Methodology for Linked Enterprise Data Quality Assessment Through Information Visualizations

Didem Gürdür[1], Jad El-khoury[1], Mattias Nyberg[2]

[1]*Department of Machine Design, KTH Royal Institute of Technology, Stockholm, Sweden*
[2]*Scania CV AB, Södertälje, Sweden*

Abstract

Today's development environments in the manufacturing industry require different development tools to work together. These complex environments are highly heterogeneous and constantly changing, and the development tools are producing a huge amount of data. As a result, these development environments must overcome a significant problem related to data integration. In this paper, we examine a case study from the automotive industry using the linked enterprise data approach to integrate data from different development tools. The study explains and applies a data quality assessment methodology as a post-integration phase for linked enterprise data.

In this study, important data quality dimensions from the literature are merged with empirical rules that have been defined by Scania CV AB employees. As a result, a comprehensive methodology is developed and introduced to assess these data quality dimensions. This paper presents the methodology, which aims to develop a data quality assessment tool—a dashboard—in addition to policies and protocols to manage data quality. Moreover, the proposed methodology includes systematic guidelines for planning the data quality assessment activity, extracting requirements for the data quality management, setting priorities to expedite the adaptation, identifying dimensions and metrics to ease the understanding, and visualizing these dimensions and metrics to assess the overall data quality.

Keywords: data quality, linked data, quality assessment, linked enterprise data, information visualization, methodology.

1. Introduction

In modern manufacturing environments, many engineering software tools are used, and a vast amount of information is produced throughout the product lifecycle – from requirements to product design, testing, quality control, and so on. Since this information is produced by different software tools, it is highly heterogeneous given that it is presumably managed using different data formats and structures. The software variability in manufacturing ecosystems is not necessarily a downside, though. In fact, having variability in software tools is an approach to safeguard against vendor lock-in, helping enterprises avoid dependence on only a limited vendors for software products and services. However, heterogeneity is an obvious obstacle to overcome for well-integrated, interoperable manufacturing environments.

One way to manage data heterogeneity is to use a *linked data* approach to create a uniform information space, across which data from the different sources can be integrated. Linked data refers to data that "is published on the Web in such way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can, in turn, be linked to for, external data sets" (Bizer, Heath, & Berners-Lee, 2009 p.17). At the same time, *linked enterprise data* (Wood, 2010) is a special form of linked data that aims to create a uniform information space within a specific enterprise.

Irrespective of the integration approach, when dealing with heterogeneous data from different sources, one must measure the quality of the data being integrated. Such an understanding of the quality is essential to using data with confidence for a given purpose. Borovina Josko and Ferreira (2017) states that data quality assessment outcomes are essential to ensure useful analytical processes results. More specifically, having high-quality data enables analytical approaches that can assess and improve key parameters, such as time, performance, cost, and so on. Moreover, assessing the quality of the data after an integration solution implemented is beneficial to improve the solution and utilize it better. Therefore, one must assess the data quality and be knowledgeable about the quality of data before using it for any further purpose.

When dealing with data quality, one needs to consider several issues: errors and inconsistencies, misspellings in data entry, missing information, or other invalid data (Rahm & Do, 2000). Moreover, integrating multiple data sources requires accessing accurate and consistent data, consolidating different data representations, and eliminating duplicate information.

During the transition from unstructured "islands of information" (El-khoury, Gürdür, & Nyberg, 2016) to integrated linked enterprise data, priority is initially given to the integration of the data. Therefore, data quality assessment becomes a secondary concern during the initial stages. This study aims to provide a data quality assessment methodology to identify the data quality dimensions and metrics from the literature that are most relevant to the specific needs of an enterprise. The methodology leads to the development of a data quality assessment dashboard that provides different information visualizations to illustrate data quality dimensions. Furthermore, the dashboard shows the status of the data quality metrics for the purpose of identifying missing, inconsistent, invalid, or non-informative data. This study is based on a case study at

Scania CV AB, a heavy truck manufacturer in Sweden, where this methodological approach was proposed and tested to assess the quality of the company's linked enterprise solution.

This paper is organized into seven sections: Section 2 describes the case study by explaining the two phases of the study—data integration and data quality assessment—in addition to giving a brief explanation of the selected framework. Section 3 defines the methodology phase by phase, explaining each step in detail. Then, Section 4 discusses the application of the methodology, explains selected data quality dimensions, and presents an example dashboard. Section 5 briefly mentions the related work. Finally, the paper concludes with a summary of the study in Section 6.

2. Case Study Description

The case study that initiated this research is related to enterprise development environments and was done in collaboration with Scania CV AB. The case is divided into two phases: data integration and data quality assessment. The next two subsections describe these phases.

2.1. Data Integration

The data integration phase was completed at Scania CV AB to overcome, primarily, the traceability need across its heterogenous data sources. Traceability is an important motivation due to the requirement of compliance with ISO 26262 (ISO, 2011). ISO 26262:2011 mandates that requirements and design components be developed at different levels of abstraction and that clear trace links exist between requirements from the different levels, in addition to the links between requirements and system components. This request for traceability implies that the development artifacts are constantly accessible, even if they reside across different development tools. The case study illustrates a typical development environment where several engineering software tools—such as a requirements management tool, an architecture design tool, and a configuration management tool—are used. Integration of these tools is achieved through a linked enterprise data solution where data from different tools are published through linked data principles. This approach enables traceability between product artifacts from different development tools.

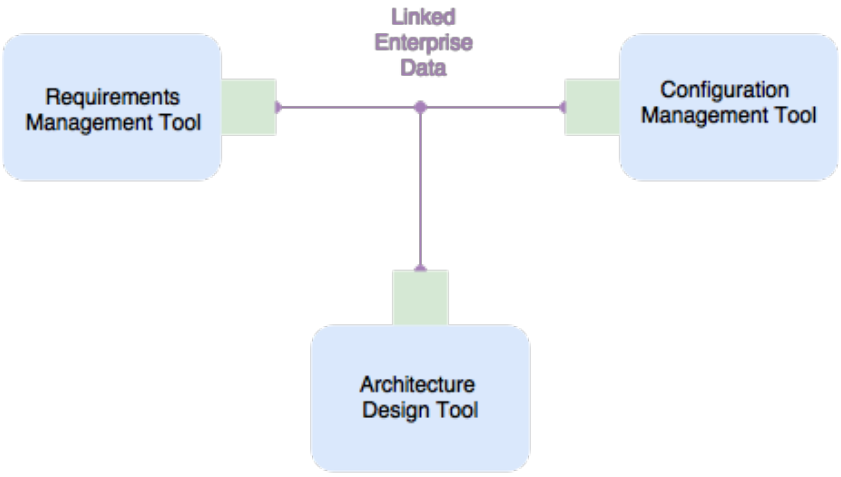

Fig. 1. Use case software tools and the integration solution.

When developing this integrated development environment, different adaptors—specialized tool extensions that allow sharing data—are developed where artifacts—across the different engineering tools—are made accessible throughout the development toolchain, as shown in Figure 1. The decision to use the linked data approach was based on its ability to migrate from a monolithic data structure to a distributed structure. This migration increases the size and scope of the potential data sources that organizations can use (Batini, Cappiello, Francalanci, & Maurino, 2009).

The integration technologies and techniques that have been used for integration are not the focus of this paper. However, interested readers can learn more about the technical-level contributions of the case in the work of El-khoury et al. (2016). The focus of this article is on the data quality assessment of the result of the data integration solution. Surely, the linked data integration solution is an important step toward better integrated engineering environment. However, the quality of the end data after the integration solution is an vital measure before starting to use this data further.

2.2. Data Quality Assessment

The second phase of the study, the data quality assessment phase, is the main focus of this paper. The proposed data quality assessment methodology was developed based on the results of this phase.

Initially, the previously described linked enterprise data was integrated with the toolchain. Later, Scania CV AB devised some rules based on experience and wanted to measure the rules. To do so, it was first necessary to understand their relationship with the existing data quality, state-of-the-art research, and data quality dimensions. For this reason, the literature was studied, interviews and meetings were conducted, an information visualization approach was proposed, and finally, the dashboard was developed. During the case study, we realized the need for a well-defined methodology and revised the whole process to form a methodology.

Since the linked enterprise data integration solution primarily aimed to integrate data across the software tools in order to solve the traceability issues, data quality was not the focus in the first phase. However, when the solution was to be used for data analysis, data quality became important, especially when making decisions based on the data. Assuring data quality is particularly challenging in linked data as the underlying data stems from a set of evolving data sources (Zaveri, Rula, Maurino, Pietrobon, & Lehmann, 2012), even though it is an efficient solution for the integration problem.

During the study, we identified data quality issues and, to improve data quality, defined a set of rules for updating the linked data resources. The rules defined constraints and conventions about the naming, linking, and structuring of the linked data resources and their properties. This exercise evocated the idea of merging the existing data quality literature with these empirical rules. By reviewing the state-of-the-art research on data quality assessment, we gathered several data quality dimensions and metrics, allowing us to merge the empirical rules with the dimensions from the literature.

Moreover, it was necessary to develop data quality policies, protocols, and monitoring mechanisms to assess the data quality. Therefore, this case study not only aims to map empirical rules with the data quality dimensions from the research but also proposes a methodology for developing data quality management policies and an interactive information visualization dashboard to aid stakeholders with assessing linked enterprise data quality.

2.3. Selected Framework

With this study, we wanted to promote the reuse of existing data quality research and adopted the suggested metrics, dimensions, and methods relevant to our case study. Therefore, this subsection presents an analysis of the most prominent data quality research, with the aim of extracting dimensions and metrics to visualize these dimensions. We based our analysis on the work of Zaveri et al. (2012) since it is itself a comprehensive survey of other work.

2.3.1. Linked Data Quality Dimensions

Zaveri et al. (2012) conducted a comprehensive survey on linked data quality assessment and identified 16 quality dimensions that have been studied in the literature. The authors classified these dimensions into four categories: (i) Accessibility, (ii) Intrinsic, (iii) Contextual, and (iv) Representational. The categories and associated dimensions are listed below and will be further detailed in the next section:

- **Accessibility:** Availability, licensing, interlinking, security, and performance.
- **Intrinsic:** Accuracy, consistency, conciseness, and completeness.
- **Contextual:** Relevancy, trustworthiness, understandability, and timeliness.
- **Representational:** Representational conciseness, interoperability, and versatility.

Detailed explanations of the selected dimensions from the work of Zaveri et al. (2012) will be presented in Section 4, where we describe the application of the methodology.

3. Methodology for Linked Enterprise Data Quality Assessment

This study presents guidance to enterprises to consider metrics for data quality assessment and use information visualizations to represent the quality of their linked enterprise data through a visual dashboard. The need to understand the current data quality appeared after integrating the different engineering tools we explained in the case study description section (Section 2). One important reason behind motivating the use of a dashboard for this purpose is the opportunities that information visualization techniques promise (Gürdür, El-Khoury, Seceleanu, & Lednicki, 2016). By visualizing the quality dimensions, stakeholders have the opportunity to monitor the specific aspects of data they identified as relevant through these dimensions (Gürdür, 2016; Andrienko & Andrienko, 2013). However, developing such a dashboard is not straightforward (Haglin, Trimm, & Wong, 2017) and requires a step-by-step process—a methodology that aids stakeholders in discussing, defining,

and identifying data quality dimensions and developing a dashboard to monitor the dimensions. The methodology presented in this section is a combination of exploratory, creative, and improvement technics.

This methodological approach mainly aims to support data quality assessment by developing an information visualization dashboard. Yet it also motivates developing protocols related to data quality management. The methodology aims to create several deliverables as a result of different phases that can be used for this purpose. Moreover, the methodology requires several stakeholders to meet, discuss, and make decisions together to build a common language. For example, discussing the data quality dimensions from the state-of-the-art research and mapping them with the needs of the company gives all stakeholders an overall understanding about the current state of the art. This would help the enterprise to adopt well-studied data quality dimensions from the literature and to ground the development of the data quality dashboard on this understanding.

We use several keywords throughout this paper—such as requirements, rules, dimensions, and metrics—the definitions of these keywords are found below:

- **Requirement:** The desired outcome that stakeholders want to have—for example, fully complete data resources.
- **Rule:** The set of explicit constraints that the data is expected to satisfy in order to fulfil the requirements—for instance, all resources must have a label. This rule ensures that the linked data resource label information is complete. One or more rules are needed to fulfill one requirement.
- **Dimension:** A data quality attribute defined by the literature and explained in the data quality literature study section of this paper—for example, the completeness dimension.
- **Assessment Metric (in short Metric):** A measure or indicator used to measure a data quality dimension—for example, to measure the completeness dimension, one can calculate the average number of resources that do not have a label property defined. This ratio can be used to calculate the percentage of completeness about the label property.

3.1. Phases and Steps

The methodology consists of the following six phases:

- Phase 1 (Plan): Identify the stakeholders who will be involved throughout the process and define the first draft of a roadmap that will guide the team members throughout the process.
- Phase 2 (Extract): Extract important requirements on the overall desired quality assessments.
- Phase 3 (Identify): Identify relevant data quality dimensions from the reference framework (Zaveri et al., 2012) and recognize important rules and metrics that will be used to develop visualizations for each dimension.
- Phase 4 (Query): Query the data to be used to calculate metrics and collect the data for analysis.
- Phase 5 (Visualize): Visualize data and develop a dashboard for assessing data quality dimensions.
- Phase 6 (Improve): Improve the resulting dashboard according to the user experience.

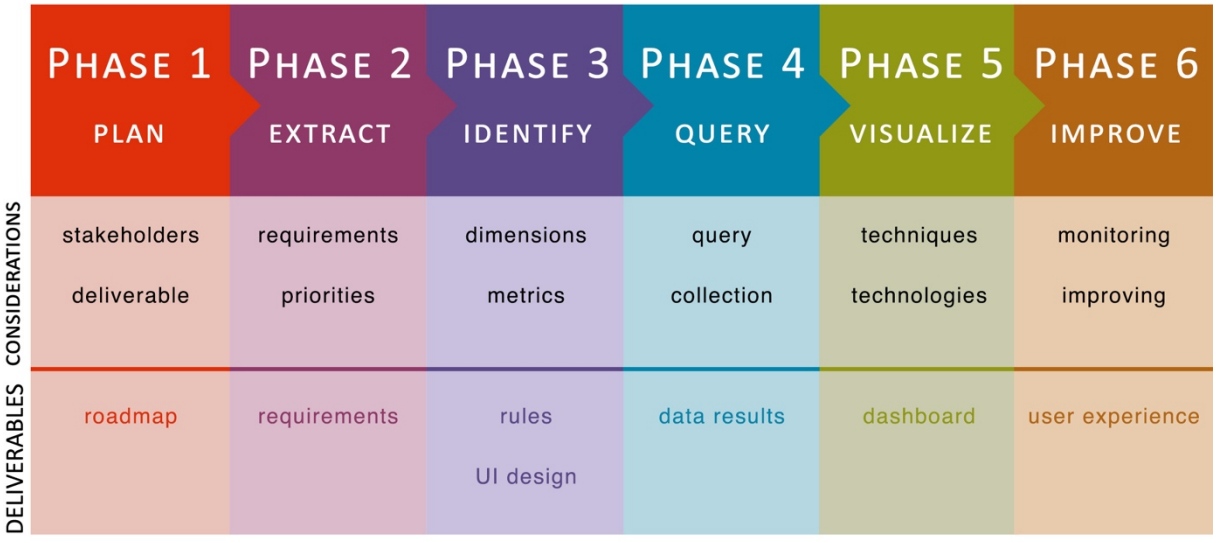

Fig. 2. Linked enterprise data quality assessment methodology phases.

Considerations and deliverables for each phase are summarized in Figure 2. Each phase of the methodology completes with an expected deliverable. These deliverables are important documents, decisions, and design or development plans that can later be used to form a data quality assessment protocol.

The methodology includes participative phases where different stakeholders take part in meetings and discussions. Moreover, it is suggested that a few deliverables be presented to these team members in different phases for the purpose of receiving feedback during the development of the final dashboard design. Stakeholders can be different for each phase due to the requirements of each particular phase. Furthermore, some phases require several iterations to establish a common understanding of data quality assessment and management.

3.1.1. Phase 1: Plan

The process starts with the planning phase. In this phase, several decisions related to the data quality assessment and its management need to be considered. Therefore, this phase aims to form a data management group (DMG), which comprises the stakeholders who will work on different phases to develop the data management protocol and data quality assessment dashboard.

Firstly, important team members, roles, and stakeholders should be identified. These stakeholders should decide how many deliverables they would like to produce throughout the process as a result of different phases, though we strongly suggest that the deliverables be used later as policies. In addition, the objective(s) of the linked enterprise data quality assessment and the roadmap to reach the objective(s) must be drafted in this phase.

One of the reasons that the process starts with finding and selecting relevant stakeholders is the need to discuss the importance of the data quality assessment. This discussion is the key to identifying the need for data quality protocols and policies, and finding team members who are willing to take responsibility and make an active decision to change the current situation. This step will guide different stakeholders to explore their main interest and their willingness to be part of different phases of the methodology.

Secondly, the interest of different team members needs to be considered, and people from different backgrounds in different groups should be included in the discussions related to data quality. For example, engineers from different departments can be part of a group for the third "Identify" phase, which will later work on data quality dimensions. At the same time, technical stakeholders who worked on the integration solutions should be involved in this first phase in order to both better understand the importance of the data quality assessment and to guide other stakeholders on the technical details of the integration solution since they know more about the architecture for an integrated development environment.

Finally, in this first phase, it is expected that DMG should agree on the form of all deliverables for each of the other phases. Phase 1 continues until the DMG reaches a consensus and completes the roadmap. The roadmap lists the stakeholders, members of the DMG, responsibilities, deliverables, and a time plan where the phases, processes, and deliverables are scheduled. This roadmap is the first deliverable of the methodology and will later be updated throughout the phases and become part of the enterprise data management protocol.

3.1.2. Phase 2: Extract

In the second phase, extraction, one of the first considerations is the list of requirements. Requirements, in this phase, encompass the compulsory or necessary end result that the team is aiming to have at the end of the assessment process. Team members of Phase 2 should define and list these requirements for the linked enterprise data in hand. The method for collecting these requirements is through meetings. One important reason for having this particular stage is to create a common understanding of the needs of different stakeholders. Conducting meetings with different stakeholders and understanding each other's expectations is, therefore, an important part of this phase.

Another important step in this phase is prioritizing the requirements. After listing all the requirements and understanding different expectations, the group should prioritize these requirements according to difficulty, time consumption, availability of the team members in the next phases, and so on. The deliverable of this phase is a document that lists high-level requirements (see Appendix).

3.1.3. Phase 3: Identify

Phase 3 aims to develop the rules based on requirements defined in the earlier phase—based on empirical needs—and map them into the LED quality dimensions, as defined in the selected framework of Zaveri et al. (2012).

Once prioritized, each requirement must then be further broken down into a set of rules that detail the LED quality. These rules are more detailed versions of requirements. Team members can create the rules by going through the requirements and further defining them as rules. For example, a requirement can be defined as having an understandable resource name for all

resources. To fulfill this requirement, a rule would state that each resource should have a property dedicated to the resource name. For instance, an understandable resource name can be achieved by having two rules: 1) having a property defined as a label and 2) making sure that this label (rdfs:label) property exists and is not null for any resources. These rules are domain-specific rules that have been identified by experience.

Based on discussions among stakeholders, each rule is then categorized as part of one of the predefined dimensions. This exercise is useful for the team members to understand both their desired rules and the relevant dimension(s) in detail for developing a common understanding of the selected framework for linked data quality. This practice will widen the discussion to include potentially new dimensions in case the dimensions that have been explained are not relevant for the specific system at hand. Some dimensions may similarly be excluded since they are not important or are not high priority for the time being. In either case, this exercise will help the enterprise to develop understanding and knowledge on data quality dimensions; this will later result in a data quality dashboard with which all team members already have familiarity. In this way, the data quality protocol can be easily introduced to new team members and other related parties.

In this phase, several metrics should be defined for each dimension. As mentioned previously, the purpose of metrics is to identify measurable concepts for each dimension. The rule from Phase 2 that all resources should have the rdfs:label property provides a good example. This rule can be used to assess the usefulness and understandability dimensions of the linked enterprise data. One way, for instance, to measure the usefulness is to check all available resources and to count how many of these resources have a label property that is not null. The ratio of the resources with a nonnull label property to the total number of resources gives a measurable metric for the usefulness dimension.

The dimensions and metrics, then, can be used in designing the user interface (UI) for the dashboard in Phases 4 and 5. This UI design should consider the rules and requirements according to the priority decided in earlier phases. The exercise of drafting the UI design aims to support Phase 5 but is also beneficial to help stakeholders who will not be active in the next phases to understand the future work that will be carried out by other stakeholders. To this end, the first draft of the UI design summarizes the most important dimensions and the rules to assess each dimension, and is the result of this phase. Another deliverable of this phase is a document that lists a well-defined set of rules and the LED quality dimensions (see Appendix) from the research.

3.1.4. Phase 4: Query

In the fourth phase, team members work on the metrics identified in the third phase and develop queries to measure each metric, for any given data set. In addition, depending on the size of the data, performance issues and query optimization techniques should be considered.

This step can be challenging since it requires some information from Phase 3 and from the later Phase 5. Most likely, a number of local iterations between these three phases will be required. For instance, the metrics from Phase 3 and the result of queries will be used in Phase 5 to create visualizations. Therefore, the team member(s) in all three phases should be in sync and inform each other about the process. Moreover, the UI design can be used to understand what additional information will be needed by the stakeholders who will use queries in the dashboard design. Therefore, meetings to discuss the implementation process are suggested within the proposed methodology.

Phase 4 should deliver the values of each metric when applying the defined queries on the available data set. These values will be later used by the information visualization team to develop the dashboard.

3.1.5. Phase 5: Visualize

In the fifth phase, team members work on developing several visualizations to create the dashboard. They should decide on the best visualization technique to present the metrics by considering different factors. One should first consider the function of the visualization:

- If the aim is to illustrate a comparison, then a bar chart, line graph, or radar chart is a appropriate choice.
- If the goal is to show the relationships, then a network diagram, tree diagram, or arc diagram can be employable.
- If the goal is to give statistics about the data set, it is suitable to use big, legible font families with bold text as headings to increase both the readability and scannability of the information.

Different technologies can be adapted when developing information visualizations. One way is to use off-the-shelf products, such as Tableau (Chabot, Stolte, & Hanrahan, 2003), Qlik Sense (Ilacqua, Cronstrom, & Richardson, 2015), Pentaho (Bouman & Dongen, 2009), and so on. Data sources for these solutions can be a relational database, flat files, XML/JSON, and reports. Some of these software tools are available as open-source platforms, which can be low cost for an enterprise. However, there are also different versions of each where more functionalities are implemented according to the licensing agreement. Dashboards can also be developed using Web visual presentation tools, such as D3.js. D3.js (Bostock,

2012) is a JavaScript library for manipulating documents based on data. D3.js uses Scalable Vector Graphics (SVG), a World Wide Web Consortium specification that defines the network vector graphics standard.

The DMG should consider different technologies and make decisions according to costs, applicability, and the availability of the specialized personnel who will work on the selected technology. The deliverable of this phase is a working dashboard that visualizes the data quality dimensions selected after the third phase and uses the data sources from the query results of Phase 4.

3.1.6. Phase 6: Improve

The main focus of the sixth and final phase is the improvement of the dimensions, metrics, and dashboard. This requires monitoring the usage of the dashboard by allowing users to interact with the dashboard and collecting information about usage patterns. Detecting the most used metric and understanding the user activity on different information visualizations are key activities in this phase.

Later, the user experience data will be used to improve the dashboard by adding or extracting more capabilities, simplifying the design, and introducing new information visualizations and statistical information throughout the dashboard.

This phase is linked with Phase 5, where the information visualizations development are repeated. Phase 6 also has direct links with Phases 3 and 4—where, if new visualizations are necessary, new metrics and new queries are designed, and new data is collected.

The overall aim of the methodology is to improve the data quality of the linked enterprise data set. A successful use of the methodology and dashboard will lead to improvements in the data quality, which in turn will require changes to the measured metrics. Another source of change is that the integrated development environment will change over time, leading to new data sources being integrated, which will also call for iterations of the methodology.

4. Application of the Methodology

This section details applying the methodology to the use case by presenting the dashboard and different information visualizations to exemplify some of the dimensions introduced in the literature study.

As previously mentioned, the methodology was inspired based on the initial application of the quality assessment case study, and the implementation of the linked enterprise data environment has been initially completed without any data quality assessments. Therefore, some recommended phases (in particular Phases 1 and 6) are not realized.

4.1. Phase 1

In the first phase, Planning, which was not carried out in this study, the stakeholders should be identified and allocated defined roles to engage in specific phases of the methodology. In this use case, we suggest defining five roles, assigning the stakeholders to one or more phases, and including them in creating the necessary deliverables through planning meetings. Brief information about these stakeholders is given below:

Lead Research Engineer: The research engineer working in the company and leading the data quality assessment work.

Senior Software Developer: The software developer who leads the implementation of the linked enterprise data solution.

Head of the Division: An experienced engineer who is managing the division and leading the linked enterprise data solution for the development environments of the company.

Software Developer: A consultant who works on improving the linked enterprise data solution and who has SPARQL experience.

Data Scientist: A research engineer who specializes in data science, works on developing the information visualizations, and is responsible for the dashboard implementation.

Table 1. Roles, their associated phases and expected deliverables.

Roles	Phases	Deliverables
Lead Research Engineer	Phase 1, 2, 3	Roadmap, rules
Senior Software Developer	Phase 1, 2, 3	Roadmap, rules

Head of the Division	Phase 1	Roadmap
Software Developer	Phase 3, 4	Query results
Data Scientist	Phase 3, 5, 6	Dashboard, UI design, UX

Table 1 illustrates the stakeholders, the phases they are taking part in, and their responsibilities regarding the deliverables. We suggest that the Lead Research Engineer and Senior Software Developer take part in Phases 1, 2, and 3. It is important to include these stakeholders in discussions about the roadmap and rules deliverables. Then, the Software Developer experienced in the SPARQL query languages for performing data extraction with regard to the specific data quality dimensions will be part of Phases 2, 3, and 4. This person will likely work on the query results deliverable. The Data Scientist with experience in information visualization with D3.js can lead Phases 5 and 6; the two phases are highly connected and require several iterations between them.

4.2. Phase 2

In Phase 2, the stakeholders defined important requirements for the purpose of data quality assessment. In the case study, the process of exposing data from each software tool as linked data resources was performed over a long period of time, with many developers involved, and where the policy of what information was exposed had changed over time. This resulted in an increased risk that not all information was being transformed correctly and completely. For this reason, "having a complete data set" was an important requirement for the stakeholders. Another important requirement was the need to ensure the data content was representable to the end-user from a search application developed by Scania CV AB as part of its linked data solution for implementing search capabilities. Stakeholders were concerned that some of the data content was not understandable since it may not have any textual properties that explain what the content is about. Thus, the second requirement was defined as "having an understandable data set." These were identified as the first two requirements, and they were assigned a high priority. As a next step, two rules were defined from the requirements:

- all resources should have complete information (properties), and
- all properties should be understandable.

These rules, as well as several more addressing the concerns of the group, were developed and formed the Phase 2 deliverable. The two rules will be used throughout this section to exemplify the application of the methodology. (See Appendix for the full list of rules.) The rules list includes more than 20 rules and is part of more comprehensive documents about rules concerning linked enterprise data. The document describes linked data resources and their definitions, relationships, formats, and so on. In this use case, we have, for the sake of brevity, only included the first two requirements and the rules related to the requirements to assess two data quality dimensions.

4.2.1. Summary of Relevant Data Quality Dimensions

As previously mentioned, the linked enterprise data does not share exactly the same difficulties as linked open data. Therefore, based on our case study, we analyzed the in total 16 dimensions and identified the ones relevant to LED quality. This section provides a summary of these relevant dimensions. We used the extensive literature review done by Zaveri et al. (2012) as the main reference framework for describing the relevant data quality dimensions for linked enterprise data. Later, these dimensions will also be considered as part of the methodology that we present in this paper.

- **Accessibility:** The accessibility dimension includes all aspects related to the "access, authenticity and retrieval of data to obtain either the entire or some portions of the data for a particular use case" (Zaveri et al., 2012).
    - **Availability:** Availability "is the extent to which data (or some portion of it) is present, obtainable and ready for use." One can measure the availability by checking the accessibility of the SPARQL end-point and the server, or by checking whether an RDF dump is provided and can be downloaded (Hogan et al., 2010).
    - **Security:** Security involves "the possibility to restrict access to the data and to guarantee the confidentiality of the communication between a source and its consumers" (Hogan et al., 2010). In this case study, we are mainly interested in the confidentiality of the company structures and the software tools.
- **Intrinsic:** The intrinsic category includes all aspects that are independent of the context within which the data is being used. In other words, it focuses on whether the information is correct, compact, and complete.

- **Accuracy:** Accuracy is defined as "the degree to which the data correctly represents the real world facts and is also free of syntax errors" (Zaveri et al., 2012). One can measure the accuracy by detecting outliers (Bizer & Cyganiak, 2009), inaccurate values, malformed datatype literals (Hogan et al., 2010), erroneous annotations (Fürber & Hepp, 2011), or inaccurate labeling, classifications (Flemming, 2010).
- **Consistency:** Consistency is defined as a knowledge base that is "free of logical or formal contradictions with respect to particular knowledge representation and inference mechanisms." Consistency can be measured by the detection of classes and properties used without any formal definition (Hogan et al., 2010).
- **Completeness:** Completeness is concerned with determining whether all required information is present in a particular dataset (Zaveri et al., 2012). This dimension can be calculated for different levels, such as schema completeness, property completeness, population completeness, and so on.

- **Contextual:** Unlike the intrinsic dimensions, contextual dimensions are those that highly depend on the context of the task at hand.
  - **Usefulness:** Usefulness is explained as understandability in the work of Flemming (2010), Hogan et al. (2010), and Zaveri et al. (2012), and defined as the comprehensibility of data or the ease with which human consumers can understand and utilize the data. One can measure the usefulness by detecting human-readable labeling of classes, properties, and entities; dereferencing representations; providing human-readable metadata, and so on (Flemming, 2010).
  - **Timeliness:** The timeliness dimension measures how up-to-date the data is. Timeliness can be measured by, for instance, a positive difference between current and expiry time of the data (Flemming, 2010). Another approach to assessing timeliness is to use methods that guarantee near real-time updates.
- **Representational:** Representational dimensions capture aspects related to the format with which the data is ultimately encoded and stored persistently.
  - **Representational Conciseness:** Representational conciseness measures whether the data representation is "compact and well formatted on the one hand and clear and complete on the other hand" (Zaveri et al., 2012). Hogan et al. (2010) discussed the benefits of using shorter URI strings and encouraged the use of a concise representation of the data. To measure the representational conciseness, one can detect the length of URIs.

These data quality dimensions not only are important to help enterprises understand the state-of-the-art research but also are a part of the methodology that requires the stakeholders with roles in data quality to learn, understand, and use the dimensions in their specific data quality management activities.

4.2.2. Excluded Data Quality Dimensions

The research on linked data dimensions took shape around the variety of linked open data published on the Web, whereas this particular study concentrates on linked enterprise data as opposed to open Web data. The interlinking, licensing, performance, relevancy, trustworthiness, and versatility dimensions are removed not because they are not important dimensions, but because the nature of linked enterprise data means some dimensions were already at an acceptable quality level. These dimensions and the reasons they have not been included are explained in Table 2.

Table 2. Dimensions and associated reason for excluding.

Dimension	Reason to exclude
Interlinking	We are interested with not only the same but also similar linked data resources (See Traceability).
Licensing	Outside the scope of this case study.
Relevancy	Outside the scope of this case study.
Performance	Outside the scope of this case study.
Trustworthiness	Linked enterprise data solution assures that the linked data resources are

coming from trustworthy sources.

Interoperability	We are interested with reuseage of the existing resources (See Reuse).

4.2.3. New Data Quality Dimensions

We have identified two new dimensions—Traceability and Reuse.

- **Traceability:** Traceability is "the identification and documentation of derivation paths (upward) and allocation or flow down paths (downward) of work products in the work product hierarchy" (IEEE, 1998). Traceability of linked enterprise data resources can be assessed by checking the quality and quantity (Neto, Kontokostas, Hellmann, Müller, & Brümmer, 2016) of the links between related data resources from different software tools.
- **Reuse:** Reuse is the degree to which the data format and structure of the information use the earlier structures or data resources again. Reuse of linked data resources can be measured by detecting whether existing terms from all relevant vocabularies for that particular domain have been reused (Hogan et al., 2010) or usage of relevant vocabularies for that particular domain (Flemming, 2010).

4.3. Phase 3

In Phase 3, the requirements linked with the dimensions from the state-of-the-art research and rules are identified (See Appendix). The first rule, for instance, is associated with the completeness dimension, whereas the second rule is associated with the usefulness dimension. For each dimension, according to the rules, several metrics are defined to measure each dimension. During the first iteration, stakeholders extracted two dimensions and five metrics. Table 3 summarizes two example rules, dimensions, and metrics. To illustrate the current data quality situation of each metric, it was decided that an information visualization would be created. The deliverable of this phase was created by drafting a sample UI design.

Table 3. Requirements, data quality dimensions, and metrics (numbered for referencing).

Requirement	Dimension	Metric
Having a complete data set	Completeness	$\frac{\#\text{ of resources with } nonnull \text{ rdfs:label property}}{\text{Total number of resources}}$ (M1)
		$\frac{\#\text{ of resources with } nonnull \text{ rdfs:label property}}{\text{Total number of resources}}$ (M2)
		(M3)
Having an understandable data set	Usefulness	(M4)
		(M5)

4.4. Phase 4

In Phase 4, SPARQL queries were developed for each of the metrics for the purpose of measuring the metrics. Some queries were designed to provide statistical numbers. For instance, to calculate the metric about the completeness listed in Table 3, a query was designed to give the average number of resources with nonnull rdfs:label property (M1). Other than metrics, resources that have rdfs:label property null were also listed by another query. This allowed for detailed information about the resources that had incomplete or un-understandable information to be made available for use in the dashboard.

4.5. Phase 5

In Phase 5, information visualization techniques were decided upon. For dimensions where relationships were important, the tree type of visualization was used. In other cases where more detail was needed, tables with direct links to resources were used. Moreover, important numbers were represented by big, bold headings throughout the dashboard design to draw attention to some statistics about the dimensions through the rules and metrics.

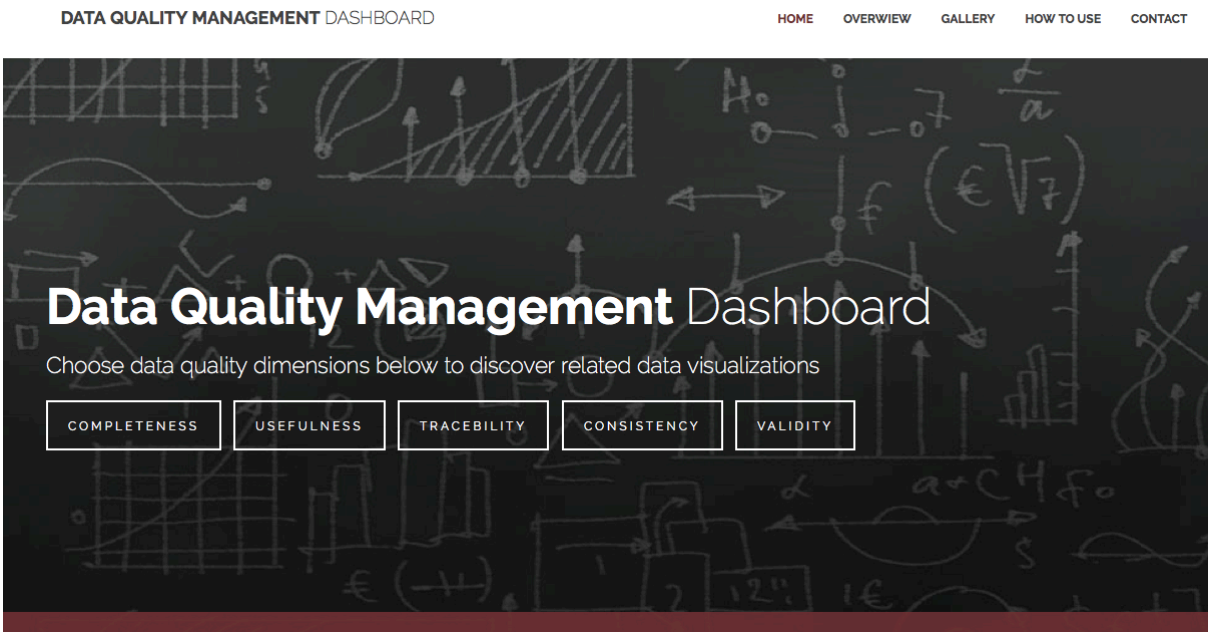

Figure 3. User interface design for the selected data quality domains.

Figure 3 shows the intro page of the designed dashboard where the data quality dimensions are accessible with buttons. These dimensions are completeness, usefulness, traceability, consistency, and validity. The completeness and usefulness dimensions are supported by several information visualizations, as part of this study, to exemplify the application of the methodology.

Figure 4 illustrates the statistical information about the first metric (M1) of Table 3. A total number of resources (2,145) is stated at top of the first rectangle in bold. On the other hand, the total number of resources that have rdfs:label property null (2,017) is stated in the second rectangle on the right. The percentage of the resources with a missing label is also shown.

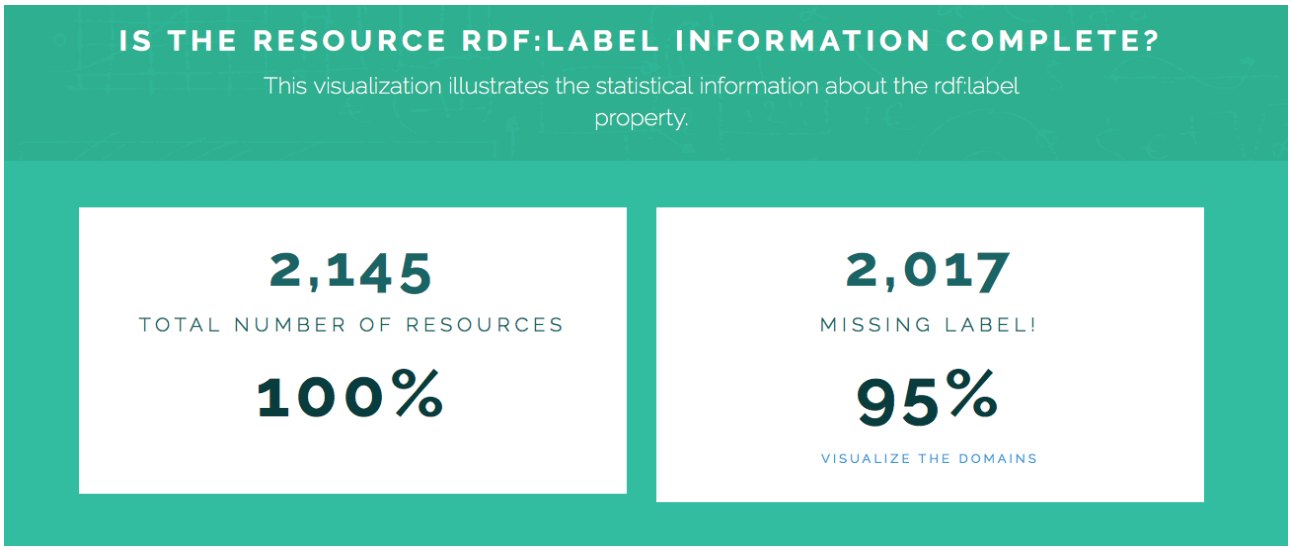

Figure 4. Visualization of statistical information.

This visualization was designed to quickly and easily give overall information about the completeness metric. The aim of the visualization is to draw the attention of the user to the fact that 95% of the resources are missing the expected label information. This information visualization is not aiming to list details about the resources, their properties, or relationships. For further details, a hyperlink was added under the percentage information that redirects the user to another view (Figure 5).

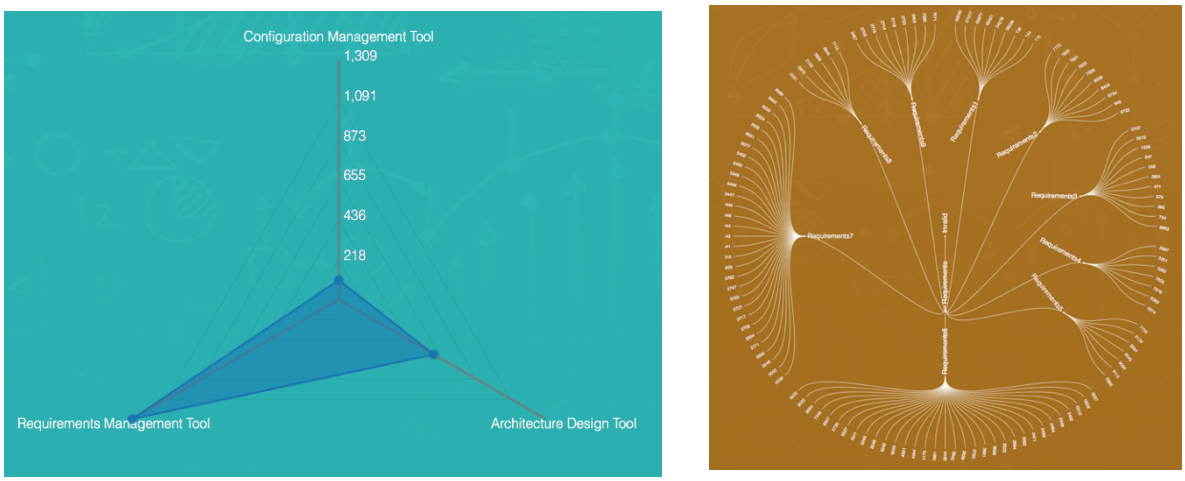

Figure 5. Radar diagram visualization of domains (left) and circular tree visualization of resources from requirements management tool with null rdf:label property (right).

Figure 5 shows more detailed information about the resources that have rdf:label information missing. A radar diagram was used (left in Figure 5) to show different tools on different axes and the number of resources with the missing information. In this visualization, one can easily realize that most of the resources with a missing label property are coming from the requirements management tool. Since the dashboard design is interactive, when the user hovers on the point on the requirements management tool axis, the exact number of missing resources (1,309) is shown. The radar diagram visualization was designed to inform the user where the effort should be put to improve the linked data quality. By only seeing this visualization, stakeholders could easily make a decision to concentrate their efforts on improving the data coming from the requirements management tool.

At the same time, the circular tree visualization (right in Figure 5) gives even more detailed information about requirements and their relationships. The circular tree visualization shows a set of requirements that belongs to the requirements management tool and has no label information.

4.6. Phase 6

In the final phase, a platform to monitor user activity was designed and implemented. This beta version of the data quality assessment dashboard is under observation where user experiences are collected through tracking the clicks and the amount of time spent on each visualization, in addition to developing a short questionnaire about the user experience in order to improve the dashboard design and future implementations.

5. Related Work

Generally, data quality assessment approaches are incorporated around data quality dimensions. These dimensions are especially beneficial for guiding the stakeholders to create a common language to facilitate conversations about different data properties. Kahn, Strong, and Wang (2002) stated that several healthcare, finance, and consumer product companies have chosen to use questionnaires to extract these well-known data quality dimensions.

Data quality may depend on various factors. Wang, Strong, and Taylor (1996), for instance, defined 20 different dimensions after surveying 355 data consumers. Some dimensions identified by this survey were accuracy, availability, believability, completeness, consistency, conciseness, relevancy, timeliness, objectivity, understandability, and verifiability.

Few researcher studied the data quality specific to linked data. For instance, Radulovic, Mihindukulasooriya, García-Castro, & Gómez-Pérez (2017) presented a quality model for LOD. The proposed quality model provides a reference for linked data quality specification and evaluation, in addition a set of quality characteristics and quality measures related to linked data. This study gives formulations to measure different quality dimensions related with data which can be useful during the Phase 3 and 4.

At the same time, Sadiq & Indulska (2017) outline the challenges in dealing with data quality of open datasets. Even though the research focused on open data several research challenges that were identified were common with our discussions. The three main challenges that authors find important and necessary to change are; shared understanding of data quality dimensions, support for quality awareness, and strengthening the quality-to-use Nexus.

Before concluding, it is important to summarize the existing tool support for data quality assessment concerning linked open data. Several existing tools assess the quality of linked open data—for example, Luzzu (Debattista, Auer, & Lange, 2016), linked open data Laundromat (Beek et al., 1996), and Loupe (Mihindukulasooriya, Poveda-Villalon, García Castro & Gomez-Perez, 2015). These tools aim to convert linked open data in a standard, compliant way, removing data stains such as syntax errors, duplicates, and blank nodes. However, these tools do not support the assessment of linked enterprise data, and for this reason, we did not discuss them in detail.

6. Conclusion

This paper introduced a linked enterprise data quality assessment methodology for the purpose of developing a dashboard for monitoring data quality, in addition to creating protocols through deliverables to ensure the quality of linked enterprise data. The essence of the methodology is to merge empirical rules with dimensions from state-of-the-art research. The methodology consists of six phases and seven deliverables as a result of each phase.

The methodology was developed after implementing the data quality assessment dashboard for the exploratory use case explained in this article. Therefore, use case implementation includes both phases that have been tested for applicability and feasibility of the methodology (Phase 2, Phase 3, Phase 4, Phase 5) and phases that have not been tested in practice but were deemed important enough to add to the methodology (Phase 1, Phase 6).

The methodology proposes a structured approach to guide enterprises in developing a dashboard to assess the quality of their linked enterprise data. Assessing the quality of the data integration solution benefits the enterprise in few different ways. Firstly, the data quality assessment helps stakeholders to identify where the integration solution falls short and guides the stakeholders to focus their efforts on to the right place to utilize their solutions further. Secondly, developing a dashboard for this purpose required different stakeholders to work together and improves the common understanding of the data. Thirdly, the methodological approach supports data quality management documentations to be created which can be use later as guidance to ensure the quality of linked enterprise data. More specifically, the methodology uses the advantages of data visualizations to show the overall information, to identify patterns, to better understand the current situation, and to realize relationships between resources. Furthermore, it helps stakeholders with developing protocols related to data quality by employing frequent meetings and creating deliverables at the end of each phase. The methodology is not only useful for developing a data quality management culture in any enterprise but also provides a practical and interactive tool for assessing the quality of linked enterprise data.

References

Andrienko, N., & Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, *12*(1), 3–24. http://doi.org/10.1177/1473871612457601

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. http://doi.org/10.1145/1541880.1541883

Beek, W., Rietveld, L., Bazoobandi, H. R., & Wielemaker, J. (n.d.). LOD Laundromat : A Uniform Way of Publishing Other People ' s Dirty Data.

Bizer, C., & Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*, 1–10. http://doi.org/10.1016/j.websem.2008.02.005

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on* .... Retrieved from http://eprints.soton.ac.uk/271285/

Borovina Josko, J. M., & Ferreira, J. E. (2017). Visualization properties for data quality visual assessment: An exploratory case study. *Information Visualization*, *16*(2), 93–112. http://doi.org/10.1177/1473871616629516

Bostock, M. (2012). D3. js.

Bouman, R., & Dongen, J. van. (2009). *Pentaho Solutions ®*. Wiley Publishing.

Chabot, C., Stolte, C., & Hanrahan, P. (2003). *Tableau Software*.

Debattista, J., Auer, S., & Lange, C. (2016). Luzzu – A Framework for Linked Data Quality Assessment, 124–131. http://doi.org/10.1109/ICSC.2016.48

El-khoury, J., Gurdur, D., & Nyberg, M. (2016). A Model-Driven Engineering Approach to Software Tool Interoperability based on Linked Data, *9*(3), 248–259.

Flemming, A. (2010). *Quality Characteristics of Linked Data Publishing Datasources*. Humboldt-Universität of Berlin.

Fürber, C., & Hepp, M. (2011). SWIQA – A Semantic Web Information Quality Assessment Framework. In *European Conference on Information Systems (ECIS) 2011 Proceedings*.

Gürdür, D. (2017). *Making Interoperability Visible: A Novel Approach to Understand Interoperability in Cyber-Physical Systems Toolchains*. KTH Royal Institute of Technology, Stockholm.

Gürdür, D., El-Khoury, J., Seceleanu, T., & Lednicki, L. (2016). Making interoperability visible: Data visualization of cyber-physical systems development tool chains. *Journal of Industrial Information Integration*, *4*, 26–34. http://doi.org/10.1016/j.jii.2016.09.002

Haglin, D., Trimm, D., & Wong, P. C. (2017). Big graph visual analytics. *Information Visualization*, *16*(3), 155–156. http://doi.org/10.1177/1473871616679013

Hogan, A., Harth, A., Passant, A., Hogan, A., Harth, A., Passant, A., … Polleres, A. (2010). Weaving the Pedantic Web. In *3rd International Workshop on Linked Data on the Web (LDOW2010), in conjunction with 19th International World Wide Web Conference*. CEUR.

IEEE. (1998). *IEEE Guide for Information Technology — System Definition — Concept of Operations (ConOps) Document* (Vol. 19).

Ilacqua, C., Cronstrom, H., & Richardson, J. (2015). Learning Qlik Sense®: The Official Guide. Packt Publishing Ltd.

ISO. (2011). 26262: Road vehicles-Functional safety. *International Standard ISO/FDIS 26262*.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information Quality Benchmarks : Product and Service Performance, *45*(4), 184–192.

Mihindukulasooriya, N., Poveda-Villalon, M., García-Castro, R., & Gomez-Perez, A. (2015). Loupe - An Online Tool for Inspecting Datasets in the Linked Data Cloud. In *International Semantic Web Conference* (pp. 4–7).

Neto, C. B., Kontokostas, D., Hellmann, S., Müller, K., & Brümmer, M. (2016). Assessing Quantity and Quality of Links Between Linked Data Datasets. *Ldow2016*, 2–6.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2017). A comprehensive quality model for Linked Data. *Semantic Web*, *9*(1).

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches, *23*(4).

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management*, *37*(3), 150–154. http://doi.org/10.1016/j.ijinfomgt.2017.01.003

Wang, R. Y., Strong, D. M., & Taylor, P. (1996). Beyond Accuracy : What Data Quality Means to Data Consumers Beyond Accuracy : What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wood, D. (2010). Linking Enterprise Data (p. 291). Springer Science & Business Media, 2010.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., & Lehmann, J. (2012). Quality Assessment for Linked Data : A Survey. *Semantic Web – Interoperability, Usability, Applicability*, *7*(1), 63–93.

APPENDIX

Requirement	Rule	Dimension
Access to status	The server handling the RDF resource MAY forward requests with non-RDF ACCEPT headers (jpeg, html, etc) to the artifact URI, using the 303 status code.	Availability
Complete representation	A term (property and/or class) SHALL have a Turtle representation.	Completeness
Complete data structure	A term SHALL have an rdf:type property whose value is one of the following: • rdfs:Class – for classes • rdf:Property – for properties	Completeness
Complete data type	The URI of a resource, with some identification instanceID, of type someType (which is defined within the domain someDomain) SHOULD/SHALL follow a specific pattern.	Completeness
Complete information on media types	If an rdfs:seeAlso property is defined, the resource SHOULD also include dcterms:format property, whose range SHOULD be dcterms:MediaTypeOrExtent, which is from the list of valid Internet Media Types [MIME].	Completeness
No information on	The URI SHALL be treated as an opaque string. The URI structure	Representational

URI about the resource properties	SHALL not be interpreted to deduce any information about the resource properties.	Conciseness
Short names for the terms	A term (class and property) SHALL be rooted at the namespace URI of its containing domain. That is, the “short name” of the term is appended to the end of the domain URI.	Representational Conciseness
Reuse of existing RDF types	A resource SHALL have at least one rdf:type property to state that the resource is an instance of the specified rdfs:Class object.	Reuseability
Turtle representation	A resource SHOULD have a Turtle representation.	Reuseability
Reuse of existing vocabularies	A resource SHOULD reuse terms from common existing vocabularies (such as Dublin Core, FOAF, OSLC domain specification terms) – instead of defining Scania-specific terms.	Reuseability
Reuse of existing namespaces	For each Scania-specific domain, a unique namespace URI SHALL be defined, with a structure that matches a specific pattern.	Reuseability
No information about the software tools	The domain/server name part of the URI SHALL NOT reflect the technology, tools, or software being used.	Security
No information about domain/server names	The same types of resources SHALL NOT be distributed across more than one domain/server name.	Security
No information about organizational structure	The URI SHALL NOT reflect the organizational structure, nor project structure, nor a work breakdown structure.	Security
Show related resources	The RDF representation SHOULD include an rdfs:seeAlso (or dcterms:source?) property, whose value is a URI that points to the real non-RDF artifact.	Traceability
Traceable resources	A resource SHOULD have a dcterms:creator (referring to a dcterms:Agent) and a dcterms:created property.	Traceability
Understandable URI	The URI SHALL NOT contain uppercase letters, nor characters that require encoding.	Understandability
Understandable vocabularies	The vocabulary itself SHOULD be described.	Understandability
Understandable labels	A resource SHALL have an rdfs:label property, whose value should provide a human-readable version of a resource's name.	Understandability
Useful URI terms	A URI term SHALL use “hash URIs,” in which the URI is constructed by appending first a hash character (“#”) and then a “local name” to the domain URI.	Usefulness
Useful details	The URI MAY contain the creation date of the resource – the date the URI is issued.	Usefulness