSTRUCT, and DESCRIBE.

Figure 2. RDFa added value in a web page (Wood et al., 2014)

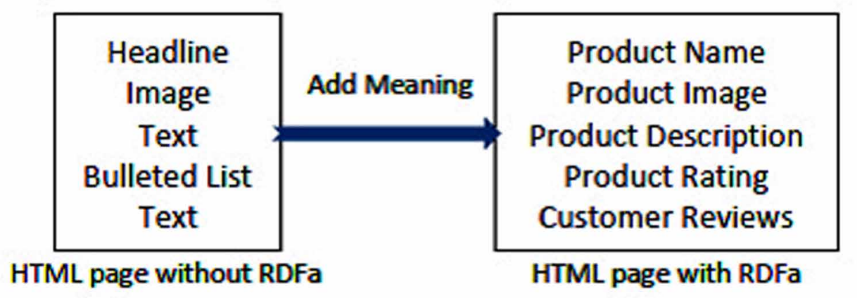


Figure 3. HTML+RDFa Code Example

```
<div vocab="http://schema.org/" typeof="Person">
 <a property="image" http://www.w3.org/
        Press/Stock/Berners-Lee/2001-europaeum-eighth.jpg ">
  <span property="name">Tim Berners Lee</span></a>,
  <span property="jobTitle">Web Inventor</span>
 <div>
  Links: <a property="url" href=" http://www.w3.org/
        People/Berners-Lee/">Tim Berners Lee's Homepage</a>
 </div>
</div>
```

- SELECT queries provide answers in a tabular form, such as for an SQL query executed against a relational database
- ASK form checks whether the SPARQL endpoint can provide at least one result; the answer to the query is YES or NO
- CONSTRUCT form provides the answer to the query as an RDF graph
- DESCRIBE form is used to retrieve information without knowing the vocabulary in use, producing an RDF graph as the result

The optional set of FROM or FROM NAMED clauses defines the dataset against which the query is executed. The WHERE clause is the core of a SPARQL query. It is specified in terms of a set of triple patterns. SPARQL also provides a set of optional query modifiers such as ORDER BY, which orders the results set, and LIMIT and OFFSET, which allow results to be obtained in a specified number. An example of a SPARQL query[10] is, '*To find landlocked countries with a population greater than 15 million, with the highest population country first*'; see Figure 4.

HOW TO PUBLISH LINKED DATA

Linked Data publishers should preferably follow some best practices in order to fully exploit the potential of Linked Data. Therefore, it is possible to optimize the discovery of the datasets in Semantic Web search results by:

- Publishing DOAP files related to projects (using DOAP A MATIC, for example)
- Publishing VoID files for describing datasets (using Ve2 editor, for example)
- Publishing semantic sitemaps for websites (using Semantic Web crawling, for example)
- Providing metadata (labels and comments) to enhance the retrieval of datasets

Figure 4. SPARQL Query Example

```
PREFIX type: <http://dbpedia.org/class/yago/>
PREFIX prop: <http://dbpedia.org/property/>
SELECT ?country_name ?population
WHERE {
   ?country a type:LandlockedCountries ;
         rdfs:label ?country_name ;
         prop:populationEstimate ?population .
   FILTER (?population > 15000000 &&
   langMatches(lang(?country_name), "EN")) .
} ORDER BY DESC(?population)
```

Techniques for web Linked Data publication[11] depend on the possible forms of data available on the web (Bizer & Heath, 2011). More explicitly:

- Datasets stored in relational databases can be published as Linked Data by using relational database-to-RDF wrappers. These tools allow mappings to be defined between relational databases and RDF graphs
- Structured data related to a custom API (such as the Flickr or Amazon web APIs). In this case, a wrapper has to be developed according to the API
- Text documents are transformed into Linked Data by using Linked Data entity extractors such as Calais[12], Ontos[13], or DBpedia Spotlight[14], which annotate documents with the Linked Data URIs of entities referenced in the documents

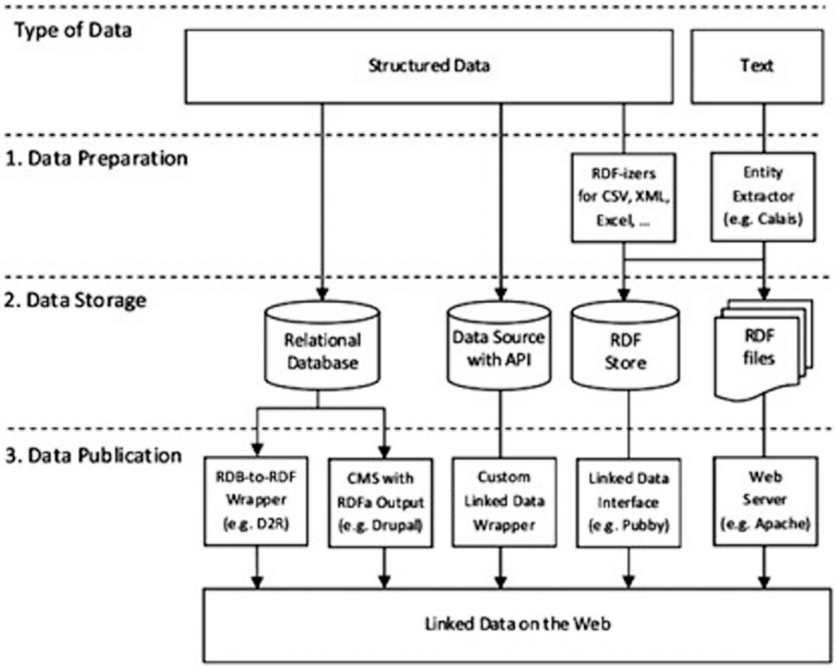
Figure 5 summarizes all possible transformation scenarios from heterogeneous web data sources to web Linked Data.

Alignment of multiple data sources using Linked Data principles (Wood et al., 2014) is possible because of:

- The use of URIs as unique identifiers
- The use of common vocabularies
- The use of *OWL:Same As* or *RDFS: See Also* properties to highlight equivalences between primitives

We want to place particular emphasis on a specific data source shown in Figure 5, i.e. RDF databases, because of its status with regard to Linked Data. RDF databases fall into the category of NoSQL (Not only SQL) databases as opposed to traditional relational databases.

Figure 5. Linked data publication scenarios (Bizer & Heath, 2011)



The distributed and changing nature of the web makes NoSQL databases more suitable than relational databases for storing web data. RDF databases are triple stores that allow the running of SPARQL queries on such data, e.g. Allegrograph, Virtuoso, and Big Data.

More generally, Tim Berners-Lee has described Linked Data sets in terms of a five-star rating scheme (Berners-Lee, 2009); in other words, he proposes that stars should be awarded to published datasets according to the following criteria:

- **1 Star:** Data is available on the web (whatever format), but with an open license
- **2 Stars:** Data is available as machine-readable structured data (e.g., Microsoft Excel instead of a scanned image of a table)
- **3 Stars:** Data is available as machine-readable structured data but in a non-proprietary format (e.g., CSV instead of Excel)
- **4 Stars:** In addition to the above principles, the use of standards from the W3C (RDF and SPARQL) to identify entities
- **5 Stars:** Data is available according to all of the above, plus outgoing links to other people's data to provide context

Figures 6 and 7 show how the number of datasets published on the web as Linked Data has increased since the foundation of the Linking Open Data project.

In the most recent LOD cloud, datasets are categorized into the following domains: geographic, government, media, libraries, life science, commerce, user-generated content, and cross-domain datasets.

A dataset can be published in this diagram if it complies with the following conditions:

Figure 6. Linked open data cloud May 2007

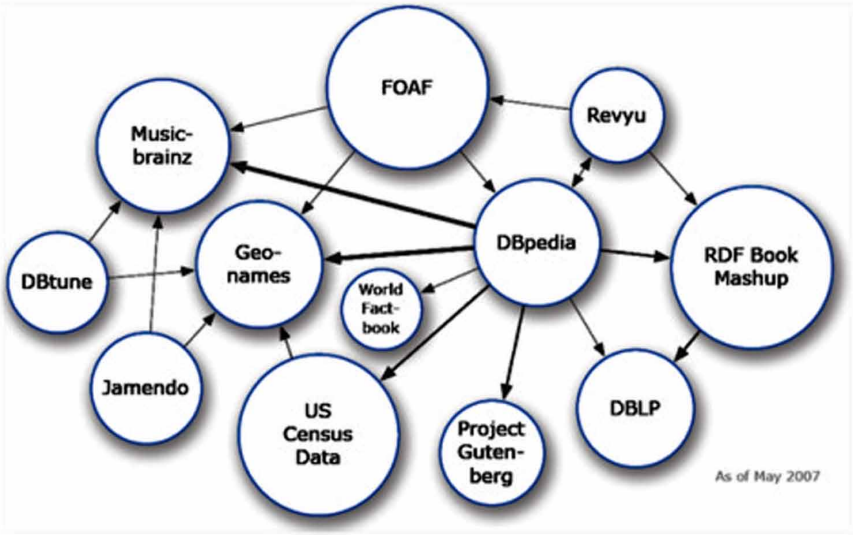
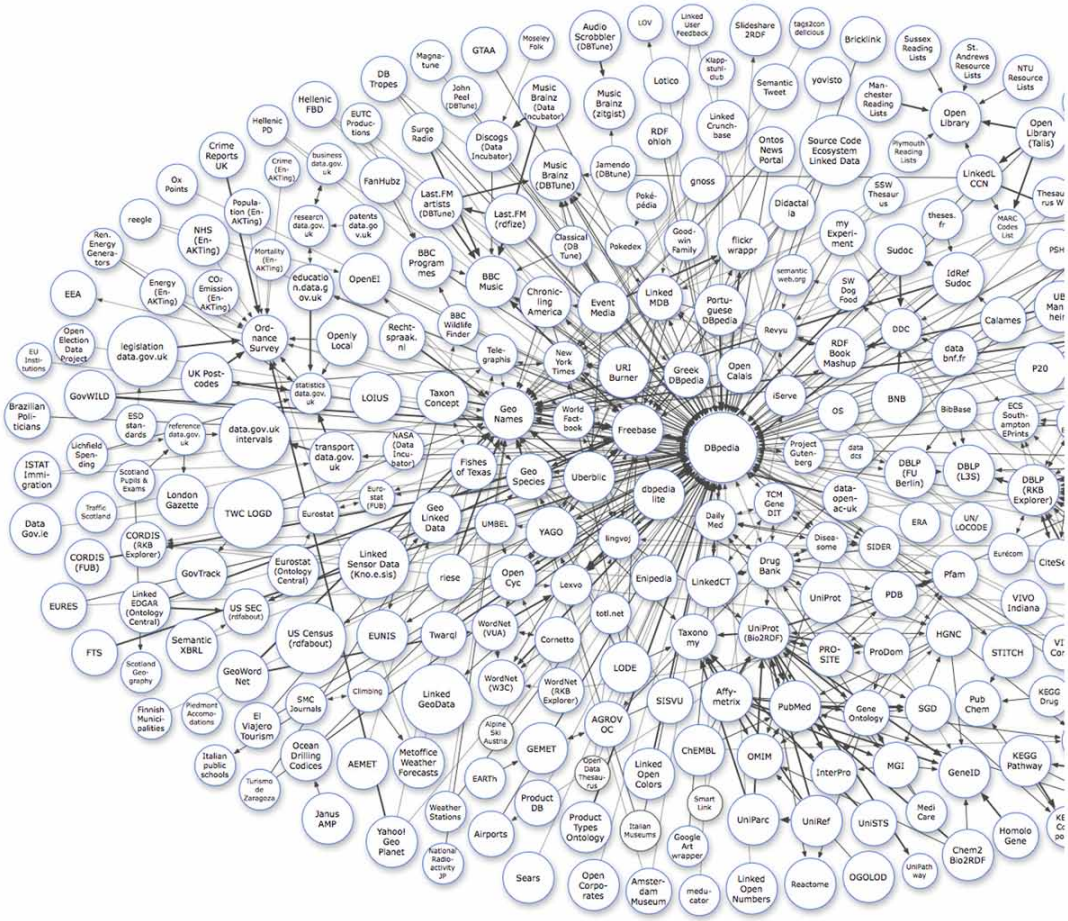


Figure 7. Linked open data cloud September 2011



- Data is published according to the Linked Data principles
- Data must resolve (with or without content negotiation) to RDF data in one of the popular RDF formats (see Section 4)
- The dataset must contain at least 1000 triples
- The dataset must be connected via RDF links to a dataset that is already in the diagram. This means that each dataset must use URIs from the other. At least 50 links are required
- Access to the dataset must be possible via RDF crawling, an RDF dump, or a SPARQL endpoint

The LOD cloud consists of more than 300 datasets from various domains, over 31 billion data items, and 500 million links between them. More information about each of these datasets can be obtained by exploring the LOD Cloud Data Catalog[15], which is maintained by the LOD community in the Comprehensive Knowledge Archive Network (CKAN)[16]. Some of the best known publication tools are Virtuoso Universal Server[17], Pubby[18], CKAN registry, and Sitemap4rdf[19].

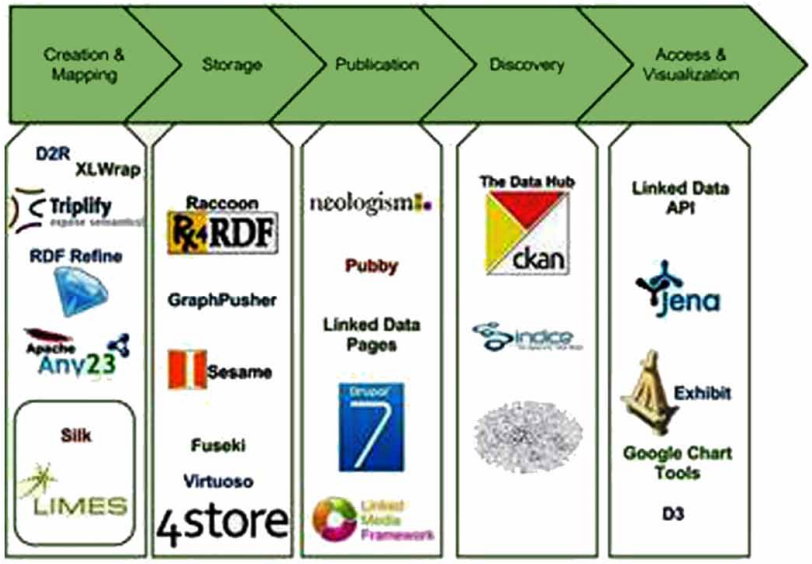
Additional tools dedicated to web Linked Data manipulation are shown in Figure 8.

HOW TO CONSUME LINKED DATA

Once the data is published, it has to be described efficiently in order to enable its discovery via search engines by means of two primary mechanisms available for publishing descriptions of a dataset:

- Semantic Sitemaps (Cyganiak et al., 2008): This is an extension of the well-established Sitemaps protocol[20], which provides search engines with hints about pages in a web site that are available for crawling. This extension allows data publishers to state where documents containing RDF data are located and web clients to choose the most efficient means of access for the task at hand.

Figure 8. Linked data tools (Deirdre, 2012)



- VoID (Alexander et al., 2009): Vocabulary of Interlinked Datasets (VoID) is the standard vocabulary for describing Linked Data sets descriptions (metadata). It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloguing and archiving of datasets.

There are several ways to consume published Linked Data on the web:

1. Linked Data browsers: These enable users to navigate between different data sources by following RDF links. An RDF link means that one piece of data has some type of relationship to another piece of data. These relationships can have different types, for example two persons who have FOAF profiles may be linked by the relation 'know'. Some of these Linked Data browsers are the Disco hyperdata browser[21], Tabulator Browser[22], Marbles[23], and OpenLink RDF Browser[24].
2. Linked Data mashups: Multiple sources can be queried and combined on the fly using Linked Data mashups, which are domain-specific applications. These mashups conduct the following phases (Bizer & Heath, 2011):
    - Discover data sources that provide data about a specific entity by following RDF links from an initial seed URI into other data sources
    - Download data from the discovered data sources and store the data together with provenance meta information in a local RDF store
    - Retrieve information to be displayed using SPARQL

Some of the Linked Data mashups are Revyu[25] (website for rating everything), DBtune Slashfacet[26] (visualizes music-related Linked Data), and DBpedia mobile[27] (geospatial entry point into the Web of Data)

3. **Search Engines:** Traditional search engines such as Google and Yahoo have also started to use structured data from the web within their applications. Google crawls RDFa and microformat data describing people, products, businesses, organizations, reviews, recipes, and events. It uses the crawled data to provide richer and more structured search results to its users in the form of Rich Snippets. Whereas SWSE[28] and Falcons[29] provide search functions for human needs, other search engines have been developed to serve the needs of applications built on top of distributed Linked Data, such as Swoogle[30] (Ding et al., 2005), Sindice[31] (Oren et al., 2008), and Watson[32] (d'Aquin et al., 2009); these provide APIs through which Linked Data applications can discover RDF documents on the web that reference a certain URI or contain certain keywords (Bizer & Heath, 2009).

LINKED DATA APPLICATIONS

Linked Data is being used beyond the LOD cloud and is becoming the basis for data sharing in many contexts. As stated by Wood, '*Linked Data successes are difficult to see because they are under the hood*' (Wood et al., 2014).

Linked Data allows concretization (Mendez & Greenberg, 2012) of the following purposes:

- Making the meaning of data and information more explicit, visible, and linkable
- Integration of data from different sources in dynamic environments such as the web (Jaffri, 2010)

- Generating search functionalities and reports on top of integrated data in a cost-efficient way
- Using the web and web technologies to publish data and/or metadata to leverage the value of assets

Some concrete applications of Linked Data are reported in the case studies of the W3C consortium[33]. These case studies are descriptions of systems that have been deployed within an organization, and are now being used within a production environment, such as Google's Rich Snippets, which provide richer and more structured search results to users. Google embedded RDFa in its webpages, which resulted in an increase of 15%–30% in the click-through rate for its global results (Wood et al., 2014).

Some other success stories are presented below:

- The BBC is the largest broadcasting corporation in the world. The BBC's team of developers builds the BBC website by integration of the available Linked Data. For example, BBC Nature aggregates data from different sources, including Wikipedia, WWF WildFinder, the IUCN Red List of Threatened Species, the EDGE of Existence program of the Zoological Society of London, and the Animal Diversity Web. BBC Wildlife Finder repurposes that data and puts it in a BBC context, linking out to program clips extracted from the BBC's Natural History Unit archive.
- In April 2010, Facebook launched the Open Graph protocol to manage users' interests and their preferences using vocabularies from the Linked Data cloud, such as FOAF and Semantically-Interlinked Online Communities (SIOC) (Bojars et al., 2008). The point is that if a person's friends recommend something, he or she will be more inclined to like it. Less than a month after the rollout of Open Graph, more than 100,000 sites had integrated the technology (Heitmann et al., 2010). In fact, every time someone clicks on a like button, two RDFa triples are generated; one of them is stored in Facebook's database and the other one is sent to the supplier of the product that is liked (Wood et al., 2014).
- The Open Government Data (OGD) is an international collaboration between the governments of the US, the UK, France, and Singapore for sharing machine-readable datasets covering government activities. Datasets are produced by governments or government-controlled entities. One significant benefit of the OGD initiative is greater governmental transparency, i.e. the creation of services that deliver social and commercial value and encourage participatory governance (Ding et al., 2011).

Linked Data technologies certainly allow large organizations to set up data integration with relatively little effort as compared to traditional data warehousing solutions that require the design of a global schema (Bizer & Heath, 2011). Gradually, these organizations can invest in Linked Data by reusing shared vocabularies or schema mappings between datasets.

DISCUSSION

In this section, we discuss open challenges related to Linked Data, their origins, and possible approaches for solving them (see Table 3).

If we have to arrange these challenges in terms of priority, we approve Needleman's (Needleman, 2011) summary of the most significant Linked Data issues, which are:

Table 3. Linked data issues – causes and potential solutions

Cause: Lack of	Effect: Difficulty of	Potential Solution(s): Development of
Metadata: Provenance, ownership, versioning (Bechhofer et al., 2013)	Reuse, reproducibility	Research objects (container infrastructure of data resources, relationships, experiments conducted on Linked Data sets) (Bechhofer et al., 2013)
Formal description of LOD Cloud datasets(Jain et al., 2010) Reference dataset per domain (Zemmouchi-Ghomari & Ghomari, 2009)	Knowledge discovery (especially systematically)	Formal descriptions of datasets LOD sets Description rules Domain Reference Ontologies
Detailed knowledge of the dataset structures Adequate number of LOD datasets to answer complex and frequent queries (Polleres et al., 2010)	Querying of the LOD cloud (Hartig et al., 2009; Naumann, 2002)	Available detailed and formal descriptions of queried datasets at the endpoints. Reuse of query results by means of query subsumption (based on analysing graph patterns of cached SPARQL queries)
Different datasets cover the same knowledge domain	Selection from a set of different results of a given query	(Toupikov et al., 2009) mechanisms for classification of datasets according to their popularity among end users
Interlinking or mappings between datasets (schema and instance level) (Biessmann & Harth, 2010) Link maintenance is expensive Publishers continue to use locally defined URIs (Polleres et al., 2010)	Link discovery (Jain et al., 2010; Bechhofer et al., 2013) in terms of recall (all links) and precision (correctness)	Development of more efficient link discovery platforms than LIMES or SILK (Jain et al., 2010) use of an upper level ontology such as SUMO to formalize relationships and descriptions of the datasets, (Reed & Lenat, 2002) use of CYC, (Bergman & Giasson, 2008) use of UMBEL
Language Expressivity used to design datasets	Knowledge reasoning, inferring new facts	Use of more expressive languages such as OWL 2 instead of RDFa, RDF Lite, or OWL Lite
Services to exploit datasets	Disinterest/underuse of available LOD datasets (Jain et al., 2010) Understanding dataset semantics	Efficient Tools to achieve data parsing, data transformation, data standardization, and data fusion.
Mechanism to identify dataset domain performance: still substantial penalties compared to relational datasets (Auer & Lehmann, 2010) Large-scale processing (usually cannot be loaded in standard OWL reasoners) (Bizer & Schultz, 2009) Data fusion mechanisms Missing end-user tools	Data consumption Data Usability	The Semantic Web community should conduct more applied research to demonstrate Linked Data cloud possibilities (Polleres et al., 2010) Large community-driven effort, such as efforts in and around SIOC and FOAF Adaptive automatic data indexing technologies (Auer & Lehmann, 2010) Existing machine learning algorithms have to be extended (from DL) and they have to be optimized for processing large-scale knowledge bases (Auer & Lehmann, 2010) Open licenses or at least clear information about licensing Very large RDF data management platforms Adaptive UI interfaces
Subjective and individual design of real-world objects Differences in contexts of represented knowledge	Contradictory facts reported about common entities from different knowledge bases	Mechanisms for entity disambiguation

continued on following page

Table 3. Continued

Cause: Lack of	Effect: Difficulty of	Potential Solution(s): Development of
No common and standardized evaluation metrics Data providers do not receive feedback on its use	Evaluation of the quality of the dataset (Harting & Zhao, 2009)	Use of upper-level ontologies (e.g. SUMO) to do reasoning and propagation of the queries from concepts of upper-level ontologies to LOD set instances User feedback mechanism Evaluation approach of the data quality integrated as part of the publication process (Hartig and Zhao, 2009) such as: • (Motro & Rakov, 1998): automatic assessment for evaluating soundness and completeness • (Bobrowski et al., 1999): use of questionnaire based on user input • (Yang et al., 2002): measurement of information quality from soundness, dependability, usefulness User input (instance and schema mappings) for machine learning techniques (inductive reasoning) and results can be assessed again by end users (iterative refinement) Services to execute constant evaluation of links between knowledge bases Precision and recall measures as more realistic evaluation metrics (Hitzler et al., 2010)
Human design errors Soundness and completeness (in terms of formal semantics) by means of existing reasoning systems (Hitzler et al., 2010)	Incoherencies and inconsistencies in the LOD sets, examples: use of the foaf:image property to relate an arbitrary resource with an image, missing the fact that foaf:person is the domain of this property (Polleres et al., 2010)	The web community can resolve inconsistencies by working with data providers (pointing out mistakes and helping to fix them) such as the pedantic web group Semi-automatic repair algorithms, such as explanation of OWL entailments[35], a protégé plugin, or model-based revision operators that remove axioms causing inconsistencies (Qi & Du, 2009)
Mechanisms to track the provenance of the published data on the web	Determining the provenance of Linked Data	(Hartig and Zhao, 2009) proposed an approach consisting of 3 steps: collecting elements of provenance information, deciding on its influence, and applying a function to calculate the quality

- **Quality and Relevance:** Data quality is a problem in every data management system. Ensuring that data most relevant or appropriate to the user's needs is identified and made available is an enduring issue.
- **Maintenance:** Keeping links valid is quite difficult, especially with the evolution of Linked Data clouds. Robust mechanisms that can automatically check and update links will be needed for this purpose.
- **User Interfaces and Interface Design:** A Linked Data browser has to dynamically integrate access to data from distributed and heterogeneous data sources. This may involve integration of data from sources not explicitly selected by the user
- **Licensing:** Some data will have licensing restrictions. How online user interfaces, such as web browsers, will tackle this type of data remains an open question.

CONCLUSION

The Linked Data project has been initiated to enable computers to understand online content in order to help users easily find, share, and combine information, i.e. to fulfil the purposes of the Semantic Web.

One of the most evident benefits of Linked Data is that it presents the web as a giant global database that can be queried using tools such as SPARQL endpoints.

Linked Data is a viable means of concretizing the idea of the Web of Data, in addition to the existing Web of Documents, if all stakeholders (web data publishers) agree to publish data according to the principles articulated by Tim Berners-Lee. These principles aim to realize standardization in terms of creation, mapping, storage, publication, discovery, access, and visualization of web Linked Data.

In spite of these restrictions, data publishers in this framework have some flexibility in their activity because they are not constrained in their choice of vocabulary to represent data, and they can freely represent disagreement and contradictory information about an entity.

Annotated data with RDF vocabularies using RDFa Lite is found more easily by search engine crawlers, contributing significantly to the publication of Linked Data. Furthermore, the Web of Data is sufficiently generic for anyone to publish any type of data. Standardization will accelerate the progress of the Linked Data community.

There are several unresolved issues concerning Linked Data in spite of its promise. These include quality, maintenance, licensing, and end-user consumption. Nevertheless, we are convinced that Linked Data will constitute a significant evolutionary step in realizing the full potential of the web.

Our future efforts will be dedicated to develop approaches to assess, monitor, maintain and improve Linked Data quality based on a combination of ontology evaluation approaches.

REFERENCES

Adida, B., & Birbeck, M. (2008). Rdfa primer - bridging the human and data webs. *W3C recommendation*. Retrieved from http://www.w3.org/TR/xhtml-rdfa-primer/

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2009). Describing linked datasets. *Proceedings of the WWW Workshop on Linked Data on the Web*, Madrid, Spain.

Allemang, D., & Hendler, J. (2011). *Semantic Web for the Working Ontologist, Effective Modeling in RDFS and OWL* (2nd ed.). Morgan Kaufmann publishers.

Auer, S., Buhmann, L., & Dirschl, C. (2012). Managing the life-cycle of linked data with the lod2 stack. *Proceedings of the International Semantic Web Conference, ISWC*, Boston, USA. doi:10.1007/978-3-642-35173-0_1

Auer, S., & Lehmann, J. (2010). Creating knowledge out of interlinked data. *Semantic Web Journal*, *1*(1), 97–104.

Bechhofer, S., Buchan, I., de Roure, D., Missier, P., Ainsworth, J., Bhagat, J., & Goble, C. et al. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599–611. doi:10.1016/j.future.2011.08.004

Beckett, D. (2004). Rdf/xml syntax specification (revised). *W3C recommendation*. Retrieved from http://www.w3.org/TR/rdf-syntax-grammar/

Bergman, M. K., & Giasson, F. (2008). *Umbel ontology*, Technical report, Structured Dynamics LLC. Retrieved from https://github.com/structureddynamics/UMBEL/blob/master/Doc/UMBEL_TR-11-2-10.pdf

Berners Lee, T. (2009). Linked data design issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.htmls

Biessmann, F., & Harth, A. (2010). Analysing dependency dynamics in Web data. *Proceedings of the Linked AI: AAAI Spring Symposium*, Palo Alto, California, USA.

Bizer, C., & Heath, T. (2011). *Evolving the Web into a Global Data Space*. Morgan and Claypool Publishers. doi:10.1007/978-3-642-24577-0_1

Bizer, C., Heath, T., & Berners Lee, T. (2009). Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Bizer, C., & Schultz, A. (2009). The berlin sparql benchmark. *International Journal on Semantic Web and Information Systems*, *5*(2), 1–24. doi:10.4018/jswis.2009040101

Bobrowski, M., Marrie, M., & Yankelevich, D. (1999). A homogeneous framework to measure data quality. *Proceeding of Information Quality workshop*.

Bojars, U., Passant, A., Breslin, J., & Decker, S. (2008). Social Network and Data Portability using Semantic Web Technologies. *Proceedings of the Workshop on Social Aspects of the Web*, Innsbruck, Austria.

Cyganiak, R., Stenzhorn, H., Delbru, R., Decker, S., & Tummarello, G. (2008). Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. *Proceedings of the 5th European Semantic Web Conference*, Tenerife, Canary Islands, Spain. doi:10.1007/978-3-540-68234-9_50

D'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., & Guidi, D. (2008). Toward a new generation of semantic web applications. *IEEE Intelligent Systems*, *23*(3), 20–28. doi:10.1109/MIS.2008.54

Deirdre, L. (2012). Linked open data. Retrieved from http://fr.slideshare.net/deirdrelee/linked-open-data-15303345

Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., & Hendler, J. A. et al. (2011). Twc logd: A portal for linked open government data ecosystems. *Web Semantics: Science, Services, and Agents on the World Wide Web*, *9*(3), 325–333. doi:10.1016/j.websem.2011.06.002

Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., & Kolari, P. (2005). Finding and ranking knowledge on the semantic web. *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland. doi:10.1007/11574620_14

Domingue, J., Fensel, D., & Hendler, J. (2011). *Handbook of semantic web technologies*. Springer. doi:10.1007/978-3-540-92913-0

Halpin, H., Hayes, P., & Mccusker, J. (2010). When owl: same as is not the same: An analysis of identity in linked data. *Proceedings of the International Semantic Web Conference, ISWC* (pp. 305-320). Springer Berlin Heidelberg.

Hartig, O. Bizer, C. and Freytag, J.-C. (2009). Executing SPARQL queries over the Web of Linked Data. *Proceedings of the International Semantic Web Conference*, Whashington, USA.

Hartig, O., & Zhao, J. (2009). Using web data provenance for quality assessment. *CEUR Workshop Proceedings*, Corfu, Greece.

Heitmann, B., Kim, J. G., Passant, A., Hayes, C., & Kim, H. G. (2010). An architecture for privacy-enabled user profile portability on the Web of Data. *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems* (pp. 16-23). ACM Publishers. doi:10.1145/1869446.1869449

Herman, I. (2010). Introduction to semantic web technologies. *Proceedings of the* Semantic Web Activity *Lead World Wide Web Consortium*. *Semantic Technology Conference*, San Francisco, California, USA.

Hogan, A. (2014). Linked Data and the Semantic Web Standards. In *Linked Data Management* (pp. 3–48). CRC Press. doi:10.1201/b16859-3

Jaffri, A. (2010). Linked data for the enterprise - an easy route to the semantic web. Retrieved from http://www.capgemini.com/blog/capping-it-off/2010/03/linked-data-for-the-enterprise-an-easy-route-to-the-semantic-web

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., & Sheth, A. P. (2010). Linked Data Is Merely More Data. *Proceedings of the AAI Spring Symposium: linked data meets artificial intelligence*, Palo Alto, California, USA.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. doi:10.1016/S0378-7206(02)00043-5

Mendez, E. and Greenberg, J. (2012). Linked data for open vocabularies and hive's global framework. *El profesional de la informacion*, 21(3), 236-244.

Motro, A., & Rakov, I. (1998). Estimating the quality of databases. In *Flexible query answering systems* (pp. 298–307). Springer Berlin Heidelberg. doi:10.1007/BFb0056011

Naumann, F. (2002). *Quality-driven query answering for integrated information systems*. Springer Verlag. doi:10.1007/3-540-45921-9

Needleman, M. (2011). Linked data: What is it and what can it do? *Serials Review*, 37(3).

Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., & Tummarello, G. (2008). Sindice.com: A document-oriented lookup index for open linked data. *Journal of Metadata*, *Semantics and Ontologies*, *3*(1), 37–52. doi:10.1504/IJMSO.2008.021204

Pinsonneault, A., & Kraemer, K. L. (1993). Survey research methodology in management information systems: An assessment. *Journal of Management Information Systems*, 10(2), 75–105.

Polleres, A., Hogan, A., Harth, A., & Decker, S. (2010). Can we ever catch up with the Web? *Semantic Web Journal*, *1*(1), 45–52.

Qi, G., & Du, J. (2009). Model-based revision operators for terminologies in description logics. *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Pasadena, California, USA.

Reed, S., & Lenat, D. (2002). *Mapping ontologies into Cyc* (Technical report). Cycorp, Inc. Retrieved from http://www.cyc.com/doc/white papers/

Staab, S., & Studer, R. (2010). *Handbook on ontologies*. Springer.

Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., & Tummarello, G. (2009). DING! Dataset ranking using formal descriptions. *Proceedings of the WWW2009 Workshop on Linked Data on the Web*, Madrid, Spain.

Wood, D., Zaidman, M., Ruth, L., & Hausenblas, M. (2014). *Linked Data, Structured Data on the Web*. Manning publishers.

Zemmouchi-Ghomari, L., & Ghomari, A. R. (2009). Reference Ontology. *Proceedings of the International IEEE Conference on Signal-Image Technologies and Internet-Based System*, Marrakech, Morocco.

ENDNOTES

1 http://linkeddata.org
2 http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/ LinkingOpenData
3 http://www.w3.org/RDF/
4 http://www.w3.org/TR/rdf-sparql-query/
5 http://rdfweb.org/foaf/
6 http://www.w3.org/2004/02/skos/
7 http://dublincore.org/
8 http://sioc-project.org/
9 http://stats.lod2.eu/
10 http://www.w3.org/2009/Talks/0615-qbe/
11 http://www.w3.org/TR/swbp-vocab-pub/
12 http://viewer.opencalais.com/
13 http://www.ontos.com/20-10-2010-ontos-links-lod/
14 https://github.com/dbpedia-spotlight/dbpedia-spotlight
15 http://datahub.io/group/lodcloud
16 http://ckan.org/
17 http://virtuoso.openlinksw.com/
18 http://wifo5-03.informatik.uni-mannheim.de/pubby/
19 http://lab.linkeddata.deri.ie/2010/sitemap4rdf/
20 http://www.sitemaps.org/fr/protocol.html
21 http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/
22 http://www.w3.org/2005/ajar/tab
23 http://mes.github.io/marbles/
24 semanticweb.org/wiki/OpenLink_RDF_Browser
25 http://revyu.com/
26 http://dbtune.org/

[27] http://dbpedia.org/DBpediaMobile
[28] http://swse.deri.org/
[29] http://ws.nju.edu.cn/falcons/objectsearch/index.jsp
[30] http://swoogle.umbc.edu/
[31] http://sindice.com/
[32] http://watson.kmi.open.ac.uk/WatsonWUI/
[33] http://www.w3.org/2001/sw/sweo/public/UseCases/
[34] http://www4.wiwiss.fu-berlin.de/lodcloud/state/#license
[35] http://owl.cs.manchester.ac.uk/explanation/

This research was previously published in the International Journal of Technology Diffusion (IJTD), 6(4); edited by Ali Hussein Saleh Zolait, pages 20-39, copyright year 2015 by IGI Publishing (an imprint of IGI Global).