



HAL
open science

An Open Platform for Quality Measures in a Linked Data Index

Pierre Maillot, Jennie Andersen, Sylvie Cazalens, Catherine Faron, Fabien Gandon, Philippe Lamarre, Franck Michel

► **To cite this version:**

Pierre Maillot, Jennie Andersen, Sylvie Cazalens, Catherine Faron, Fabien Gandon, et al.. An Open Platform for Quality Measures in a Linked Data Index. WWW '24: The ACM Web Conference 2024, May 2024, Singapore, Singapore. pp.1087-1090, 10.1145/3589335.3651443 . hal-04575211

HAL Id: hal-04575211

<https://inria.hal.science/hal-04575211v1>

Submitted on 14 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

An Open Platform for Quality Measures in a Linked Data Index

Pierre Maillot

Univ. Cote d’Azur, Inria, CNRS, I3S
Sophia-Antipolis, France
pierre.maillot@inria.fr

Jennie Andersen

Univ. Lyon, INSA
Lyon, CNRS, UCBL, LIRIS, UMR5205
Lyon, France
jennie.andersen@insa-lyon.fr

Sylvie Cazalens

Univ. Lyon, INSA
Lyon, CNRS, UCBL, LIRIS, UMR5205
Lyon, France
sylvie.cazalens@insa-lyon.fr

Catherine Faron

Univ. Cote d’Azur, Inria, CNRS, I3S
Sophia-Antipolis, France
catherine.faron@inria.fr

Fabien Gandon

Univ. Cote d’Azur, Inria, CNRS, I3S
Sophia-Antipolis, France
fabien.gandon@inria.fr

Philippe Lamarre

Univ. Lyon, INSA
Lyon, CNRS, UCBL, LIRIS, UMR5205
Lyon, France
philippe.lamarre@insa-lyon.fr

Franck Michel

Univ. Cote d’Azur, CNRS, Inria, I3S
Sophia-Antipolis, France
franck.michel@inria.fr

ABSTRACT

There is a great diversity of RDF datasets publicly available on the web. Choosing among them requires assessing their “fitness for use” for a particular use case, and thus, finding the right quality measures and evaluating data sources according to them. However, this is not an easy task due to the large number of possible quality measures, and the multiplicity of implementation and assessment platforms. Therefore, there is a need for a common way to define measures and evaluate RDF datasets, using open standards and tools. IndeGx is a SPARQL-based framework to design indexes of Knowledge Graphs declaratively. We extend it to support more advanced data quality measures. We demonstrate our approach by reproducing two existing measures, showing how one can formalize and add measures using such an open declarative framework.

CCS CONCEPTS

• **Information systems** → Information systems applications; **Semantic web description languages**; • **Social and professional topics** → *Digital rights management*; *Copyrights*; *Licensing*; **Intellectual property**;

KEYWORDS

Quality, Quality metric, Linked Data, FAIR, SPARQL, RDF

ACM Reference Format:

Pierre Maillot, Jennie Andersen, Sylvie Cazalens, Catherine Faron, Fabien Gandon, Philippe Lamarre, and Franck Michel. 2024. An Open Platform for Quality Measures in a Linked Data Index. In *Companion Proceedings of the ACM Web Conference 2024 (WWW ’24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3589335.3651443>

1 INTRODUCTION

There is a wide variety of RDF datasets publicly available on the web, which is a great asset for research and business. To find which dataset is adapted for a specific use case, data portals build indexes of dataset descriptions and support metadata-based search features. Yet, simple metadata is not always sufficient to identify relevant datasets, and various measures may need to be evaluated on each candidate dataset. The size of datasets, their number, and the complexity of the measures create challenges for such evaluation. Moreover, different use cases call for different measures: different users may value different characteristics in choosing their datasets (e.g. update frequency, availability, provider’s identity, etc.). As a result, many measures are available to evaluate the general quality of a dataset. They can be broken down into many kinds of measurement, such as FAIRness, accountability, consistency, etc.

In this context, there is a need to consider open and extensible approaches to RDF dataset evaluation, where users can describe in a declarative format what they are looking for. In this paper, we propose a model, a method, and a tool to support the declarative definition of custom measures. These measures are to be included in a unified index of dataset descriptions. Using this framework, one can extract the data necessary to apply different measures whose definitions and evaluation results are stored as part of the index. This makes them easy to compare and reuse. To illustrate our proposition, we evaluate 81 different datasets from 18 publicly available SPARQL endpoints with two quality measures for dataset descriptions: FAIRness as defined by FAIR-Checker [4] and Accountability [2]. We show that implementing both measures in our framework enables their clear comparative study and their evaluation in a real setting.

2 BACKGROUND

Evaluation and monitoring of dataset quality is an active field of research [10]. For publicly available datasets in particular, several communities are pushing for the adoption of common good practices, e.g. FAIR principles [9]. This push has led to the implementation of methods to evaluate the compliance of datasets with them.

Ever since the FAIR principles [9] were stated in very general terms, academia and industry have been working to clarify and evaluate them. They are centered around four concepts: Findability, Accessibility, Interoperability, and Reusability of data. As a result, multiple approaches to evaluate the FAIRness of digital objects have been proposed. For the evaluation of knowledge graphs alone, many approaches exist. O'FAIRe [1] and FOOPS! [5] evaluate ontologies according to their definitions and the metadata they are annotated with. FAIR-Checker [4] and F-UJI [6] are online tools to evaluate RDF resources according to the metadata that are either dereferenceable or included in HTML pages documenting the resources. The life science community has designed and uses O'FAIRe, FOOPS!, and FAIR-Checker. To this end, their interpretation of the FAIR principles notably expects the evaluated objects to be referenced in specialized databases for the life science community. F-UJI is oriented toward scientific research data objects and relies on databases such as DataCite.

Andersen et al. [2] propose a measure of a specific quality aspect: accountability. It applies the LiQuID [8] metadata model on a dataset description to define a measure for knowledge graphs. The model emphasizes the description of each step of the life cycle of the dataset using a "WH" questioning approach (who, when, where, how, and what). For example, one of the "who" questions for the usage step checks for the intended audience of the dataset. Accountability is closely related to the notion of transparency of data. It is designed for open data portals and to understand how data were produced and for which purposes, for traceability and trust. Other existing quality measures provide insight into limited and very specific aspects of accountability. Accountability goes beyond the meta-information of the FAIR principles while being more detailed in the metrics that they share [2].

IndeGx [7] is both (1) a framework to profile KGs accessible through SPARQL endpoints using SPARQL-based rules declared in RDF, and (2) a queryable index of 339 SPARQL endpoints of the Linked Open Data. IndeGx can be easily extended by adding new rules to augment the description generated as an RDF dataset. It also generates and stores, in RDF, traces of the rules applied to each dataset and their results. Being based solely on SPARQL, IndeGx is limited due to the maximum time allowed by each endpoint for query evaluation, the support by each endpoint of some complex queries and some SPARQL clauses, e.g. VALUES, and the general availability of each endpoint.

3 DECLARATIVE QUALITY MEASURES

In this article, we use the IndeGx framework [7] to declare a set of rules for two KG quality measures: accountability [2] and FAIRness [4], enabling to evaluate datasets wrt. these measures and to compare the measures. For the FAIRness we chose the formalization proposed by Gagnard et al. [4] in FAIR-Checker, as it applies to general-domain KGs with minimal loss, as detailed in section 3.1.

We now present our implementation of the FAIRness and accountability measures using IndeGx, and the results of their comparison. The rules used in our experiments are publicly available¹ and can serve as a model for expressing new measures.

3.1 FAIR-Checker^I

Our implementation of FAIR-Checker's measure of FAIRness, called FAIR-Checker^I, relies on two sets of declarative rules. The first set extracts, from the targeted SPARQL endpoint, all the triples involving resources explicitly typed as a dataset. Hence, the FAIRness metrics can take into account all the classes and properties used in dataset descriptions. The second set measures the FAIRness of the datasets based on their extracted descriptions. These rules are the adaptation of each FAIR-Checker metric to the IndeGx framework. In its original implementation² FAIR-Checker uses the dereferenceable RDF data and the markup data included as JSON-LD in the webpage at the URI of a dataset. Since SPARQL queries cannot easily retrieve the JSON-LD content of a webpage, this content is not accessible to the IndeGx framework. Thus, our implementation of FAIR-Checker uses the dereferenceable RDF data and the description of the dataset URI in its corresponding SPARQL endpoint. The original implementation of FAIR-Checker also uses three websites as external sources, the Ontology Lookup Service³ (OLS), the Linked Open Vocabularies⁴ (LOV), and BioPortal⁵, to determine if classes and properties are part of a "known vocabulary", i.e. publicly shared on dedicated websites. Among these three sources, only LOV offers a SPARQL endpoint that can be exploited with IndeGx, while the other two are REST APIs that cannot be used in the current version of IndeGx.

This difference (web page vs. SPARQL endpoint) leads to slight differences in the FAIRness scores computed by FAIR-Checker and FAIR-Checker^I. Besides, the three rules of FAIR-Checker involving OLS and Bioportal may produce lower values in FAIR-Checker^I, i.e. the evaluation of shared metadata vocabularies (FAIR principle F2B), shared ontologies usage (I2), and the use of community standards (R1.3). However, OLS and BioPortal are specialized in bioinformatics ontologies. Since our implementation is domain-agnostic, the absence of these vocabulary sources is negligible when comparing FAIR-Checker^I and FAIR-Checker scores on a broader scope of KGs.

Debattista et al. [3] have shown that on average only 35% of the URIs are dereferenceable in the LOD cloud. This makes it less likely for the datasets found in SPARQL endpoints to have dereferenceable descriptions while they certainly have a description in the SPARQL endpoints they were found in. Thus, for the same dataset URI, FAIR-Checker is less likely to find a dataset description to evaluate than FAIR-Checker^I, introducing a bias in our comparison. This explains that the majority of datasets shown in Figure 1 have a higher FAIRness score with the IndeGx implementation than with the original implementation. The datasets on the left side of Figure 1, with a higher FAIRness score with FAIR-Checker than with

¹<https://doi.org/10.5281/zenodo.10787028>, <https://doi.org/10.5281/zenodo.10787430> and <https://doi.org/10.5281/zenodo.10794312>

²<https://fair-checker.france-bioinformatique.fr/>

³<https://www.ebi.ac.uk/ols/index>

⁴<https://lov.linkeddata.es/dataset/lov/>

⁵<https://biportal.bioontology.org/>

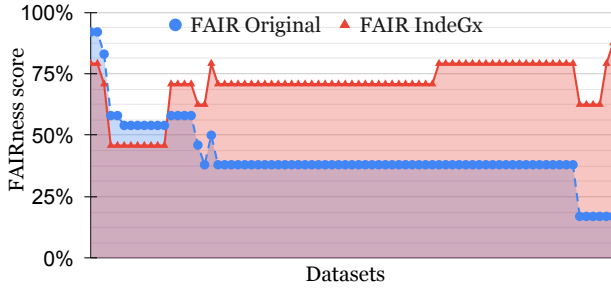


Figure 1: Comparison of the score of the original FAIR-Checker implementation and our IndeGx-based implementation, in ascending order of the difference in the two scores.

FAIR-Checker^I, almost all come from the same endpoint⁶ for which the provenance information required by the R1.2 FAIR principle is available in the metadata of the HTML page of the dataset’s URI but not in the SPARQL endpoint.

3.2 Accountability^I

The original implementation [2] of the accountability measure is based on the IndeGx framework. In our experiment, we name this implementation Accountability^I. Accountability^I proceeds in three steps, i.e. in the subsequent application of three kinds of rules. First, extraction rules restrict the target to the datasets related to the URL of their endpoint: they extract the description of any dataset that is linked to the URL of its endpoint. Then, a set of equivalence rules augments the extracted descriptions. They capture the knowledge that some properties are equivalent when describing datasets, as has been stated by Andersen et al. Finally, thirty evaluation rules are applied to measure the accountability of each extracted and augmented dataset description. The results of the evaluations are generated as RDF data added to the IndeGx KG.

3.3 Comparison FAIR-Checker^I-Accountability^I

Using a common framework enables the comparison of the methods underpinning the measures. Moreover, the scores obtained by evaluating several KGs according to different measures emphasize their similarities and differences in a complementary way.

Comparison of the methodologies for the extraction of dataset descriptions. Both FAIR-Checker^I and Accountability^I start by extracting the information required for the evaluation, and then query it locally. FAIR-Checker^I considers all the datasets of each endpoint, while Accountability^I limits itself to those linked to the URL of the endpoint under study. Both extract all triples with the selected dataset as the subject and do other specific extractions according to their respective needs, as explained above.

The difference in the source of the dataset descriptions between the two measures underlines a common problem with the delimitation of the metadata of a KG. Indeed, there are five possible operational methods for retrieving a KG’s metadata: extracting the metadata in the webpage associated to the KG’s URI, dereferencing the KG’s URI in RDF, querying the SPARQL endpoint of a KG to

retrieve the description of the latter, dereferencing the endpoint URL of a KG, dereferencing the “well-known” URI associated with the endpoint of a KG. These five methods may output different descriptions of a dataset, hence discrepancies between measures.

The major difference between FAIR-Checker^I and Accountability^I stems from the extra step added by the latter to enrich the data with property or class equivalences. This makes it easier to write measurement rules and it simplifies the introduction of new metrics with new vocabularies. However, this choice prevents the computation of statistics on the usage of vocabularies, as FAIR-Checker does.

Comparison of the methodologies for dataset evaluation. FAIRness and accountability quality measures are designed according to a specific understanding of the concept of “quality”. The conceptual difference between them appears clearly when both measures are implemented in the same framework.

The FAIRness measure according to FAIR-Checker and the accountability measure have some overlap in the details of their declarations. The findability dimension of the FAIRness measures the presence of at least one triple in the description of the dataset (F2.A) and to what extent known shared vocabularies are used in the description. The accountability measure of a dataset will be zero if it does not contain any dataset description using known shared vocabularies. Therefore, any dataset with a zero findability score will also have a zero accountability score. Similarly, the interoperability dimension of FAIRness measures the application of some basic requirements, i.e. shared vocabularies and a machine-readable format, and these requirements are necessary for the accountability of a dataset to be non-zero. Two FAIRness metrics – the inclusion of authorization procedure or access rights (A1.2) and license (R1.1) – look for the same triples as those used in the rights, location, and license requirements of the usage dimension of the accountability measure. The provenance evaluation metric (R1.2) of FAIR-Checker^I checks the presence of at least one property among a list of 30., whereas Accountability^I only gives full marks to a dataset description if most of the 30 properties describing provenance are present.

The overlaps and differences that we just highlighted allow us to conclude that datasets are more likely to have a high FAIRness than a high accountability as any dataset with a low FAIRness is likely to have a zero accountability score.

Comparison of the scores obtained by both measures. In our experiment, starting from the 339 endpoints usually evaluated by IndeGx, we were able to obtain a FAIRness score for 199 datasets from 25 endpoints and an accountability score for 99 datasets from 22 endpoints. The difference in endpoint coverage is attributed to the fact that the accountability evaluation is restricted to only datasets whose description contains any triple linking them to their endpoint. At the intersection, we evaluated the FAIRness and accountability of 81 datasets from 18 endpoints.

Figure 2 shows the FAIRness and accountability scores of the datasets of our experiment. The first peak of accountability compared to FAIRness corresponds to three datasets that contain extensive information about their creators and their sources in their descriptions. Conversely, there is a drop in accountability compared to FAIRness for datasets that contain no information about their creation whatsoever. Both measures capture different aspects of

⁶<http://ldf.fi/ww1lod/sparql>

