5 Data Quality in the Semantic Web

The Semantic Web is an initiative of the World Wide Web Consortium (W3C) with the vision to evolve the traditional Web, which is essentially a graph of interlinked documents, into a “Web of Data” (Berners-Lee et al., 2001; cf. W3C, 2013). One of the major goals of the Semantic Web is the supply of machine-interpretable data at Web scale to gain a higher degree of automation and to facilitate more complete processing of information (cf. Berners-Lee et al., 2001). For example, if the prices of all consumer products were published in a machine-readable format and structure throughout the whole Web, then more complete price comparisons at global scale would be possible with minimal manual effort. While the traditional Web is mainly used to publish information in a form that empowers a Web browser to render the contents in a form suitable for human consumption, the Semantic Web shall additionally allow computer-based devices to extract and process the meaning of the contents (cf. Berners-Lee et al., 2001). To facilitate the publication and use of structured data at Web scale, Semantic Web formalisms such as RDF (Manola & Miller, 2004), RDFS (Brickley & Guha, 2004), and OWL (Bechhofer et al., 2004; Hitzler et al., 2012) have been developed to support the publication of data. Semantic Web applications can then extract and use the published data, e.g. to derive decisions to automate tasks or to answer complex queries (cf. Berners-Lee et al., 2001). However, Semantic Web-based applications have a high risk to fail if the processed data is of insufficient quality.

In this chapter, we give an overview of existing data sources on the evolving Semantic Web vision and discuss data quality problems and their impact.

5.1 Data Sources of the Semantic Web

As already explained, data on the Semantic Web is mostly published according to the RDF data model (cf. Heath & Bizer, 2011; Manola & Miller, 2004, see also section 4.2.2), which represents graphs of information in the form of simple statements known as triples with the basic structure of subject, predicate, object (cf. Manola & Miller, 2004). The Semantic Web already provides billions of such triples with data about several different domains such as geography, media, health care, life sciences, linguistics, and e-commerce (cf. Bizer, Heath, et al., 2009, p. 5f.; Heath & Bizer, 2011;

C. Fürber, *Data Quality Management with Semantic Technologies,*
DOI 10.1007/978-3-658-12225-6_5, © Springer Fachmedien Wiesbaden 2016

Mühleisen & Bizer, 2012). Figure 24 shows the well-known linking open data (LOD) cloud diagram[22] which represents a large part of available data on the Semantic Web (Cyganiak & Jentzsch, 2011a).

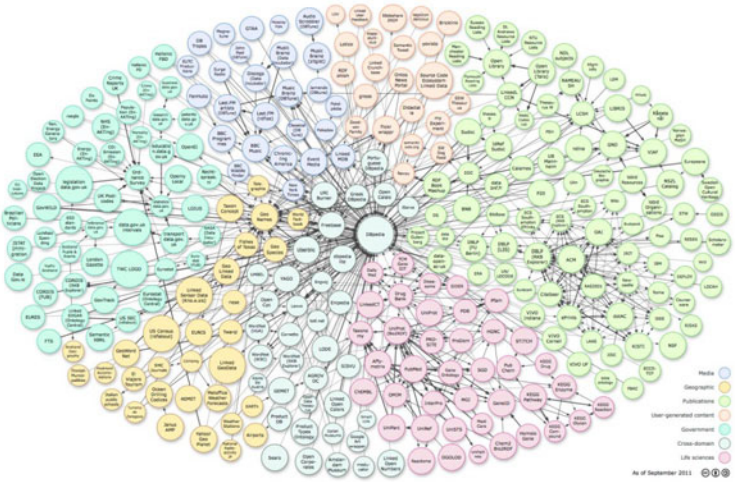

Figure 24: Linking Open Data (LOD) cloud diagram[22] (Cyganiak & Jentzsch, 2011a)

The amount of triples of the LOD cloud was estimated to be around 31 billion triples in September 2011 (Cyganiak & Jentzsch, 2011b). But the LOD cloud only represents part of the Semantic Web, since the latest available version of the diagram was created on September 19th 2011, and data sources have to meet certain criteria to be included in the diagram. For instance, a data source must contain at least 1000 triples and have at least 50 RDF links to other data sets in the diagram (cf. Cyganiak & Jentzsch, 2011a). Hence, a large amount of data that is not linked to data sets in the LOD cloud is not part of the diagram and its statistics. For example, a lot of product data published via the GoodRelations ontology[23], a popular vocabulary for publishing E-Commerce data (Hepp, 2008a), lack explicit links to the LOD cloud and is, therefore, not visible in the diagram despite its significance for the practical application of the Semantic Web.

In addition to the intended usage of data published in the LOD-cloud, like intelligent information processing (cf. Bizer, Lehmann, et al., 2009) or entity recognition in natural language processing (cf. Kobilarov, Scott, et al., 2009, p. 732; Reuters, 2013), the data

[22] Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/ (Last accessed on April 2nd 2012)
[23] http://purl.org/goodrelations (Last accessed on April 12th 2012)

can also be a relevant source for data quality management. Several data quality management heuristics use reference data sets to identify data quality problems (cf. Apel et al., 2010, p. 74; English, 1999, p. 166; Loshin, 2001, p. 161). In (Fürber & Hepp, 2010a), we have shown that Semantic Web data can particularly be useful for the identification of illegal values or functional dependencies between attribute values in the geographic domain with minimal effort. To proof its practical usefulness for DQM, we performed a data quality analysis of real address data from BestBuy stores, a popular North American retailer for consumer electronics (cf. Fürber & Hepp, 2011a). The address data contained addresses of BestBuy stores which were published on the Web via the GoodRelations ontology and the vCard ontology[24], a vocabulary for publishing business card data. We compared the BestBuy data with data from Geonames[25], a Semantic Web data source for geographical information, and identified several data quality problems such as mistyped values and a few illegal city / country combinations. We only used the reference data as provided by Geonames for the data quality analysis which contained all valid city / country combinations and, therefore, saved the tremendous manual effort that would have to be invested for the manual creation and maintenance of this data. Despite these promising first results, it must be stressed that the Semantic Web data sets should be also frequently monitored for data quality errors, when used as a trusted reference. Otherwise, data quality problems in the reference data will be spread to other data sources without being noticed.

In near future, the Semantic Web will most likely further grow and expand its data diversity to additional domains. Therefore, we can expect that more useful data will be published that will open further possibilities for DQM. On the other hand, the number of individuals and organizations who publish data will grow, which may make it more difficult to evaluate the reliability of data from the Semantic Web as reference data for data quality management.

5.2 Semantic Web-specific Quality Problems

In section 3.3, data quality problems types have been shown that are typical for data in relational databases. While most of the illustrated problems may also occur in

[24] http://www.w3.org/2006/vcard/ns-2006.html (Last accessed on April 12th 2012)
[25] http://www.geonames.org (Last accessed on April 12th 2012)

Semantic Web data, there are some quality problems that are specific for Semantic Web data. In the following, we enumerate and describe several Semantic Web-specific quality problems based on findings by (Hogan et al., 2010; Lei & Nikolov, 2007; Lei et al., 2007). We thereby use the term “conceptual elements” to refer to classes and properties. Moreover, we sort the different types of errors into problems related to (1) document content, (2) data format, (3) data definitions and semantics, (4) classification, and (5) hyperlinks. The following representation of Semantic Web data quality problems does not claim to be complete. In fact, due to missing research in this area, additional quality problem types of Semantic Web data will most likely be discovered in future.

5.2.1 Document Content Problems

Missing structured data: In the Semantic Web, it is often expected that machine-processable data is returned when looking up links. But in many cases, the returned content type indicates unstructured data which is not as useful for Semantic Web agents (cf. Hogan et al., 2010).

Imprecise / misreported content types: Although Web documents on the Semantic Web are published in one of the various syntaxes for RDF, like RDF/XML, the content type as returned by the Hyper Text Transfer Protocol (HTTP) response header may be incompatible or more generic than the actual type of the content (cf. Hogan et al., 2010).

5.2.2 Data Format Problems

Document syntax errors: Semantic Web data is usually encoded according to W3C standards for the syntactical representation or formal semantics, such as RDF, RDFS, or OWL (cf. Hogan et al., 2010). These standards provide syntactic and structural requirements which may sometimes be violated. The W3C provides validation

applications which test documents for compliance to the syntax rules of such standards[26].

Misplaced conceptual elements: As stated in section 4.2.2, triples consist of subjects, predicates, and objects. Properties should only be used in the predicate position and classes should usually be the only objects of an `rdf:type` property. Therefore, the URIs of classes and properties may be considered as misplaced, if they do not obey these position rules (cf. Hogan et al., 2010). However, it must be stressed that in OWL Full knowledge bases, properties may also be in subject position of a triple. In OWL Full, it depends on the conceptual model whether the appearance of a class or property URI in another position of a triple is a data quality problem or an intended form of meta-modeling.

Violation of datatype syntax: In RDF documents, it is possible to define XML datatypes for literal values. Such datatypes indicate syntactic rules for literal values of such datatype properties without strictly enforcing them (cf. Hogan et al., 2010). E.g. the datatype `xsd:date`[27] requires date values in the syntax YYYY-MM-DD.

Missing language tags: In RDF documents, it is possible to define so called language tags for literal values indicating the language in which the literal is written (Heath & Bizer, 2011). Language tags are especially useful for multilingual support. However, if language tags are not assigned, then automated multiple language support is obviously not possible. Therefore, some applications may assume missing language tags as a data quality problem.

5.2.3 Problems of Data Definitions and Semantics

Undefined conceptual elements: In RDF documents, it is best practice to publish definitions of all conceptual elements, i.e. classes and properties with a formalism like RDFS (Brickley & Guha, 2004) or OWL (Bechhofer et al., 2004; Hitzler et al., 2012), within the data set, so that they are retrievable and reusable on the Web. However, a significant amount of conceptual elements are still undefined in Semantic Web data (cf. Hogan et al., 2010).

[26] See http://www.w3.org/RDF/Validator/ for the W3C RDF Validation service (Last accessed on April 12th 2012)
[27] See http://www.w3.org/TR/xmlschema-2/#date for a full description of the required syntax (Last accessed on July 20th 2014)

Ontology hijacking: Ontology hijacking is "the redefinition […] of external classes/properties" by third parties (Hogan et al., 2010). In other words, conceptual elements of existing ontologies are reused in a way that conflicts with the initial definition, e.g. by adding additional axioms to the URI of the original element that are incompatible with the original meaning.

Ambiguous inverse functional property values: In OWL, the objects of inverse functional properties uniquely identify an individual (Bechhofer et al., 2004). The use of ambiguous values in the object position of inverse functional properties may cause that reasoners assume two or more individuals to be identical, although they are different individuals. Thus, ambiguous functional property values represent a severe data quality problem when reasoning shall be applied (cf. Hogan et al., 2010).

Misuse of `owl:DatatypeProperty` and `owl:ObjectProperty`: Datatype properties usually contain a resource in subject position and a literal value in object position (cf. Bechhofer et al., 2004). Object properties usually relate two resources (cf. Bechhofer et al., 2004). Cases where datatype properties connect resources to each other and object properties contain literal values in subject or object positions may be considered as misuse of these two property types (cf. Hogan et al., 2010). However, it must be stressed that datatype properties with datatype range `xsd:anyURI` may also contain literal values that look like resources (cf. Biron & Malhotra, 2004).

5.2.4 Problems of Data Classification

Imprecise classification: Imprecise classification occurs when instances are not classified to the most specific available class (cf. Lei et al., 2007, p. 139). E.g. `Peter Miller` belongs to the class `foo:Agent` and not to the class `foo:Person`.

Missing classification: Sometimes instances may not be classified at all, i.e. do not belong to a class more specific than `owl:Thing` or `rdfs:Resource` (cf. Lei & Nikolov, 2007; Lei et al., 2007). E.g. the individual `Peter Miller` does not belong to a class, although it should be member of the class `foo:Person`.

Incorrect classification: Instances are incorrectly classified when they belong to a wrong class, i.e. they actually cannot be a member of this class due to their real-world

semantics (cf. Lei & Nikolov, 2007). E.g. the individual `Peter Miller` is member of the class `foo:PopulatedPlace`.

Spurious conceptual elements: Sometimes not all conceptual elements of an ontology are used, i.e. not all classes have instances or not all properties have values. Unused conceptual elements may, therefore, be considered as spurious (cf. Lei et al., 2007, p. 139).

Membership in disjoint classes: With the OWL property `owl:disjointWith` two classes can be connected that do not share the same individuals. Hence, an individual cannot be member of two or more disjoint classes or their subclasses at the same time (cf. Hogan et al., 2010; Lei & Nikolov, 2007).

Membership in deprecated conceptual elements: In OWL, classes and properties may be flagged as deprecated via the classes `owl:DeprecatedClass` and `owl:DeprecatedProperty` when they are shall not be used anymore (Bechhofer et al., 2004). In OWL 2, alternatively the annotation property `owl:deprecated` with value "`true`" annotates deprecated classes and properties (Bao et al., 2012). Hence, the usage of such deprecated conceptual elements may be considered as a quality problem, although it may not be as severe as other quality problems (cf. Hogan et al., 2010).

5.2.5 Problems of Hyperlinks

Dereferencability problems: In Semantic Web environments, it is recommended to use HTTP URIs to represent individuals, properties, and classes in order to be able to look up names and link data (cf. Berners-Lee, 2006). Sometimes the links may not be dereferencable, i.e. we receive an error when looking up the URI on the Web. In most of these cases the target data source of the link address is missing (cf. Hogan et al., 2010).

5.3 Distinct Characteristics of Data Quality in the Semantic Web

There are major differences between data quality in business information systems (BIS) and data quality in open environments such as the Semantic Web. The World Wide Web and the Semantic Web architecture facilitates that anyone that has an internet connection and Web space can publish anything about anything (cf. Berners-Lee, 1998b). In other words, anyone with access to a Web server can publish any data on the Semantic Web, even non-sense data. In opposite to the Web, traditional business information systems usually put control upon the creation and maintenance of data, e.g. via constraints or role and authorization systems to avoid the creation of heterogeneous and willfully conflicting data. These different policies are driven by different needs. While in BIS it may be necessary to establish a common way to create, update, and publish information in order to manage and control business processes, the Web relies on an open architecture to use the creativity and intelligence of the crowd and to serve as an open platform for information exchange (cf. Berners-Lee & Fischetti, 2000). In fact, the large-scale introduction of firm constraints and authorization systems in the Semantic Web would violate freedom of speech and other human rights. Moreover, while large BIS may have a couple of 100.000 users, the Web has most likely several billion users. Thereby, the amount of users also raises the level of heterogeneity. Consequently, the diversity of quality perceptions and data requirements is likely much bigger on the World Wide Web than in BIS. Furthermore, not existing information underlies different interpretations in the Web and in BIS. The Semantic Web assumes an open world, i.e. everything that we do not know is not defined, yet, and, therefore, is neither wrong nor right (cf. Hebeler et al., 2009, p. 103f.). Traditional BIS follow the opposite interpretation, i.e. they close the world and assume that everything that is not represented can be assumed as false (cf. Hebeler et al., 2009, p. 103f.). In other words, a missing instance in BIS would be assumed to not exist, while in the Semantic Web it would be assumed that additional instances may exist, but are currently not member of the class. During the interpretation of data, especially aggregated data, it is important to be aware that knowledge may be incomplete and, therefore, information may be missing. While data quality metrics typically assume a closed world, human interpretation of data quality assessment results can assume an open world, even for traditional BIS, since it is unlikely that all data requirements are known at all times. E.g. an accuracy score of 97 % should be interpreted with special regard to the assumed data requirements. Thus, the score may

be higher or lower, when further knowledge about data requirements is added or different data requirements apply.

However, the Web's openness must be respected by data quality management systems for the Semantic Web, especially with regard to the large diversity of data requirements. But data quality management systems can be a good support to identify and monitor deficient data according to specific quality perspectives and thereby help to improve processing of heterogeneous data for specific tasks, even for the open Semantic Web.