Chapter 5
Open Data Interoperability

"Semantic technologies enable open data interoperability beyond the point of pure format and structure alignment."

5.1 Interoperability in a Highly-Dynamic Open Data Ecosystem

The rapid growth of information technology during the last decade has put governments and businesses alike in front of a number of barriers to overcome in order to tap the full potential of this new digital era. One of the most challenging, but also most potential developments, comes with the web of data (Auer et al., 2007) and the inherent mass of freely-available information, i.e., open data (Zeleti, Ojo, & Curry, 2016). Especially open government data (OGD) holds the power to unlock innovation in both sectors, government and business, regarding the development of new, better, and more cost-effective services for citizens (Zuiderwijk & Janssen, 2014a). This interaction of actors forms a highly-dynamic ecosystem of data (Hammell et al., 2012), yet has to be re-evaluated with the increasing voluntary contribution of data by citizens, e.g., through citizen science initiatives (Lampoltshammer & Scholz, 2016) and open science data initiatives in general (Karmanovskiy, Mouromtsev, Navrotskiy, Pavlov, & Radchenko, 2016). Thus, approaching this ecosystem of open data from a quadruple helix (Carayannis & Rakhmatullin, 2014) approach is the next logical step. Figure 5.1 shows such an extended version of the ecosystem.

1. **Open Government Data** – this refers to data that was collected or produced within the public administration and the public sector in general. However, data affected by legislation, such as data privacy or national security, are not included.
2. **Open Business Data** – this refers to data that was collected or produced within the private sector, e.g., by organizations or companies. Its degree of openness

Y. Charalabidis et al., *The World of Open Data*, Public Administration and Information Technology 28, https://doi.org/10.1007/978-3-319-90850-2_5

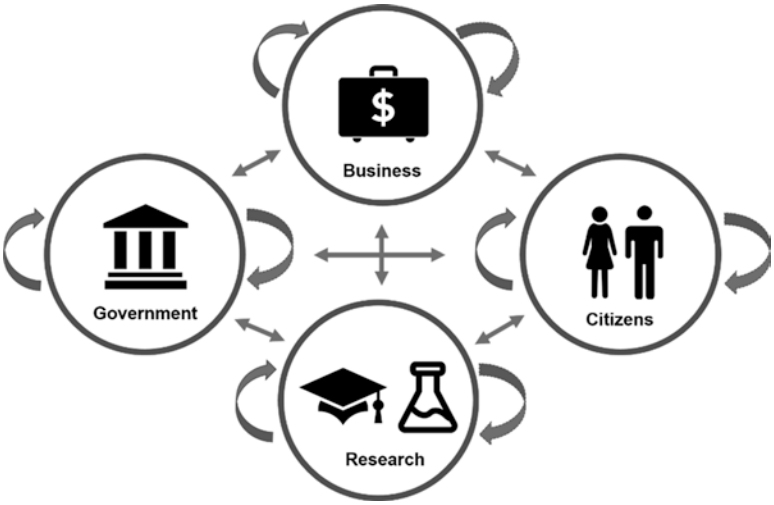

Fig. 5.1 Quadruple Helix-based open data ecosystem

and availability strongly depends on contract-based or sector-specific restrictions, put on the data by their producers.
3. **Open Citizen Data** – this refers to data regarding personal and non-personal related data of individual citizens. Examples can be found in the area of social media platforms or citizen science projects.
4. **Open Research Data** – this refers to data, which was collected or produced within academia and research sectors. It includes, e.g., publications or raw research data originating from interviews or experiments.

Obviously, this ecosystem introduces a certain level of complexity regarding the exchange and therefore also the interoperability of open data from the involved stakeholders. When discussing interoperability of open data, several levels can be distinguished in order to approach this issue via a technology-oriented, holistic way. According to Janssen, Estevez, and Janowski (2014b), the following four main levels of interoperability can be defined:

1. **Technical** – this level refers to a network-based interconnectivity between systems in order to be able to exchange data, e.g., on a per-transaction basis or via real-time streaming. By employing X-as-a-Service (XaaS) approaches, incompatibilities such as different operating systems or programming languages can be resolved.
2. **Syntactic** – this level refers to the use of standards in terms of exchange formats, e.g., XML or JSON, on a web interface level, i.e., for web services to exchange data.
3. **Semantic** – this level refers to reducing ambiguity in terms of data interpretability. This in turn requires semantic technologies and well-defined metadata, e.g., via ontologies.

4. **Pragmatic** – this level refers to quality and trust from an overall organizational perspective, including, e.g., service level agreements (SLAs) or context sensitivity in terms of meaning and involved stakeholders.

While all four levels are important to achieve a holistic approach towards the interoperability of open data, this chapter focusses on two of these levels, the semantic level and the pragmatic level, i.e., linking data as well as metadata and data quality.

5.1.1 A Semantic View on Data Interoperability

The World-Wide-Web (WWW) literally contains billions of pieces of information, spread out over a plethora of websites and information silos. This situation becomes challenging, when we are considering the search and retrieval of particular pieces of information. Thus, this unstructured way of storing information, e.g. as HTML pages, will – on the long run – not be sustainable. To counter this issue, the Linked Data paradigm arose, striving to interlink data on the web, pushing a new way of data handling towards the establishment of a semantically-enabled version of the WWW.

A way of describing this new version of WWW was originally provided by Berners-Lee via his Semantic Web Stack. The stack has become to some degree a blueprint for numerous implementations along the principles of the Semantic Web. Yet, the stack also visualizes the web from a high-level point of view, leaving open some important aspects and technology-related challenges yet to overcome. It is therefore no surprising that the stack has undergone several changes since it was first proposed. Figure 5.2 depicts a contemporary, but not necessarily comprehensive and final version of the stack. To provide a better understanding of the semantic stack, the following part introduces and describes the core layers, together with the core components of the stack (Hogan, 2013):

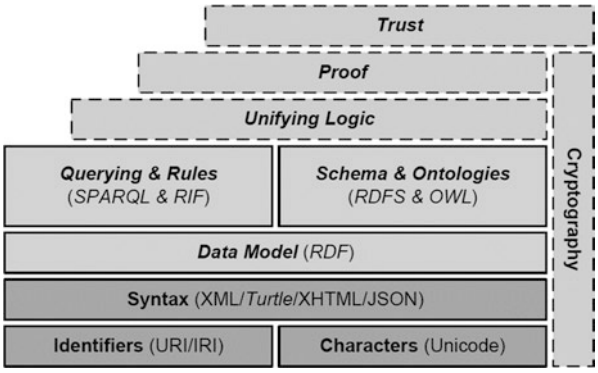

Fig. 5.2 Semantic Web Stack (Hogan, 2013)

The fundament of the stack is comprised of two elements. The first element is represented by mapping streams of data and external storage to actual textual information via the utilization of **characters** out of the Unicode char-set. The second element presents the ability to provide unique **identifiers**, which is imperative, considering the requirement for search, retrieval, and interlinking of resources in a machine-comprehensible manner. For the realization of provision of identifiers, the original stack foresaw the application of the Uniform Resource Identifier (URI), while current implementations shift towards a more general and flexible representation via the Internationalized Resource Identifier (IRI), based on Unicode. The next layer focusses on **syntactical** aspects, in particular, the provision of automatically parse-able elements, i.e., a common syntax in form of XML and JSON. While these classical forms are widely-adopted, custom syntaxes via, e.g. the TURTLE syntax (associated to RDF), are also possible.

On top of the syntax layer resides the **data model**. To provide the necessary means of data exchange, a common and machine-readable data model must be defined. This data model needs to be generic in that sense that it allows for the adoption of any content, originating from any given domain, while at the same time it must be usable without the need of proprietary technology. During the design of the Semantic Web, the Resource Description Framework (RDF) (Pan, 2009) has been chosen to serve as core data model.

Within the next layer, two components reside, which are required to introduce semantics into the Semantic Web. As RDF is only handling the structure of the content, but adds no semantic description to it, a formal way of additive modification to the existing model must be provided. This modification comes in form of formal languages, including meta vocabulary. The two basic variants contained within the stack are either the RDF Schema (RDFS) (McBride, 2004) or the Web Ontology Language (OWL) (Horrocks, Patel-Schneider, & Van Harmelen, 2003).

As it is the entire purpose of Linked Data to increase access and availability of data, there must be a way to search for these data by formulating **queries**, filters, and to design and apply search patterns in order to be able to identify data, as well as associated data, of interest. To realize this functionality, complementary to RDF, the SPARQL Protocol and RDF Query Language (SPARQL) (Quilitz & Leser, 2008) developed. In order to also be able to define certain sets of rules, the Semantic Web currently builds on the Rule Interchange Format (RIF) (Kifer, 2008), which covers numerous rule-based languages and therefore provides a high level of flexibility and compatibility in terms of different stack implementations.

For following layers on top, as well as the vertically-reaching layer, an increasing amount of technologies emerges to handle associated issues and tasks within these elements. Yet, there is no defined standard available so far. The **unifying logic** layer strives to provide an overarching compatibility, unifying all query languages and knowledgebases via the application of a comprehensive and unifying language. While there have been several research works addressing these challenges (Gyawali, Shimorina, Gardent, Cruz-Lara, & Mahfoudh, 2017; Krötzsch, Maier, Krisnadhi, & Hitzler, 2011; Polleres, 2007; Straccia & Bobillo, 2017) none of them was able to achieve a “one size fits all” solution up till now. The concept of a layer of **proof** is

dedicated to the idea that the combination of various and externally-hosted data sets is a complex process and therefore has to provide some way of re-assurance for potential users of a stack implementation. This also holds true regarding applied reasoning processes, filters, or task completion. The **trust** layer is directly-connected to the layer of proof. Potential users or machine clients should be able to evaluate, if and to what degree they are able to trust certain agents providing data as well as resources and results, based on issued queries. Classical approaches use white-listing or black-listing, which in turn triggers the question, who is going to be responsible for maintaining these lists and therefore keeping them up-to-date. This again would push the issue of a central authority, which to some degree might compromise the entire idea of a distributed resource network. Finally, the **cryptography** layer is envisioned to integrate security and controlled access as cross-cutting concern throughout the entire stack. Aspects to be covered by this layer include the possibility to establish encrypted connections via secure protocols or the application of crypto algorithms such as RSA or AES to guaranty protection and privacy of data, information, the requests and search queries respectively. Furthermore, the layer also provides means of controlling, who can find, query, and finally access linked resources.

5.1.2 A Schema View on Data Interoperability

Besides approaching the topic of data interoperability from a semantic point of view, one can also refer to it through an architectural point of view, expressed by metadata schemata. Zuiderwijk, Jeffery, and Janssen (2012a) suggested the following three layer-based metadata architecture approach, as shown in Fig. 5.3.

The first layer enables to initiate queries for Linked Open Data, while the second layer provides enriched information regarding the dataset of interest, such as

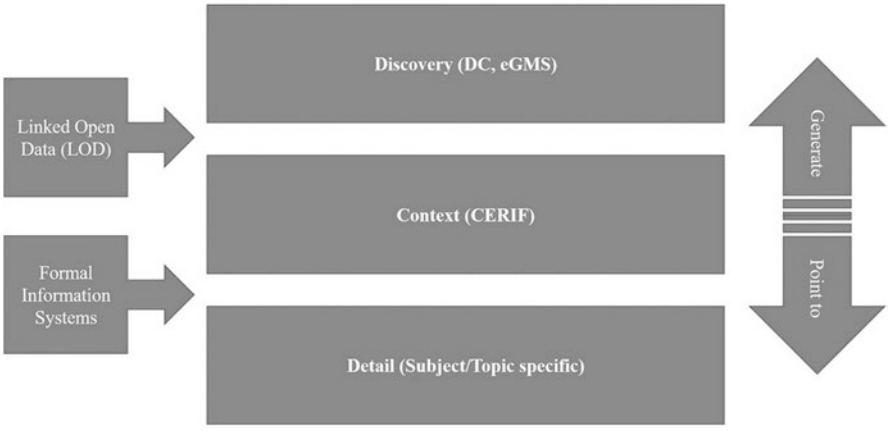

Fig. 5.3 Three layer-based metadata architecture. (Adapted from Zuiderwijk et al. (2012a))

involved persons, organizations, publications etc. At the same times, this layer is also responsible for the identification and generation of common metadata information to achieve a high-level of congruence. The third layer features metadata information which is specific to a domain, such as the Infrastructure for Spatial Information in the European Community (INSPIRE) (Directive, 2007). Within the first layer, several types of metadata standard descriptions can be applied, such as Dublin Core (DC)[1], the e-Government Metadata Standard (e-GMS)[2], or the Comprehensive Knowledge Archive Network (CKAN)[3]. The level of reduced complexity in these standards allow for an eased mapping process. Yet, this comes at a cost, namely, the used vocabulary not meeting necessarily the real-world demands, and compromises have to be made, which could after all results in poor query results or datasets not being discovered at all. It is due to this reason, why the second layer incorporates a layer of contextual metadata, expressed by the use of CERIF[4]. By doing so, the establishment of relationships between entities becomes possible. In addition, CERIF is the recommended metadata standard by the EC to be used by its Member States. Finally, the third layer allows for the attachment of highly-specific metadata, e.g., information about the domain, in-depth descriptions of the actual data, about the data collection process, etc. It is due to their important task of providing interoperability that metadata schemata play a significant role within the process of setting up a data infrastructure. For more information regarding data infrastructures, please refer to Chap. 6.

5.2 The Data Life-Cycle Within the Semantic Web

According to Auer, Lehmann, Ngomo, and Zaveri (2013), the following steps are required to form a complete data life-cycle (see Fig. 5.4) in the domain of Linked Data. It has to be noted though that while the cycle forms a kind of sequential order of steps, these steps may also occur in different combinations, depending on the current status of the resources under observations.

To begin with, any unstructured representation in form of, e.g., data sets have to be transformed in order to be compatible and map-able via the RDF data model (EXTRACTION). This process continuous until a critical mass of RDF-based data has been accumulated. In the next step, it is then necessary to not only provide sufficient storage for the collected data, but to provide features such as indexing and the possibility to formulated and apply search queries on to the data as well (STORAGE & QUERY). While current systems are already capable of interlinking data semi- or even fully automatically (LINKING), based on defined criteria and attributed features within data sets, it is essential that manual link creation as well as the possibil-

[1] http://www.dublincore.org/

[2] http://www.agls.gov.au/links/

[3] https://ckan.org/

[4] https://www.eurocris.org/cerif/main-features-cerif

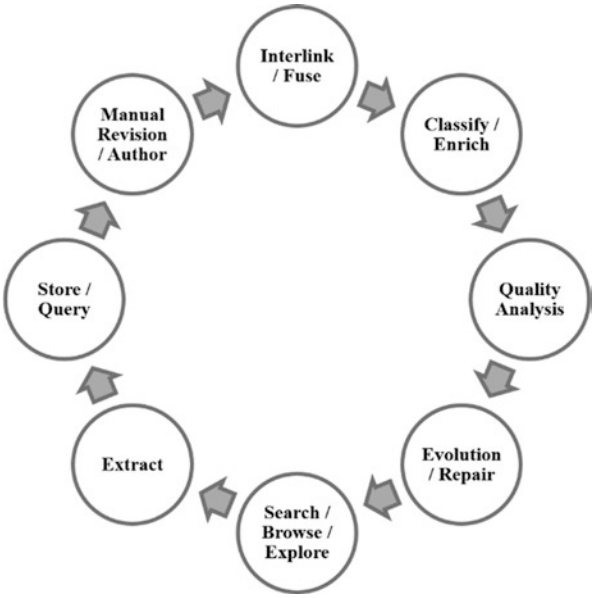

Fig. 5.4 Linked data life-cycle. (Adapted from Auer (2011))

ity to modify existing links is provided to further improve and refine the growing network between the data resources (AUTHORING). Yet, linking existing data sets and resources is not enough. Theses established links are per se not revealing any additional information regarding the classification of data sets or resources, nor are they providing knowledge about inherent structure as well as associated schemata. Therefore, the enrichment of data with high-level information and semantics is imperative (ENRICHMENT), to be able to increase the level of efficiency regarding aggregation and, in turn, towards searching and querying the growing semantic network. While identification and retrievability of data sets and resources is important, the results as such do not provide any information regarding the actual quality of the data or the associated metadata. Therefore, functionalities and services must be established to analyze the linked data and to identify potential errors or missing pieces of information within these data sets. Hitherto, for the services to work effectively, they require a well-defined set of quality metrics, describing what the term data quality implies for the given type of data (QUALITY ANALYSIS) – a detailed overview of such metrics can be found in Chap. 8. Once open issues are identified, smart algorithms can then be applied to correct these errors or, in some cases, even to reconstruct missing data pieces and therefore information (EVOLUTION & REPAIR). The last step then covers the usability of the entire system and Linked Data network by potential users (SEARCH, BROWSING & EXPLORATION). The best and most refined data corpus is of no use, if users are not able to efficiently browse through the data structure, intuitively formulate questions in form of queries and patterns, as well as to retrieve the desired information. Furthermore, smart

systems will not only detect results that match user queries 1:1, but also allow for a certain form of fuzzy queries, providing users with potentially interesting alternative search paths and therefore leveraging the full potential of Linked Data.

As the presented cycle is of iterative nature, it is per se never completed and thus continuously leads to the improvement of Linked Data and in the long run, offers several benefits such as (Auer et al., 2013):

- *Uniformity*: as all data sets have undergone the transformation process from non/semi-structured data towards structure data into the RDF data model, the benefits of the RDF structure can be exploited. As all facts within this data model are formulated as triples formed by subjects, predicates, and objects, theses directly correspond to the applied unique identifiers (i.e., URI/IRI) and therefore reduce ambiguity.
- *De*-referenceability: via the application of the afore-mentioned unique identifiers, entities within data sets cannot only be precisely defined, but at the same time, serve as links between resources on the web, similar to URLs used to navigate between HTTP resources.
- *Coherence*: the core data model RDF supports the use of so-called namespaces. These namespaces allow for multiple use of identifiers without causing conflicts in terms of ambiguity. For example, the subject-predicate-object structure allows the establishment of links of entities between different namespaces via their URIs.
- *Integrability*: as the RDF data model provides uniformity across all transformed data sets, it becomes possible to build upon this unified structure to attach additional schema information or semantics in terms of ontologies. By doing so, the level of expressiveness of queries and answers can be significantly increased, which in turn enables and improves a more sophisticated matching process.
- *Timeliness*: the underlying process of publishing Linked Data is, due to the existing tools and technologies, relatively straightforward. In addition, once a linked data set has been updated, the process of accessing the newly-added information is easier, compared with the alternative way involving complex procedures in course of ETL (extract, transform, load) task.

An in-depth discussion regarding the single steps of the cycle, including the required tools and methods can be found in Chap. 2, paired with a comprehensive overview of different use-cases of the data life-cycle.

5.3 Ontologies as Means of Providing Semantics

The term “ontology” takes different meanings throughout different disciplines. Approaching the origin of this term from a philosophical point of view – the “big O” ontology – it can be described as a set of types and associated structures of objects, combined with properties, processes, all in relation towards every aspect of reality (Smith, 2003). Within the domain of computer science, one of the most

referenced definition is provided by Gruber (1995), who sees ontologies as a formal way to explicitly specify a conceptualization and share it with others as a simplified representation of the real world for a specific purpose.

Ontologies haven been applied in a variety of application domains, such as the automated generation of user interfaces based on Linked Data (Hitz, Kessel, & Pfisterer, 2017), the detection of discriminatory language (Salguero & Espinilla, 2018), the classification of objects in satellite imagery (Lampoltshammer & Wiegand, 2015), the implementation of content management systems in the field of curricula development (Olteanu, Ionita, & Solomon, 2017), for the purpose of requirements engineering (Dermeval et al., 2016), as well as for data management in general (Daraio et al., 2016). Yet, this plethora of potential application domains also comes along with some drawbacks. Firstly, one of the most significant issues during the design and development of ontologies can be found in the so-called "semantic gap" (Smeulders, Worring, Santini, Gupta, & Jain, 2000). This term describes the difficult situation of providing detailed and concise description of visual interpretations. Although this example is strongly-related to the image interpretation domain, it well exemplifies the challenge of formalizing an objective view on reality, which is discussed in philosophy since decades, also known as the paradigm of "constructivism" (Jonassen, 1991). Besides this hurdle, ontology design and development suffer from the same issue, already known from knowledge modelling, such as overfitting (Hawkins, 2004). Overfitting occurs, if the knowledge model includes more features than necessary to describe a certain concept properly. This situation can arise, if the data set, which is used for the modelling, contains attributes and features, which are not representative for the kind of data at hand, but are present, e.g., due to errors within the actual data.

Yet, not only the process of designing and modelling of ontologies is a challenging task, the process of integrating and joining ontologies on different levels within one domain, or across domains, generates pitfalls as well. In addition to the beforementioned challenges, the following problems have also to be considered (Zhao & Ichise, 2014):

- **Ontology heterogeneity problem:** As data sets are published within a Linked Data environment, one part of the publishing process is to interlink these newly published data sets with already existing data sets. Yet, there is no existing "jack of all trades" ontology, meaning that the controlled vocabulary is nowhere close to completely cover all aspects of the interlinked data sets at once. Amongst other dimensions, two particular aspects increase the level of difficulty during the integration process. The first aspect addresses terminological issues. For example, one particular entity is modelled and described differently between ontologies foreseen to be integrated ("startingDate" vs. "beginningDate"). The second aspect focusses on conceptual issues, namely, entities differ in their hierarchical position within the ontology, as they were modelled in each of the ontologies as children of different parents, and therefore originate from different core concepts.

- **Identification of core ontology entities:** real-world entities based on the class descriptions including their attributes and properties within an ontology are called individuals. If the ontology and the included instances are of high volume, the identification of essential core properties of a specific class becomes increasingly difficult. To tackle this issue, the observation and notation of commonly-used core classes can support developers in their task to describe instances of particular data resources. Via these core entities and their associated attributes and properties, it becomes possible to design and construct suitable SPARQL queries, closing the gap regarding missing pieces of information within data sets.
- **Missing domain or range information:** the underlying relation between classes and properties within an ontology is expressed via domain information in the RDF core data model. This information describes the suitability of properties to be used for instances of certain classes. In addition, range information, also within the RDF core data model, help to better comprehend data sets in terms of the included values. Yet, in a real-world environment, ontologies are often missing this crucial information regarding domain and range, which in turn renders the process of integrating different ontologies based on their classes and properties more difficult.

The research community currently works towards potential solutions to the aforementioned challenges. For example, Lampoltshammer and Heistracher (2014) proposed a workflow for classification of data instances with use of a dedicated plugin for the ontology modelling environment Protégé (Gennari et al., 2003), called OWLET. This plugin enables ontology modelers to import external data as instances into their ontology model for classification of these data items. Furthermore, the suggest approach can also be used for testing newly design ontologies, by using gold-standard test data and evaluating the classification results as well as the level of coverage regarding the included classes as well as associated properties. In addition, this evaluation approach enables designers to also verify the existing range and domain information, which is an essential step towards lowering the bar of integrating other existing domain ontologies.

Another research work comes in form of the Framework for InTegrating Ontologies (FITON) by Zhao and Ichise (2014). It also addresses the heterogeneity issue, as well as the difficult task regarding the identification of core entities as well as to provide the crucial information considering domain and range for ontology properties. The authors achieve this via the combination of three approaches (see Fig. 5.5):

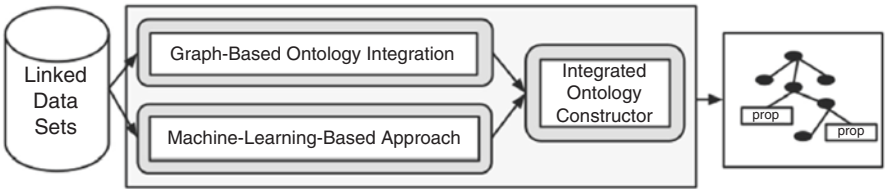

Fig. 5.5 Core components of FITON. (Adapted from Zhao and Ichise (2014))

- Step 1 – Ontology Similarity Matching on the SameAs Graph Pattern: during the process of integrating ontologies, 2:n ontologies are merged to deliver one unified model. Yet, in cases of small numbers of links regarding classes or properties, alignment becomes a challenging task. The authors therefore apply a WordNet-based (Pedersen, Patwardhan, & Michelizzi, 2004) approach, to establish undirected graphs between linked instances, which in turn provides valuable information regarding forming patterns between concepts over different data resources. These patterns can then be used to identify matching concepts to foster and speed-up the overall integration process.
- Step 2 – Machine Learning for Core Ontology Entity Extraction: to identify core entities within a given ontology, the authors apply machine learning algorithms. These algorithms comprise different approaches, starting out from rule-based classification via a priori knowledge, up to learning entirely new rules based on a data-driven approach.
- Step 3 – Automatic Ontology Enrichment: to be able to comprehend and understand the relationships between entities in the ontologies of observation, the domain and range information has to be seen crucial. Consequently, it is the next logical step to include this information during the integration process. The authors therefore take random samples out of the entire set of instances within the ontology and analyze their range and domain information via inspecting the associated properties and values. These results, paired with available standard range and domain information, is then used for annotating the resulting integrated ontology.

Considering the before-discussed complexity and depth of creating and maintaining linked data sets, the results will only be as good as the quality of the provided (meta) data, used to construct the actual links between the data sets. If the overall (meta) data quality is poor, linking of data sets may be not possible or might end up in erroneous links. Therefore, the next section will discuss the importance of quality aspects of Open Data and means to assess and evaluate quality of (meta) data.

5.4 Quality Aspects of Open Data

The overall quality of data sets is of upmost importance for several reasons. One reason is that without proper meta data and data quality, it is hard for experts to design and construct suitable ontologies for the domain the data set belongs to, due to missing information. Furthermore, this missing information, paired with potential errors within the data and the meta description itself can lead to false classification and therefore false linking or even no linking at all, as no common denominator as basis for the linking process could be identified.

The study conducted by Vetrò et al. (2016) identified several generic issues that can negatively affect the quality of Open Data (see Table 5.1). The first

Table 5.1 Potential data quality issues in open data sets

Incomplete data	Format not compliant to well-known standards	Lack of data source traceability
Incongruent data	Out-of-date data	Lack of metadata
Errors	High time to understand data	Lack of modification traceability

Adapted Vetrò et al. (2016)

issue is related to the data being **incomplete**. This leads to the metadata not matching, e.g., the time range of the actual data, which in return would deliver no matching data to search results of users. In addition, with the data being incomplete, analyses on this data is prone to produce wrong or misleading results.

The second issue comes in form of the actual data format **not being compliant** to well-known standards. This can cause problems from several directions. On the one hand side, automated data extraction, transformation, and loading (ETL) processes become difficult, if not impossible, due to the data not adhering to known and well-define structures and schemata. On the other side, the data as such might require special software to work and to incorporate them into existing data infrastructure and therefore acts as impediment for adopting the data. This manifests itself through additional costs for users as well as potential issues for long-time preservation of data, as proprietary software might not be available in the future. The third issue is present through the **lack of traceability** regarding the origin of the data at hand. This is not only a problem regarding potential licensing issues, but also in terms of contacting the original author(s) of the data, in case errors or gaps in the data have been identified and could be reported back to fix these. The next issue comes along in terms of **incongruent data**. This problem usually arises when data is merged, and the particular data set was not aligned to use the same format or schema. Thus, data items can have mixed data representations such as different date formats (Linux timestamp vs. date-time format). In consequence, filtering and/or sorting of data, as well as providing statistics regarding the actual content of data set becomes burdensome and only possible, after an additional step of type conversion. Next issue on the list is present by the data being **out-of-date**. An example would be a data set containing scheduling information regarding a certain type of public transportation, e.g., bus lines. Such public transportation information often changes slightly from 1 year to the next, thus, if the data set called “bus schedule Vienna” is not updated accordingly, this leads to issues regarding the use of this data in, for instance, customer apps for public transportation. Further issues are present in the lack of metadata. In cases, where **no metadata** is available at all, mapping and interconnecting of data becomes only possible, after going through the data themselves, which can be a time-consuming and costly operation. Also, an assessment regarding schema or format compliance, as well as the application of other metrics is not straightforward, same goes for indexation of datasets. Another common issue is found in **errors** directly within

the data themselves, or within the associated metadata. Of course, if the data at hand are incorrect, analyses of these data will produce erroneous results as well. An often neglected but still important issue comes with a **high time to understand the data**. While the data themselves can be complex, the understanding of them can be eased via meaningful descriptions and annotations by a complete set of metadata. If this description is missing, it is sometimes not even possible to determine, what the data is about, what is their range, and what details are included in the data set at hand. Finally, there is the issue that comes along with a **lack of modification traceability**. While the origin of the data as well as their producer can probably be determined via the associated metadata, changes within the data are not obvious. If not provided with a set of history or changelog, detecting modifications, additions, or removal of a single datum or even complete sequences of data are impossible. Thus, manipulation or unintended data loss cannot be detected or proven.

As all of these issues can fairly impact the usability and adoptability of open data, numerous research projects are focusing on assessing the quality of open data via the introduction of metrics as well as approaches to fix some of the identified issues automatically or at least provide support during the manual process of data cleaning and repair. Thus, the next section provides an overview of ongoing activities in that regard.

5.5 Quality Assessment and Improvement of Open Data

To identify suitable data sets for a particular application, their quality has to be assessed first. This assessment is usually performed via the use of so-called data quality dimensions and associated metrics (for an in-depth discussion see Chap. 8). According to Heinrich, Kaiser, and Klier (2007), well-defined metrics should match the following criteria:

1. **Measurability** – being defined quantitatively, normalized, at least interval-scaled
2. **Interpretability** – specific focus to increase comprehensibility
3. **Aggregation** – quantification on attribute level, while keeping semantic consistency across all levels, to enable cross-level aggregation
4. **Feasibility** – clearly defined input parameters, while at the same time providing a high level of automation

Alongside these basic preconditions, researchers have developed various approaches regarding the assessment of data quality. The work by Borovina Josko and Ferreira (2017) presents a case study regarding the use of visualization approaches to enable data quality assessment to identify defects in the structure of the observed data. Debattista, Auer, and Lange (2016) introduced the Luzzu framework as a generic approach to assess the quality of linked open data. Luzzu consists out of four main components, namely a flexible interface to enrich the

framework with new assessment metrics if required, an ontology-driven backend regarding metadata quality representation, a scale-able stream processor as endpoint for, e.g., SPARQL endpoints, and a user-defined ranking algorithm. Kontokostas et al. (2014a) adopted the idea of test-driven evaluations out of the software engineering domain into the task of assessing the quality of Linked Open Data. The authors leverage a large collection of test patterns, derived from SPARQL queries to conduct their test runs. Acosta et al. (2018) applied an innovative solution towards the quality assessment of Linked Data via a crowdsourcing approach. Crowdsourcing in this case means that a large group/network of people, which are not pre-defined, are working towards a common task or goal. Crowdsourcing has established itself in many different areas, starting from microtask working (e.g., Amazon Mechanical Turk), to funding projects of common interests (e.g., Kickstarter). Usually, the tasks put towards the crowd are single-iteration based, yet there are also approaches building on multiple iteration to assess and evaluate the results from the crowd by the crowd itself. Acosta et al. describe three main ways of crowdsourcing on a given task:

Contest-based Crowdsourcing follows the idea of handing over a particular task or problem to solve to the crowd and in consequence to reward the best, most efficient, most effective, or most innovative solution (Leimeister, Huber, Bretschneider, & Krcmar, 2009). The approach leverages on the exploitation of intrinsic motivational factors, triggered by competition and intellectual stimuli. The contests are usually held open for an extended period of time – depending on the complexity of the task – to allow for enough time to submit a solution to the described problem. While there are several ways of stating a reward to the best solution, usually a main prize is provided by the entity that issues the challenge. While these challenges have been around for years to attract experts to work on a given problem, they are increasingly used towards working with citizens as well, and in consequence, also contribute towards the entire citizen science movement (Lampoltshammer & Scholz, 2016).

Microtask Crowdsourcing applies the approach of splitting a given problem into chunks, thus called microtasks (Howe, 2006). This approach works best if the abilities for solving these microtasks are either based on basic audio or visual comprehension, or towards the understanding and interpretation of language-related issues, rather than towards the necessity of a priori expertise in the related topic. In order to be handled in an efficient way, microtask crowdsourcing requires a high level of parallelization, and in consequence a large number of participants. Thus, this decentralized method results in faster responses, in conjunction with the possibility to validate the proposed solutions to the posed problem based on, e.g., majority voting or other consent-finding methodologies. Typical awards issued for successfully solving microtasks are provided in micropayments.

Crowdsourcing Pattern Find-Fix-Verify (Bernstein et al., 2015) similar to the microtask crowdsourcing splits a more complex task into a set of tasks of less complexity, which are then processed throughout three consecutive stages. In the

first stake, the individuals within the crowd are to **find** data, which is of interest to solve the given task. In the following second stage, the outcomes of the first stage are corrected/amended (**fixed**) if required to match the given task in a better way. Then, in the third stage, the final results are verified one last time to conclude the overall quality assessment. This pattern does not only exploit the benefits of the before-described microtask, but also gains within each step of the negotiation process between all involved crowd members. Furthermore, alongside the three different stages, different compositions of crowds can be used to even more increase the likelihood of high quality output.

As discussed before, not only the linking of data supports interoperability, data quality does as well. Regarding the later, promising approaches have been found in regarding to the assessment of data quality via metrics as well as via leveraging the knowledge and the abilities of the crowd. From the given point of view, it is the next logical step to combine these two approaches to make use of advantage of both side in a synergistic way. The following section therefore presents two research projects and initiatives, which also build heavily upon the crowdsourcing aspect for the identification of data issues, paired with automated assessment and correction abilities for data quality and thus going towards the improvement of open data interoperability.

5.5.1 ADEQUATe Project

The ADEQUATe project was initiated to develop innovative approaches towards the measurement, monitoring, and improvement of date quality and to demonstrate these concepts via two pilot use-cases in Austria, i.e., data.gv.at and opendataportal.at (see Fig. 5.6). To achieve this ambitious goal, the project tackles the four main issues identified during its initial requirements elicitation phase (Höchtl & Lampoltshammer, 2016):

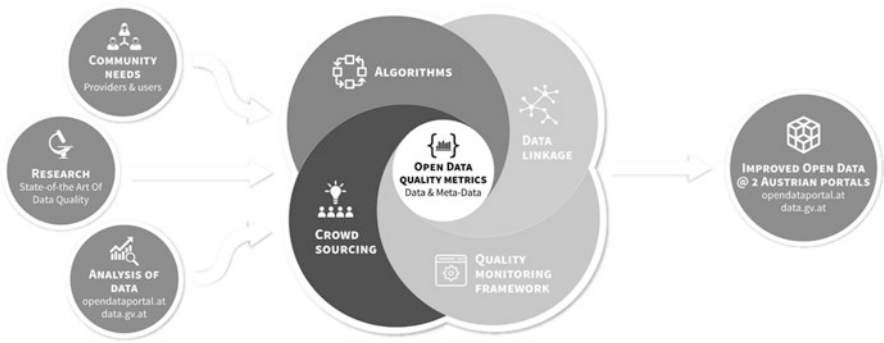

Fig. 5.6 The overall conceptual model of the ADEQUATe project. (https://www.adequate.at/)

1. **Issue – Defining suitable quality metrics targeted for open data**: as already discussed in the sections before, there do exist numerous metrics to assess data quality. Yet, often they do lack, besides still fulfilling the basic criteria of well-define metrics, the specific characteristics required by open data as well as the target platform and audience. Furthermore, applying all available metrics to a given data set may introduce an unjustified bias by falsifying the assessment results due to, e.g., important metadata fields missing, which results in a reduction of the overall quality score of the assessed dataset.
2. **Issue – Providing (semi-) automated improvement of metadata and data quality**: while identifying issues regarding metadata and the data as such is one aspect, the overall big picture would be incomplete without considering the automated correction of potential issues as well as further improvements towards the dataset and its associated metadata. Yet, this part is challenging in particular, as the algorithm itself has to decide what to change in order to improve the overall quality scoring. At the same time, improvements expressed by quality metrics do not necessarily reflect the possible introduction of content-wise errors by the system.
3. **Issue – Coping with CSV-based data sets**: one of the biggest challenges within the existing datasets of the two pilot portals are represented by data in the CSV format, as these data present the majority of datasets on the portals at this point in time. CSV files are known for their issues regarding proprietary formats, such as delimiters (depending of their source language. e.g., German vs. English), nested tables, or non-present metadata.
4. **Issue – Foster open data community engagement**: while algorithms may assess and correct potential errors within data, without the continuous feedback and expertise of the community, i.e., the end-users of the data, data providers, as well as service provider, building their services on top of the existing open data, no sustainable development can be realized.

To deal with these four main challenges, the ADEQUATe project combines community-driven solutions with state-of-the-art technologies in the domains of data quality assessment, correction, as well as monitoring. In a first step, the project continuously monitors the quality of open data being published at the two uses-cases, namely data.gv.at and opendataportal.at. This is achieved via a set of well-defined dimensions and metrics, specifically designed to match the data within the two data portals being observed. In the next step, data quality algorithms are applied to (semi)-automatically correct identified issues within the observed (meta)data. In addition, the ADEQUATe platform provides a community component, based on the well-established technology git, to fork data sets of interest and to resubmit fixed and/or enhanced versions of this particular data set. Furthermore, these suggested changes can then be discussed with other members of the open data community, making full use of the intended crowdsourcing approach. Finally, the semantic enrichment component of ADEQUATE, based on tools such as Odalic (Knap, 2017), tackles the open issue of existing legacy data and transforms them into Linked Data.

5.5.2 *Openlaws*

The linking of data provides increased access, transparency, and availability of information. This fact does not only hold true within the business and research domain, but also for public administration, which have an obligation and responsibility towards their citizens. In case of public administrations and governments, the distribution, availability and access towards legal information is imperative. Yet, there exist some severe issues at the moment regarding this access. One of them is found in form of available APIs, which are not always up and running on a 24/7 basis, paired with slow systems and often non-compliant data towards standard or even self-issued schemata. This in turn makes the use of automated crawling and analysis more than difficult. Translating this situation into a cross-border context, the problem becomes even bigger, as each member state within the European Union are providing their open legal data in different formats, often with metadata in their own language (e.g. the Netherlands) and not towards a better understanding in a common language such as English. To overcome these issues, the EU research project openlaws[5] and the resulting spin-off are built around three core pillars, namely open legal data, open source software, and open innovation towards the establishment of Open Justice in Europe through open access to legal information (Lampoltshammer, Guadamuz, Wass, & Heistracher, 2017). The project's main goal is to increase the level of access towards legal information by supporting users in organizing and sharing their respective information (Wass et al., 2013). Nowadays, a small number of organizations and companies sign responsible for publishing and distributing legal information. Yet, this distribution occurs in somewhat restrictive and non-transparent ways, e.g., through public governance bodies or through public-private-partnerships with certain established publishing houses. Due to this fact, the important access to metadata of legal data is also restricted, which hinders automated processing of these data. Within this often-commercialized ecosystem, legal experts publish their research work and knowledge, with little to none free information flow towards the public and wider research community. This stands in sharp contrast to other research areas, where open research data and knowledge is shared increasingly.

Openlaws tries to break this restricted circle and therefore supports citizens in accessing, working with, and finally understanding legal information and in consequence, their rights and responsibilities towards the state and society. But not only citizens can profit from the project's outcomes, companies and organizations do as well. Supporting them with the required information and knowledge regarding necessary legal compliance according to their field of business, the experts within these organizations and companies can contribute to the sustainability of their business model as well as demonstrate proficiency towards their customers and clients. In comparison to the existing environment, the newly established platform is all-inclusive, meaning that publishing house can as well offer and integrate their premium content, enriching the data at hand even more.

[5] https://openlaws.com/

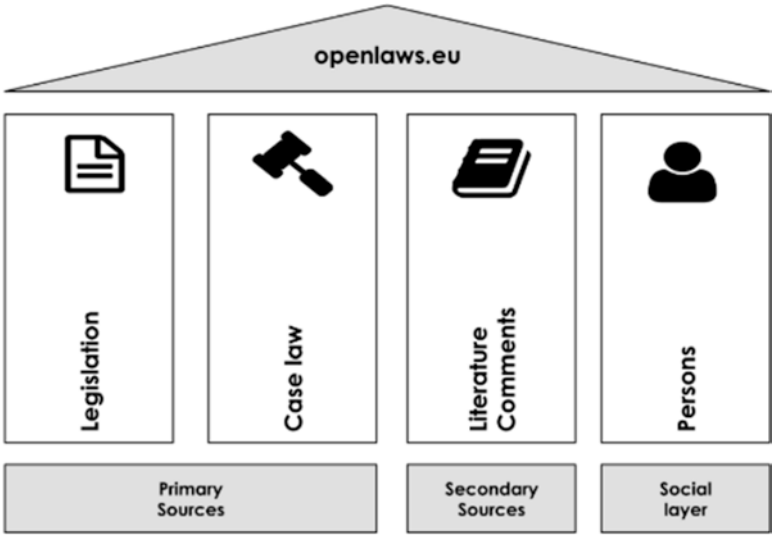

Fig. 5.7 Core components of the openlaws platform (Lampoltshammer et al., 2017)

Finally, public bodies and governments can push more than ever open legal information towards the community, following the idea and legal context of the public-sector information (PSI) directive.

To achieve this ambitious goal, the project provides the following services to its users, based on the core components shown in Fig. 5.7:

- The possibility to conduct a meta-search across several national legal databases and therefore provides cross-border and also cross-language access to legal information
- The amount of legal information is increased, providing additional possibilities for legal scholars and researchers to distribute their work, in direct context with the legal basis their working on and the audience there are targeting, who is affected respectively.
- An improvement of legal data and information quality, as experts can evaluate and curate the data within the platform, as well as the hosted publications in a new way of peer-review
- The existing network of legal scholars, experts, and practitioners is further extended and is also made available and searchable for citizens
- Finally, the access to, e.g., case law can provide a better understanding of laws, regulations and associated consequences for all affected stakeholders. Thus, the availability of open legal data and therefore the derived open legal information contributes towards better democracy and policy-making in the long run

To provide these services, the openlaws platform builds upon existing open data sources across the Union, such as national legal databases and EUR-Lex. These information are aggregated into Big Open Legal Database (BOLDbase), based upon

an innovative graph database approach (Lampoltshammer, Sageder, & Heistracher, 2015). This new way of interlinking previously disconnected open legal data generates a new way of working with and providing legal information for all interested stakeholders. In addition, while experts and citizens are interacting on the platform with each other and with the legal data in openlaws, the platform makes full use of these interaction via integrated analytics, e.g., creating recommendations for individuals in regard to potentially-interesting legal information as well as additional benefits such as automated update services to broadcast important changes within legal domain of particular interest for each individual user.

5.6 Conclusion

Open data interoperability is imperative to drive the movement of Linked Open Data and therefore to increase not only the level of discovery and accessibility of data, but also the possibility to fuse data in order to create new application scenarios. These application scenarios can cover various stakeholders in a transdisciplinary way, including businesses, academia, public administrations, and citizens alike. Data interoperability is also the key for the exchange of data in different types of infrastructure (see Chap. 6), which can be seen as the key to enable the vision of the European Commission regarding the Digital Single Market. But interoperability is not only expressed by the application of common data formats and standards, the overall quality of the data itself and the associated metadata is also important, as these factors do not only impact processing of the data but usability of the data in general (see Chap. 8 for more about quality metrics and overall assessment). Overall, it can be stated that although the barriers of open data adoption have been known for a while, the "golden solution" is still missing to fulfil the high expectations that were expressed when the Public-Sector Information Directive (PSI) was put into place. Interdisciplinary research projects such as the openlaws project and the ADEQUATe project are an important step forward to increase accessibility of open data, focusing especially on data quality, as well as the semantic linkage of data to increase awareness on the one hand, but also adoption of available data on the other hand. The second aspect is crucial, if sustainable data-driven business models (see Chap. 7) shall push the European Union back into the international "game of data".