



EU Project No:601043 (Integrated Project (IP))

DIACHRON

Managing the Evolution and Preservation of the Data Web DIACHRON

Dissemination level:

Type of Document:

Contractual date of delivery:

Actual Date of Delivery:

Deliverable Number:

Deliverable Name:

Deliverable Leader:

Work package(s):

Status & version:

Number of pages:

WP contributing to the deliverable:

WP / Task responsible:

Coordinator (name / contact):

Other Contributors:

EC Project Officer: Federico Milani

Keywords:

Abstract:

Some meaningful abstract here.



Grant Agreement No. 601043

Document History			
Ver.	Date	Contributor(s)	Description
0.1	24.06.2014	Jeremy Debattista	Created TOC and LaTeX Setup

TABLE OF CONTENTS

1	INTRODUCTION	4
1.1	SCOPE AND OBJECTIVES	4
1.2	CONTEXT OF THIS DOCUMENT	4
1.3	DOCUMENT STRUCTURE	4
2	DATA QUALITY FRAMEWORK	4
2.1	HIGH-LEVEL ARCHITECTURE	4
2.1.1	SEMANTIC SCHEMA LAYER	5
2.1.2	PROCESSING UNIT	6
2.1.3	QUALITY ASSESSMENT LAYER	7
2.1.4	SEMANTIC ANNOTATION UNIT	7
2.1.5	VISUALISATION LAYER	8
2.2	SEQUENTIAL STREAM PROCESSOR	9
2.2.1	THE INITIALISATION PROCESS - <code>setUpProcess()</code>	10
2.2.2	THE PROCESSING OF TRIPLES - <code>startProcessing()</code>	11
2.3	THE DATASET QUALITY ONTOLOGY	11
2.3.1	EXTENDING DAQ FOR MULTI-DIMENSION REPRESENTATION AND STATISTICAL EVALUATION	12
2.3.2	ABSTRACT CLASSES AND PROPERTIES	14
2.3.3	EXTENDING DAQ FOR CUSTOM/SPECIFIC QUALITY METRICS	14
2.3.4	A TYPICAL QUALITY METADATA GRAPH	14
2.4	QUALITY RESTFUL API DESIGN	16
3	LIBRARIES USED	16
4	RANKING SERVICE	16
4.1	DATA QUALITY ASSESSMENT PROCESS	16
4.2	DATA QUALITY METRICS	16
4.2.1	ACCESSIBILITY CATEGORY	16
4.2.2	AVAILABILITY	16
4.2.3	INTRINSIC CATEGORY	17
4.2.4	ACCURACY	17
4.2.5	CONSISTENCY	17
4.2.6	CONCISENESS	19
4.2.7	REPRESENTATIONAL CATEGORY	19
4.2.8	REPRESENTATIONAL CONCISENESS	19
4.2.9	UNDERSTANDABILITY	19
4.3	VISUALISATION OF QUALITY ASSESSMENT	20
4.4	RANKING OF QUALITY-COMPUTED DATASETS	20
5	CRAWLING SERVICE	20
6	CONCLUSIONS	20

TABLE OF FIGURES

1	Quality Framework High Level Architecture Design	4
2	Quality Assessment Layer Class Diagram - A Quality Framework as a Pluggable Platform	8
3	Horizontal Bar Chart	8
4	Vertical Bar Chart	8
5	Radar Chart	9
6	Lines Plot	9
7	Closer look at the Quality Assessment process	10
8	The extended Dataset Quality Ontology (daQ)	12
9	Venn Diagram depicting Definition 1	13
10	Extending the daQ Ontology TBox and ABox	15

LIST OF TABLES

1 Introduction

1.1 Scope and Objectives

1.2 Context of this Document

1.3 Document Structure

2 Data Quality Framework

2.1 High-Level Architecture

The purpose of the Quality Framework is to provide an integrated platform that:

1. assesses RDF datasets and triple stores in a scalable manner;
2. provides queryable quality metadata on the assessed datasets;
3. provides visualisations on the quality

Furthermore, we aim to create an infrastructure and a platform that (i) can be easily extensible by different third party by creating their custom and more specific pluggable metrics required to assess their particular dataset domain, and (ii) having the necessary ontology framework to represent the metadata about the quality of the assessed linked datasets.

Currently, there is no uniform infrastructure to address the quality assessment problem, allowing the extension or redefinition of custom-specific metrics such as those required by the DIACHRON use cases. Tools such as Trellis ??, WIQA ?? and Sieve ?? implement a number of metrics but lacked flexibility wrt the level of automation, and user friendliness ??.

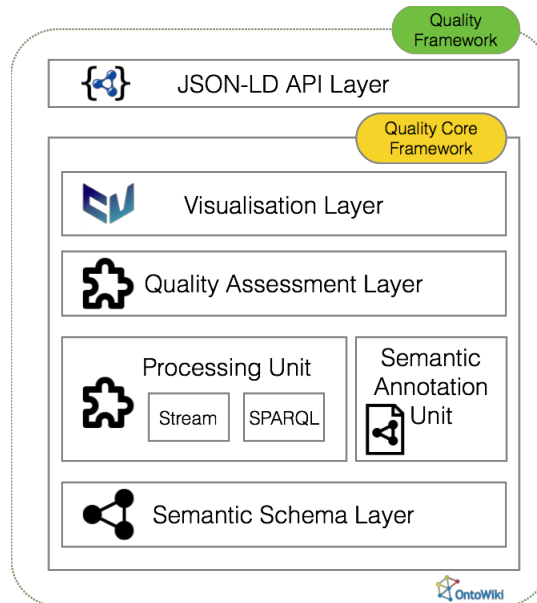


Figure 1: Quality Framework High Level Architecture Design

Figure 1 illustrates the high level architecture of the Quality Framework. The two main components are the API layer (cf. Section 2.4) and the Core framework. The core framework is made up of five modules: *Semantic Schema Layer*, *Processing Unit*, *Semantic Annotation Unit*, *Quality Assessment Layer*, and *Visualisation Layer*.

The quality framework is envisioned to run on top of OntoWiki¹, a wiki based on semantic technologies. The current technologies and extensions available in OntoWiki allow us to reach both the main objectives in this deliverable, and eventually to commence our initial prototypes of the quality framework (e.g. as an enterprise add-on tool to linked data publishers). Further investigation upon the usability of OntoWiki is still necessary in order to discover if the mentioned application can fulfil further the quality framework's ambitions, beyond the initial prototypes.

2.1.1 Semantic Schema Layer

The Quality Framework is based on semantic technologies and thus has an underlying semantic vocabulary layer which currently is made up of two ontologies: (i) the Dataset Quality Ontology (daQ)²; and (ii) the Quality Problem Report Ontology (qr)³. The former describes the quality metadata representation whilst the latter describes quality problems found in the dataset itself. The semantic schema layer is meant to be domain independent, where it could be reused in other similar frameworks. The daQ ontology (cf. Section 2.3) is the core vocabulary of this schema layer, and any other ontology part of this layer builds upon it.

The daQ ontology is a comprehensive generic vocabulary framework, based on three abstract concepts (Category, Dimension and Metric). Any newly implemented specific metric should have its representation in RDF, extending the daQ ontology. In DIACHRON, all metrics are defined in the Diachron Quality Metric vocabulary (dqm)⁴. Such vocabularies are easily integrated in the Quality Framework, since they adopt and extend the generic daQ vocabulary (by inheriting class and properties) as the way quality metadata is represented (cf. Section 2.3.3). The Quality Problem Report Ontology (qr) is made up of two classes a qr:QualityReport and qr:QualityProblem. The former represents a report on the problems detected during the assessment of quality on a dataset, whilst the latter represents a quality problem detected during the assessment of quality metrics on triples. Four properties are also defined in the ontology. The qr:computedOn represents the dataset URI on quality assessment has been made. This property is attached to a qr:QualityReport. qr:hasProblem links a qr:QualityProblem to a qr:QualityReport. The mentioned property identifies problem instances in a report. Each qr:QualityProblem isDescribedBy an instance of a daq:Metric⁵. The property qr:problematicThing represent the actual problematic instance from the dataset. This could be a list of resources (rdf:Seq) or a list of reified statements. Listing 1 represents an excerpt from a typical dataset showing the instance of ex:JoeDoe who is a foaf:Researchers working for ex:UniBonn. In these two instances there are three problematic triples:

- (A) $\langle \text{ex:JoeDoe a foaf:Researcher} \rangle$ - The problem in this triple is caused by the usage of an undefined class, in this case foaf:Researcher;
- (B) $\langle \text{ex:JoeDoe rdfs:label "JoeDoe"} \rangle$ - The literal ("JoeDoe") in the triple causes the malformed capitalisation metric to point out a problem in this triple;
- (C) $\langle \text{ex:UniBonn rdfs:label "UniBonn"} \rangle$ - The literal ("UniBonn") in the triple causes the malformed capitalisation metric to point out a problem in this triple.

Listing 2 represent these three problems using the Quality Problem Report ontology.

```
ex:JoeDoe a foaf:Researcher ;
  rdfs:label "JoeDoe" ;
  ex:worksFor ex:UniBonn .

ex:UniBonn rdfs:label "UniBonn" ;
  foaf:name "University Bonn" .
```

Listing 1: An excerpt of a typical Dataset

¹<http://ontowiki.eu/Welcome>

²<http://purl.org/eis/vocab/daq>

³<http://purl.org/eis/vocab/qr>

⁴<http://purl.org/eis/vocab/dqm>

⁵refer to Section 2.3

```

ex:QualityReport a qr:QualityReport ;
qr:computedOn <uri:datasetResearchers> ;
qr:hasProblem <#prob1>,<#prob2>,<#prob3> .

<#prob1> a qr:QualityProblem ;
qr:isDescribedBy <urn:metric/UndefinedClasses123> ;
qr:problematicThing [
  diachron:hasSubject ex:JoeDoe ;
  diachron:hasPredicate rdf:type ;
  diachron:hasObject foaf:Researcher ;
] .

<#prob2> a qr:QualityProblem ;
qr:isDescribedBy <urn:metric/Capitalisation789> ;
qr:problematicThing [
  diachron:hasSubject ex:JoeDoe ;
  diachron:hasPredicate rdfs:label ;
  diachron:hasObject "JoeDoe" ;
] .

<#prob3> a qr:QualityProblem ;
qr:isDescribedBy <urn:metric/Capitalisation789> ;
qr:problematicThing [
  diachron:hasSubject ex:UniBonn ;
  diachron:hasPredicate rdfs:label ;
  diachron:hasObject "UniBonn" ;
] .

```

Listing 2: An corresponding Quality Report for Listing 1

2.1.2 Processing Unit

The Processing Unit is an integral part of the Quality Framework. In this framework, we provide two main scalable processing units: a sequential stream processor (cf. Section 2.2) and SPARQL processor⁶. The former streams triples from RDF data dumps one by one in a sequential fashion. The latter allows the framework to assess quality on data that is available only in SPARQL endpoints. This unit is one of the two extensible modules (the other being Quality Assessment Layer) in the Quality Framework. For DIACHRON, the plan is to extend the sequential stream processor, enabling the de-reification of RDF statements into RDF triples.

Typically, an initialised processor has 2 inputs: the dataset URI (for the sequential stream processor) or the dataset SPARQL endpoint (in the case of the SPARQL processor), and a metric configuration file. Listing 3 shows an example of a typical metric configuration file.

```

@prefix diachron: <http://www.diachron-fp7.eu/diachron#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2004/03/trix/rdfg-1> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

diachron:ebi a diachron:Usecase ;
diachron:metric "accessibility.availability.SPARQLAccessibility" ;
diachron:metric "accessibility.availability.RDFAccessibility" ;
diachron:metric "accessibility.availability.Dereferencibility" ;
diachron:metric "accessibility.performance.DataSourceScalability" ;
diachron:metric "accessibility.performance.HighThroughput" ;
diachron:metric "accessibility.performance.LowLatency" ;
diachron:metric "intrinsic.accuracy.DefinedOntologyAuthor" ;
diachron:metric "intrinsic.accuracy.SynonymUsage" ;
diachron:metric "intrinsic.accuracy.POBODefinitionUsage" ;
diachron:metric "intrinsic.consistency.EntitiesAsMembersOfDisjointClasses" ;
diachron:metric "intrinsic.consistency.HomogeneousDatatypes" ;
diachron:metric "intrinsic.consistency.MisplacedClassesOrProperties" ;
diachron:metric "intrinsic.consistency.ObsoleteConceptsInOntology" ;
diachron:metric "intrinsic.conciseness.OntologyVersioningConciseness" ;
diachron:metric "dynamicity.timeliness.TimelinessOfResource" ;
diachron:metric "dynamicity.currency.CurrencyDocumentStatements" ;
diachron:metric "dynamicity.currency.TimeSinceModification" ;
diachron:metric "representational.understandability.HumanReadableLabelling" ;

```

⁶This processor is still being investigated and will not be ready by the deliverable deadline.

```
diachron:metric "representational.understandability.LowBlankNodeUsage" .
```

Listing 3: An typical metric configuration file

Each data processor in the Quality Framework has a defined 3-stage procedure (Listing 4): (i) processor initialisation; (ii) processing; and (iii) memory clean up. In the first process (processor initialisation), the processor create the necessary objects in memory to process data and load the required metrics that are instructed in the configuration file. Once the initialisation is ready, then processing is done by passing the streamed triples into the metrics. Finally, memory clean up ensures that no unused objects are using unnecessary computational power.

```
public interface IOProcessor {
    // Initialise the io processor with the necessary in-memory objects and metrics
    void setUpProcess();

    // Process the dataset for quality assessment
    void startProcessing() throws ProcessorNotInitialised;

    // Cleans up memory from unused objects after processing is finished
    void cleanUp() throws ProcessorNotInitialised;
}
```

Listing 4: IO Processor Interface

2.1.3 Quality Assessment Layer

The Quality Assessment Layer is unarguably the most important layer in this Quality Framework. The framework can be extended by any third party providing their own custom specific metric. This is already done in the DIACHRON project, where a number of metrics (cf. Section 4.2) required to assess the various use cases specified in Deliverable ??? are implemented. The Quality Assessment Layer provides two interfaces and an abstract class (cf. Figure 2) which facilitate the quality framework to be a pluggable and extensible platform. The interface `QualityMetric` is the core interface class which describes the metric classes. Each metric implementing this interface, must implement the following classes:

- compute** - This method assess the quad/triple which is passed by the stream processor by the defined metric;
- metricValue** - This method returns the value computed by the quality metric;
- toDAQTriples** - This method will return a list of daQ triples, containing quality metadata about the assessed metric, which will be stored in the dataset as a new named graph (quality graph);
- getMetricURI** - This method returns the URI of the Quality Metric from the ontology description (e.g. `http://purl.org/eis/vocab/dqm#DereferenceabilityMetric`);
- getQualityProblems** - This method returns a typed (`List<Resource>` or `List<Quad>`) `ProblemList` which will be used to create a quality report of the metric;

Furthermore, a metric might require some pre-processing or post-processing. Therefore, an interface (`ComplexQualityMetric`) extending `QualityMetric` was developed. This interface allow metric developers to perform such processing using the `void before()` and `void after()` methods.

In order to facilitate further such development of pluggable metrics, the `AbstractQualityMetric` class was developed, implementing the `QualityMetric` interface. In this abstract class, the method `List<Statement> toDAQTriples()` is implemented, generating daQ observation instances (cf. Section 2.3) for the metric being assessed.

2.1.4 Semantic Annotation Unit

The Semantic Annotation Unit takes the generated triples (from the `toDAQTriples()` method) in order to create the quality metadata in a dataset. The unit provides a number of helper classes that provide inferencing queries on

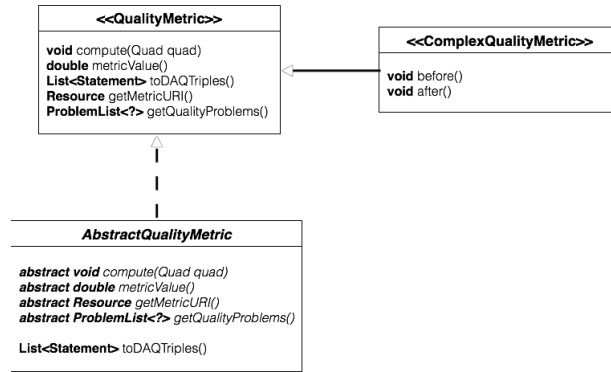


Figure 2: Quality Assessment Layer Class Diagram - A Quality Framework as a Pluggable Platform

vocabularies that describe metrics (such as DQM) based on the core ontology daQ. Therefore, RDF descriptions of metrics extending the daQ (cf. Section 2.3.3) ontology is absolutely required. These inferencing queries enable the framework to create a complete metadata description (cf. Section 2.3) of an assessed quality metric.

2.1.5 Visualisation Layer

CubeViz is an OntoWiki extension for visualising data cubes (observation instances). Figures 3, 4, 5, and 6 depicts four different CubeViz chart visualisations from computed quality metadata⁷.

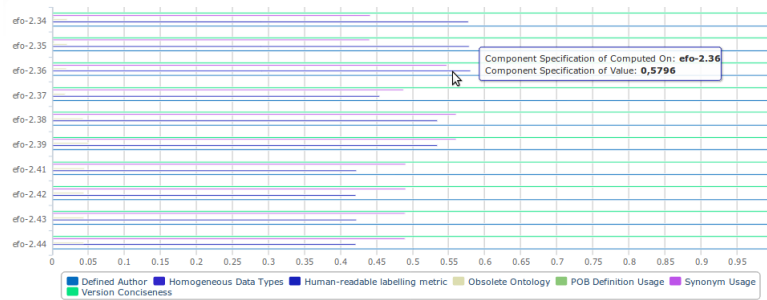


Figure 3: Horizontal Bar Chart

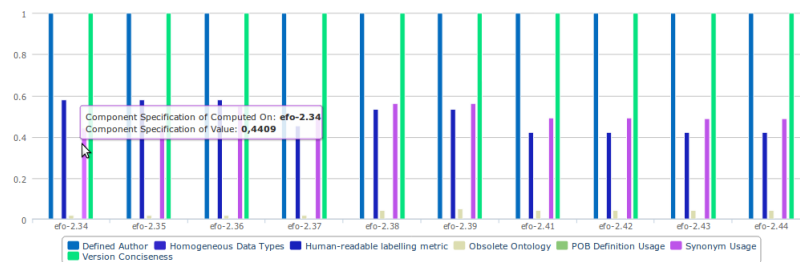


Figure 4: Vertical Bar Chart

⁷The quality metadata used can be found in https://raw.githubusercontent.com/diachron/quality/master/src/test/resources/cube_qg.trig

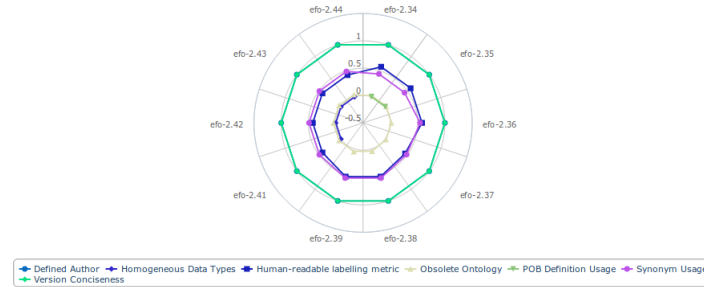


Figure 5: Radar Chart

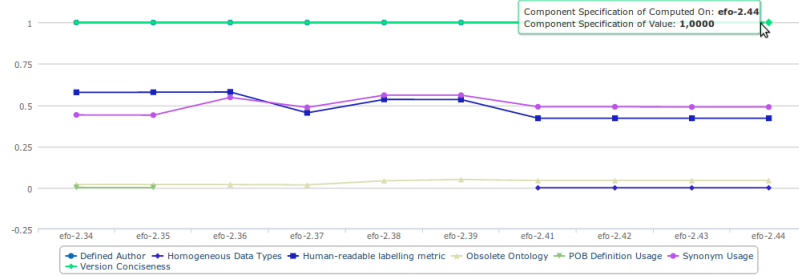


Figure 6: Lines Plot

A *horizontal bar* represents each metric (Figure 3) and shows its value (x-axis) with respect to the dataset (y-axis). Here, the different datasets analysed are actually successive revisions of one dataset. This chart provides a clear view of how the value associated to each one of the measured metrics changes as the dataset evolves. The horizontal layout is appropriate when the range of metric values is wide, and the number of different datasets is relatively small.

Similar to the horizontal bars chart, the *vertical bar chart* (Figure 4) allows the user to compare the values computed for each of the metrics (y-axis), with respect to the dataset (x-axis). In contrast with its horizontal counterpart, this chart is more appropriate when there are many datasets analysed but the range of metric values is not so wide.

In the *radar chart* (Figure 5), the datasets are represented as slices of a circle and the values corresponding to the metrics are depicted as points and lines of a particular color. This chart provides a clear view of how the values of the metric differ from each other for each particular dataset. Furthermore, it allows one to assess the overall quality of a dataset, by showing whether the values of the metrics are concentrated around sections of the circle regarded as good or bad.

The lines plot (Figure 6), lists the different datasets against the values of the metrics. Here, where different datasets are actually different revisions in the evolution of one dataset, this plot provides a comparison of the evolution of the quality of the dataset, with respect to each metric. The lines emphasise the points where the values of the metrics changed noticeably from one version to the next.

2.2 Sequential Stream Processor

In order to accurately assess linked dataset for quality measures, the assessment should on all triples in the assessed datasets. One must keep in mind that the computation of metrics on large datasets might be computationally expensive; thus, such stream processors computing dataset's quality must be scalable. In Figure 7, a closer look towards the quality assessment process is illustrated. A user first choose a dataset and the metrics which are required for the assessment of quality. The submitted information is passed to the Quality Framework via its API

and initialise the processing unit (stream processor) in the core framework. The stream processor is then initialised by: (1) creating the necessary objects in memory, and (2) initialise the chosen metrics. In Figure 7, “Metric 1” is shadowed out - to illustrate that it was not chosen by the user for this particular use case. Once the objects created, the stream processor fetches the dataset and in a sequential fashion it starts streaming quads one by one to all initialised metrics in parallel. After all statements are assessed, the semantic annotation unit requests the metric value for each metric and creates (or updates - in case the dataset already has one) the Quality Graph. This named graph represents the quality metadata of a dataset using the representation defined in Section 2.3, and it is stored in the dataset itself. Having this metadata, it will allow us to rank and crawl datasets based on different quality attributes.

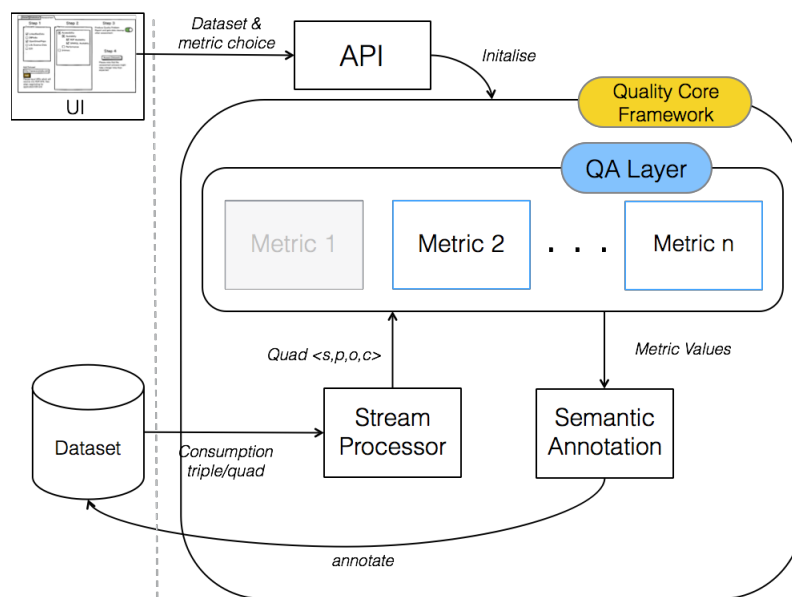


Figure 7: Closer look at the Quality Assessment process

Apache Jena⁸ (cf. Section ??) provides a dedicated module for the reading and writing of RDF Data (RDF I/O Technology - RIOT⁹). The RIOT API functionality provides a number of classes. Typically the `RDFDataMgr` is used, which contains the main set of functions to read and load models and datasets. For the sequential stream processor, the Jena RIOT API was used.

2.2.1 The Initialisation Process - `setUpProcess()`

The first operation on the initialisation is the execution of the `setUpProcess()` method. Listing 1 is the pseudocode of the the process. The sequential stream processor starts its initialisation by first trying to identify the serialisation used by the available RDF data dump. The method `guessRDFSerialisation` analyses the file serialisation by mapping the file name to one of the Jena’s in-built RDF languages (e.g. RDF/XML, NTriples, Turtle, NQuads, etc...) According to the file’s serialisation, the process then assign different types of `PipedRDFIterator`¹⁰ and either `aPipedQuadsStream`¹¹ or `PipedTriplesStream`¹² These two objects are required for the scalable execution of the sequential stream processor as together they act as the “producer”¹³ of sequential RDF triples

⁸<http://jena.apache.org>

⁹<http://jena.apache.org/documentation/io/rdf-input.html>

¹⁰<https://jena.apache.org/documentation/javadoc/arq/org/apache/jena/riot/lang/PipedRDFIterator.html>

¹¹<https://jena.apache.org/documentation/javadoc/arq/org/apache/jena/riot/lang/PipedQuadsStream.html>

¹²<https://jena.apache.org/documentation/javadoc/arq/org/apache/jena/riot/lang/PipedTriplesStream.html>

¹³As in the producer in the “Producer-Consumer problem”http://en.wikipedia.org/wiki/Producer-consumer_problem. The consumer is the on a separate thread, feeding the metrics.

from the RDF data dump. Once these are initialised, a flag is set to true to signal that the processor unit is in progress. Finally, the chosen metrics are loaded into memory. The loading of metrics is done dynamically during runtime, using the Java specific `newInstance()`¹⁴ method.

Algorithm 1 The Initialisation of the Sequential Stream Process

```
1: procedure SETUPPROCESS
2:   rdfSerialisation = guessRDFSerialisation(datasetURI) ;
3:   if rdfSerialisation is Quads then
4:     iterator = new PipedRDFIterator<Quad>() ;
5:     rdfStream = new PipedQuadsStream((PipedRDFIterator<Quad>) iterator) ;
6:   if rdfSerialisation is Triple then
7:     iterator = new PipedRDFIterator<Triple>() ;
8:     rdfStream = new PipedTriplesStream((PipedRDFIterator<Triple>) iterator) ;
9:   set initialised boolean to true ;
10:  loadMetrics() ;
```

2.2.2 The Processing of Triples - `startProcessing()`

After the initialisation process, the method `startProcessing()` is invoked. The `RDFStream rdfStream` object starts parsing the RDF dump and producing triple or quad statements in the `iterator`. On a different thread, the “consumer” - the sequential stream processor - consumes these statements from the `iterator`, converts them into quads of $\langle s, p, o, c \rangle$, and passes them to all initialised metrics. The consumption process is repeated until all statements are exhausted from the `iterator`. The semantic annotation unit is then signalled to start its annotation. Listing 2 describes this process in pseudocode.

Algorithm 2 Processing Triple/Quad Statements

```
1: procedure STARTPROCESSING
2:   if initialised == false then
3:     throw exception ;
4:   create new producer thread for rdfStream ;
5:   while (iterator has another statement) do
6:     quad = Object2Quad(iterator.next()) ;
7:     pass quad to all metrics and compute metric ;
8:   invoke semantic annotation unit ;
```

2.2.3 Clean Up - `cleanUp()`

The final process is to clean up the objects from memory. The processor follows a simple approach by assigning `null` to all objects and shutting down all running threads.

2.3 The Dataset Quality Ontology

The idea behind the Dataset Quality Ontology [?]¹⁵ (daQ) is to provide a comprehensive generic vocabulary framework, allowing a uniform definition of specific data quality metrics and thus suggest how quality metadata should be represented in datasets. This metric definition would then allow publishers to attach data quality metadata with quality benchmarking results to their linked dataset. Figure 8 depicts the current state of the daQ vocabulary.

¹⁴<http://docs.oracle.com/javase/8/docs/api/java/lang/Class.html#newInstance-->

¹⁵<http://purl.org/eis/vocab/daq>

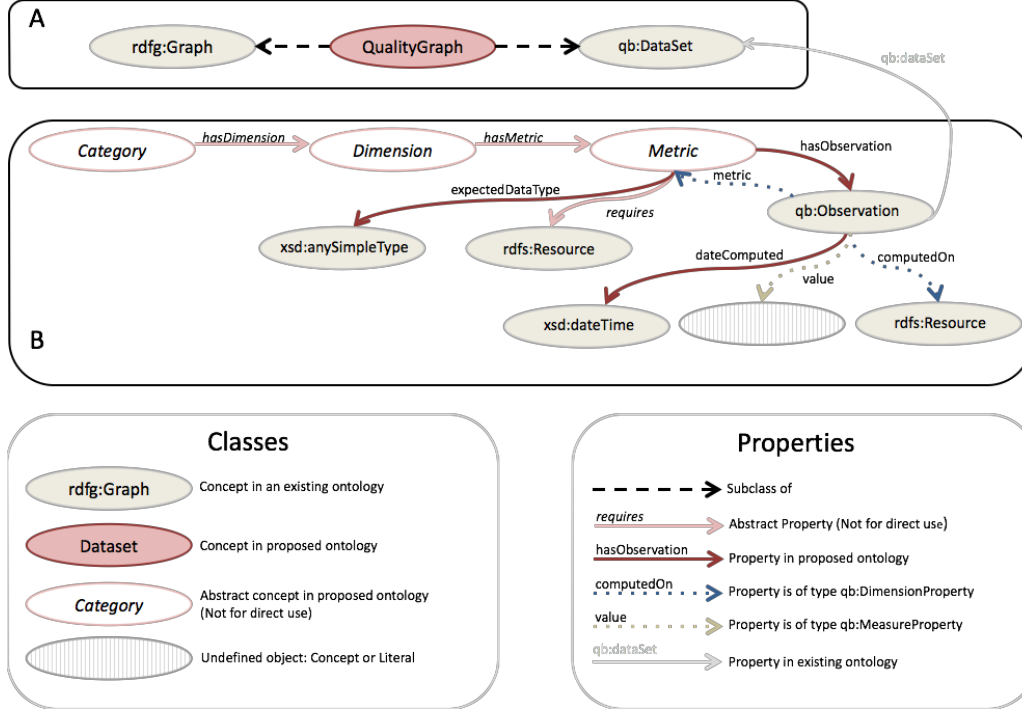


Figure 8: The extended Dataset Quality Ontology (daQ)

Using daQ, the quality metadata is intended to be stored in what we defined to be the *Quality Graph*. The latter concept is a subclass of `rdf:Graph` [?]. This means that the quality metadata is stored and managed in a separate named graph from the assessed dataset. Named graphs are favoured due to

- the capability of separating the aggregated metadata with regard to computed quality metrics of a dataset from the dataset itself;
- their use in the Semantic Web Publishing vocabulary [?] to allow named graphs to be digitally signed, thus ensuring trust in the computed metrics and defined named graph instance. Therefore, in principle each `daq:QualityGraph` can have the following triple `:myQualityGraph swp:assertedBy :myWarrant`.

The daQ ontology distinguishes between three layers of abstraction, based on the survey work by Zaveri et al. [?]. As shown in Figure 8 Box B, a quality graph comprises of a number of different *Categories*, which in turn possess a number of quality *Dimensions*¹⁶. A quality dimension groups one or more computed quality *Metrics*. To formalise this, let G represent the named Quality Graph (`daq:QualityGraph`), $C = \{c_1, c_2, \dots, c_x\}$ is the set of all possible quality categories (`daq:Category`), $D = \{d_1, d_2, \dots, d_y\}$ is the set of all possible quality dimensions (`daq:Dimension`) and $M = \{m_1, m_2, \dots, m_z\}$ is the set of all possible quality metrics (`daq:Metric`); where $x, z, y \in \mathbb{N}$, then:

Definition 1

$$\begin{aligned} G &\subseteq C, \\ C &\subseteq D, \\ D &\subseteq M; \end{aligned}$$

Figure 9 shows this formalisation in a pictorial manner using Venn diagrams.

¹⁶In this deliverable we will refer to these as quality dimensions, in order to distinguish between the data cube dimensions

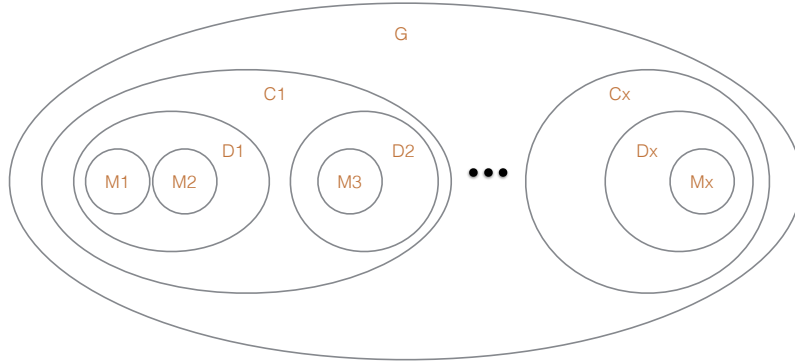


Figure 9: Venn Diagram depicting Definition 1

2.3.1 Extending daQ for Multi-Dimension Representation and Statistical Evaluation

The Data Cube Vocabulary [?] allows the representation of statistics about observations in a multidimensional attribute spaces. Multidimensional analysis of these observations, e.g. across the revision history of a dataset, would thus have required complex querying. Extending daQ with the standardised Data Cube Vocabulary allows us to represent quality metadata of a dataset as a collection of *Observations*, dimensions being the different quality metrics computed, the resources whose quality is assessed, revisions of these resources, and arbitrary further dimensions, such as the intended application scenario. It also permits applying the wide range of tools that support data cubes to quality graphs, including the CubeViz visualisation tool¹⁷.

A *Quality Graph* is a special case of `qb:DataSet`, which allows us to represent a collection of quality observations complying to a defined dimensional structure. Each observation represents a quality metric measured out against a particular resource (e.g. a specific revision of a dataset). daQ defines the structure of such observations by the `qb:DataStructureDefinition` shown in Listing 5.

```
daq:dsd a qb:DataStructureDefinition ;
# Dimensions: metrics and what they were computed on
qb:component [
  qb:dimension daq:metric ;
  qb:order 1 ; ] ;
qb:component [
  qb:dimension daq:computedOn ;
  qb:order 2 ; ] ;
# Measures (here: metric values)
qb:component [ qb:measure daq:value ; ] ;
# Attribute (here: the unit of measurement)
qb:component [
  qb:attribute sdmx-attribute:unitMeasure ;
  qb:componentRequired false ;
  qb:componentAttachment qb:DataSet ; ] .
```

Listing 5: The Data Structure Definition (Turtle Syntax)

The `daq:QualityGraph` also defines one restriction that controls the property `qb:structure` and its value to the mentioned definition, thus ensuring that all *Quality Graph* instances make use of the standard definition. Having a standard definition ensures that all *Quality Graphs* conform to a common data structure definition, thus datasets with attached quality metadata can be compared. Listing 6 describes the definition of `daq:QualityGraph`.

```
daq:QualityGraph
a rdfs:Class, owl:Class ;
rdfs:subClassOf rdfg:Graph, qb:DataSet,
[ rdf:type owl:Restriction ;
  owl:onProperty qb:structure ;
  owl:hasValue daq:dsd ] ;
rdfs:comment "Defines a quality graph which will contain all metadata about quality metrics on the dataset." ;
```

¹⁷<http://cubeviz.aksw.org>

```
rdfs:label "Quality Graph Statistics" .
```

Listing 6: The Quality Graph Definition (Turtle Syntax)

2.3.2 Abstract Classes and Properties

This ontology framework (Figure ??) has three abstract classes/concepts (`daq:Category`, `daq:Dimension`, `daq:Metric`) and three abstract properties (`daq:hasDimension`, `daq:hasMetric`, `daq:requires`) which should not be used directly in a quality instance. Instead these should be inherited as parent classes and properties for more specific quality metrics. The abstract concepts (and their related properties) are described as follows:

daq:Category represents the highest level of quality assessment. A category groups a number of dimensions.

daq:Dimension – In each dimension there is a number of metrics.

daq:Metric is smallest unit of measuring a quality dimension. Each metric instance is linked to one or more observations. Each observation has a value (`daq:value`), representing a score for the assessment of a quality attribute. This attribute is defined as a `qb:MeasureProperty`. Since this value is multi-typed (for example one metric might return true/false whilst another might require a floating point number), the value's `daq:hasValue` range is inherited by the actual metric's attribute defined by the property `daq:expectedDataType`. An observation must have the Dimension Properties (`qb:DimensionProperty`) `daq:computedOn` and `daq:metric`, which defines the assessed resource and the metric the mentioned resource was assessed by respectively. A metric might also require additional information (e.g. a gold standard dataset to compare with). Therefore, a concrete metric representation can also define such properties using subproperties of the `daq:requires` abstract property. Another important attribute for any observation is the `daq:dateComputed`, where it records the date of the observation's creation.

2.3.3 Extending daQ for Custom/Specific Quality Metrics

The classes of the core daQ vocabulary can be extended by more specific and custom quality metrics. In order to use the daQ, one should define the quality metrics that characterise the “fitness for use” [?] in a particular domain. We are currently in the process of defining the quality dimensions and metrics described in Deliverable 5.1. **Extending** the daQ vocabulary means adding new quality protocols that inherit the abstract concepts (Category-Dimension-Metric). Custom quality metrics do not need to be included in the daQ namespace itself; in fact, in accordance with LOD best practices, we recommend extenders to make them in their own namespaces. In Figure 10 we show an illustrative example of extending the daQ ontology (TBox) with a more specific quality attribute, i.e. the RDF Availability Metric as defined in [?], and an illustrative instance (ABox) of how it would be represented in a dataset.

The Accessibility concept is defined as an `rdfs:subClassOf` the abstract `daq:Category`. This category has five quality dimensions, one of which is the *Availability* dimension. This is defined as an `rdfs:subClassOf` `daq:Dimension`. Similarly, *RDFAvailabilityMetric* is defined as an `rdfs:subClassOf` `daq:Metric`. The specific properties *hasAvailabilityDimension* and *hasRDFAccessibilityMetric* (sub-properties of `daq:hasDimension` and `daq:hasMetric` respectively) are also defined (Figure 10).

2.3.4 A typical Quality Metadata Graph

The excerpt listing in 7 show a typical quality graph metadata in a dataset.

```
# ... prefixes
# ... dataset triples

ex:qualityGraph1 a daq:QualityGraph ;
  qb:structure daq:dsd .

ex:qualityGraph1 {
```

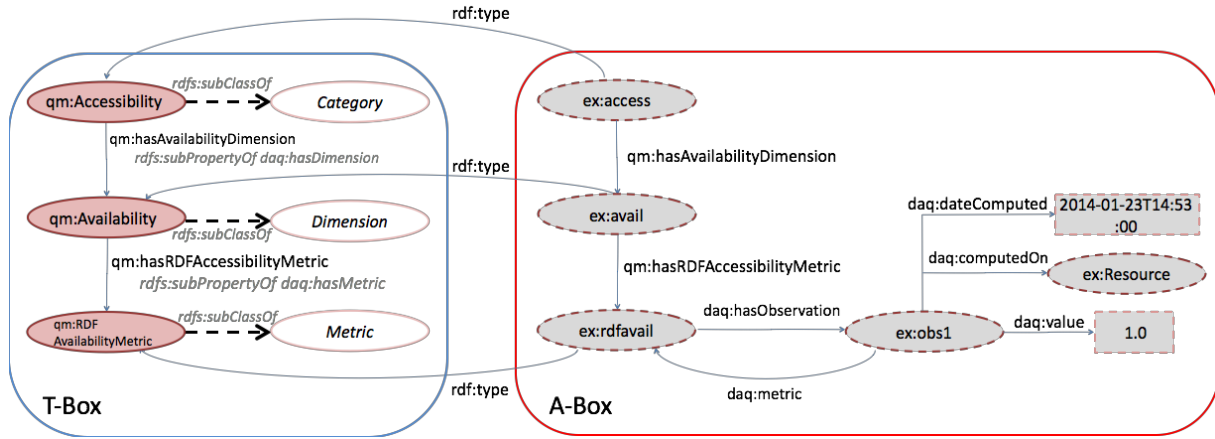


Figure 10: Extending the daQ Ontology TBox and ABox

```
# ... quality triples
ex:accessibilityCategory a dqm:Accessibility ;
  dqm:hasAvailabilityDimension ex:availabilityDimension .

ex:availabilityDimension a dqm:Availability ;
  dqm:hasEndPointAvailabilityMetric ex:endPointMetric ;
  dqm:hasRDFAvailabilityMetric ex:rdfAvailMetric .

ex:endPointMetric a dqm:EndPointAvailabilityMetric ;
  daq:hasObservation ex:obs1, ex:obs2 .

ex:obs1 a qb:Observation ;
  daq:computedOn <efo-2.43> ;
  daq:dateComputed "2014-01-23T14:53:00"^^xsd:dateTime ;
  daq:value "1.0"^^xsd:double ;
  daq:metric ex:endPointMetric ;
  qb:dataSet ex:qualityGraph1 .

ex:obs2 a qb:Observation ;
  daq:computedOn <efo-2.44> ;
  daq:dateComputed "2014-01-25T14:53:00"^^xsd:dateTime ;
  daq:value "1.0"^^xsd:double ;
  daq:metric ex:endPointMetric ;
  qb:dataSet ex:qualityGraph1 .

ex:rdfAvailMetric a dqm:RDFAvailabilityMetric ;
  daq:hasObservation ex:obs3, ex:obs4 .

ex:obs3 a qb:Observation ;
  daq:computedOn <efo-2.43> ;
  daq:dateComputed "2014-01-23T14:53:01"^^xsd:dateTime ;
  daq:value "1.0"^^xsd:double ;
  daq:metric ex:rdfAvailMetric ;
  qb:dataSet ex:qualityGraph1 .

ex:obs4 a qb:Observation ;
  daq:computedOn <efo-2.44> ;
  daq:dateComputed "2014-01-25T14:53:01"^^xsd:dateTime ;
  daq:value "0.0"^^xsd:double ;
  daq:metric ex:rdfAvailMetric ;
  qb:dataSet ex:qualityGraph1 .

# ... more quality triples
}
```

Listing 7: A Quality Graph Excerpt (Turtle Syntax)

The instance *ex:qualityGraph1* is a named `daq:QualityGraph`. The defined graph is automatically a `qb:DataSet`, and due to the restriction placed on the `daq:QualityGraph` (see Listing 6), the value for the `qb:structure` property is defined as `daq:dsd` (see Listing 5). In the named graph, instances for the `daq:Accessibility`, `daq:Availability`, `daq:EndPointAvailabilityMetric` and `daq:RDFAvailabilityMetric` are shown. A metric instance has a number of observations. Each of these observations specifies the metric value (`daq:value`), the resource the metric was computed on (`daq:computedOn` here: different datasets, which are actually different revisions of one dataset), when it was computed (`daq:dateComputed`), the metric instance (`daq:metric`) and finally to what dataset the observation is defined in (`qb:dataset`).

2.4 Quality RESTful API Design

3 Libraries Used

4 Ranking Service

4.1 Data Quality Assessment Process

4.2 Data Quality Metrics

- Metric input is a quad `?s, ?p, ?o, ?g` -

4.2.1 Accessibility Category

4.2.2 Availability

Dereferenceability Metric

HTTP URIs should be dereferenceable, i.e. HTTP clients can retrieve the resources identified by the URI. A typical web URI resource would return a 200 OK code indicating that a request is successful and 4xx or 5xx if the request is unsuccessful. In Linked Data, a successful request should return a document (RDF) containing the description (triples) of the requested resource. In Linked Data, there are two possible ways which allow publishers make URIs dereferenceable. These are the 303 URIs and the hash URIs¹⁸. Yang et. al [?] describes a mechanism to identify the dereferenceability process of linked data resource.

Calculates the number of valid redirects (303) or hashed links according to LOD Principles.

This metric (listing ??) will count the number of valid dereferenceable URI resources found in the subject (?s) and object (?o) position of a triple. The `isDereferenceable(resource)` method uses the rules defined in [?]. The metric will return a ratio of the number of dereferenced URIs (`deref`) against the total number of triples in a

Algorithm 3 Dereferenceability Algorithm

```

1: procedure INIT
2:   totalTriples = 0 ;
3:   deref = 0 ;
4: procedure DEREFERENCE(?s, ?p, ?o, ?g)
5:   if (isURI(?s)) && (isDereferenceable(?s)) then deref++ ;
6:   if (isURI(?o)) && (isDereferenceable(?o)) then deref++ ;
7:   totalTriples++;
```

dataset (`totalTriples`). The expected range is [0..1], where 0 is the worst rating and 1 is the best rating.

¹⁸<http://www.w3.org/TR/cooluris/>

4.2.3 Intrinsic Category

4.2.4 Accuracy

Malformed Datatype Literals

Literals are nodes in an RDF graph, used to identify values such as numbers and dates. The RDF specifies two types of literals: plain and typed. A plain literal is a string combined with an optional language tag. A typed literal comprises a string (the lexical form of the literal) and a datatype (identified by a URI) which is supposed to denote a mapping from lexical forms to some space of values. In the Turtle syntax typed literals are notated with syntax such as: `"13"8sd:int` This Malformed Datatype Literals metric intends to check if the value of a typed literal is valid with regards to the given xsd datatype.

Calculates the ratio of typed literals not valid regarding its given xsd datatype to all literals

Algorithm 4 Malformed Datatype Literals Algorithm

```

1: procedure INIT
2:   totalLiterals = 0 ;
3:   malformedLiterals = 0 ;
4: procedure COMPUTEMALFORMEDDATATYPELITERALS( $\langle ?s, ?p, ?o, ?g \rangle$ )
5:   if (isLiteral(?o)) then totalLiterals++ ;
6:     if (isTypedLiteral(?o)) && (!hasValidDatatype(?o)) then malformedLiterals++ ;
7:   metric=malformedLiterals/totalLiterals;

```

Literals Incompatible with Datatype Range

Defined Ontology Author

POBO Definition Usage

Synonym Usage

4.2.5 Consistency

Entities As Members of Disjoint Classes

Homogeneous Datatypes

The literals assigned to a certain property should be the same datatype. E.g. `lite` This metric reports outliers or conflicts of literal datatypes. Since it is * meant to measure the homogeneity and not the validity possible `rdfs:range` * restrictions are not evaluated. * * This metric makes two distinctions as far as homogeneity is concerned: * 1) if there are just a few triples of one predicate having a datatype * different from all other triples, they are considered

as outliers and * are all reported * 2) if there is no obvious ratio of possibly wrong and possibly right triples * the conflicting view candidates are reported *

Misplaced Classes or Properties

Misused Owl Datatype or Object Properties

Obsolete Concepts in Ontology

Ontology Hijacking

Undefined Classes Metric

Oftentimes a terms which is used in the object position of a triple and is not a literal is not formally defined as being a class. 'Being defined' means that the term is defined either in some external ontology or at an earlier position in the given dataset. Regarding to Hogar [?] to the most used undefined classes belong `foaf:UserGroup`, `rss:item`, `linkedct:link`, `politico:Term`. The probability for undefined class in the subject position is very low, because the subject of a quad never references classes or properties in external vocabularies. Therefore they is no need to analyse the subject for this metric. For the most LOD data sets is sufficient to check object by the predicate `rdf:type`. In the case when LOD data set defines its own vocabulary the following predicates indicate that the object must be a defined class: `rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`, `owl:allValuesFrom`, `owl:someValuesFrom`, `owl:equivalentClass`, `owl:complementOf`, `owl:onClass`, `owl:disjointWith`. The undefined classes problem occurs due to spelling or syntactic mistakes resolvable through minor fixes to the respective ontologies. The missing classes should be define in corresponding ontology or in a separate namespace.

Calculates the ratio of undefined classes to all classes in object position in a dataset

Algorithm 5 Undefined Classes Metric Algorithm

```

1: procedure INIT
2:   totalClasses = 0 ;
3:   undefinedClasses = 0 ;
4: procedure CLASSNOTDEFINED( $\langle ?s, ?p, ?o, ?g \rangle$ )
5:   if (isClassProperty(?p) && (isURI(?o))) then ;
6:     if (!isDefined(?o)) undefinedClasses++ then;
7:   totalClasses++ ++;
```

Undefined Properties

Similar to the Undefined Classes metric 4.2.5 the Undefined Properties metric identifies terms in the predicate position that are used without any formal definition. Hogan [?] identified the following properties that are often used without being defined: `foaf:image`, `cycann:label`, `foaf:tagLine`. The following list of predicates indicate that the object of the quad must be a defined property: `rdfs:subPropertyOf`, `owl:onProperty`, `owl:assertionProperty`, `owl:equivalentProperty`, `owl:propertyDisjointWith`.

Calculates the ratio of undefined properties to all properties in the given data set

Algorithm 6 Undefined Properties Algorithm

```
1: procedure INIT
2:   totalProperties = 0 ;
3:   undefinedProperties = 0 ;
4: procedure PROPERTYNOTDEFINED( $\langle ?s, ?p, ?o, ?g \rangle$ )
5:   if (isURI(?p)) totalProperties++ then ;
6:     if (!isDefined(?p)) undefinedProperties++ then;
7:     if (isFromList(?p)) && !isDefined(?o) undefinedProperties++ then;
8:   totalProperties ++;
```

4.2.6 Conciseness**Duplicate Instance****Extensional Conciseness****Ontology Versioning Conciseness****4.2.7 Representational Category****4.2.8 Representational Conciseness****Short URIs****4.2.9 Understandability**

Empty Annotation Value The metric Empty Annotation Value consider the following widely used annotation properties (labels, comments, notes, etc.) and identifies triples whose property is an annotation property and whose object is an empty string.

Calculates the ratio of annotations with empty objects to all annotations in the data set.

Algorithm 7 Empty Annotation Value Algorithm

```
1: procedure INIT
2:   totalAnnotations = 0 ;
3:   emptyAnnotations = 0 ;
4: procedure ISEMPYANNOTATION( $\langle ?s, ?p, ?o, ?g \rangle$ )
5:   if (isAnnotation(?p)) then
6:     if (isEmpty(?o)) then emptyAnnotations++ ;
7:   totalAnnotations++;
```

Human Readable Labelling**Labels Using Capitals****Low Blank Node Usage****Whitespace in Annotation**



Grant Agreement No. 601043

- 4.3 Visualisation of Quality Assessment**
- 4.4 Ranking of Quality-Computed Datasets**
- 5 Crawling Service**
- 6 Conclusions**