Should you link (Linked)Data?
A quality assessment methodology

Jesse Bakker
Vrije Universiteit
De Boelelaan 1105
Amsterdam 1081 HV
j9.bakker@student.vu.nl

ABSTRACT

Quality of Linked Open Data is often approached with scepticism and this impairs the transition to a global data space. As quality can be assessed from different point of views, it is difficult to establish a distinction between good data and faulty data in the LOD cloud. There is need for justification of interlinking when information quality is fundamental. This research proposes a methodology by which such a justification can be formulated, based on a quality assessment. The methodology is tailored such that the assessment takes a comparative stance, with which a dataset, owned by a third party, can be assessed from the scope of another dataset. The resulting justification is in the form of a new quality measurement dataset, containing individual measurements and thorough documentation on related concepts. This dataset is intended to be used by both the assessor and the users of the linkset, to gain insights.

1 INTRODUCTION

The Semantic Web (SW) is an evolution of the current web of documents, to a web of data. At the heart of this lies Linked Data. Linked Data is a means to structure data in a way that facilitates interoperability and reuse of data. This is based on four principles, namely:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things

The grand idea of Linked Data, and what makes the SW a 'web' is the linking of data, especially between datasets. But to whom should you link to and how do you know if they actually fulfil your needs, be it a content or a quality need? Today, with the Linked Open Data (LOD) cloud ever growing, we are moving towards a global data space. The LOD cloud already contains over 38 billion triples, from numerous publishers, each with different methods for curation. Trust is an integral part of the SW[1]. The entire concept of linking to others is based on Trust. However, not all organisations have the luxury of making decisions solely on trust. Kadaster, the Netherlands' Cadastre, Land Registry and Mapping Agency, is one of those organisations. Kadaster, with its data, is to protect legal certainty and therefore, data quality is of the utmost importance. As an information broker, Kadaster advocates Linked Data and is widely implementing it on a Geo platform called PDOK, where the Kadaster and other governmental agencies can publish geospatial (Linked) Open Data. However, the ambition of PDOK is not just to facilitate data publication for governmental agencies, they want to stimulate innovative use of this data, by developers. Providing richer datasets might be a way to do so. Under the motto of Semantic Web, one can enrich a dataset by interlinking with an external Linked Data. If this were to happen, under no circumstance can it lower the quality of the platform such that it harms its reputation.

Interlinking linked data is done by means of linksets. Linksets contains triples that map one dataset to another. These linksets can either be dynamically created during query-time, where the linkset de-materialises after the query session. Or be created in a static form, independent from query sessions, where the linkset remains persistent. The latter can be offered as a service to consumers. First of al, the advantage of linking data is that information from multiple sources can be aggregated such that, previously unobtainable (or hard to obtain), insights can be obtained. The advantage of persistent linksets are (1) time saving in query execution time, the time it takes to set up the linkset is removed from the query execution time, (2) query efficiency and conciseness, as fewer lines are required for queries over multiple sources. These advantages provide an incentive for platforms to offer these linksets as a service. However, aimlessly creating linksets with every seemingly interesting dataset might do more harm than good as the quality of LOD has often been approached with scepticism. Therefore, the creation of a persistent linkset should be justified and substantiated by some means. Over the years, certain datasets (such as DBpedia[13]) have grown into the role of a 'linking hub', for more specific oriented datasets. Due to their size, their general applicability and the amount of links pointing towards them, the threshold to refer to such datasets lowers[3]. It is easy to make the assumption that a dataset is 'good' if a lot of other publishers refer to it. This research argues that, decisions based on trust is not enough, especially for datasets of less prominent publishers, and propose the use of a thorough methodology, based on quality metrics, to assess a dataset.

Quality is a fuzzy term and very task dependent, data which is considered to be dirty for one task, might be of ideal quality for another. In order to solidify the notion of quality to some extent, this research focusses on the use case of Kadaster. This research aims to investigate ways of assessing whether one dataset should be linked with another by addressing quality. Quality is often described as 'fitness for use'. For an open data platform a single use might be too narrow, as innovative use is advocated instead of a single use. Therefore, a more suitable definition for quality is 'fitness for purpose' where the context of quality is broadened to incorporate

[1] https://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html

this innovative use. By this notion, literature on quality is reviewed, transformed and extended to forge an extensive disambiguation of quality.

This disambiguation of quality helps to understand and structure metrics and constitutes the framework for this research. This framework is what gives meaning to the assessment and both are tied together in a single linked dataset as the primary artefact of this methodology. In this dataset, one can easily traverse from quality measurements, to metrics, to descriptions of quality dimensions and categories. In addition, this methodology includes a 'flagging feature' with which erroneous instances can be directly identified. The resulting graph is a form of metadata for the linkset. This fosters trust, by improving the utility of the linkset and reduce the potentiality of errors[8].

1.1 Research Questions

To the extend of the author's knowledge, no literature addresses the interlinking paradigm as mentioned in the introduction. Related research areas, such as Linked Data interlinking and quality assessment do address aspects such as linkset quality, dataset quality and matching techniques. This research argues that a crucial step in the interlinking process has been disregarded. Namely, the formal justification of interlinking. In order to capture this gap, the following research question has been posed.

How can a quality assessment help justify the interlinking of Linked Data and foster trust?

The research question is partitioned into four sub questions.

- What is the state of quality assessment methodologies for Linked Data?
- How can quality assessment for Linked Data be improved?
- How can a quality assessment be tailored to interlinking?
- How can a quality assessment be used to foster trust?

Interlinking two dataset is inherently, semantically, complex. This research aims to reserve the time and effort, required to complete this complex task, for only datasets proven to be worthwhile.

2 LITERATURE REVIEW

The notion of quality is present in nearly all research and is commonly referred to as "fitness for use" or "fitness for purpose". This definition is ambiguous for a reason, that is that, quality is not only context dependent, but also user dependent. Zaveri et al. reviewed twenty-one papers on quality assessment, published between 2002 and 2012, in Linked Data and identified a set of recurring quality dimensions [23]. In the following section, quality dimensions are described. Then, several approaches and methodologies mentioned.

2.1 Quality dimensions

Quality dimensions can be seen as the characteristics of a dataset, and a metrics are a means to measure such a characteristic. In Linked Data, six types of quality dimensions are identified and each consists of several quality dimension which can in turn be applied in the assessment.

Accessibility. The accessibility type group involves dimensions related to the measure of ease, with which, the dataset is accessed, traversed and retrieved. Availability, a dimension assigned to this group, concerns the extend to which information is present, obtainable and ready for use. Availability is measured by means of the presence of SPARQL-endpoints and RDF dumps. It also takes into account the dereferencability of URIs and whether the URIs resolve. Finally, availability can also be subjectively measured by detecting whether the content, and its metadata, is suitable for human consumption. Licensing, another quality dimension, concerns under which terms data can be (re)used. Licensing is a dimension, usually not taken into account for relational databases. Even more for Open Data, it is important to inquire under which terms the data is published. Information about the licensing of the dataset should be provided both in machine and human readable form. One of the most noteworthy facets of Linked Data is interoperability. Interoperability is exploited by means of interlinking. Interlinking involves creating triples where the object and subject refer to the same concept and the object and subject can either originate from different datasets, or the same dataset. The interlinking dimension refers to the extend where this the case. Several network measures can be employed to calculate the interlinking degree. Security concerns the extent to which (part of) the data is restricted from use, for instance when dealing with sensitive information. Security also shields the data from illegal alterations and the importance of this dimension therefore depends on the type of data, and its sensitiveness. The security dimension is measured based on access methods and or whether the data has a proprietor. A final dimension under Accessibility is Performance. How the source performs in terms of response and uptime depends on several factors such as: network traffic, server workload, server capabilities and query complexity.

Trust. The trust type group contains dimensions concerning the perceived trustworthiness of a dataset. Reputation is a dimension measured subjectively and often expressed as a score between 0 and 1. Reputation is the results of direct experience or word-of-mouth. Reputation is calculated by means of either a centralised authority or decentralised voting-system. Believability reflects the degree to which the data is considered true and correct. Methods to measure this include several versions of computing trust values and or using existing trust annotations. Verifiability is referred to as the degree of ease with which the correctness of the data can be assessed. This is measured by the availability of provenance information, digital signatures or the presence of judgements by (trusted) third-parties. Objectivity is defined as the degree to which the interpretation and usage of data is unbiased, unprejudiced and impartial. This dimension impossible to quantify and thus, only measured subjectively. Ways to do so include looking into the data and judge whether opinions or biases are embedded in the data, whether independent sources can confirm a specific fact or by checking whether the publisher is neutral.

Intrinsic. The intrinsic type group contains task independent dimensions. Accuracy, concerns the degree of correctness and precision with which information represents state of the real world. This can be divided into (1) Semantic accuracy, correctness in regard to real world values, (2) Syntactic accuracy, correctness in regard to its data model. Accuracy can be measured by comparing the data

to a defined gold standard. Another way to measure accuracy is by checking for violations of functional dependency rules. A third method is by comparing a fact in the data against other sources. These methods are classified as semantic accuracy. Methods classified as syntactic accuracy are the following. Checking whether literals adhere to their data type ranges or the lexical syntax of said data type. Lastly, the labels and annotations can be measured in terms of accuracy. Consistency concerns the presence of conflicting information with respect to particular knowledge representation and inference mechanisms. With the revelation of implicit knowledge, by means of inference and reasoning strategies, contradictions can be pinpointed. Metrics employing such strategies focus on violations of entities and their class membership, homogeneous data types, usage of classes and properties, ambiguous annotations and ontology hijacking (as described in [9]). Conciseness concerns the presence of redundant information in the dataset, either at schema level or instance level. This is measured by taking the number of unique attributes or objects, compared to its total. This way the presence of duplicates can be indicated.

Contextual. The contextual type group contains task dependent quality dimensions. Completeness concerns the extend to which the required information is present. Completeness can be measured on four levels, schema, property, population and interlinking. This dimension assumes the presence of a gold standard to which the dataset can be compared. Amount of data dimension compares the actual amount of data in the dataset, to the preferred, or appropriate amount of data for the task. One can measure this in terms of triples, instances and/or links. Appropriateness is expressed through (1) coverage, in terms of scope (no. of entities) and level of detail (no. of properties) (2) volume of data, in terms of size of the dataset and the amount of triples. Relevancy of data is defined as the extend to which information is applicable and helpful for a given task, and more specifically a query. Relevancy is highly correlated with conciseness dimension of data, since with the absence of redundant information, inadvertently implies that the dataset is either empty, or highly relevant. Relevancy is measured subjectively by looking at the usage of meta-information attributes and the retrieval of relevant resources. Furthermore, whether the dataset includes exemplary SPARQL-queries and an overview of the data models as well as the vocabularies used. It is also taken into account whether there is support for the dataset in terms of a forum and/or mailing lists.

Representational. Quality dimensions grouped under the Representational type involve the model over, and structure of, the data, primarily focussing on best-practice compliance and usability. Representational-conciseness concerns the extent to which the data is compact, clearly formatted and complete. Metrics belonging to this dimensions look at URIs and whether these contain superfluous information as well as prolix RDF features. Representational-consistency concerns the presentation of data as well as the (re-)use of well-known terms and established formats. This dimension contain both subjective and objective metrics. The assessor is to subjectively assess whether established vocabularies are used and objectively measure whether existing terms have been reused. The understandability dimension refers to the ease with which the data can be comprehended and used. Metrics focus on the use of labels and annotations attached to classes, properties and entities. Other metrics take into account the extend to which the meta-data is human-readable and facilitates comprehension. Interpretability is highly related to understandability, but covers a more technical aspect of information. It should also be noted that the more interpretable the data is, the easier it is to integrate it with other datasets. This dimension is measured by taking checking the use of globally unique identifiers for resources and the use of various schema languages.

Dataset Dynamicity. Quality dimensions grouped under Dataset Dynamicity type, concerns the temporal aspects of a dataset. Freshness of the dataset, how recently it has been updated/published is an important topic. Three dimensions are identified of this type. The currency dimension concerns how current the data is. It measures the speed with with data values are updated after real-world values change. Metrics in this dimension, measure the currency of document, age of values and the amount of outdated information in the dataset. The volatility dimension concerns the volatility of the dataset, or the duration in which all information is valid. It takes into account, the frequency of change in real-world values over time. Metrics measure this two aspects, the length of time in which values are valid, and the frequency of change over time. The timeliness dimension concerns how up-to-date the dataset is. The dimensions Currency and Volatility form the basis for this and metrics are partially reused. Metrics compute a value based on the difference between two dates, such as, current date and expiry date of information.

2.1.1 Quality Assessment. To this point, the results of ten years of research on Linked Data Quality dimensions have been presented. We can conclude, based on this, that there is relative consensus over a core set of dimensions, which are mentioned in at least three approaches, namely (and ranked in descending order of frequency)

(1) Believability
(2) Consistency
(3) Completeness
(4) Availability
(5) Currency
(6) Accuracy
(7) Conciseness
(8) Performance
(9) interpretability
(10) Representational-consistency
(11) Amount of data

The quality dimensions presented in Linked Data research has common grounds with relevant ISO standards. Precision is defined in the ISO standard ISO/IEC 25012[11] and is defined as 'the measure with which information is exact or discriminative enough for a specific task'. Precision seems to overlap with the Accuracy dimension found in Linked Data approaches as well as the Relevancy dimension. Accuracy is categorised as part the Intrinsic group, which implies that Accuracy is seen as context independent, while precision is the opposite. Furthermore this dimension is used to measure the extend to which the data is correct, whereas Precision entails data aspects such as resolution, rather than correctness. Therefore, a more 'accurate' association can be made with Relevancy. Relevancy is a Contextual dimension and is loosely defined as 'how well does the data fill a user's needs', in other words, 'how relevant is the data for a given task'. Whether or not information is relevant, among other things, depends on the precision with which

the data is obtained. High standard deviation of measurements might have large implications when dealing with parcel ownership, but the same standard deviation is less perilous when, for instance, plotting monuments on a world map. We therefore, assign Precision to the Contextual type group and acknowledge its association with Relevancy. A recurring issue in Linked Data Quality Assessment is the lack of maturity, as expressed earlier. This is reflected in the metrics, where across approaches, different measures and descriptions are provided. moreover, metrics are often only clarified by means of a short textual description without having provided any means to measure and/or interpret the metric. This research argues that, on a conceptual level, quality assessment of data should comprise of a single set of dimensions, regardless of the format of the data. The metrics, however, are tailored to the data, such that the format is leveraged to optimise the measurement.

2.2 Methodologies and frameworks

Few approaches in literature consider the entire scope of quality dimensions and instead only consider a small subset. This is mostly due to the fact that the quality dimensions are interrelated and its importance is dependent on the given task. Moreover, subjective preferences of the assessor highly influence the selection of quality dimensions[12, 14].

Several frameworks for Quality assessment have been presented and analysed in literature. Fewer full-fledged methodologies have been presented. In this section several methodologies and approaches for quality assessment of information are presented, both from the Linked Data domain and others.

One way to assess the quality of a dataset is to leverage the power of the crowd. Crowdsourcing has been proven useful for several other use cases such as tagging resources in the cultural heritage domain[17] and the creation of taxonomies[21]. Acosta et al. operated on the premise that crowdsourcing might also be a viable option for a quality assessment[1]. Their methodology gathers the crowd by financial and competitive incentives. The methodology follows a fix-and-verify[2] procedure where the crowd is (1) asked to identify problematic elements, (2) asked to fix identified elements and, (3) asked to verify fixed elements. A more intuitive approach for Linked Data is explored by Kontokostas et al. [18]. This research leverages SPARQL to measure quality indicators of a dataset. The key value in this approach, lies in the fact that even datasets of third parties can be, automatically assessed, given that the dataset is openly available over a SPARQL endpoint. This research defined a set of SPARQL patterns by which specific metrics can be automatically measured. With this approach, the authors were able to identify a substantial amount of errors in a repeatable fashion.

An issue with quality assessments as identified by Debattista et al.[4] is, that many quality assessment methodologies and frameworks create an output that is not machine readable. This makes it very difficult to compare quality assessments for multiple datasets. To solve this, LUZZU was described in the same paper. LUZZU is a tool that produces quality measurements in an interoperable format, understandable by machines, and uses a scoring function to rank quality assessment results. Both the research of [4, 18] have been identified as especially relevant for the research question of this research. The key features explored further in this research are (1) measuring metrics with use of SPARQL and (2) presenting assessment results in a machine-readable format.

Zaveri et al.[23], in their survey paper, presented basic steps a methodology for quality assessment should include. These were later clarified in [19]. The structure of the methodology is as follows. First a requirement analysis is performed, then the quality assessment, and finally a quality improvement step is performed. Folmer[6] discusses his own methodology consisting out of three phases. Only minor differences, on a conceptual level, exist with the methodology of[19]. The core parts comprise some preparation, an assessment, some interpretation and some quality improvement. The latter is deemed of less importance since, this methodology deals with data from third parties. This means that the data cannot be amended by the assessor, although suggestions can be made.

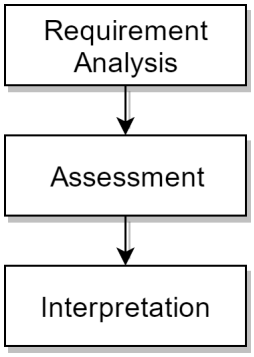
3 APPROACH

This research distances itself from the commonly accepted definition of quality as 'fitness for use', and instead advocates 'fitness for purpose' as the definition for quality. This research argues that with open data, such as the data published at PDOK, there often is no single task and the data is instead to facilitate information for a range of tasks. The approach this methodology embodies is different from traditional quality assessment methodologies, in the sense that assessing quality is *not* the primary goal. The end goal is to aid decision making and the quality assessment is a means to that end. The quality is not assessed within the context of a specific use or task, as one would argue for 'fitness for use'. Instead, possible uses are the outcome of the quality assessment. The goal of the assessment is not to measure whether the dataset upholds certain quality criteria, even though it is an important aspect. The main goal is to decide whether the dataset is qualified for interlinking, and with this provide insight to its possible uses. For instance, highly accurate geographical information has different applications than a less accurate geographical information. This research understands the following under fitness for purpose: "The extent to which a dataset adheres to specified standards and is applicable in selected domain(s)". This notion, captures not what quality is, but rather what we need to measure to be able to discern possible uses. If this set of possible uses is too small, due to unsatisfactory measured quality, this would imply the dataset is not suitable for interlinking.

After reviewing literature on Linked Data quality assessment, Zaveri et al. proposed a methodological framework for quality assessment in Linked Data. This framework comprises the following six steps: (1) Requirement analysis, (2) Data quality check list, (3) Statistics and low-level analysis, (4) Aggregated and higher level metrics, (5) Comparison and (6) Interpretation. The authors did not provide a thorough clarification as how to complete these steps, but this preposition does provide an interesting starting point for a methodology. Three constitutional parts of a quality assessment methodology are identified and depicted in the Figure 1. The first step in an assessment is always a requirement analysis, or sometimes also called 'preparation'. In this step, the assessor is to lay the foundation of the assessment by selecting a dataset, identifying relevant quality indicators and setting requirements or identifying a target quality. The second step comprises the actual

assessment of selected metrics and quality indicators. During the final step the measurements are interpreted and validated according to the requirements set in the first step.

Figure 1: Methodology Skeleton



3.1 Methodology

As identified in previous sections, a methodology consists out of three basic steps. Namely, a preparation, or requirement analysis, the actual assessment or measurement of metrics and lastly some interpretation of results. By tailoring metrics, which will be clarified in section 3.3, a quality assessment can be orchestrated to answer questions related to the interlinking of two datasets. This alone could answer part of the research question for this paper. In order to answer the second part of the research question, the assessment not only has to be leveraged to answer interlinking questions, but also to foster trust. For this, a publication step has been added to the methodology. This entails more than just attaching a seal of approval, or a grade to the meta data of the linkset. Each step in the methodology contributes to the construction of a dataset, in the form of linked data, with which not only the assessors can delve into the measured quality of the dataset, but also users of the (potentially) resulting linkset. Having the quality explicitly available, and queryable, alongside the linkset should not only foster trust, but also help identify strengths and weaknesses of the data. After completing every step in the methodology, the following artefacts have been created (1) A quality graph, containing definitions and quality measurements, and (2) A judgement on the creation of a linkset. In this section the methodology is briefly walked through and, for each step a rationale is offered. Afterwards, this paper will delve into the specifics.

The methodology is presented in Figure 2. During the initialisation of the methodology, four resources have to be gathered, for later use. These resources constitute (1) Metric Table, (2) Graph Template, (3) Shape Template and (4) Measurement Template

3.1.1 Preparation. The preparation phase commences with the selection of a primary dataset. The primary dataset is a dataset that is held by the issuer of the assessment and is not of questionable quality, for its curation process is known. The primary dataset is selected with the intended to be interlinked with an arbitrary externally held dataset. The second activity is the selection of a target dataset. The target dataset is a dataset that is not owned by the issuer of the assessment, but by any third party. The methodology assumes that there is some overlap between the primary and target dataset and that they can be interlinked, as actual interlinking is out of the scope for this methodology. The assessor is tasked with making sure this is the case and it is advised that resources other than this methodology are employed, a great starting place is [16, 20].

Now that both the primary and the target dataset have been selected, the assessor selects the metrics it wants to measure. For this methodology, 79 metrics have been documented as 'standard metrics'. These metrics are applicable to most datasets without much tinkering. It is advised to select as many metrics as possible as these metrics are used for more than just the quality assessment. A single metric says little about the overall quality of a dataset. Instead, they offer a single piece of a puzzle and, the more pieces available, the clearer the puzzle becomes. Even if the puzzle contains gaps, the picture can be distinguished, given that enough pieces are on the table. Having a comprehensive set of quality measurements, not only helps the assessor with completing the puzzle, it also offers more awareness to data consumers and diminishes uncertainty. The metrics are not only used for the quality assessment, but also as a form of metadata properties. However, these metrics cannot capture any unique properties of a dataset. Often, entities in a dataset cannot be accurately described by using standardised vocabularies such as FOAF[2] or SKOS[3] and more specific vocabularies and ontologies have to be created. The semantics within these specific vocabularies and ontologies cannot be leveraged in the quality assessment by standard metrics. With this, more gaps in the puzzle are identified. Now, that a set of standard metrics has been selected, and we know what is being measured (or not being measured), custom metrics can (optionally) be created. The process of doing so is clarified in section 3.3.3. Custom metrics are brought into life to be able to assess highly dataset, or domain specific quality indicators.

The penultimate step in the preparation phase is defining quality targets. Creating Quality targets help the assessor contemplate what measure of quality is satisfactory. Furthermore, quality targets aid during the interpretation steps in the Quality Assessment phase. With SHACL[4] the assessor can formality define quality targets as a shape. Such a shape then targets the measurement of a specific metric and automatically validates whether the measurement 'fits' the shape of the quality target and violations are reported. The creation of such shapes is clarified in section 3.3.2. Quality targets should not necessarily be defined for each metric, but rather for metrics where certain possible values are highly unfavourable. An example of such a case could be when dealing with a primary dataset where its validity is bound to a specific period of time and interlinking with data, where there is no temporal overlap would be a deal breaker. In such, and even in less extreme, cases a quality target could automatically evaluate whether any violations are present.

This methodology iteratively measures the quality of a dataset. As will be clarified shortly, the methodology contains three measurement iterations. Before the assessment starts, the graph is prepared for later use. In the graph compose step, the graph template (which

[2] http://xmlns.com/foaf/spec/
[3] https://www.w3.org/2004/02/skos/
[4] https://www.w3.org/TR/shacl/

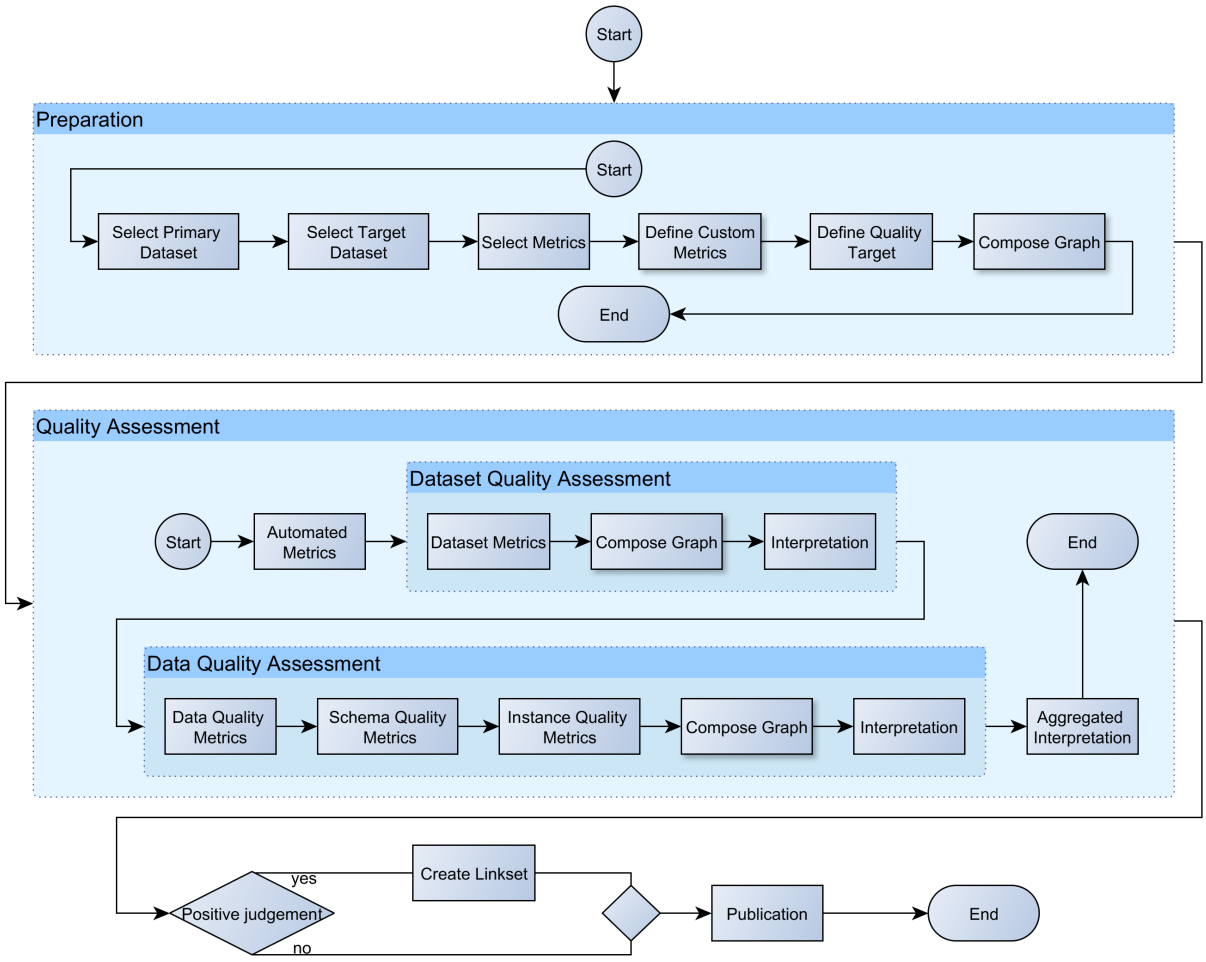


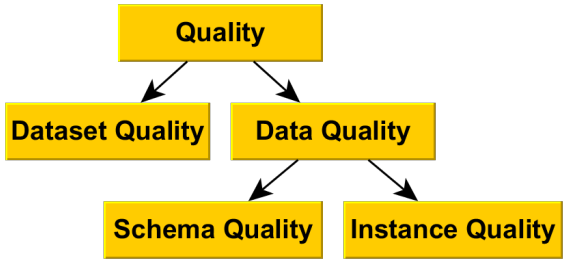
Figure 2: Methodology

is clarified in section 3.2) is extended with the information (selected datasets, custom metrics and shapes) generated in the previous steps. The composed graph, is then stored in an accessible manner for the assessor. The assessor should at least be able to perform SPARQL queries over the graph. With this the preparation phase ends and the Quality assessment phase commences.

3.1.2 Quality Assessment. Manually analysing large amounts of data is prone to errors. The quality assessment phase is, therefore, partitioned into two sub-phases, each with an interpretation step. An aggregated interpretation is also included where the result of each interpretation step is aggregated and interpreted as a whole. The partitions are created based on a taxonomy of quality, presented in Figure 3. The partitions are constructed by attributing terms from the taxonomy to metrics by means of a specified property *ex:hasMetricType*. This can then later be used to retrieve metrics of a specific type using a simple SPARQL query.

Quality is measured on two facets of a dataset. Dataset Quality, which concerns every aspect of the dataset except for the actual information it contains and Data Quality, which only concerns the information in the dataset. This is aligned with the Data Catalogue Vocabulary [5], where the notion of Dataset is reflected by *dcat:Distribution* and the notion of Data is reflected by *dcat:Dataset*. In Linked Data, data consists of two parts, the schema and the underlying instances. Based on that, data quality is specified even further, with Schema Quality and Instance Quality.

Figure 3: Taxonomy of Quality



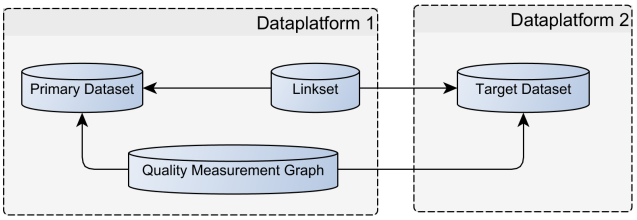
[5] https://www.w3.org/TR/vocab-dcat/

The assessment is ordered in such a way that the assessor starts from a broad scope and slowly narrows down to specific instances. This is beneficial for the interpretation steps since the broader the scope, the less circumstantial knowledge is required to be able to interpret the data. Many metrics can be automatically measured, as shown in section 4. These are "precomputed", before the following interpretation phases start, during the automated metrics step. This way, during the subsequent metric steps, only the metrics that have to be manually computed, are to be measured. During both phases within the Quality Assessment phase, one (1) measures the manual metrics, (2) retrieves relevant measured metrics and (3) interprets them as a set. During the "Dataset Metrics" step, the previously composed graph is utilised. The assessor can easily retrieve the metrics to be measured manually by means of a SPARQL query, such as A.1. In this case, the assessor retrieves all metrics of type "Dataset Quality" that are *not* already automatically measured in a previous step. Furthermore, these returned metrics can now also be dereferenced to find additional information about the metric. By composing the graph alongside the assessment, the assessor has access to an interactive reference guide with the information required to perform each task. The recurring 'compose graph' step is performed prior to steps that require newly created data, not yet present in the queryable graph. During the interpretation step, the assessor needs access to all the, not interpreted, information processed thus far. Since, this is all stored in the newly composed graph, the assessor can retrieve this by means of a SPARQL query as presented in A.1. Here, in the context of the "Dataset Metrics" step, one retrieves every measurement of metrics of type "Dataset Metric". Metrics of type Dataset Quality, mostly target the representation and the accessibility of the dataset, again, without considering any of the data included in the dataset. After a short summary of the interpretation step is created, the data quality assessment phase commences. Again, manual metrics are computed, while retaining this breadth-first approach. An interpretation step marks the end of this phase. Concluding, an aggregated interpretation step is performed where the results of both previous interpretation steps are combined and a judgement is passed on whether the two datasets should be interlinked. After this step, the assessor should be able to provide substantiated arguments on why, in terms of quality, the target dataset is deemed of sufficient quality.

During the interpretation steps, the assessor takes into account all measurements, relevant for the specific interpretation step (either Dataset Quality of Data Quality type metrics). The assessor attempts uncover the implications in terms of possible uses facilitated by the measured quality. The assessor also attempts to identify how the measurement aligns with the primary dataset (not in a sense that they can be semantically interlinked, but rather how the information is presented and constructed). All this, should be documented to allow for reuse during the aggregated interpretation step. This document is informal of nature. An example document for the Dataset Quality interpretation could include a description of how both datasets include a turtle[6] serialization, which is considered a human-friendly notation, but the target dataset lacks a JSON-LD serialization. This implies that JSON-LD based applications created for the primary dataset will have more difficulty using the potentially resulting linkset . A similar document will be created for the Data Quality metrics. During the aggregated interpretation step, the assessor has two documents describing inferred information about each metric. Here, the assessor assesses whether there is sufficient alignment with the primary dataset and if that, in combination with the identified possible uses, is adequate to propose the creation of a linkset.

Based on the judgement, a linkset is either created or not. Regardless, the publication step should be completed. After publication, the following situation, as presented in Figure 4, is established. Data platform 1 contains the primary dataset, the *potential* linkset and the quality measurement graph. The linkset and the quality measurement graph, both point to the target dataset, published on data platform 2. There are several reasons as to why the graph should be created (and published), regardless of whether a linkset is created. As a data platform, a goal is to serve users and support them in their usage of the data on the platform. For Linked Data, an important usage is the interlinking with other datasets. It is likely, that users combine data on the platform, with other data. When a negative judgement is passed on the combination of two datasets, a quality graph can be used to indicate to users, that creating a linkset themselves is a slippery endeavour, with a clear indication of why. Furthermore the quality graph can be utilised to mend the faulty dataset since, it indicates where the dataset lacks quality and which specific instances are deemed erroneous.

Figure 4: Graph Storage



3.2 Graph Template

One of the artefacts created during the execution of this methodology is a graph, containing quality measurements, definitions of metrics, quality dimensions, quality categories and requirements. In order to aid in the creation of such a graph, a template is provided. As identified during the literature study, the notion of the higher level concepts of quality are roughly domain independent. Furthermore, the same holds for metrics as a concept. Across use-cases, the principle behind the metrics hold, whereas the measurement procedures might differ. Precisely these concepts are captured in the quality graph. Since, the graph captures this domain independent concept of quality, it can be included in the template. Even when for specific use-cases concepts have to be tailored, these alterations can be easily included in the graph. In addition, the graph template also features example measurement instances and requirement instances. The graph consists of two parts, the metadata and the data. A representation of the graph is provided in figure 7.

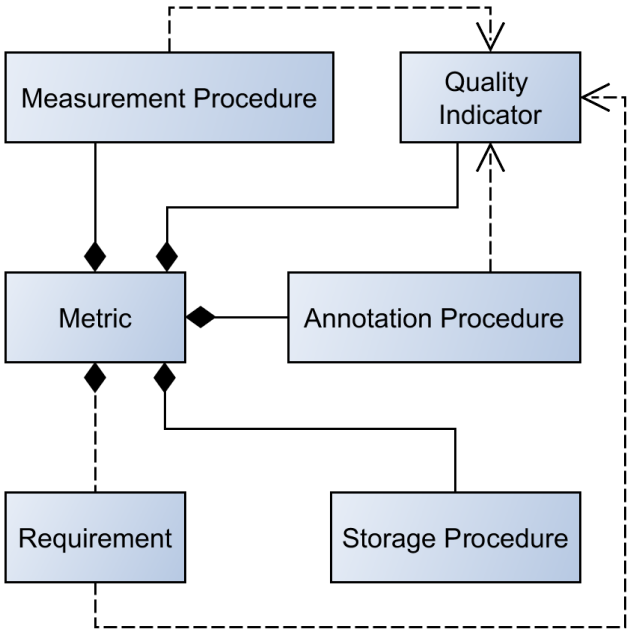
[6] https://www.w3.org/TR/turtle/

3.2.1 MetaData. The metadata of the graph contains a structured representation of what is in the graph and provides a means to link both the primary and the target dataset to the quality measurement graph. The graph is represented by the instance *ex*:*QualityMeasurementDataset* (qmd) of type *dqv*:*QualityMeasurementDataset*. The Data Quality Vocabulary[7] (DQV) is used to semantically enrich the quality graph. DQV is a vocabulary, published as a best practice by the w3c[8]. DQV a framework which allows for describing quality, metrics and measurement of metrics. The DQV vocabulary is highly compatible with what is required for the quality graph as defined for this methodology. The instance qmd has a target dataset and a primary dataset, defined by *ex*:*hasTargetDataset* and *ex*:*hasPrimaryDataset*, respectively. Each qmd has exactly one primary dataset, but one or more target datasets.

The primary dataset governs the notions of 'fit for use' and what it entails. This could lead to tailored definitions used in the quality graph. Since the purpose of the quality assessment is the interlinking of the primary dataset to the target dataset, the target dataset is to adhere to the primary dataset and its notion of quality. Each target dataset is measured according to a fixed notion of quality when each use case contains the same primary dataset. Therefore, only one quality graph is created for each primary dataset. It could then, include multiple target datasets. In order to differentiate between, and keep track of the provenance of the different target datasets and their respective measurements, the *ex*:*Assessment* instance is introduced. This instance, is of type *prov*:*Activity* and describes who, measured what, on which dataset.

3.2.2 Data. The content of the dataset is represented in the data section of the graph. The data largely comprises definitions of metrics, quality dimensions and quality categories. Which can be tailored at will. These definitions are to aid in the understanding of both the measured value and the goal of a metric. Measurements are instances of the class *dqv*:*QualityMeasurement*. Through the property *dqv*:*isMeasurementOf* the respective metric can be retrieved. This metric instance contains a definition, an expected datatype and a metric type. The latter is used to retrieve each metric relevant for a specific interpretation step. Each metric is also attributed to one or more dimension instances, of type *dqv*:*Dimension*. Each dimension is included a definition and is attributed to a category, of type *dqv*:*Category*. Categories, again, include a definition. With this, each measurement can be thoroughly described and understood. Lastly, the data includes requirement shaped, defined in SHACL. These shapes are of type *sh*:*NodeShape* and target a specific metric. This means, that the requirement holds only for that specific metric. The requirement also contains a property where the actual restraints are specified. The *sh*:*property* makes use of the *sh*:*path* property in order to retrieve the *dqv*:*value* (the attribute containing the measured value) property of the measurement related to the targeted metric. The *sh*:*path* property specifies the route to be taken to reach a specific resource. The graph Template features example shapes in order to validate integer value ranges, boolean values and IRI values for *dqv*:*value*. Each measurement instance also includes some provenance. This helps to differentiate between different quality assessments on the same data. Each measurement is linked to a specific *prov*:*activity* instance by means of the property *prov*:*wasGeneratedBy*. Each *prov*:*activity* depicts an entire quality assessment of a target dataset. Multiple assessments on the same target dataset should each be attributed to their own unique *prov*:*activity*. This also allows for keeping track of the time line of the entire graph and when, which target dataset was assessed for a specific primary dataset. Lastly, assessors can, optionally, include a *skos*:*note* in order to clarify some aspect of a measurement.

Figure 5: Metric Decomposition



3.3 Metrics

Metrics capture the essence of the concept of 'fitness for purpose'. In previous sections, the notion of quality has been dissected into quality categories, and those in turn, into quality dimensions. This allowed for a guided assessment where specific aspects of a dataset could be targeted and partitioned into interpretable chunks. These quality dimensions are applicable to almost all datasets. Metrics, however, are more difficult to generalise such that they are universally applicable. Even if a metric can be described in a universally applicable manner, on a conceptual level, the actual measurement of the metric might still be dependent on the given use case. Other, highly specific, metrics are only applicable for a given use case. These are so-called custom metrics and will be described, as well as the process of creating them, in section 3.3.3.

Zaveri et al, in their survey paper, gathered 96 metrics. Most of which are briefly described or explained by means of a formula. This resulted in a silhouette, of sorts, of a metric. It provided handles with which a user can create a manifestation of such a silhouette. The silhouettes however, were often blurry and lacked explicit

[7] https://www.w3.org/TR/vocab-dqv/
[8] https://www.w3.org/

Table 1: Metric Table

	Component	Description
1	Name	Name of the Metric.
2	Source Reference	List of relevant resources.
3	Dimension	List of Dimensions under which the metric
4	Tags	List of Tags applicable to the metric. Examples are: Dataset quality, Data Quality, Schema Quality, Instance Quality, Automatic, Semi-Automatic, Manual, Objective, Subjective.
5	Description	A Description of the metric. This should at least cover the goal and roughly the measurement process.
6	Value Type	Value type of the result of the metric. An example is 'Boolean'.
7	Value Structure	For more complicated Value Types, a structure can be provided.
8	Measure Function	A Measurement function for the metric. Pseudo-code, Formulae and SPARQL queries can be used to describe the Measure Function.
9	Measure Element	Individual elements used in the Measure Function, should be disclosed here.
10	Example	Illustration of a usage of this metric.
11	Annotation Procedure	Description of the Annotation Procedure. Should cover why, when and what to annotate.
12	Identifier	A unique identifier for the Metric, in the form of a URI.

details. By consulting the original resource, where Zaveri et al. retrieved the metric, one could often expose the silhouette's details, as is expected of a survey paper. However, often, references contained broken links and others were shielded by a pay-wall. This obstructed the usage the respective metrics. Nonetheless, this research has attempted to improve the silhouette of the metrics, as well as extending the list of metrics. By reviewing not only papers presented under the Linked Data flag, as mentioned in section 2, but also from other domains, the list of metrics is extended. In addition, based on the ISO standard[10], providing a thorough description of each metric.

Figure 5 provides a decomposition of a metric. As pictured, a metric consists of three procedures, a quality indicator and an optional requirement. A metric constitutes more than just a quality indicator and a description on how to measure it. Apart from formally describing both aspects, a metric is reinforced with a description on how to store this information in the graph, and a description on how to enrich the target dataset with annotations based on the measurement. Including, an optional, definition of a requirement. The Metric Table, presented in table 1, allows for a structured and extensive description of a metric. Two facets are added to the original table, one which introduces the annotation procedure and another to introduce tags.

The table provides a definition of the metric and its quality indicator and groups it to a dimension. Furthermore one or more tags are selected, with which, the methodology partitions the quality assessment. The table describes the measurement procedure by describing the expected value and its structure, and a method for computing that value. The measure function row is usually described using pseudo-code or formulae. Furthermore, an illustration of a usage of the metric is provided to enhance understanding. The metric table also includes a description of the annotation procedure. Here it is explained when resources should be annotated, or flagged and how this should be done.

3.3.1 Flags. Key, in this research, is fostering trust of users in the data platform and understanding what is linked to by the data platform. This methodology offers more than numbers in its quality assessment. Due to the Linked Data principles, especially having URIs as names for things, it can quickly be picked up which explicit resources the assessment is dealing with. During the measurement procedure, it can easily be tracked, precisely which "things" negatively affected the metric and work with them. A flag constitutes the following triple pattern:

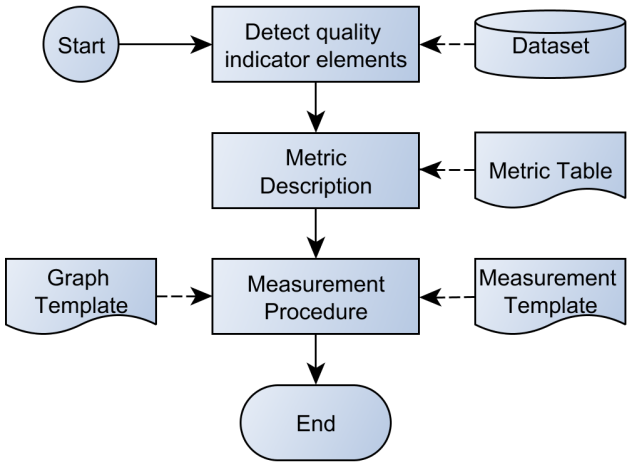
$$?metric\ ex{:}flags\ ?IRI$$

By annotating the metric instead of the resource in question, ontology hijacking[9] is avoided. In the metric table, it is described when a resource is to be flagged and this is metric specific. Additionally, it describes how resources should be flagged in accordance with the measurement procedure. Flags facilitate two use cases. One, where the user of the data platform can identify discrepant resources and can create a more robust application based on this knowledge. The user can, for instance, identify incomplete instances and even for instances missing a crucial property, given that a custom metric is provided for specific property. Knowing that the application, created by the user, will not work for the flagged instances, the user can act appropriately. The other use case includes the original owner of the target dataset. Flags identify flawed resources. Since, these specific resources are retrievable, the data owner can execute a targeted amelioration of the data.

3.3.2 Quality Targets. Even when metrics are clearly defined, it can be difficult to interpret the measurement of a metric in any given context. There is a need for a definition of what constitutes a satisfactory value for a measurement. This methodology defines quality targets to take on this role. Quality targets are to denote the minimal satisfactory value for a measurement. It is up to the assessor to define such values and can consult with other stakeholders and domain experts. When the quality targets are defined, the assessor should be able to quickly identify which measurements have not passed their quality targets and use these targets during the interpretation to assess to what degree the measurements adhere to the quality targets. Quality targets are included in the graph.

The quality target have three tasks, as mentioned in the previous paragraph, namely to be able to validate, compare and be retrievable in the graph. Each of these tasks van be facilitated by SHACL. SHACL allows for representing the requirements in RDF and is able to validate requirements to resources it targets. In SHACL,

Figure 6: Metric Creation Process



requirements are referred to as shapes. Each shape roughly consists of three components, namely: (1) *rdf*:*type*, (2) *sh*:*targetNode* and (3) *sh*:*property*. A shape in SHACL is of type *sh*:*NodeShape* or *sh*:*PropertyShape*, depending on what the shape is targeting. *sh*:*PropertyShape* is used when the shape is to target a property(or "edge" in terms of graph theory[22]) *sh*:*NodeShape* is used when the shape is to target resources other than properties. In other words, nodes. Shapes defined in the quality measurement graph are always of type *sh*:*NodeShape*, for the shapes target metrics. Each shape targets a resource to which the respective shape applies by means of the *sh*:*targetNode* property. For this methodology, shapes always target a metric. *sh*:*property* is a blank node and contains the attributes considered requirements and each shape can contain several. In this blank node a path is given, by means of the property *sh*:*path*, and with this path SHACL can point to specific attributes of the target node. More elaborate paths allow the shapes to traverse the graph with the target node as starting point. The following triple pattern is used to traverse the graph and retrieve the value of the measurement of the metric targeted by the respective shape.

ex:*requirement sh*:*path* (*ex*:*hasMeasurement dqv*:*value*)

Values for measurements are commonly integers, floats, booleans or IRIs. Each of these can be assessed with SHACL. Furthermore, the shape can also validate datatypes with the use of the *sh*:*datatype* property. Illustrative usage is provided in the shape template.

3.3.3 Custom Metrics. Standard metrics fail to identify case specific quality indicators. Custom metrics are created to target such case specific quality indicators and are able to leverage non-standardised schemata and constructions. The process of creating a custom metric is illustrated in Figure 6. The construction of a custom metric starts with the identification of a quality indicator where the target dataset is used as input. Then a description of the metric is written based on the metric table. After the table is completed, the measurement procedure is written. This should be done based on the specification in the metric table and in code. Measurement Procedures are likely to consist of a combination of a script and a SPARQL query as will be shown in the experiment section of this paper. The measurement procedure also takes as input the graph template and the measurement template which depict how the computed result is to be stored.

In Figure 8 the previously explained process of creating a metric is more thoroughly described. Here, the "Fill Metric description" activity is broken down and ordered in such a way that the assessor gradually solidifies the metric. This is done by, first addressing what and how the quality indicator should be measured, and ending with a concrete definition. The measurement procedure is also broken down. Furthermore, the dependencies between activities are shown. By using this structured approach to create unambiguous custom metrics, re-usability is improved. Possibly, with minor adjustments, custom metrics can be tailored to apply on other dataset as well. Given the artefacts created during the process, understanding of the metric is retained (and/or regained) more easily.

4 EXPERIMENT

For the experiment the entire methodology was executed. 25 standard metrics were measured, three custom metrics were created and measured, and three quality targets were defined and validated. Measurements were conducted in Python[9] and used a library[10] to fire SPARQL queries to an endpoint from the code.

As mentioned in the introduction, Kadaster, the Netherlands' Cadastre, Land Registry and Mapping Agency, is taken as the use case of this research. Kadaster concerns itself largely with geographical data. The Basisregistraties, Adressen en Gebouwen (BAG) dataset[11] is chosen as the primary dataset and a cbs dataset[12] as the target dataset. The BAG contains base registries about buildings and addresses in the Netherlands and is published on PDOK as linked data. The cbs dataset contains statistics about "buurten", or neighbourhoods, in the Netherlands during 2015. Following the methodology, some overlap has to be identified to be able to assume the two datasets can be interlinked. Transforming the assumption into fact, is beyond this methodology. Both datasets deal with geographical data in the Netherlands. Geometries of specific buildings and their locations are known in the BAG and geometries of neighbourhood are known in the CBS dataset, including their location in the Netherlands. Since, both of these geometries are known, they can be implicitly linked by checking for inclusion of a building in a neighbourhood with GeoSPARQL. With this, we can assume that the two datasets can be interlinked and the methodology continues. Preferably, as many metrics as possible are measured during the assessment. For the experiment the goal was set at thirty metrics, of which ten percent should be custom metrics. This number is substantial enough to be able to draw some conclusions and small enough to be executable within the time frame of this project.

Many metrics can be automatically measured. For instance, a reasoner can pick up semantic inconsistencies and syntactical errors. Tools such as [7] attempt to automatically compute trust metrics and other tools such as Sieve[15] take a more complete approach

[9] https://www.python.org/
[10] https://rdflib.github.io/sparqlwrapper/
[11] https://data.pdok.nl/datasets/bag
[12] https://data.pdok.nl/cbs/2015/

to quality assessment. Both operate semi-automatically as several metrics are very difficult to automate. This experiment follows a similar approach, where the assessment is semi-automatically executed.

```
Data: Target Dataset
Result: Measurement Instance
initialization;
if Uses SPARQL then
    set up SPARQLwrapper;
    fire SPARQL query;
    parse results;
end
if Manually measured then
    request user input;
end
else
    obtain data;
end
compute value;
parse value;
input value to measurement template;
if Annotation Procedure then
    create annotation;
    aggregate annotations to measurement instance;
end
```

Algorithm 1: Pseudo-code single Measurement Procedure

For the experiment, several metric have had their measurement procedures formalised in Python. This resulted in a measurement script that can compute values for metrics and automatically transform them according to the graph and measurement templates. This script returns valid RDF triples which can be instantly pushed to the graph. The Measurement script consists out of three parts. (1) initiation, where endpoints and other variables are set. Including a description of the graph. (2) execution of measurements, where each defined metric is computed and parsed according to templates, and lastly (3) input of created instances to the graph. Measurement Procedures are described in pseudo-code in Algorithm 1. Three types of measurements are identified based on the way data is obtained. Metrics either require information within the target data, which is acquired with SPARQL queries. Or, information not in the target data, which also cannot be automatically obtained, require manual efforts. Lastly, metrics can require information not present in the dataset, but can be measured automatically. Examples would be, measuring usage of specific metadata, population completeness or measuring whether resources are dereferencable, respectively.

Each artefact, metric tables, quality measurement graph, templates and measurement script, is published on a github page[13] as an appendix. The measurement script was created based on the metric tables. Well-defined metrics resulted in a relatively easy conversion to actual code. What was not captured by the table was how to reach only the target dataset behind a SPARQL endpoint containing multiple graphs. By using the SPARQL operator GRAPH, this issue could easily be resolved.

From the interpretation steps, several useful insights are gathered. The custom metric, geometryBlankNode, depicts the percentage of entities with a geometry attribute, which store this geometry as a blank node. The target dataset scored 100%. Furthermore, the geometries were consistently presented and only one percent of the entities were identified as incomplete. A representative amount spatial entities were included in the dataset in order to be comparable to the primary dataset. None of the requirements were violated. The judgements for the remaining metrics were positive and none could be used to argue against interlinking of the two datasets. Based on this, the assessor formulates a judgement, advocating that the two datasets be interlinked.

5 DISCUSSION

Creating linksets is inherently difficult. This largely has to to with the fact that determining whether two terms are synonyms, in such a way, that *owl:sameAs* correctly defines their relation is difficult. For instance, are *bag:Pand*[14] and *top10nl:Gebouw*[15] semantically equivalent? Both describe buildings, but whether both datasets ascribe the identical interpretation to the term, is difficult to identify. Many attempts have been made to automate the creation of linksets, with some success, by the OAEI community[16]. Never have perfect linksets been created by such tools. Consulting domain experts (for both datasets) is likely an expensive endeavour and in many cases not feasible. This complexity, and the attached financial restrictions, provides an incentive to utilize this methodology. For, it allows to reserve this expensive process of interlinking only for qualitatively satisfactory datasets. As well as abstaining from investing, both time and resources, in datasets of unsound quality. This research does acknowledge that it makes no sense to measure the quality of a target dataset that is not linkable to the primary dataset. Again, the fact that the assessor should briefly assess whether the two datasets can be interlinked after the target dataset has been selected, is emphasized.

This methodology is based on quality assessment literature. Many aspects of this methodology are, therefore, also very similar to "regular" quality assessments. This allowed for the reuse of the structure of other quality assessment methodologies and vocabularies created for that purpose. There are, however, several aspects in which this methodology differentiates itself from others. First of all, the mindset expected for this methodology is that the assessor is not actually solely interested in the quality of the target dataset. The assessor should clearly understand the goal of the assessment, namely to identify its suitability for interlinking to the primary dataset. This mindset should then be reflected in the interpretation steps. However, this change of mindset could also be facilitated with slightly tailored existing methodologies, even though it is an important aspect of this methodology.

The larger part of metrics identified in literature were far from effortlessly implementable, due to the lack of documentation and often complex semantics. Substantial effort of this research has

[13] https://github.com/BakkerJesse/MasterThesis

[14] http://bag.basisregistraties.overheid.nl/def/bag#Pand

[15] http://brt.basisregistraties.overheid.nl/def/top10nl#Gebouwd

[16] http://oaei.ontologymatching.org/

gone into formalising metrics and this has been done in a structured approach. More than formalising, the metrics, if possible, have also been tailored for the given scope of this project. Tailoring metrics resulted in metrics taking into account not only the target dataset, but also the primary dataset. The use of SHACL also fits this purpose. Not only can requirements embody a baseline of sorts, such as a minimal required integer value. But SHACL shapes can also be used to validate whether a quality indicator of the target dataset is conform some value based on the primary dataset. For instance, it can easily be checked whether the target dataset is described in the same language as the primary dataset. Flemming[5] identified versatility in available languages in a dataset as a quality metric. This methodology is more interested in whether both datasets are described in the *same* language, such that they can more easily be interlinked. This can be measured by firing a SPARQL query to both dataset, retrieving all languages used and computing the intersection. This could get rather complex very quickly. Another approach, using SHACL, would require only retrieving the language(s) of the target dataset, and automatically validating whether it is conform the set of languages specified in the respective SHACL Shape.

Often, metrics offer little insights by themselves. For instance, the metric ”amount of triples”, where its name is indicative to what is measures, says little about the quality of the target dataset. Extreme values could still provide insights (such as a dataset containing one triple). But, knowing a dataset contains 100.000 triples offers little by itself (knowing the dataset contains 100.000 entities, in addition, would make a world of difference). This metric serves as a tiny building block of which the interpretation constitutes. Such metrics also help measure, and give meaning to other metrics. For instance, we can measure the ”descriptiveness” of the dataset, by dividing the amount of triples, by the amount of entities in the data. This tells us, on average and approximately, how many properties are provided for each entity. This gives us insight, but hardly touches the quality of the dataset. Instead, it is the combination of numerous metrics, and the insights they provide, which allow the assessor to make an educated guess of the quality of the target dataset. Without any information about the procurement and curation of the data, nor availability of a ground truth, it is very difficult to assess a dataset. This methodology uses as many quality indicators as it can measure to formulate a quality judgement. The metric ”indicativeness” has also been introduced by this methodology in order to measure, in a way, the completeness of the assessment. This metric assumes that, the more standard metrics are measurable, the higher the quality of the dataset. This is based on the fact that metric tend to require specific information to be available in the dataset (such as ”attribute accuracy”, which denotes the precision with which measurements were made). If such information is not present, the respective metrics cannot be measured (it cannot be uncovered how precise measurements were done, without the data owner explicitly disclosing this). This would then negatively affect the metric ”indicativeness”.

The quality graph is constructed in such a way that, users can easily discover the underlying meaning of each measurement, by consulting optional notes and adjacent nodes. From any measurement, the user can travel to the respective metric and inform about the meaning of said metric. Then, the same could be done for its quality dimension and in turn the quality category. Each are supplemented with definitions and optional notes. The user can take any route, or request any information with SPARQL. The same holds for the assessor. In order to aid in the interpretation of numerous measurements (as is the goal), several annotations are provided by which the assessor can partition the metrics. In addition to the metric types, as required by the methodology, the assessor can also group metrics on other facets such as dimension, category, expected datatype, whether it has a shape, which metrics were *not* measured, which were measured on a particular distribution ect. The graph offers many edges which both the user and the assessor can make use of.

The quality graph has been designed for extendibility. An important part of the methodology is the inclusion of custom metrics, and the versatility it offers. With the custom metrics, the assessor can virtually measure any quality indicator, relevant for the given use case. Custom metrics are constructed with the same components as standard metrics and therefore fit the graph template perfectly. With the guided process of creating custom metrics, the quality measurement graph can easily be extended. At the heart of the quality measurement graph lies the primary dataset. Each dataset, given the role of primary dataset, can only do so in a single quality graph. Each combination of a specific primary dataset and an arbitrary target dataset is stored in a single quality graph. The provenance activity, which is a container for a single quality assessment, is used to link target datasets and measurements computed based on the respective target dataset. Furthermore, measurements point directly to a dataset or its distribution to denote on what the measurement is based. Instances of *prov*:*Activity* always point to the dataset as an entity, whereas measurements can point to either the dataset as an entity, or one of its distributions.

The provenance is used to specify the temporal aspects of the assessments and differentiate measurements between assessments and target datasets. Moreover, it is in compliance with the prov-o[17] specification. In Figure 7, it is stated that every shape (*ex*:*Requirement*) is created during an *ex*:*Assessment*. During the first assessment with a given primary datasets this offers no issues, shapes are can be created for both standard and custom metrics and this holds no integration issues. But, what happens when another target dataset is included in the graph? Following the methodology, when new shapes are created and measurements are conducted, they are linked to a new *ex*:*Assessment*. However, the shapes are designed to target the metrics directly, which are independent of the target dataset. This means that every shape, constrains *all* measurements of the given metric, including future and past measurements based on every target dataset. This reflects the core of this methodology and how it differentiates from traditional quality assessments. Although the quality of the target dataset is measured, the primary dataset is what governs the assessment. The shapes should therefore not be created for a target dataset (with the exception shapes targeting custom metrics), but rather, reflect what is required for the primary dataset. This way, the created shapes remain relevant, regardless of the target dataset in question. It could occur measurements from previous assessments, do not comply with a newly created shape. Regardless, the outcome can be used to enrich the quality graph,

[17] https://www.w3.org/TR/prov-o/

even for previous assessments. Still, if the affected assessments were properly conducted, the inclusion of the new shapes should not have major implications. Since shapes are independent of target datasets, and created solely for the primary dataset, the most important, and influential, shapes are likely to be identified and created during the first or second assessment.

Usability of the methodology has been key during the design of the graph template. The metrics and the requirement shapes could have been merged. Such that, there would be a single instance, both of type *dqv*:*Metric* and of type *sh*:*NodeShape*. This would have been possible since, both classes are not disjoint. This way, a single instance would define the metric *and* denote restrictions posed on the desired outcome of the measurement. However, these have been kept separated. Having separate instances for both metrics and requirements allows us to easily work with them independently from each other. Requirements can be added without having to alter an existing metrics, and more importantly, they can easily be deleted if found obsolete. Had both been merged, these operations would have been more cumbersome. Metrics are intended to be static, and only rarely be subject to change. Whereas, shapes are dynamically created prior to the assessment to provide handles for the interpretation.

The measurement script, used for the experiment, has been designed in a generic fashion. Semi-automatically, measurement instances and flags can be created. Furthermore, it is possible to target specific named graphs, which can be specified in the script. The creation of the actual instances is done according to a measurement template, to which the script does string insertions with specific information for a given metric. Good practice would dictate the use of tools, to create valid RDF that is sustainable and more easily editable. This could be a topic for future work. The creation of such a measurement script, requires expert knowledge. The assessor should be able to write both code and SPARQL queries. Preferably, a more user-friendly approach would be taken to transition from descriptive measurement procedures, to an actual script or tool to perform the measurements.

The methodology and the resulting artefacts are designed to not only help users choose datasets to link to, but also grow trust with users of the data platform. This is the added value that is achieved by the quality measurement graph. The added value brought by the methodology is twofold. First, value is generated for the business, by providing insights on the quality of the data platform. Businesses are often well-versed in curating their own data, but now a means to assess data of other's datasets, which inadvertently are connected to the data platform after interlinking, is also presented. Furthermore, the quality measurement graph is a service for the data consumer (users of the data platform). Data consumers obtain added value from the quality measurement graph, for it can be used to identify possible uses of the data, by identifying weaknesses, strengths and characteristics.

As future work, the following items are proposed. First, the resoluteness of individual metrics tend to be insignificant, as has been counterbalanced by increasing the amount of metrics to be measured. The author proposes more effort to be invested in the investigation and implementation of more elaborate metrics. These should by themselves, offer more significant insights, but not replace less elaborate metrics. Instead, the list of metrics should merely be extended, and the assessor should still strive to measure as many as possible. This is because any metric could hold key information for the data consumer. Even trivial information provided by a metric, such as amount of instances, could be of importance to a data consumer.

Ease of use, has been an imports factor in the creation of this methodology. The author has attempted to facilitate users, of the methodology, as much as possible. This lead to the inclusion of several process descriptions and templates. These templates provide examples of all facets of the resulting quality measurement graph. Furthermore, metrics have been thoroughly described using the metric table. However, how to convert the descriptions of metrics into executable pieces of code is left rather ambiguous. The metric tables do provide all the components of a metric and how they can be leveraged in a measurement, but the actual implementation can still be rather challenging. Therefore, as future work, I propose that code snippets be created for each metric, in an accessible repository such that they can easily be reused. This should further improve the ease of use of the methodology.

The resulting Quality Measurement Graph is designed to be highly understandable and interpretable. To do this, definitions have been provided for every metric, their quality dimension and quality category, as well as the possibility to include a *skos*:*note* at measurements and requirements of choice. However, this context, might still be too laconic for specific users. Therefore, a user study is proposed where the information need identified and investigated where the graph template is lacking in descriptiveness.

6 CONCLUSION

During the preliminaries of this project, a disambiguation of quality was sought after. After consulting literature from various domains and sources a set of recurring quality categories and dimensions are identified and described. Metrics were identified to be lacking in formality and documentation. By generating a metric table for each metric an understandable and easily measurable representation was created. Next, input from multiple quality assessment methodology was obtained and this lead to the construction of the methodology as presented in this research. The methodology has been validated in terms of executability by means of an experiment. This resulted in an implementation of the proposed methodology of which the created artefacts are found in the appendix on the following github page[18].

A large portion of the metrics proved difficult to measure, especially for data from third parties. Furthermore, this difficulty is amplified by the fact that making the metrics measurable, requires knowledge about both programming and SPARQL. Furthermore, the metrics used only offer little resolution by themselves. This methodology compensates for this by measuring *more* metrics. The combination of multiple metrics help form a picture of the state of the quality of the target dataset. This in turn made the interpretation more complex which lead to the partitioning of the interpretation. The experiment showed that the assessment can be semi-automatically, and repeatably computed. This is due to the fact that metrics use generic SPARQL patterns (if applicable) in order to retrieve the necessary information needed, to compute a value

[18] https://github.com/BakkerJesse/MasterThesis

for the metric. The metric tables proved to be beneficial during the creation of the measurement script. Although, specific metrics can be inherently difficult to measure. The resulting graph, can be seen as metadata to the linkset and provides extensive information for both the assessors and the data consumer.

This research was conducted with the following research question in mind: *How can a quality assessment help justify the interlinking of linked data and foster trust?*The short answer to this question states that with the aid of a tailored quality assessment, one can devise a more established judgement on whether two datasets should be interlinked. This, by addressing the alignment with the primary dataset and the possible uses (allowed by the measured quality) of the target dataset. Furthermore, by storing the measurements with supplementary context in an accessible manner, data consumers are able to do their own investigation which cultivates their trust in the data platform *and* allows them to make better use of the data. The experiment provided a wealth of facts about the target dataset, which would otherwise have remained untapped. Now, these facts can be easily exploited and help with the argumentation behind interlinking two specific datasets. The major research contributions of this paper include a structured documentation of quality metrics and an implementable methodology with which the quality of a dataset can be assessed and documented according to the Linked Data principles.

REFERENCES

[1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. 2013. Crowdsourcing linked data quality assessment. In *International Semantic Web Conference*. Springer, 260–276.

[2] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.

[3] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts* (2009), 205–227.

[4] Jeremy Debattista, Sören Auer, and Christoph Lange. 2016. Luzzu–A Framework for Linked Data Quality Assessment. In *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*. IEEE, 124–131.

[5] Annika Flemming. 2010. Quality characteristics of linked data publishing datasources. *Master's thesis, Humboldt-Universität of Berlin* (2010).

[6] Erwin Johan Albert Folmer. 2012. *Quality of semantic standards*. Universiteit Twente/CTIT.

[7] Jennifer Golbeck, Bijan Parsia, and James Hendler. 2003. Trust networks on the semantic web. *Cooperative information agents VII* (2003), 238–249.

[8] Olaf Hartig and Jun Zhao. 2010. Publishing and Consuming Provenance Metadata on the Web of Linked Data. *IPAW* 6378 (2010), 78–90.

[9] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the Pedantic Web. *LDOW* 628 (2010).

[10] ISO 19157:2013 2013. *Geographic information – Data quality*. Standard. International Organization for Standardization.

[11] ISO/IEC 25012:2008 2008. *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*. Standard. International Organization for Standardization.

[12] Ian Jacobi, Lalana Kagal, and Ankesh Khandelwal. 2011. Rule-based trust assessment on the semantic web. In *International Workshop on Rules and Rule Markup Languages for the Semantic Web*. Springer, 227–241.

[13] Georgi Kobilarov, Christian Bizer, Sören Auer, and Jens Lehmann. 2009. Dbpedia-a linked data hub and data source for web and enterprise applications. *Programme Chairs* 16 (2009).

[14] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. 2002. AIMQ: a methodology for information quality assessment. *Information & management* 40, 2 (2002), 133–146.

[15] Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. 2012. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. ACM, 116–123.

[16] Lorena Otero-Cerdeira, Francisco J Rodríguez-Martínez, and Alma Gómez-Rodríguez. 2015. Ontology matching: A literature review. *Expert Systems with Applications* 42, 2 (2015), 949–971.

[17] Mia Ridge. 2013. From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing. *Curator: The Museum Journal* 56, 4 (2013), 435–450.

[18] Jonah E Rockoff. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review* 94, 2 (2004), 247–252.

[19] Anisa Rula and Amrapali Zaveri. 2014. Methodology for Assessment of Linked Data Quality.. In *LDQ@ SEMANTICS*.

[20] Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering* 25, 1 (2013), 158–176.

[21] Matthias Stevens and Ellie D'Hondt. 2010. Crowdsourcing of pollution data using smartphones. In *Workshop on Ubiquitous Crowdsourcing*.

[22] Douglas Brent West and others. 2001. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River.

[23] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.

A APPENDIX

A.1 SPARQL queries

Interpretation

```
PREFIX dqv: <http://www.w3.org/ns/dqv#>
PREFIX sh: <http://www.w3.org/ns/shacl#>
PREFIX ex: <http://example.org/>
SELECT ?metric ?measurement ?requirement
WHERE { GRAPH <http://example.org> {
    bind(<http://example.org/def#DatasetQuality> as ?metrictyp
        ?metric ex:hasMetricType ?metrictype .
    ?measurement dqv:isMeasurementOf ?metric .
    ?requirement sh:targetNode ?metric .
        }
}
```

Retrieve Manual Metrics

```
PREFIX dqv: <http://www.w3.org/ns/dqv#>
PREFIX ex: <http://example.org/>
SELECT ?metric
WHERE { GRAPH <http://example.org> {
        #alter bind statement to switch metric types.
    bind(<http://example.org/def#DatasetQuality> as ?metrictyp
        ?metric a dqv:Metric ;
        ex:hasMetricType ?metrictype .
        } FILTER NOT EXISTS {?measurement dqv:isMeasurementOf
}
```

Retrieve Resources for interpretation

```
PREFIX dqv: <http://www.w3.org/ns/dqv#>
PREFIX ex: <http://example.org/>
SELECT ?metric ?measurement
WHERE { GRAPH <http://example.org> {
    bind(<http://example.org/def#DatasetQuality> as ?metrictyp
        ?metric a dqv:Metric ;
        ex:hasMetricType ?metrictype .
    ?measurement dqv:isMeasurementOf ?metric .
        } FILTER {?metric ex:hasMetricType ?metricType}
        #invert filter for Data quality assessment's interpret
        #remove filter for Aggregated Interpretation step
}
```

A.2 Figures

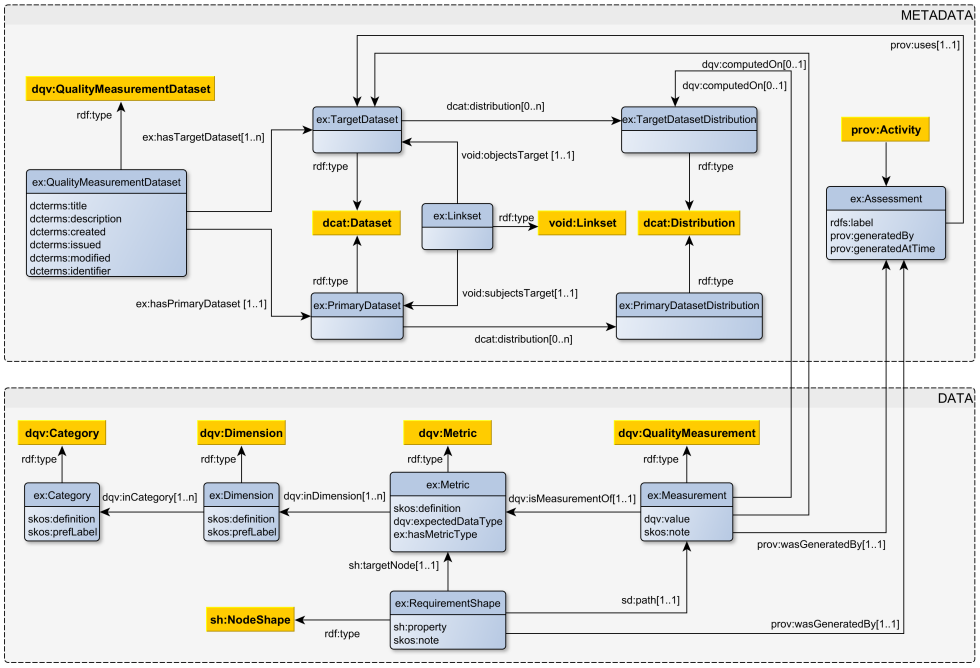


Figure 7: Graph Template

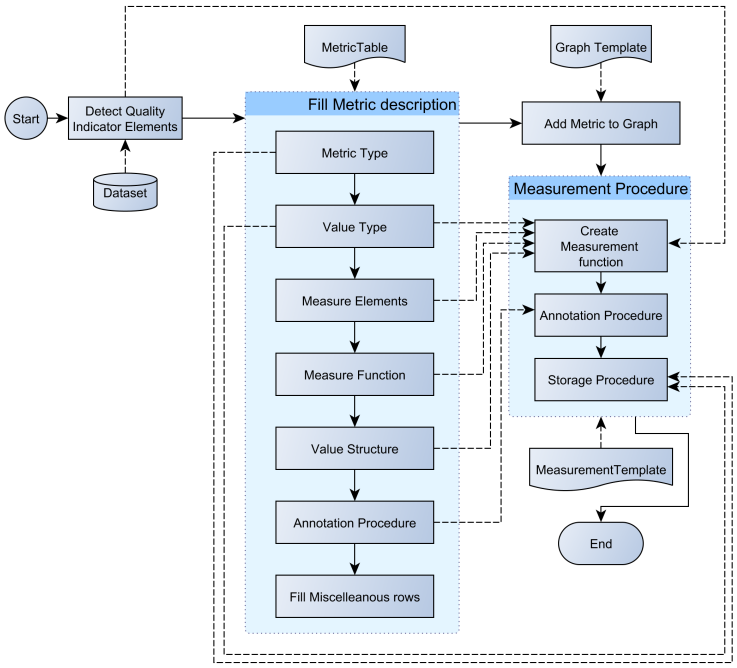


Figure 8: Custom Metric Creation Process