An Extensive Methodology and Framework for Quality Assessment of DCAT-AP Datasets

Bianca Wentzel[1(✉)], Fabian Kirstein[1,2], Torben Jastrow[1], Raphael Sturm[1], Michael Peters[1], and Sonja Schimmler[1,2]

[1] Fraunhofer FOKUS, Berlin, Germany
{bianca.wentzel,fabian.kirstein,torben.jastrow,raphael.sturm,michael.peters,sonja.schimmler}@fokus.fraunhofer.de
[2] Weizenbaum Institute for the Networked Society, Berlin, Germany

Abstract. The DCAT Application Profile for Data Portals is a crucial cornerstone for publishing and reusing Open Data in Europe. It supports the harmonization and interoperability of Open Data by providing an expressive set of properties, guidelines, and reusable vocabularies. However, a qualitative and accurate implementation by Open Data providers remains challenging. To improve the informative value and the compliance with RDF-based specifications, we propose a methodology to measure and assess the quality of DCAT-AP datasets. Our approach is based on the FAIR and the 5-star principles for Linked Open Data. We define a set of metrics, where each one covers a specific quality aspect. For example, if a certain property has a compliant value, if mandatory vocabularies are applied or if the actual data is available. The values for the metrics are stored as a custom data model based on the Data Quality Vocabulary and is used to calculate an overall quality score for each dataset. We implemented our approach as a scalable and reusable Open Source solution to demonstrate its feasibility. It is applied in a large-scale production environment (data.europa.eu) and constantly checks more than 1.6 million DCAT-AP datasets and delivers quality reports.

Keywords: Open Data · DCAT-AP · Data Quality

1 Introduction

Open Data constitutes a global movement to make data of public interest openly available without any restrictions. Popular providers of Open Data are public administrations, governments, and nonprofit and research organizations. Typically Open Data is published and managed through Web portals and aggregated into central portals. A well-known example for an aggregator is the official portal for European data[1], that provides access to more than 170 individual data catalogs, containing more than 1.6 million datasets.

[1] https://data.europa.eu/.

I. Lindgren et al. (Eds.): EGOV 2023, LNCS 14130, pp. 262–278, 2023.
https://doi.org/10.1007/978-3-031-41138-0_17

In order to efficiently disseminate, aggregate, and reuse Open Data a harmonized, standardized, and machine-readable metadata model is paramount. A widely adopted and powerful standard is the DCAT Application Profile for Data Portals (DCAT-AP). It is based on the W3C (World Wide Web Consortium) Resource Description Framework (RDF) Data Catalogue Vocabulary (DCAT)[2] and therefore follows Linked Data and Semantic Web principles. DCAT-AP provides a plethora of properties, vocabularies and guidelines to extensively express information about Open Data. However, currently many published DCAT-AP datasets are affected with quality issues, such as sparse use of properties, wrong or no use of vocabularies, application of incorrect data types or unavailable data. This is caused by several aspects: (1) The DCAT-AP is fuzzy to a certain extent and precise requirements for some properties are missing. (2) Only a few properties are declared as mandatory, allowing datasets with little expressiveness. (3) DCAT-AP only represents the metadata, the actual data is linked and its availability depends on external resources. (4) There does not exist an extensive quality baseline for DCAT-AP, making it difficult for providers to ensure the quality of their DCAT-AP datasets.

In this paper, we present a methodology, framework and software implementation to address these issues and support the iterative improvement of the quality and expressiveness of DCAT-AP datasets. We mainly address two research questions in our work. Firstly, how can we measure and represent the quality, completeness, and validity of DCAT-AP datasets? Secondly, how can we present and communicate these quality assessments to data providers? The main contributions of our work are:

- We designed concrete metrics and a data model to describe and store quality measurements about DCAT-AP datasets based on the Data Quality Vocabulary (DQV) and the FAIR and 5-star principles for Linked Open Data (LOD).
- We implemented a highly scalable processing pipeline to determine the indicators for sets of DCAT-AP catalogs and a reporting tool to browse and download the current and past results.
- We tested and evaluated our approach with a corpus of more than 1.6 million datasets to demonstrate its feasibility and added value.

In Sect. 2 we introduce related work and related projects that deal with quality assessment of Open Data and that act as a foundation for our work. Our qualitative quality metrics and our data model are described in Sect. 3. Our implementation is illustrated in Sect. 4. Our approach is evaluated in Sect. 5 with a feature comparison and a practical use case. Section 6 summarizes our work and gives an outlook for future developments.

2 Related Work

Our work is based on several related standards, specifications, and technologies from the domains Open Data, research data, Linked Data and Semantic Web, as

[2] https://www.w3.org/TR/vocab-dcat-2/.

well as data quality. In the following, we present a brief overview of the relevant foundations.

According to the DIN, *quality* is defined as the "totality of characteristics (and characteristic values) of a unit with regard to its suitability to fulfill specified and presupposed requirements"[3]. When referred to data, and more specifically to data and datasets in a scientific context, it can be assumed that there are general requirements for *data quality*, such as verifiability, reusability, relevance or completeness. To ensure high quality standards, improving data quality has been enforced by various public authorities in the past years, e.g. by the Information Quality Act[4]. Such efforts help mitigate the problem that insufficient data quality leads to higher costs and time loss in science projects[5].

DCAT is a mature and popular standard for expressing metadata about data catalogs and foster interoperability between them. The standard consists of multiple classes, where the most relevant ones are dataset and distribution. The first one represents a collection of data, and the second one represents the actual access to the data, e.g. a downloadable file. [12] *DCAT-AP* is a practical extension of *DCAT* which introduces additional metadata fields and mandatory ranges for certain properties. These ranges are provided as a Simple Knowledge Organization System (SKOS)[6] controlled vocabulary, published by the Publications Office of the European Union. For example, properties like language, spatial information or MIME type can be harmonized by applying the provided vocabularies [1].

The *FAIR principles* describe guidelines for data, broken down into "Findability", "Accessibility", "Interoperability" and "Reusability". The principles are intended to ensure a uniform presentation of collected data. Within the four principles, there are 15 sub-principles, as shortly described in the following. *Findability* summarizes that records are tagged with globally unique and persistent identifiers (F1) as well as rich metadata (F2). The data should also be present inside of a searchable resource (F4). In addition, the metadata must specify the data identifier (F3). *Accessibility* describes the need for a simplified access of datasets through standardized communication protocols (A1). Those protocols should be open, free, universally implementable (A1.1) and should allow an authentication/ authorization procedure (A1.2). Furthermore the metadata should be accessible even when the data is no longer available (A2). *Interoperability* ensures that datasets have rich metadata and provide a formal, accessible, shared and broadly applicable language to represent the information (I1). Also metadata (and data as well) should use vocabularies that follow the FAIR principles (I2) and include qualified references to other data (I3). *Reusability* requires datasets to be provided with a variety of descriptive attributes (R1) and a clear and accessible data usage license (R1.1). In addition, datasets should have a traceable provenance (R1.2) and comply with community standards (R1.3) [14].

[3] DIN EN ISO 8402 (1995), p. 212.
[4] https://sgp.fas.org/crs/RL32532.pdf.
[5] https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.
[6] https://www.w3.org/TR/skos-reference/.

The FAIR principles overlap with the *5-star principles for LOD* defined by Tim Berners Lee[7]. While this model also aims at improving FAIRness, it is more focused on interlinking the data to enable the use of Semantic Web technologies. Furthermore, this model refers to Open Data which means that data can be used, modified and shared freely by users[8] while the FAIR principles do not primarily target openness. Based on this model, a dataset is evaluated according to five criteria resulting in a star rating indicating its openness. If the dataset does not comply with any criteria of the model it receives a zero star rating. One star is earned by using an open license, receiving a second star requires the provision of data using a structured format. A dataset has three stars if its format is open and non-proprietary. The fourth star honors the use of URIs to describe properties, and the fifth star evaluates the interlinking of data to provide context.

The *Data Quality Vocabulary (DQV)* is an extension of the DCAT vocabulary, forming the basis for defining and interpreting dataset quality. When interpreting quality, the *DQV* takes into account factors such as the constant updating of data, the possibility of corrections by the user, and persistence obligations[9].

The *Shapes Constraint Language (SHACL)* is used to validate RDF graphs against predefined rules, which are described as RDF graphs as well. These rules specify, for example, specific formats, cardinalities, or relations for properties of an RDF graph. The results are also rendered as an RDF graph and contain detailed information about any errors or violations for the affected properties[10].

2.1 Related Projects

Based on research that is primarily centered around developing standards and metrics to evaluate the quality of Open Data, various solutions have been created to automatically measure the quality of Open Data. A well-known example is the work by Vetrò et al. [11] which resulted in the creation of the Open Data Quality Measurement Framework.

Langer et al. [6] describe the quality assessment tool *SemQuire*, which enables the assessment of Linked Open Data sources based on DQV. Building on a semantic literature review by Zaveri et al. [16] they developed a list of metrics segmented into four dimensions: Accessibility, Contextual, Intrinsic, and Representational. The implementation consists of a user interface, a RESTful API, a set of implemented metrics and the graph database Stardog. The input data can be specified via a direct upload or by fetching a URL or a SPARQL endpoint. Afterwards, the desired metrics can be selected by the user, and the analysis is executed. The measurement results are then shown in the user interface and can be exported into DQV with an overall score.

Another implementation developed by Neumaier et al. [8] is the *Open Data Portal Watch* framework. The metrics used are based on previous work by Reiche

[7] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data.
[8] https://opendefinition.org/.
[9] https://www.w3.org/TR/vocab-dqv/.
[10] https://www.w3.org/TR/shacl/.

et al. [9] and refer to the existing metadata keys of DCAT. They are divided into five dimensions: Existence, Conformance, Retrievability, Accuracy, and Open Data. In contrast to SemQuire, the Open Data Portal Watch framework contains an additional harvesting component that enables the aggregation of data from different Open Data Portals based on various technologies (CKAN, Socrata, OpenDataSoft). The measurement results are shown in a user interface and can be downloaded as a CSV or PDF report. The work of Neumaier et al. [8] serves as the basis for further implementations such as the ODPQ Dashboard [5] and the ADEQUATe platform [7] which enhance the feature set of the Open Data Portal Watch framework, mainly by providing a more advanced dashboard.

In addition to the solutions mentioned above, current research offers implementations with a dedicated focus on evaluating the FAIR principles. These include the *FAIR Evaluator* by Wilkinson et al. [15], which analyzes open datasets using 15 metrics based on the FAIR principles and presents the results in a user interface, as well as the *FAIR Checker* by Rosnet et al. [10] and the FAIR data assessment tool *F-UJI* by Devaraju and Huber [2,3].

3 DCAT-AP Quality Metrics

To determine metadata quality we defined a set of quality metrics called DCATAP Quality Metrics (DCAT-AP-QM) for metadata sets of catalogs, datasets and distributions using DCAT-AP. These metrics are based on the FAIR and 5-star principles and their application results in measurements with qualitative values as well as (aggregated) scores.

3.1 Designing the DCAT-AP Quality Metrics

Inspired by Wang and Strong [13], so-called dimensions are used to categorize different aspects of metadata quality within our DCAT-AP-QM. For each of these abstract classes, metrics are defined that test certain criteria (e.g. timeliness). For each of these metrics, a qualitative or quantitative value describes the metadata quality by directly measuring it (e.g. "yes" in case of timeliness).

For our quality metrics, we define four dimensions, which are in line with the FAIR principles: Findability, Accessibility, Interoperability and Reusability. Additionally, a fifth dimension that emphasizes contextual usability is added: Contextuality. For each of these dimensions metrics are defined adapting the FAIR and the 5-star principles. In the following, for each dimension, the defined metrics are described and the corresponding FAIR (sub-)principles and 5-star principles are detailed. Some metrics are not checked by Piveau Metrics, as they are either out of scope or are always fulfilled due to the way Piveau is implemented. A detailed overview of unchecked principles is provided at the end of this section giving details on why this is the case.

For each metric defined, the name of the tested metadata property is assigned (e.g. Keyword Availability). The semantic representation of this metadata property is added in brackets (e.g. dcat:keyword) consisting of the short version of

the respective namespace[11] and the property name. Among the metrics defined, most test either the presence of a certain property ("Availability Metrics", e.g. Keyword Availability) or the matching of certain metadata with values of controlled vocabularies ("Vocabulary Alignment Metrics", e.g. License Vocabulary Alignment). In case of multidimensional properties for Availability Metrics the number of instances is not taken into account, only the sheer presence is measured. Both metric types store their results as boolean values. Additionally, there are some special metrics whose functionality and values will be described in more detail in the following.

Findability. [*Keyword Availability* (dcat: keyword), *Category Availability* (dcat: theme, dct: subject), *Spatial Availability* (dct: spatial) and *Temporal Availability* (dct: temporal)] The metrics cover F2 and R1 of the FAIR principles stating that data should be described with rich metadata.

Accessibility. [*Access URL Status Code* (dcat: accessURL), *Download URL Availability* (dcat: downloadURL) and *Download URL Status Code* (dcat: downloadURL)] The first and third metric describe whether the two specified endpoints can be reached via an HTTP request. The status code returned is used as the value of the measurement. Unlike the download URL, the existence of the access URL is not checked, since it is a mandatory property of DCAT-AP and therefore must be available. The metrics fulfill A1 and A1.1 of the FAIR principles, which state that metadata should be retrievable by their identifier using a standardized, open, free and universal protocol.

Interoperability. [*Format and Media Type Availability* (dct: format, dct: mediaType), *Format and Media Type Vocabulary Alignment* (dct: format, dct: mediaType), *Format and Media Type Non Proprietary* (dct: format, dct: mediaType), *Format Machine Interpretable* (dct: format) and *DCAT-AP Compliance*] The first two metrics check the presence of media type and format information and its vocabulary alignment according to the controlled vocabularies of DCAT-AP. The metrics that deal with the non-proprietary nature and the machine-readability of the format also check against controlled vocabularies but these vocabularies have been defined especially for this purpose[12]. The fifth metric, DCAT-AP compliance, is tested by validating the metadata against the DCAT-AP SHACL shapes[13]. As soon as at least one issue occurs, the metadata is not compliant. This check covers I1 of the FAIR principles demanding the use of a formal, accessible, shared and broadly applicable language for knowledge representation. All vocabulary checks cover I2 of the FAIR principles and the four star level of the 5-star principles requiring the representation of resources using a vocabulary (URIs). Checking the given format for non-proprietary and machine-readability applies to the two star level as well as the three star level of the 5-star principles.

[11] dct: http://purl.org/dc/terms/, dcat: http://www.w3.org/ns/dcat#.
[12] https://gitlab.com/dataeuropa/vocabularies/.
[13] https://github.com/SEMICeu/DCAT-AP/tree/master/releases/2.1.1.

Reusability. [*License Availability* (dct: license), *Known License (License Vocabulary Alignment)* (dct: license), *Access Rights Availability* (dct: accessRights), *Access Rights Vocabulary Alignment* (dct: accessRights), *Contact Point Availability* (dct: contactPoint) and *Publisher Availability* (dct: publisher)] Testing the presence of a license covers R1.1 of the FAIR principles demanding the declaration of one. R1 of the FAIR principles, namely a rich description of data using a plurality of accurate and relevant attributes is tested by checking the presence of an access rights description. The usage of controlled vocabularies aligns to I2 of the FAIR principles and the four star level of the 5-star principles which demand the linking of resources using URIs. Additionally, the metrics that look at the contact point and the publisher meet the requirement of R1.2 of the FAIR principles demanding a detailed provenance description.

Contextuality. [*Rights Availability* (dct: rights), *Bytesize Availability* (dcat: byteSize), *Date Issued Availability* (dct: issued) and *Date Modified Availability* dct: modified)] These metrics cover I2 and R1 of the FAIR principles which state that data should be described by rich and relevant metadata.

While the metrics described above cover large parts of the FAIR and the 5-star principles, some (sub-)principles were not considered. F1 and F3 of the FAIR principles demand the usage and integration of a global identifier within the metadata. In order to even be accessible by our tool each DCAT-AP dataset has to have a URL which serves as global identifier, and hence, is a prerequisite. F4 of the FAIR principles, requiring that data should be indexed in a searchable resource, is also not tested, as our tool is designed for an environment that already offers such an index. A2 of the FAIR principles cannot be tested, since (meta)data that is examined is necessarily available. I3 of the FAIR principles as well as the five star level of the 5-star principles are not tested because checking the interlinking of data is out of scope for our tool. R1.3 of the FAIR principles demanding tests against domain-specific standards is also not covered. Our tool is intended for Open Data portals that store data of any discipline, and therefore testing against specific standards is not reasonable.

3.2 Applying the DCAT-AP Quality Metrics

The previously defined metrics serve as a basis to make measurements for each specific metadata set. Statements about the quality of the respective metadata record can then be derived from the totality of the values obtained in this way.

The results are purely qualitative and in most cases only describe the presence of properties in the metadata or the use of controlled vocabularies. While it is necessary to know these details to evaluate and improve the metadata set, they do not provide a quickly ascertainable indication of its overall quality. For this reason, a quantification of each metric in the form of a score is derived to describe the fulfillment of this metric in a quick and easy way.

Aspects that are considered particularly important are given a higher score than less important aspects. An essential criterion for assigning individual maximal achievable scores is the classification of metadata properties into relevance

classes according to DCAT-AP (mandatory, recommended, optional). In addition, the importance of the metric in the context of the FAIR and the 5-star principles influences the individual maximal achievable scores.

Each score is computed based on the test results of each metric. Most metric values are boolean, so a value of true receives the maximum score and false a score of zero. There are three metrics that test aspects different to the presence of a property or the use of a controlled vocabulary: In case of DCAT-AP Compliance, the agreement of the metadata with DCAT-AP is tested. As soon as at least one issue is found, the test is considered negative and the score of the metric is zero. In case of the access URL (Access URL Status Code) and the download URL (Download URL Status Code), the returned status must be a HTTP success code 2xx to indicate a successful request, and to receive the maximum score.

Aggregated quality scores enable the comparison of the metadata quality of different catalogs, datasets and distributions. Overall quality scores are computed by summarizing the individual scores. These are provided for each catalog, each dataset and each distribution. Aggregated quality scores are available for each of the five dimensions as well as overall. The higher the ratings, the higher the quality of the metadata set.

3.3 DCAT-AP Quality Metrics Data Model

Table 1. Quality Measurement

DQV Quality Measurement
rdfs:type dqv:QualityMeasurement
dqv:isMeasurementOf
dqv:value
dqv:computedOn
prov:generatedAtTime

Table 2. Quality Annotation

DQV Quality Annotation
rdfs:type dqv:QualityAnnotation
oa:hasBody
dqv:inDimension
oa:motivatedBy
dc:isVersionOf
oa:hasTarget
prov:generatedAtTime

Both the measurements of the metrics as well as the calculated scores are stored using our custom DQV data model. All measurements are stored as *DQV Quality Measurements* except for the DCAT-AP Compliance metric, i.e. the SHACL validation report, which is persisted as *DQV Quality Annotation*. The resulting quality metrics graph contains a set of those classes, one for each metric defined, including additional properties providing detailed information.

The Quality Measurement (see Table 1) describes the metric tested (dqv: isMeasurementOf), the test result (dqv: value), which resource was tested (dqv: computedOn) and when the result was measured (prov: generatedAtTime). The Quality Annotation (see Table 2) includes the SHACL validation (oa: hasBody) property, the dimension this metric is part of (dqv: inDimension) as well as

the description of the motivation for the creation of the annotation (oa: motivatedBy). The DCAT-AP version (dc: isVersionOf), the resource the test was performed on (oa: hasTarget) and the point in time those results were generated (prov: generatedAtTime) are also described.

4 A Scalable Metrics Pipeline

This section presents our practical implementation of the DCAT-AP-QM, that we call Piveau Metrics. It consists of four major layers: A persistence layer where the quality data is stored, a pipeline layer, that periodically creates the quality measurements and assessments for a given corpus of DCAT-AP datasets, a service layer that processes the generated data and provides an API for further usage and a UI layer that presents the results to the end user. Each layer consists of several sub-components, where each one is implemented as an individual Web service. Piveau Metrics follows a microservice architecture making the solution highly scalable and extendable. All services were developed in Java and Kotlin and support a container-based cloud deployment. Figure 1 illustrates the layers and their respective services.

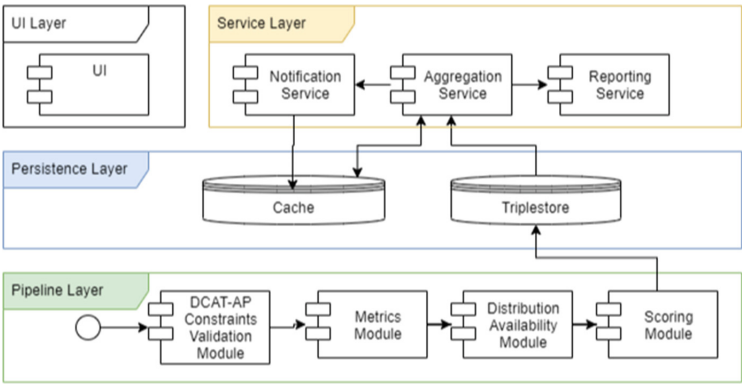


Fig. 1. Overview of Piveau Metrics Layers

4.1 Persistence Layer

The Persistence Layer comprises two different storage solutions: a triplestore graph database and a document database serving as cache. The triplestore is used to store the metric measurements and scores as RDF encoded with DQV. For each DCAT-AP dataset a dedicated named graph is generated. An important feature is that existing graphs are not overwritten, but new ones are created each time the pipeline is triggered providing a history of all previous measurements. The cache is used to store (aggregated) scores that are derived from the metrics graphs - for instance catalog quality scores. This allows for much faster access to this information compared to retrieving them on-the-fly from the DQV data.

4.2 Pipeline Layer

The Pipeline Layer consists of four modules determining the actual scores for each dataset. Ideally, the pipeline is triggered when a dataset is created or an existing one gets updated. Each DCAT-AP dataset passes the services of this layer in a predefined order, where the measurements for each metric and the (aggregated) quality scores are determined. The resulting DQV graph is sent to the triplestore. This pipeline is flexible and easily extendable, so that there is a straightforward path to adding new validation services.

- The *DCAT-AP Constraints Validation Module* constitutes the entry point of the pipeline. It validates the dataset against the official SHACL rules of DCAT-AP. It can manage multiple versions of SHACL shapes to evolve with the standard and support domains beyond DCAT-AP.
- The *Metrics Module* applies the main part of our DCAT-AP-QM to the dataset and returns the result as a DQV encoded payload. It iterates over all properties and applies the defined validations as described in Sect. 3.1.
- The *Distribution Availability Module* checks the availability of each distribution by validating if the access and download URLs are reachable and downloadable. To save resources an HTTP head request is used for the check. If this check fails, an HTTP GET request can also be utilized.
- The *Scoring Module*, as final service of the pipeline, takes all results from previous services and builds the metrics graph incorporating all measurements. In addition, it calculates quality scores for each dataset, each dimension and one overall quality score and adds them to the metrics graph. The complete metrics graph is then stored in the triplestore.

4.3 Service Layer and UI Layer

The Service Layer consists of a set of services that uses the metrics results and shows certain result items in specific formats to the end user. Each service has a scheduling component that can be configured individually.

- The *Aggregation Service* retrieves the current quality scores of datasets and aggregates them into catalog scores and an overall score. These aggregated metrics are stored, ordered by type and aggregation date, in the cache. The Aggregation Service provides an API to return the most recent aggregates as well as averages for specific time frames. Apart from that, an API is offered to retrieve other results from the triplestore that were generated by the Pipeline Layer, e.g. specific measurements for datasets and distributions. These APIs are used by other services in the Service Layer and by the UI Layer.
- The *Reporting Service* provides human-readable and processed reports of the measurements in PDF, ODS or XLSX format. It uses the Aggregation Service API to retrieve the current measurements and generates these reports on a predefined schedule.

– The *Notification Service* requests the measurements and scores for a catalog from the Aggregation Service and sends a notification to the data provider in case of a score deterioration for that specific catalog. The schedule for this service can be activated individually for each catalog.

The *UI Layer* provides an easy-to-use access to the measurements and scores for end users. It is a Web frontend providing diagrams and detailed information about each metric and their evolution in terms of quality. Next to the Notification Service it serves as an access point for data providers.

5 Evaluation

We evaluated our approach on two levels. Firstly, we compared the features of Piveau Metrics with existing and similar approaches to validate the novelty and relevance for the domain of open DCAT-AP datasets. Secondly, we performed a long-term test in a production environment to validate the practical feasibility and impact of the software.

5.1 Feature Comparison

The feature evaluation is based on a set of indicators that cover the theoretical concept and the practical applicability of each solution. The main focus is to evaluate the benefit for the domain of Open Data.

The *Data Check* feature allows to check the accessibility and validity of resources. The *FAIR*, *5-star*, and *DCAT-AP Support* features indicate the consideration of these principles in the different solutions. The *User Interface (UI)* and the *Application Programming Interface (API)* feature ensure that both is available to the user. The *Export* feature allows users to export the measurements as a report (e.g. PDF, JSON) and the *Notification* feature allows users to receive notifications in case a score decreases. The *Score Comparison* feature enables the ranking and comparison of catalogs, datasets and/or distributions (Table 3).

Table 3. Feature Comparison of Open Data Quality Tools

Feature	Piveau Metrics	Sem-Quire	Open Data Portal Watch	FAIR Evaluator	FAIR Checker	F-UJI
Data Check	x	x	x	x	x	x
FAIR Support	x	-	-	x	x	x
5-star Support	x	-	-	-	-	-
DCAT-AP Support	x	-	x	-	-	x
UI/API	x/x	x/x	x/x	x/x	x/x	x/x
Export	x	x	x	x	x	x
Notification	x	-	-	-	-	-
Score Comparison	x	-	-	-	-	-

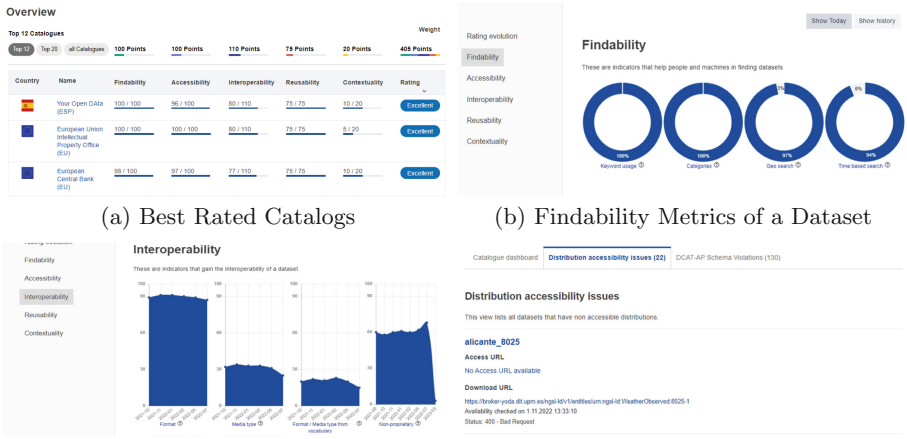
The evaluation reveals that the related projects utilize distinct standards and principles as a basis for their quality assessment. Similar to our approach, SemQuire utilizes the DQV, and the Open Data Portal Watch framework suggests a set of metrics within the scope of the DCAT specification. The authors of the FAIR Evaluator, the FAIR Checker, and the F-UJI solutions focus on representing the FAIR principles in their implementations. In contrast to our approach, most other solutions restrict the analysis to a single resource at a time. Also, none of the related projects provides a dashboard that presents the quality measurements and quality scores in a comparative view.

5.2 Use Case: data.europa.eu

We applied our solution in a large-scale real-world production system to evaluate its feasibility, scalability, and possible impact on the data quality. Therefore, Piveau Metrics was tightly integrated into the metadata registry and acquisition components of data.europa.eu[14]. The portal is provided by the European Commission and constitutes the central aggregation point for European Open Data. As of March 2023 it lists more than 1.6 million datasets, gathered from more than 170 regional, national, and pan-national data catalogs. It applies DCAT-AP as core data model and storage format. A detailed overview of the underlying software architecture of data.europa.eu can be found in [4]. The Pipeline Layer to create the actual quality information is integrated in the harvesting process, where the metadata is retrieved from the various data sources regularly. Each DCAT-AP dataset is forwarded to the Pipeline Layer, processed and the resulting DCAT-AP-QM is stored alongside the actual dataset in the triplestore of data.europa.eu. The Service Layer is retrieving the quality information from the triplestore to feed the Aggregation Service, the Notification Service and the Reporting Service. (cf. Fig. 1) A dedicated user interface (UI Layer), the Metadata Quality Dashboard (MQD)[15] acts as comprehensive interface providing multiple aggregations (provided by the Aggregation Service) of the scores for the different dimensions and metrics (e.g. the scores for a specific point in time and/or catalog). Figure 2 shows a selection of views. For each catalog users can access a dedicated view and download the reports (provided by the Reporting Service) in multiple formats. Furthermore, for readability the score is transformed into a simple rating with four ranges: excellent, good, sufficient, and bad.

The system is in place since September 2021, when it monitored 74 catalogs. Since then the portal has grown and Piveau Metrics is constantly monitoring the plethora of datasets and catalogs. As of March 2023 the triplestore holds more than 64 million discrete quality values for the current corpus of datasets. The historic data sums up to more than 1.4 billion quality values. Hence, from a technical point of view, including feasibility and scalability, our solution can be successfully applied in a production use case.

[14] https://data.europa.eu.
[15] https://data.europa.eu/mqa.



(a) Best Rated Catalogs (b) Findability Metrics of a Dataset

(c) Interoperability Metrics of a Dataset (d) Accessibility Issues of a Dataset

Fig. 2. Overview of MQD in data.europa.eu

One objective of our work is the lasting quality improvement of Open Data. Therefore we examined, how the quality evolved over time. The Aggregation Service allows to retrieve scores for specific points in time[16]. We retrieved the overall scores and scores for each of the five dimensions for each catalog and calculated the average across all catalogs. In order to include as many catalogs as possible we chose January 2022 as baseline and the current month March 2023 as comparison. Within this period complete measurements for 164 catalogs are available[17]. Table 4 shows the results and indicates a slight tendency towards better data quality. The overall score has improved and the average rating moved from sufficient to good. Accordingly, the values for the five dimensions improved, with the exception of findability, that dropped minimally. The table also shows the percentage of catalogs that have good/excellent ratings and bad ratings. The increase in the first category and the decrease in the latter reveals a positive progress. In general, we do not claim a correlation between the application of our tool and the improved quality, since many aspects can contribute to this and datasets are constantly added and removed. However, our solution supports a transparent and fine-grained evaluation of the metadata quality. Successive quality improvements can be more evidence-based by both, portal operators and data providers.

[16] https://data.europa.eu/api/mqa/cache/.
[17] The raw data can be found here: https://doi.org/10.5281/zenodo.8016840.

Table 4. Average Scores, Values and Ratings between 2022 and 2023

	2022-01	2023-03
Overall Score	218	225
Overall Rating	Sufficient	Good
Findability Value	72	71
Accessibility Value	53	54
Interoperability Value	38	42
Reusability Value	51	52
Contextuality Value	4	7
Good and Excellent Ratings	54%	56%
Bad Ratings	17%	13%

6 Conclusions and Future Work

In this paper we have presented our methodology, data model and practical implementation for assessing and reporting the quality of DCAT-AP datasets. DCAT-AP is a widely adopted RDF-based specification for describing metadata of Open Data. Although DCAT-AP defines many expressive properties and vocabularies to be used, a qualitative and accurate implementation by Open Data providers is challenging. Therefore, we designed quality metrics for DCAT-AP datasets based on a practical view on the FAIR and 5-star principles. We propose a set of specific metrics within the five dimensions findability, accessibility, interoperability, reusability and contextuality. In essence, these metrics cover the valid assignment of critical properties, the compliance with the DCAT-AP specification based on SHACL, and the availability of the actual data. Based on the values of these metrics, we determine overall scores allowing to assess and compare the quality of datasets. The results of the quality evaluations are stored in a custom RDF model based on the Data Quality Vocabulary (DQV), called DCAT-AP-QM data model. We implemented our approach as a scalable and reusable solution to demonstrate its feasibility and implications. Our software is called Piveau Metrics and mainly divided into two processing layers: a pipeline layer to constantly calculate the metrics over a corpus of DCAT-AP datasets and a service layer to provide the results and aggregation reports to applications and users. We compared our approach with existing work in the field of Open Data quality assessment and showed that Piveau Metrics offers the broadest set of features for our application scenario. In addition, with data.europa.eu, we applied our solution in a large-scale production environment. It constantly checks more than 1.6 million DCAT-AP datasets and provides quality reports to the data providers. The DCAT-AP-QM data model and Piveau Metrics is available as Open Source[18].

[18] https://gitlab.com/piveau/metrics.

With our work we have shown, that the FAIR principles, the 5-star principles and established RDF standards, such as SHACL and DQV constitute an appropriate foundation to measure and report the quality, completeness and validity of DCAT-AP datasets. This effectively can close the gap between the formal specification and the practical difficulties in applying DCAT-AP.

With data.europa.eu we have built a showcase to demonstrate how quality reports, scoring and rating can be communicated to data providers and interested users. We believe that such an open communication is crucial to increase the quality of Open Data in the future. It introduces a certain degree of gamification and can nudge data providers to improve their data. However, the evolution of the scores illustrated in Sect. 5.2 are only showing a slight improvement. This indicates that data providers need to engage more with the reports and incorporate the insights into their publication processes. Therefore, we aim to improve the feedback loop of our approach and introduce more notification and alert features towards the data providers.

The service-based architecture allows to integrate Piveau Metrics into a variety of management solutions for DCAT-AP. Piveau Metrics is under active development and constantly adapted to changes around the DCAT-AP specification, such as the introduction of Data Services. We want to refine and broaden our quality metrics and include additional aspects, such as the CARE principles[19] and other best practices for Open Data publication. We also intend to extend our quality metrics to data itself, supporting file-type-specific metrics to evaluate the quality of the actual data.

Acknowledgements. This work has been funded by the European Commission under framework contract 10801 (European Data Portal Managed Services - data.europa.eu), by the Federal Ministry of Education and Research of Germany (BMBF) under grant number 16DII138 (Weizenbaum-Institut) and by the German Research Foundation (DFG) under project numbers 441926934 (NFDI4Cat) and 460234259 (NFDI4DataScience).

The authors would like to thank our colleagues Benjamin Dittwald, Fritz Franzke, and Simon Dutkowski for contributing to the development and architecture of Piveau Metrics.

References

1. European Commission: About DCAT application profile for data portals in Europe | Joinup (2021). https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/about
2. Devaraju, A., Huber, R.: F-UJI - an automated FAIR data assessment tool (2020). https://doi.org/10.5281/zenodo.4063720
3. Devaraju, A., Huber, R.: F-UJI : An Automated Assessment Tool for Improving the FAIRness of Research Data (2020). https://doi.org/10.5281/zenodo.4068347
4. Kirstein, F., Stefanidis, K., Dittwald, B., Dutkowski, S., Urbanek, S., Hauswirth, M.: Piveau: a large-scale open data management platform based on semantic web

[19] https://www.gida-global.org/care.

technologies. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12123, pp. 648–664. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49461-2_38
5. Kubler, S., Robert, J., Neumaier, S., Umbrich, J., Le Traon, Y.: Comparison of metadata quality in open data portals using the analytic hierarchy process. Gov. Inf. Q. **35**(1), 13–29 (2018). https://doi.org/10.1016/j.giq.2017.11.003. https://hal.science/hal-01672652
6. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: SemQuire - assessing the data quality of linked open data sources based on DQV. In: Pautasso, C., Sánchez-Figueroa, F., Systä, K., Murillo Rodríguez, J.M. (eds.) ICWE 2018. LNCS, vol. 11153, pp. 163–175. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03056-8_14
7. Neumaier, S., Thurnay, L., Lampoltshammer, T.J., Knap, T.: Search, filter, fork, and link open data: the adequate platform: data- and community-driven quality improvements. In: Companion Proceedings of the The Web Conference 2018, WWW 2018, pp. 1523–1526. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3184558.3191602
8. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. J. Data Inf. Q. **8**(1), 1–29 (2016). https://doi.org/10.1145/2964909
9. Reiche, K.J., Höfig, E., Schieferdecker, I.: Assessment and visualization of metadata quality for open government data. In: Proceedings of the International Conference for E-Democrazy and Open Government, CeDEM 2014 (2014)
10. Rosnet, T., Lefort, V., Devignes, M.D., Gaignard, A.: FAIR-Checker, a web tool to support the findability and reusability of digital life science resources (2021). https://doi.org/10.5281/zenodo.5914307
11. Vetro, A., Canova, L., Torchiano, M., Minotas, C., Iemma, R., Morando, F.: Open data quality measurement framework: definition and application to open government data. Gov. Inf. Q. **33**, 325–337 (2016). https://doi.org/10.1016/j.giq.2016.02.001
12. W3C: Data Catalog Vocabulary (DCAT). https://www.w3.org/TR/vocab-dcat/
13. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996). https://doi.org/10.1080/07421222.1996.11518099
14. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
15. Wilkinson, M.D., et al.: Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci. Data **6**, 174 (2019). https://www.nature.com/articles/s41597-019-0184-5
16. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semantic Web **7**, 63–93 (2015). https://doi.org/10.3233/SW-150175