Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Quality of Metadata in Open Data Portals

JAVIER NOGUERAS-ISO[1], JAVIER LACASTA[1], MANUEL A. UREÑA-CÁMARA[2], AND F. JAVIER ARIZA-LÓPEZ[2]

[1]Aragón Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain
[2]Grupo de Investigación en Ingeniería Cartográfica (GiiC), Universidad de Jaen, Jaen, Spain

Corresponding author: Javier Nogueras-Iso (e-mail: jnog@unizar.es).

ABSTRACT During the last decade, numerous governmental, educational or cultural institutions have launched Open Data initiatives that have facilitated the access to large volumes of datasets on the web. The main way to disseminate this availability of data has been the deployment of Open Data catalogs exposing metadata of these datasets, which are easily indexed by web search engines. Open Source platforms have facilitated enormously the labor of institutions involved in Open Data initiatives, making the setup of Open Data portals almost a trivial task. However, few approaches have analyzed how precisely metadata describes the associated datasets. Taking into account the existing approaches for analyzing the quality of metadata in the Open Data context and other related domains, this work contributes to the state of the art by extending an ISO 19157 based method for checking the quality of geographic metadata to the context of Open Data metadata. Focusing on metadata models compliant with the Data Catalog Vocabulary proposed by W3C, the proposed extended method has been applied for the evaluation of the Open Data catalog of the Spanish Government. The results have been also compared with those obtained by the Metadata Quality Assessment methodology proposed at the European Data Portal.

INDEX TERMS ISO 19157, Metadata, Metadata Quality Assessment, Quality, Open Data

I. INTRODUCTION

WITH the increasing interest in facilitating government transparency or public participation, many governments have launched Open Data initiatives to release their data on the web [1]. This trend towards Open Government has also reactivated the publication of large volumes of data in other domains like science [2] or even the private sector, which considers Open Data as an enabler of innovation [3].

In general, the main mechanism to disseminate this availability of data has been the deployment of Open Data catalogs exposing metadata of these datasets, which are easily indexed by general web search engines or specialized dataset search engines like Google Dataset Search.[1] Additionally, these Open Data catalogs are designed to be highly interoperable as they must consume and be harvested from other catalogs with minimal technical agreements to allow the federation of contents.

One of these minimal agreements is the metadata schema used in these catalogs. DCAT is the "de facto" metadata standard in the Open Data context. DCAT [4] is the acronym for W3C's Data Catalog vocabulary, a W3C recommendation for describing open data. DCAT is a Dublin Core metadata profile based on RDF vocabulary that has been designed to facilitate interoperability between data catalogs published on the Web. More focused to the European context, the European Union proposed in 2013 DCAT-AP [5], a specification based on DCAT for describing public sector datasets in Europe. Compared to DCAT, DCAT-AP provides stricter definitions of catalogs, datasets, distributions and other objects. Its basic use case is to enable a cross-data portal search for datasets and make public sector data better searchable across borders and sectors. DCAT has been also adopted at national level in different regulations aiming to promote the reuse of information resources [6], [7].

Open Source platforms have facilitated enormously the labor of institutions involved in Open Data initiatives, making the setup of Open Data portals almost a simple task. For instance, CKAN [8], based on Python technology, is the most widely used Open source platform to support Open Data portals and includes the necessary plugins to exchange DCAT metadata in RDF format (*ckanext-dcat* plugin) and harvest or be harvested (*ckanext-harvest* plugin) by other Open Data catalogs. DKAN [4] is another Open Source alternative for Open Data portals, based on PHP and Drupal technology, which also provides support for DCAT-based metadata.

[1]https://datasetsearch.research.google.com/

However, as the main purpose of Open Source platforms for Open Data catalogs has been the easy and fast publication of datasets, there are currently few approaches devoted to the analysis of the quality of descriptions contained in metadata [9], [10], i.e. evaluate how precisely metadata describe the associated datasets. If metadata are not properly defined with enough quality, this hampers the discoverability and accessibility of resources through Open Data portals. The objective of this work is to extend the ISO 19157 based method for checking the quality of geographic metadata [11] to the context of Open Data metadata and study its differences with respect to other existing approaches for assessing the quality of Open Data metadata.

The ISO 19157 based method proposed by Ureña-Cámara et al. [12] adapts the ISO 19157 standard for geographic information quality to the metadata case. Apart from completeness and consistency, this quality standard reviews exhaustively the correctness of temporal, positional, and attribute information. However, this method cannot be directly applicable in the Open Data context because the geographic metadata models (together with their serialization formats) differ completely from the Open Data models. In addition, it must be noted that Open Data pays much more attention on metadata properties describing the distribution mechanisms of Open Data. Therefore, the extension of the ISO 19157 based method proposed in this work is far beyond cosmetic changes. Taking into account Open Data industry standards, we have adjusted the quality elements and associated measures to adequate them for an RDF-based metadata model. Besides, we have increased the number of automatic controls and they can be computed on-line accessing the SPARQL end-point of an Open Data catalog. Another difference with respect to the original ISO 19157 based method is that we propose to represent results in compliance with a formal vocabulary, the Data Quality Vocabulary (DQV) [13].

The rest of this paper is structured as follows. The state of the art is described in section II. Section III provides the necessary background on DCAT metadata and associated problems to understand the remainder of this paper. Section IV describes the extension of the ISO 19157 based method for metadata quality checking in the context of Open Data metadata. Section V shows the experiments performed with the Open Data metadata harvested from the Open Data catalog of the Spanish Government (*datos.gob.es*). The results of these experiments are discussed in Section VI. Finally, Section VII provides some conclusions and depicts future work.

II. STATE OF THE ART

Within the application domain of Digital Libraries, the professionals recognize the importance of counting on high quality metadata in order to allow users to discover and access resources [14]. In addition, as stated by Park [15], in this context it is commonly assumed that completeness, consistency and accuracy are the most common facets for evaluating metadata quality. This view is also shared by Gonçalves et al. [16], which propose a quality model for digital libraries based on a formal framework of five main concepts (Streams, Structures, Spaces, Scenarios, and Societies) with quality dimensions and indicators for these concepts. In this model, metadata are defined in terms of structures of atomic values where the main quality dimensions to be analyzed are accuracy, completeness, and conformance.

With the rising interest on Open Data portals to publish any type of electronic resource, the previous research on the metadata quality of Digital Library repositories has found a new research niche for evaluating the quality of metadata in Open Data portals. For instance, Veljković et al. [17] propose a five-indicator model to asses Open Government Data portals. Apart from evaluating the coverage of basic data themes, the participation of other governmental bodies and the collaboration of users, this model includes data openness and transparency indicators, which are directly linked to the availability of high quality metadata. Focusing on the quality dimensions that should be assessed on Open Data metadata, Neumaier et al. [18] propose an initial set of 18 metrics that are later extended by Kubler et al. [9] to define a total number of 21 metrics classified in five dimensions: *existence* of access, discovery, contact, rights, preservation, date, temporal and spatial properties; *conformance* of access URL, contact e-mail, contact URL, date format, license and file format; *retrievability* of datasets and resources; *accuracy* of format and file size; and an *Open Data* dimension, which checks the availability of open and machine readable formats with open license. In addition, Kubler et al. [9] propose a global measure to rank Open Data portals that is based on the Analytic Hierarchy Process [19], a technique commonly used in multi-criteria decision making to quantify the weights of decision criteria that must be combined into a single value. In this case, the criteria are the metrics that must be linearly combined to generate a global score.

It is also worth noting the existence of the Metadata Quality Assessment (MQA) methodology [10]. This methodology supports the development of a dashboard used within the context of the European Data Portal to provide an overview of the contents harvested from the different catalogs that contribute to this European portal. Inspired on the FAIR principles [20], which provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets, MQA proposes the use of 23 metrics classified in five dimensions: *findability*, which checks the availability of keywords, categories, spatial information and temporal information; *accessibility*, which checks the accessibility of access and download URLs (including their existence); *interoperability*, which checks the compliance with the DCAT-AP metadata model and the availability of well-known formats (if possible, non-proprietary and machine readable); *reusability*, which checks the description of license and access rights information as well as contact points and publishers; and *contextuality*, which checks the availability of information related to distribution rights, the file size of distributions and the dates of issue or modification. Each of these metrics can

be assigned a maximum number of points according to the percentage of metadata records verifying the check. The total points obtained for all metrics are used to rank catalogs in excellent, good, just enough or bad point ranges.

The research done with respect to metadata quality in the context of Spatial Data Infrastructures (SDI) is also relevant in this work. SDIs were defined in the nineties as the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data [21]. Therefore, we could consider SDIs as Open Data portals specialized in spatial data. With respect to the analysis of metadata quality in SDI catalogs, Ureña-Cámara et al. [12] have proposed a method based on ISO 19157 to evaluate different quality dimensions of ISO 19115 geographic metadata [22]. The ISO 19157 standard for geographic information quality [11] proposes a general data quality description framework for spatial data. Ureña-Cámara et al. [12] redefine ISO 19157 to evaluate metadata instead of data by proposing 16 metrics (which can be particularized for different types of metadata properties) associated with 12 quality dimensions in 6 different categories: *completeness*, with commission and omission dimensions; *logical consistency*, including conceptual, domain, format and topological dimensions; *temporal quality*, containing temporal consistency and temporal validity; *thematic accuracy*, which includes thematic classification correctness and non-quantitative attribute correctness; and two separate categories, out of ISO 19157, including *positional correctness* and *quality of free text*.

Last, for the analysis of the quality dimensions of metadata, we think that it is also important to take into consideration the bibliography related to quality dimensions in other more general contexts. In the case of software products, the ISO/IEC 25012 standard [23] proposes 15 quality dimensions that can be adopted to evaluate the quality. These quality dimensions are grouped into three categories: *inherent data quality*, which comprises accuracy, completeness, consistency, credibility and currentness; *inherent and system-dependent data quality*, which consists of accessibility, compliance, confidentiality, efficiency, precision, traceability and understandability; and *system-dependent data quality*, which includes availability, portability and recoverability. With respect to dataset quality, it is also remarkable the survey performed by [24] to gather a comprehensive list of 18 quality dimensions and 69 metrics applicable to the assessment of Linked Data. These dimensions are classified in four groups: *accessibility dimensions*, which includes availability, licensing, interlinking, security and performance; *intrinsic dimensions*, which includes syntactic validity, semantic accuracy, consistency, conciseness and completeness; *contextual dimensions*, which contains relevancy, trustworthiness, understandability and timeliness; and *representational dimensions*, which consists of representational-conciseness, interoperability, interpretability and versatility.

After reviewing the different approaches, we have found an important overlap between the dimensions considered in different approaches. Table 1 shows a broad matching between these dimensions from the perspective of evaluating the quality of metadata describing open resources. All the definitions of dimensions and associated metrics provide an interesting and complementary insight. Moreover, it is interesting to note that the name of a dimension may represent a different concept according to each approach: ISO 25012, Zaveri et al. [24], ISO 19157, Kubler et al. [9] and MQA. For instance, whereas the concept of *Accessibility* in ISO 25012 is related to the support provided for impaired people, MQA *Accessibility* is related to the network reachability of URLs.

This work describes an extension of the ISO 19157 based method (for analyzing geographic metadata) to the context of Open Data metadata. ISO 19157 provides a flexible framework that allows the adaptation to cover the dimensions considered in other quality models. In fact, the comparison with other models, even with those not focused on metadata (ISO 25012 and Zaveri et al. [24]), has remarked that some of the ISO 19157 quality elements (quality dimensions according to the ISO 19157 methodology) should take into account the issues related to the accessibility (reachability of URLs), the reusability (information related to licenses and rights) and the interoperability (a detailed analysis of formats). This can be achieved in ISO 19157 defining specific non-quantitative attribute correctness or conceptual consistency quality elements that put the focus on evaluating the metadata properties that refer to online resources for data download, license/right information or formats. In addition, the results obtained with our proposed method have been compared with the results obtained with MQA, which also has a great overlap with the dimensions and indicators proposed by Kubler et al. [9].

III. BACKGROUND INFORMATION ON METADATA MODELS BASED ON DCAT

The vocabularies based on DCAT are mainly focused on providing information about three main entities: *Catalogs*, *Datasets* and *Distributions*. *Catalog* properties inform about the institutional body in charge of publishing Open Data datasets. *Dataset* properties provide the main information for discovery and characterization of datasets. *Distribution* properties are mainly focused on the mechanisms for ordering or downloading the datasets.

As already mentioned in the introduction, DCAT-AP is an application profile of DCAT for describing the datasets and distributions published on Open Data portals that adds additional constraints on the metadata properties: minimum and maximum multiplicity of properties and stricter ranges. These constrains are important if we aim to evaluate metadata quality aspects such as completeness or consistency.

For the sake of simplicity and in order to facilitate the readability of the remainder of the paper, our proposed method for analyzing the quality of Open Data metadata is focused on the Spanish proposal for DCAT-based metadata. In the case of Spain, the "Technical Interoperability Standard for the Reuse of Information Resources" [6], usually known with the acronym NTI according to the first 3 initial letters in the

TABLE 1. Matching between dimensions of different approaches.

ISO 25012	Zaveri et al.	Kubler et al.	MQA	ISO 19157
Accuracy, Credibility, Precision	Semantic accuracy, Trustworthiness, Relevancy	Accuracy		Thematic classification correctness, Non-quantitative attribute correctness, Positional correctness
Completeness	Completeness, Conciseness, Representational-conciseness	Existence	Interoperability, Findability, Contextuality	Completeness omission, Completeness commission
Consistency, Compliance	Consistency, Syntactic validity, Interoperability	Conformance	Interoperability	Conceptual consistency, Domain consistency, Topological consistency, Temporal consistency, Format consistency
Currentness	Timeliness			Temporal validity
Availability	Availability, Interlinking	Retrievability	Accessibility	Non-quantitative attribute correctness
Accessibility, Confidentiality, Traceability	Licensing, Security	Open Data	Reusability	Non-quantitative attribute correctness
Portability, Efficiency, Recoverability	Interpretability, Performance	Open Data	Interoperability	Conceptual consistency
Understandability	Understandability, Versatility			Quality of free-text

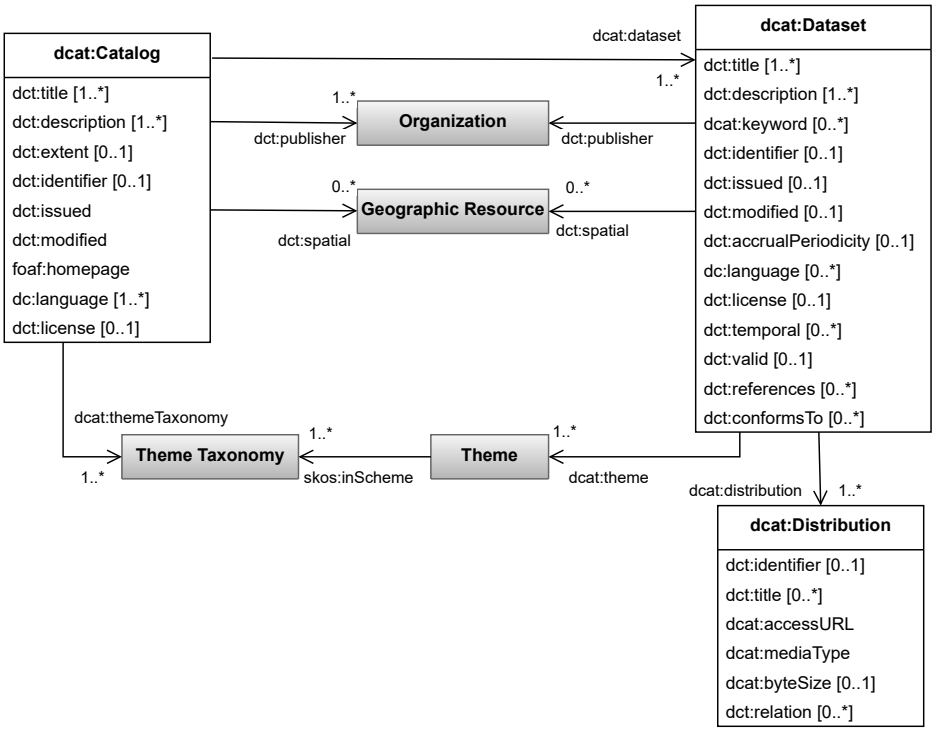

FIGURE 1. Main classes, properties and relations of the NTI metadata model [6]

Spanish name of the standard ("Norma Técnica de Interoperabilidad de Reutilización de Recursos de Información"), includes a DCAT-based metadata schema that must be used to describe public information resources. The NTI metadata model is a subset of DCAT-AP that contains representative properties of *Dataset* and *Distribution* entities. Figure 1 shows a UML diagram with the main classes, properties, and relations considered in NTI.

In order to illustrate the problems that may arise in metadata records published in Open Data portals, Figure 2 shows a NTI metadata record in Turtle format. It describes an artificial dataset of air quality observations in the urban area of Zaragoza (a city within the province of Zaragoza, Spain) that contains several problems:

- Completeness commission problem: The *Dataset* contains 2 *identifiers*, but it must have 1 *identifier* at maximum.
- Completeness omission problem: The *Dataset* does not contain the mandatory *publisher* property.
- Domain consistency problem: The *language* of the *Dataset* should be a value from a well-known linguistic system (e.g., "en" code of ISO 639-1) instead of a free-text literal ("English").
- Conceptual consistency problem: The *accessURL* of the *Distribution* is linked to a potential RDF file on the web,

```
<http://datos.gob.es/catalogo/real-time-air-quality-observations>
  a  dcat:Dataset ;
  dct:title "Real time air quality observations in the urban area of Zaragoza"@en ;
  dct:identifier "https://opendata.aragon.es/datos/catalogo/dataset/oai-zaguan-unizar-es-96845" ;
  dct:identifier "https://opendata.aragon.es/datos/catalogo/dataset/oai-zaguan-unizar-es-96846" ;
  dct:description "This dataset provides real time air quality observation data, for all the pollutants,
    at the location of each sensor."@en ;
  dc:language "English" ;
  dct:issued "2020-12-11T05:43:39.043550"^^xsd:dateTime ;
  dct:modified "2020-05-15T12:36:38.291865"^^xsd:dateTime ;
  dct:spatial <http://datos.gob.es/recurso/sector-publico/territorio/Provincia/Huesca> ;
  dcat:distribution <http://datos.gob.es/catalogo/real-time-air-quality-observations/resource/7d6a569502a8> ;
  dcat:theme <http://datos.gob.es/kos/sector-publico/sector/educacion> .

<http://datos.gob.es/catalogo/real-time-air-quality-observations/resource/7d6a569502a8>
  a  dcat:Distribution ;
  dct:format <http://datos.gob.es/catalogo/real-time-air-quality-observations/resource/7d6a569502a8/format>;
  dct:identifier
    "http://zaguan.unizar.es/record/96845/files/open_data%3Areal_time_air_quality_observations.rdf" ;
  dct:title "Real time air quality observations"@en ;
  dcat:accessURL
    "http://zaguan.unizar.es/record/96845/files/open_data%3Areal_time_air_quality_observations.rdf" .

<http://datos.gob.es/catalogo/real-time-air-quality-observations/resource/7d6a569502a8/format>
  a  dct:IMT ;
  rdf:value  "application/json" .
```

FIGURE 2. Metadata record in Turtle format containing several problems (the metadata content has been intentionally modified to remark potential problems).

but the indicated *format* seems to be JSON.

- Temporal consistency problem: The creation date (*issued* property) of the *Dataset* is older than the modification date (*modified* property).
- Thematic classification problem: Apparently, the *Dataset* is more related to the environment theme than to the education theme (*educacion* value in *theme* property).
- Positional correctness problem: If the *Dataset* compiles air quality observations in the city of Zaragoza, the spatial coverage should be the province of Zaragoza (instead of the adjacent province of Huesca).

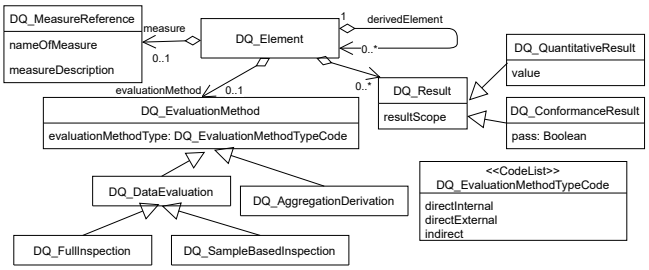

FIGURE 3. Components of a quality element in ISO 19157.

IV. EXTENSION OF ISO 19157 BASED METHOD FOR OPEN DATA METADATA

The method proposed by Ureña-Cámara et al. [12] analyzes a wide range of metadata aspects such as completeness, accuracy, and consistency according to the quality elements proposed by ISO 19157. 'Quality element' is the expression used by ISO 19157 to refer to quality dimensions, an expression more commonly used in other approaches cited in section II. Columns *Quality category* and *Quality element* in table 2 show the hierarchy of quality elements considered for metadata. This hierarchy was already described in section II.

According to ISO 19157, a quality element is also a part of a quality report. As depicted in Figure 3, a quality element is described by three components: a measure (or metric in other approaches), which is the system to measure something; an evaluation method, i.e. the procedure to evaluate the measure; and one or more results obtained as the output of the evaluation focused on a specific part (scope) of the dataset to be evaluated.

Using this definition of quality elements, the main goal of the ISO 19157 based method for metadata quality analysis is to provide quality controls. A quality control determines whether a parameter (quality dimension) of a product satisfies a specific requirement expressed as a quality level (e.g. no more than 5% of errors). In quality control, two different situations can occur. The first situation is when the automation of the control process is possible and the complete population can be checked (full inspection) for the type of errors that exist. The second situation occurs when automation is not possible and a sample-based control is used to derive a decision involving limited risks. The mechanism to implement quality controls in ISO 19157 is to define two related quality elements: a first quality element containing a quantitative result of a metadata-related measure; and a second quality element derived from the first one that contains a conformance result certifying whether the quantitative result of the first element satisfies the expected quality level or not.

Table 2 enumerates the measures proposed by Ureña-Cámara et al. [12] to provide the framework for defining quality elements containing quantitative results on a set of metadata records. Most of them are derived from the measures described in the tables of annex D in ISO 19157. There are also some measures introduced as "similar to" denoting that the mathematical construction of the referenced

TABLE 2. Quality elements of ISO 19157 based method and associated measures.

Quality category	Quality element	Measure	Measure description
DQ_Completeness	DQ_Completeness Commission	D.3 (ISO 19157)	Rate of records with excess items.
	DQ_CompletenessOmission	D.7 (ISO 19157)	Rate of records with missing items.
DQ_LogicalConsistency	DQ_ConceptualConsistency	D.13 (ISO 19157)	Rate of records compliant with the conceptual schema.
		Similar to D.22 (ISO 19157)	Number of records with inconsistent information in metadata elements.
	DQ_DomainConsistency	D.17 (ISO 19157)	Value domain conformance rate.
	DQ_FormatConsistency	D.21 (ISO 19157)	Physical structure conflict rate.
	DQ_TopologicalConsistency	D.23 (ISO 19157)	Rate of records having faulty relationships with other records in the catalog.
		Topological contradiction	Rate of records having faulty relationships between two metadata elements of the same record.
DQ_TemporalQuality	DQ_TemporalConsistency	Similar to D.62 (ISO 19157) using a rate	Rate of records with conflict time sequences.
	DQ_TemporalValidity	D.18 (ISO 19157)	Value domain non-conformance rate.
DQ_ThematicAccuracy	DQ_ThematicClassificationCorrectness	D.63 (ISO 19157)	Number of incorrectly classified records.
	DQ_NonQuantitativeAttribute Correctness	D.69 (ISO 19157)	Rate of incorrect attribute values.
		D.68 (ISO 19157)	Number of records with incorrect attribute values.
DQ_PositionalCorrectness		Similar to D.33 (ISO 19157)	Rate of records with positional errors: no overlapping between direct and indirect georeferences.
DQ_QualityOfFreeText		Overall quality of free text	Number of records using text values with a bad quality level.
		Readability of free text	Rate of records using text values considered readable with a readability index (e.g. Flesch) above a threshold.

measure is identical to the cited measure. It must be noted that the measures defined in terms of rates are expected to be applied in automatic evaluations making a full inspection of the population. In contrast, the measures defined in terms of number of correct/incorrect items are expected to be applied in manual evaluations making a sample-based inspection.

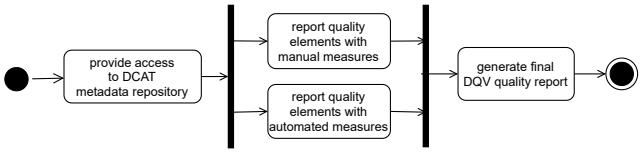

FIGURE 4. General workflow for reporting quality elements.

The description of measures in Table 2 has been written in an abstract way in terms of records and metadata elements. However, the proposed definition of quality elements in the original method [12] designed for ISO 19115 geographic metadata cannot be directly applied for Open Data metadata. For each quality element type, we need to decide whether it is pertinent in the context of Open Data metadata and identify the particular scopes that must be analyzed, i.e. the entities (*Dataset* and *Distribution*) and the specific metadata properties that must be reviewed. Then, for each quality element type and scope, we need to define a pair of quality elements: a quality element containing a quantitative result and explaining the manual or automatic procedure followed to evaluate the measure; and a related quality element with the conformance result.

Figure 4 shows the general workflow for reporting the quality elements. After accessing the metadata repository that must be evaluated, we can work in work in parallel with the quality elements linked to measures evaluated automatically and the quality elements linked to measures evaluated manually by experts. Finally, all the quality elements are compiled in a single DQV quality report. The following subsections describe in detail these tasks related to manual measures (section IV-A), automated measures (section IV-B) and the final report (section IV-C).

A. QUALITY ELEMENTS ASSOCIATED WITH MANUAL MEASURES

In order to implement quality controls and define conformance results, the ISO 19157 based method for metadata quality analysis uses the concept of Acceptance Quality Limit (AQL) for establishing a demanded metadata quality level. As defined by ISO 2859-1, the AQL represents the worst or poorest level of quality that would be considered acceptable as a process average (e.g., 5%). This parameter (AQL) is the key element of the series of ISO 2859 and ISO 3951 international standards, and it is adopted in our proposed method as 5%.

However, the AQL is only directly applicable for measures evaluated automatically on a full inspection basis. In the case of measures evaluated manually over a sample of the population, the limiting quality (LQ) must be used. This concept is statistically related to the AQL, and the ISO 2859-2 international standard offers this relation. Considering all metadata records as a whole (a unique set or lot), ISO 2859-2 provides the rules for quality control according to "Table A" (see an excerpt in Table 3). Thus, Table 3 is the one that provides the sample size to be taken for quality control. In a standardized quality control environment, the sample size is also crucial: it carries costs (more sample size is more cost) and risks from a statistical point of view (type I error and type II error). Thus, the use of this table and its values is mandatory. The input required to use this table is the size of the lot under control and a limiting quality (LQ) index that is thrice the AQL ($LQ \approx 3 \times 5\% = 15\%$). "Table A" of the international standard outputs the sample size to be randomly extracted and the maximum number of errors that can be allowed in this sample to ensure a 5% producer's risk and a

TABLE 3. Excerpt of "Table A - Single sampling plans indexed by limiting quality (LQ) (procedure A)" from ISO 2859-2 [25].

Lot size		Limiting quality in percent (LQ)									
		0.5	0.8	1.25	2.0	3.15	5.0	8.0	12.5	20	32
...											
51 to 90	n	→	→	90	50	44	34	24	16	10	8
	Ac			0	0	0	0	0	0	0	0
91 to 150	n	→	150	90	80	55	38	26	18	13	13
	Ac		0	0	0	0	0	0	0	0	1
151 to 280	n	200	170	130	95	65	42	28	20	20	13
	Ac	0	0	0	0	0	0	0	0	1	1
...											
3201 to 10000	n	450	315	315	200	200	200	125	80	80	80
	Ac	0	0	1	1	3	5	5	5	10	18
10001 to 35000	n	500	500	315	315	315	315	200	125	125	80
	Ac	0	1	1	3	5	10	10	10	18	18
35001 to 150000	n	800	500	500	500	500	500	315	200	125	80
	Ac	1	1	3	5	10	18	18	18	18	18
...											

10% consumer's risk. Applying this table means establishing the AQL, and consequently the LQ, and depending on the size of each isolated lot being worked with, determining the sample size (n) and the maximum number of errors that may appear (Ac). If the Ac value is exceeded in the sample taken from the lot under consideration, it is considered that the lot does not have sufficient quality. It is important to notice also that Table 3 does not include LQ=15%. For this reason, the assumed input LQ in this proposed method will be 12.5%.

Taking into account the rationale to implement manual controls, Table 4 presents the quality element types that are analyzed, the scope (entities and properties) on which they are focused, and the pair of quality elements that are required. In the case of the quality elements containing the quantitative results (the identifier of the quality element uses a *_QR* suffix), we indicate the associated measure and that they are evaluated using a sample-based inspection (*SI* value refers to the use of *DQ_SampleBasedInspection* in Figure 3). In the case of the quality element containing the conformance result (the identifier of the quality element uses a *_CR* suffix), it is remarked that the evaluation is derived from a quantitative result (*AD* value refers to the use of *DQ_AggregationDerivation* in Figure 3).

Figure 5 shows the workflow for the reporting of quality elements. Once the population size and the size of the sample have been defined, the random sample of the corpus is selected by means of a random number generator. Then, the random sample is analyzed by several experts independently and if there is a disparity in the consideration of a case, a decision is made by consensus. In this way, it is possible to work with an even number of experts, and the difference of one vote does not mark the decision because a consensus must be reached. Last, the quality elements with the quantitative and conformance results are defined.

With respect to the specific details for evaluating the measures, the thematic classification correctness quality elements check whether the semantic information of the selected theme(s) (*dcat:theme* property) is coherent with the description and title of the *Dataset*. Although a *Dataset* may have multiple themes, *DQ_TheClaDatThe_QR* only accounts a failure if none of the themes is related to the *Dataset*. There are also manual controls on the non-quantitative attribute correctness. Although the analyzed properties containing URIs are also analyzed automatically from both the domain consistency perspective (to check whether the values are valid URIs) and the non-quantitative attribute correctness (to check whether the URIs are accessible), these manual controls make more emphasis on checking that the content of the accessible resources comply with the expected semantics of the property. In addition, the quality of free text is manually reviewed for the title and description of *Datasets*. The text values not appropriate or incomplete are annotated as errors.

B. QUALITY ELEMENTS ASSOCIATED WITH AUTOMATED MEASURES

Considering the case of automatic controls, Table 5 presents the quality element types that are analyzed, the scope (entities and properties) on which they are focused, and the pair of quality elements that are required. In the case of the quality element containing the quantitative results (the identifier of the quality element uses a *_QR* suffix), we indicate the associated measure and that full inspection is performed (*FI* value refers to the use of *DQ_FullInspection* in Figure 3). In the case of the quality element containing the conformance result (the identifier of the quality element uses a *_CR* suffix), we remark that the evaluation is derived from a quantitative result.

Figure 6 shows the workflow for reporting the quality elements associated to automated measures. The first step is to identify the granularity of the population. In general, the population is the count of relevant entities (*Datasets* or *Distributions*) analyzed by the quality element. However, the domain consistency and the automated non-quantitative attribute correctness is focused on the distinct values of specific metadata properties. The second step is to count the number of the correct/erroneous items according to the type of rate imposed by the measure associated to the quality element that will contain the quantitative result. The third step is the computation of the rate of correct/erroneous items. Last, the quality elements with the quantitative and conformance

TABLE 4. Quality elements associated with manual measures. Notes: EM=Evaluation Method; SI=DQ_SampleBasedInspection; AD=DQ_AggregationDerivation.

		Quality element with quantitative result			Qual. el. with conformance result	
Quality element type	**Scope**	**Identifier**	**Measure**	**EM**	**Identifier**	**EM**
DQ_ThematicClassification Correctness	Dataset (theme)	DQ_TheClaDatThe_QR	D.63 (ISO 19157)	SI	DQ_TheClaDatThe_CR	AD
DQ_NonQuantitative AttributeCorrectness	Dataset (references)	DQ_TheNQADatRef_QR	D.68 (ISO 19157)	SI	DQ_TheNQADatRef_QR	AD
	Dataset (conformsTo)	DQ_TheNQADatCon_QR	D.68 (ISO 19157)	SI	DQ_TheNQADatCon_CR	AD
	Distribution (accessURL)	DQ_TheNQADisAcc_QR	D.68 (ISO 19157)	SI	DQ_TheNQADisAcc_CR	AD
	Distribution (license)	DQ_TheNQADisLic_QR	D.68 (ISO 19157)	SI	DQ_TheNQADisLic_CR	AD
DQ_QualityOfFreeText	Dataset (title)	DQ_QFTDatTitO_QR	Overall quality of free text	SI	DQ_QFTDatTitO_CR	AD
DQ_QualityOfFreeText	Dataset (description)	DQ_QFTDatDesO_QR	Overall quality of free text	SI	DQ_QFTDatDesO_CR	AD

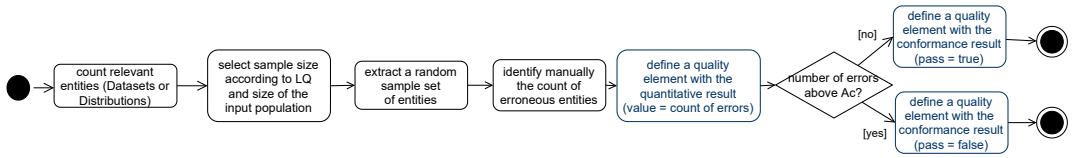

FIGURE 5. Workflow for reporting quality elements associated with manual measures.

results are defined. In order to annotate the conformance results of automatic controls, we must remind that given that Table 3 does not include LQ=15%, our final LQ proposed for manual controls is 12.5%, and this implies that AQL is equal to 4.16% for automatic controls. In the case of using measures based on error rates, the conformance is passed if the rate is below AQL. In the case of rates of correct items, the conformance is passed if the rate is above $(100 - AQL)$.

With respect to the implementation of the automated measures, a core decision has been the storage of metadata on a RDF triplestore that can be accessed through a SPARQL end-point. The selection and count of the relevant population (entities or properties) for each measure can be defined in terms of a SPARQL query. Python, the programming language used in our implementation, but also other common programming languages, include libraries for dealing with SPARQL. In addition, the count of correct/erroneous items can be also directly expressed in terms of SPARQL queries for the following quality element types of Table 5: *DQ_CompletenessCommission*, *DQ_CompletenessOmission*, *DQ_DomainConsistency* and *DQ_ConceptualConsistency* (except for the quality elements that check consistency between *format* and *accessURL*).

In order to illustrate the potential of SPARQL to automate the evaluation of measures, Figures 7 and 8 show two representative examples. Figure 7 shows an example for an automatic control on the completeness omission of *Distributions* (*DQ_ComOmiDis_QR* and *DQ_ComOmiDis_CR* quality elements). The first query retrieves the number of *Distribution* instances. The second query counts the *Distributions* without mandatory fields. Figure 8 shows an example of an automatic control on the domain consistency of the *dcat:theme* property of *Datasets* (*DQ_LogDomDatThe_QR* and *DQ_LogDomDatThe_CR* quality elements). This is also a representative example of quality elements using directly the properties as population. The first query retrieves the number of distinct values for *dcat:theme*. The second query retrieves the number of distinct values having a correct data domain (i.e., *skos:Concept*). Instead of SPARQL, we could have also used the Shapes Constraint Language (SHACL) to implement some checks, but we wanted to identify and count instances of entities and properties in different scenarios, and not only violations of the metadata model.

Anyway, there are other measures that require additional procedures. For instance, the *DQ_LogConDisFor_QR* and *DQ_LogConDisFor_CR* quality elements check the consistency between *dcat:format* and *dcat:accessURL* in *Distribution* instances. We have developed a specific function to count the *Distribution* instances where the expected format matches the file extension of *dcat:accessURL*.

With respect to the *DQ_PositionalCorrectness* quality elements, we have implemented an algorithm to detect matches between the values of *dcat:spatial* and the textual location references in *dct:title* and *dct:description* properties. The spatial references in *dcat:spatial* are aligned to a model of administrative divisions, from which we can infer the corresponding spatial polygons. Then, the *dct:title* and *dct:description* are processed using the GeoNames geographical database to identify spatial references and their corresponding coordinates. If at least a spatial reference in the descriptive properties is contained in the polygon corresponding to the *dcat:spatial* administrative division, we consider that the reference is correct. When there are not *dcat:spatial* content or the *dct:title* and *dct:description* do not contain spatial references, they are considered correct by default.

With respect to the *DQ_TemporalConsistency* quality elements, the original measure proposed by Ureña-Cámara et al. [12] has been adapted to the properties of a *Dataset*. Figure 9 shows the three general properties that must be checked in order to assure that the time sequence is correct. Because none of the temporal properties (*dct:issued*, *dct:modified*,

TABLE 5. Quality elements associated with automated measures. Notes: EM=Evaluation Method; FI=DQ_FullInspection; AD=DQ_AggregationDerivation.

		Quality element with quantitative result			Qual. el. with conformance result	
Quality element type	**Scope**	**Identifier**	**Measure**	**EM**	**Identifier**	**EM**
DQ_Completeness Commission	Dataset (identifier, modified, issued, accrualPeriodicity, license, valid, publisher)	DQ_ComComDat_QR	D.3 (ISO 19157)	FI	DQ_ComComDat_CR	AD
	Distribution (identifier, accessURL, mediaType, byteSize)	DQ_ComComDis_QR	D.3 (ISO 19157)	FI	DQ_ComComDis_CR	AD
DQ_Completeness Omission	Dataset (title, publisher, description, theme, distribution)	DQ_ComOmiDat_QR	D.7 (ISO 19157)	FI	DQ_ComOmiDat_CR	AD
	Distribution (accessURL, mediaType)	DQ_ComOmiDis_QR	D.7 (ISO 19157)	FI	DQ_ComOmiDis_CR	AD
DQ_Conceptual Consistency	Dataset	DQ_LogConDat_QR	D.13 (ISO 19157)	FI	DQ_LogConDat_CR	AD
	Distribution	DQ_LogConDis_QR	D.13 (ISO 19157)	FI	DQ_LogConDis_CR	AD
	Distribution (format vs accessURL)	DQ_LogConDisFor_QR	D.13 (ISO 19157)	FI	DQ_LogConDisFor_CR	AD
DQ_DomainConsistency	Dataset.title (string domain)	DQ_LogDomDatTit_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatTit_CR	AD
	Dataset.description (string domain)	DQ_LogDomDatDes_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatDes_CR	AD
	Dataset.theme (skos:Concept domain)	DQ_LogDomDatThe_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatThe_CR	AD
	Dataset.keyword (string domain)	DQ_LogDomDatKey_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatKey_CR	AD
	Dataset.identifier (URI domain)	DQ_LogDomDatIde_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatIde_CR	AD
	Dataset.issued (date-time domain)	DQ_LogDomDatIss_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatIss_CR	AD
	Dataset.modified (date-time domain)	DQ_LogDomDatMod_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatMod_CR	AD
	Dataset.accrualPeriodicity (frequency domain)	DQ_LogDomDatAcc_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatAcc_CR	AD
	Dataset.language (linguistic system domain)	DQ_LogDomDatLan_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatLan_CR	AD
	Dataset.publisher (foaf:Agent domain)	DQ_LogDomDatPub_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatPub_CR	AD
	Dataset.spatial (resource URI from a predefined list of province resources)	DQ_LogDomDatSpa_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatSpa_CR	AD
	Dataset.temporal (period of time domain)	DQ_LogDomDatTem_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatTem_CR	AD
	Dataset.valid (date-time domain)	DQ_LogDomDatValid_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatValid_CR	AD
	Dataset.references (URI domain)	DQ_LogDomDatRef_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatRef_CR	AD
	Dataset.conformsTo (URI domain)	DQ_LogDomDatCon_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatCon_CR	AD
	Dataset.distribution (Distribution domain)	DQ_LogDomDatDis_QR	D.17 (ISO 19157)	FI	DQ_LogDomDatDis_CR	AD
	Distribution.identifier (URI domain)	DQ_LogDomDisIde_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisIde_CR	AD
	Distribution.title (string domain)	DQ_LogDomDisTit_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisTit_CR	AD
	Distribution.accessURL (URL domain)	DQ_LogDomDisAcc_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisAcc_CR	AD
	Distribution.format (IMT domain)	DQ_LogDomDisFor_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisFor_CR	AD
	Distribution.byteSize (decimal number domain)	DQ_LogDomDisByt_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisByt_CR	AD
	Distribution.license (URI domain)	DQ_LogDomDisLic_QR	D.17 (ISO 19157)	FI	DQ_LogDomDisLic_CR	AD
DQ_TemporalConsistency	Dataset (issued, modified, valid)	DQ_TemDatIss_QR	Similar to D.62 using a rate	FI	DQ_TemDatIss_CR	AD
DQ_TemporalValidity	Dataset (*harvest date* vs issued, modified, valid)	DQ_TemDatHar_QR	D.18 (ISO 19157)	FI	DQ_TemDatHar_CR	AD
DQ_NonQuantitative AttributeCorrectness	Dataset.references	DQ_TheNQADatRef_QR	D.69 (ISO 19157)	FI	DQ_TheNQADatRef_CR	AD
	Dataset.conformsTo	DQ_TheNQADatCon_QR	D.69 (ISO 19157)	FI	DQ_TheNQADatCon_CR	AD
	Distribution.accessURL	DQ_TheNQADisAcc_QR	D.69 (ISO 19157)	FI	DQ_TheNQADisAcc_CR	AD
	Distribution.license	DQ_TheNQADisLic_QR	D.69 (ISO 19157)	FI	DQ_TheNQADisLic_CR	AD
DQ_PositionalCorrectness	Dataset (spatial)	DQ_PosCorrDatSpa_QR	Similar to D.33 (ISO 19157)	FI	DQ_PosCorrDatSpa_CR	AD
DQ_QualityOfFreeText	Dataset (title vs language)	DQ_QFTDatTitR_QR	Readability of free text	FI	DQ_QFTDatTitR_CR	AD
DQ_QualityOfFreeText	Dataset (description vs language)	DQ_QFTDatDesR_QR	Readability of free text	FI	DQ_QFTDatDesR_CR	AD

dct:valid) is mandatory, the control assumes the most favorable case in the comparison. Following this assumption, *DQ_TemporalConsistency* is applied to all *Datasets* and in case a *Dataset* does not contain any date property, this dataset will be annotated with a correct consistency.

Considering that datasets are constantly updated, Fig-

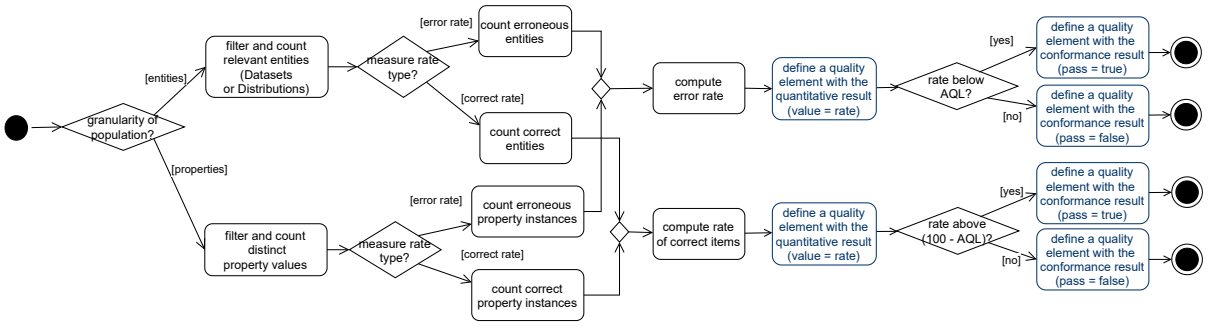

FIGURE 6. Workflow for reporting quality elements associated with automated measures.

```
PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT (count(?resource) AS ?resources) WHERE {
    ?resource rdf:type dcat:Distribution
}

PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT (count(?resource) AS ?resources) WHERE {
{
    SELECT ?resource WHERE {
      ?resource rdf:type dcat:Distribution .
      FILTER NOT EXISTS { ?resource dcat:accessURL ?value1}
    }
  } UNION {
    SELECT ?resource WHERE {
      ?resource rdf:type dcat:Distribution .
      FILTER NOT EXISTS { ?resource dcat:mediaType ?value2}
    }
  }
}
```

FIGURE 7. Checking completeness omission on *Distributions*

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT (count(DISTINCT ?value) as ?values) WHERE {
    ?resource rdf:type dcat:Dataset .
    ?resource dcat:theme ?value .
}

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT (count(DISTINCT ?value) as ?values) WHERE {
    ?resource rdf:type dcat:Dataset .
    ?resource dcat:theme ?value .
    ?value rdf:type skos:Concept .
}
```

FIGURE 8. Checking domain consistency on *dcat:theme* property of *Datasets*

ure 9 presents the temporal scenario in which datasets can be downloaded or harvested (by an automatic process). This scenario serves to illustrate the other temporal quality control related to validity (*DQ_TemDatHar_QR* and *DQ_TemDatHar_CR* quality elements). The datasets may have been created or modified at any time and date just before the harvest time. Therefore, we must check the harvest date with respect to the properties containing temporal stamps

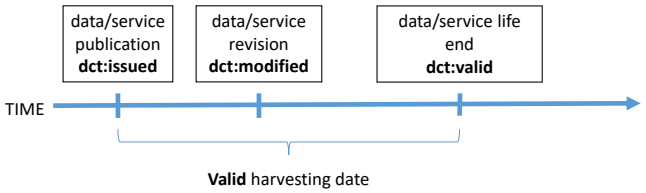

FIGURE 9. Timeline representing the properties of a Dataset until corpus was harvested.

(*dct:issued*, *dct:modified*, *dct:valid*). The harvest time must always be later, or at least the same as the time of creation / modification that is included in the temporary stamp. In addition, if there is a validity deadline (*dct:valid*), the download should happen before that date. Otherwise (harvest date after *dct:valid*), the dataset will be considered directly invalid. Last, with this quality element it also applies the same assumption as the one applied to *DQ_TemporalConsistency*: if no date properties are provided, the dataset is considered valid with respect to this control.

In the case of automatic controls related to the non-quantitative attribute correctness, we have developed a specific function to assert that the URLs used for properties of *Datasets* and *Distributions* are reachable (the response to an HTTP request has a valid status code). This function allows us to automate the computation of error rates.

Finally, the controls for checking automatically the quality of free text are implemented by means of two readability indexes: the one developed by Fernández-Huerta [26] (see Equation 1) using the syllables division of Hernández-Figueroa et al. [27]; and the Perspicuity index developed by Szigriszt Pazos [28] (see Equation 2). From these two indexes, the best value of both is selected for the *Datasets*. Using this best value, the *Dataset* will pass the test if it has an easy readability (index value > 50). However, this methodology has the drawback that it can only be applied to the languages having some readability indexes, e.g. English [29], [30], Spanish [28], French [31] or Italian [32].

$$Flesch_{FH} = 206.835 - \frac{1.015 \times N_{word}}{N_{sentence}} - \frac{60.0 \times N_{syllable}}{N_{word}} \quad (1)$$

$$Flesch_{Pers} = 206.835 - \frac{N_{word}}{N_{sentence}} - \frac{62.3 \times N_{syllable}}{N_{word}} \quad (2)$$

In Equations 1 and 2, $N_{sentence}$ represents the number of sentences, N_{word} is the number of words, and $N_{syllable}$ indicates the total number of syllables in the text.

C. REPRESENTATION OF RESULTS

With respect to the mechanism for reporting metadata quality in machine-readable formats, several options could be considered. A first option could have been the use of ISO 19157-2 [33] because it provides a specification for encoding ISO 1957 data quality reports in XML. However, as we are analyzing metadata based on semantic vocabularies, it seems more appropriate to express quality results also with semantic vocabularies. For this purpose, we have chosen the Data Quality Vocabulary (DQV) proposed by W3C [13] .

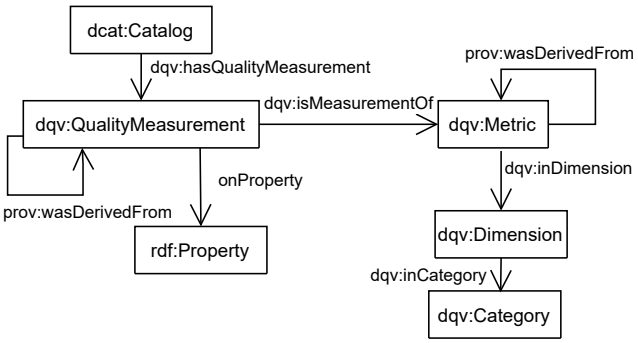

FIGURE 10. Subset of the DQV vocabulary extended to describe catalog quality.

DQV is implemented as an RDF vocabulary which extends the Data Catalog Vocabulary (DCAT) with properties and classes suitable for expressing the quality. It is defined to assess the quality of *Dataset* or *Distribution* resources by means of five different observed properties: quality annotations (*dqv:QualityAnnotation* class) about feedback and quality certificates; the standards (*dcterms:Standard* class) the resource conforms to; policies or agreements (*dqv:QualityPolicy* class) related to quality; measurements (*dqv:QualityMeasurement* class) with qualitative or quantitative information about the resource; and entities (*prov:Entity class*) involved in the provenance of the resource.

Although not stated in the DQV specification, the quality assessment is also applicable to the quality analysis of the metadata contained in an Open Data catalog. Therefore, we have extended the vocabulary to represent the quality of catalog metadata contents. This is directly done by extending the domain of the *dqv:hasQualityMeasurent* property so that it can also be applied to *dcat:Catalog* classes. Figure 10 shows an excerpt of the DQV vocabulary with this extension. In addition, it can be observed that it is possible to define the quantitative and conformance results of ISO 19157 quality elements in terms of DQV measurements, i.e. instances of the *dqv:QualityMeasurement* class. Each measurement refers to a quality metric (the concept equivalent to the ISO 19157 'measure') that is considered in a quality dimension. The DQV mechanism for annotating measurements, metrics, quality dimensions and quality categories adapts the *daQ* quality framework [34], which is an ontology for dataset quality information proposed by Debattista et al. for the assessment of Linked Data quality [35]. Furthermore, the own DQV specification document proves its feasibility for representing any type of quality models integrating the representation of the dimensions and categories proposed by ISO/IEC 25012 [23] or Zaveri et al. [24].

Figure 11 shows an example of the representation of the quality report of a *dcat:Catalog* represented according to the DQV model. In the example a *dcat:Catalog* instance is associated with two measurements to inform about the completeness commission evaluation of the metadata describing the datasets of this catalog. A measurement is the closer concept in DQV to represent jointly the results of an ISO 19157 quality element and the reference to the measure used in this quality element. We use two measurements to separate the representation of the quality element containing the quantitative result from the representation of the derived quality element containing the conformance result.

The measurement with the quantitative result (*:DQ_ComComDat_QR* resource using the same identifier as the one used to identify this quality element in Table 5) refers to the D.3 metric of ISO 19157 (*:D.3.ISO.19157* resource), which belongs to the *DQ_CompletenessCommission* dimension. The example also shows that the *DQ_CompletenessCommission* dimension belongs to the *DQ_Completeness* quality category of ISO 19157. In addition, the measurement indicates the metadata property on which the metric has been applied: the *dcat:dataset* property that links a catalog with its datasets. This is indicated through the *:onProperty* relation, which is proposed in the DQV specification as an extension feature to indicate mandatory or optional parameters on measurements.

In the case of the measurement with the conformance result (*:DQ_ComComDat_CR* resource using the same identifier as the one used to identify this quality element in Table 5), we also refer to a metric that has been created specifically to indicate whether the ISO 19157 D.3 error rate is below the AQL or not (*:D.3.ISO.19157_conformance* resource). The *prov:wasDerivedFrom* property indicates that both the conformance measurement and the conformance metric are derived from the corresponding quantitative versions.

V. EXPERIMENTS

In this section we describe the results of applying the ISO 19157 based evaluation method and the MQA methodology to a chosen corpus.

A. CORPUS

For the purpose of our experiments, we have used the contents of the Open Data catalog of the Spanish Government. This catalog is hosted at *datos.gob.es*, the Open Data portal

```
#Quality report
:myCatalog a dcat:Catalog ;
    dcterms:title "datos.gob.es" ;
    dqv:hasQualityMeasurement :DQ_ComComDat_CR, :DQ_ComComDat_QR .

:DQ_ComComDat_QR a dqv:QualityMeasurement ;
    dqv:computedOn :myCatalog ;
    dqv:isMeasurementOf :D.3.ISO.19157 ;
    dqv:value "0.0"^^xsd:double ;
    dct:date "2020-03-01"^^xsd:date ;
    :onProperty dcat:dataset .

:DQ_ComComDat_CR a dqv:QualityMeasurement ;
    dqv:computedOn :myCatalog ;
    dqv:isMeasurementOf :D.3.ISO.19157_conformance ;
    prov:wasDerivedFrom :DQ_ComComDat_QR ;
    dqv:value "true"^^xsd:boolean ;
    dct:date "2020-03-01"^^xsd:date ;
    :onProperty dcat:dataset .

#definition of categories, dimensions and metrics
:DQ_Completeness a dqv:Category ;
    skos:prefLabel "Completeness"@en ;
    skos:definition "Completeness refers to the degree in which the metadata elements are present or absent."@en .

:DQ_CompletenessCommission a dqv:Dimension ;
    dqv:inCategory :DQ_Completeness ;
    skos:prefLabel "Completeness commission"@en ;
    skos:definition "Completeness commission refers to the degree in which there are excess instances of metadata
    elements in a metadata record."@en .

:D.3.ISO.19157 a dqv:Metric ;
    skos:definition "Rate of records with excess items."@en ;
    dqv:inDimension :DQ_CompletenessCommission ;
    dqv:expectedDataType xsd:double .

:D.3.ISO.19157_conformance a dqv:Metric ;
    prov:wasDerivedFrom :D.3.ISO.19157 ;
    skos:definition "Checks if the rate of records with excess items is below AQL (statistical error level)."@en ;
    dqv:inDimension :DQ_CompletenessCommission ;
    dqv:expectedDataType xsd:boolean .

#Parameters of the measurements (derived from the DQV specification)
:onProperty a qb:DimensionProperty, rdf:Property ;
    rdfs:comment  "Property on which the quality measure is assessed."@en ;
    rdfs:domain  dqv:QualityMeasurement ;
    rdfs:label   "Label assessment property"@en ;
    rdfs:range      rdf:Property .
```

FIGURE 11. Fragment of a metadata quality report defined according to DQV in Turtle format.

providing a common access point to the Spanish Open Data Initiative ("Iniciativa Aporta"). This initiative was launched in 2009 to promote the openness of public sector information and the development of advanced services based on data in open formats that everyone can use, reuse or share. The national catalog of *datos.gob.es* is the most visible deliverable of this Spanish initiative because it acts as the meeting point for public institutions, enterprises and citizens interested in public sector information and associated services.

The Spanish Open Data catalog compiles metadata in compliance with the NTI metadata model that was explained in section III. The metadata contents are not created directly in this catalog, but harvested from the Open Data catalogs of the entities that are federated in this Spanish Open Data initiative. In 2021 *datos.gob.es* integrated more than 35,000 metadata records describing the datasets (and more than 150,000 associated distributions) originated by more than 300 different public administration offices (at local, regional or national level), universities, or research institutions. In addition, it must be noted that although most of the metadata records are available in Spanish, there are also some records using the other official languages of Spain (i.e., Catalan, Basque or Galician languages) or even English.

Technically, the Spanish Open Data catalog stores the harvested metadata records in a RDF triplestore and can be accessed through different protocols: a specialized REST API[2] to filter and download specific datasets; a SPARQL end-point;[3] or an RDF end-point for bulk download.[4] Using this third protocol the metadata contents of the Spanish Open Data catalog were downloaded on 12 June 2019 for the purpose of our quality evaluation experiments. The downloaded contents included 22,406 *Datasets* and 112,874 *Distributions* originated by 133 different publishers in 5 different languages. In addition, it must be noted that the evaluation

[2]https://datos.gob.es/es/accessible-apidata
[3]https://datos.gob.es/es/accessible-sparql
[4]http://ondemand2.redes.ondemand.flumotion.com/redes/ondemand2/Datosabiertos/datosgobes.rdf

experiments were run in April 2020.

B. RESULTS WITH THE ISO 19157 BASED METHOD

The following subsections describe the results obtained with the quality elements associated to manual measures and the quality elements associated with automated measures.

1) Results of quality elements associated with manual measures

Table 6 shows the results of the quality elements. In the case of quantitative results, the table indicates: the population size (*pop.* column); the size of the sample (*sam.* column) according to the LQ and the input population size (see the relationship between AQL, LQ and sample size in section IV-A); the maximum number of accepted errors (*Ac* column); and the number of erroneous items in the sample (*errors* column). In the case of conformance results, the *pass* column indicates whether the conformance is true (*T* value) of false (*F* value).

The evaluation of the random sample was carried out independently by two experts, each one with more than 20 years of experience in metadata. First, an approach to the problem was established to define possible cases and how to evaluate them. Subsequently, each expert carried out his evaluation independently and, finally, a pooling was carried out to resolve the discrepant cases or with doubts in their evaluations.

With respect to the thematic accuracy evaluated manually, as described in section IV-A, this quality category includes two types of controls. On the one hand, the thematic classification correctness has not been passed. Additionally, it is interesting to note that there are some *Datasets* with a total number of 22 themes, which is the maximum number of different themes in NTI. On the other hand, 3 out of 4 manual controls on the non-quantitative attribute correctness of properties containing pointers to licenses, standards or related resources are not passed either. Only the manual control on access URLs is passed.

With regards to the results obtained for the manual part of *DQ_QualityOfFreeText*, the *Dataset* titles have a bad quality according to the sample: more than 50% of the titles are not appropriate or incomplete. In contrast, the *Dataset* descriptions are more accurate, i.e. the test was passed with only 6 errors for the same sample.

2) Results of quality elements associated with automated measures

The results of the quality elements associated with automated measures are presented in tables 7 (measures based on error rates) and 8 (measures based on correct rates). In the case of quantitative results, the table indicates: the population size (*pop.* column); the number of items that pass or fail the proposed test (*correct items* or *errors* columns); and the rate of items passing or not the quality control (*correct rate* or *error rate* columns). In the case of conformance results, the *pass* column indicates whether the conformance is true (*T* value) of false (*F* value). As already indicated in section IV-B, to pass a quality control the metrics based on error rates must not surpass an AQL of 4.16%. In coherence with this, the metrics based on rates of correct items must be above 95.84% to be passed.

Table 7 presents the results of the commission and omission measures applied to the *Dataset* and *Distribution* instances. There are two cases where the error rate is not 0%. The first case, without consequences to pass the control, is the existence of one *Dataset* instance without a mandatory *dcat:distribution* property. The second case is more problematic as all *Distribution* instances are erroneous. Instead of using the mandatory *dcat:mediaType* property imposed by NTI, distribution resources are annotated with *dct:format* property. Formally, *dcat:mediaType* is a subproperty of *dct:format*. Subproperties can substitute parent properties, but not inversely.

Table 8 presents the results of the conceptual and domain consistency quality elements. With respect to the conceptual consistency, there are no problems with *Dataset* instances. The only incidence is that *Datasets* use the *dct:language* property instead of the *dc:language* property proposed by NTI. In this case, no error is reported because *dct:language* is a subproperty of *dc:language* and, as being a specialization, is compatible. Besides, as stated in NTI standard [6], "the basic entities or properties can be enriched with additional metadata considered relevant to improve the quality of information". However, all *Distribution* instances fail the test. As already reported for completeness omission, all instances lack for the mandatory *dcat:mediaType* property (the *dct:format* property cannot replace it).

In addition, it is worth noting that also within the conceptual consistency, we have identified contradictions between *dcat:format* and *dcat:accessURL* in *Distribution* instances. First, we have checked that only 71,723 instances had a *dcat:format* property properly encoded with a right correspondence between its label (*rdfs:label*) and its value (*rdf:value*). Secondly, we have checked that only 59,387 instances had a format compatible with the file extension of *dcat:accessURL*.

With respect to the domain consistency quality elements, several incidences can be reported. In the case of *dcat:publisher* properties, instead of finding URIs linking to *foaf:Agent* resources, *skos:Concept* resources are found. In the case of *dct:spatial* properties, several pointers to Region resources instead of Province resources are found. In addition, it must be noted that we have not verified the domain of the *dct:license* property of *Datasets* because no *Dataset* instance includes this property. Similarly, we have not verified the domain of the *dcat:relation* property of *Distributions* because no *Distribution* instance includes such property. On the opposite, it must be noted that although the *dct:format* property has been used instead of the *dcat:mediaType* property proposed by NTI, the domain values have been analyzed to check if they refer to valid IANA Internet Media Type (IMT) values.

TABLE 6. Results of quality elements associated with manual measures.

		Quality element with quantitative result					Qual. el. with conformance result	
Quality element type	**Scope**	**Identifier**	**pop.**	**sam.**	**Ac**	**errors**	**Identifier**	**pass**
DQ_ThematicClassification Correctness	Dataset (theme)	DQ_TheClaDatThe_QR	22,406	125	10	15	DQ_TheClaDatThe_CR	F
DQ_NonQuantitative AttributeCorrectness	Dataset (references)	DQ_TheNQADatRef_QR	5,473	80	5	10	DQ_TheNQADatRef_CR	F
	Dataset (conformsTo)	DQ_TheNQADatCon_QR	176	20	0	3	DQ_TheNQADatCon_CR	F
	Distribution (accessURL)	DQ_TheNQADisAcc_QR	104,119	200	18	12	DQ_TheNQADisAcc_CR	T
	Distribution (license)	DQ_TheNQADisLic_QR	207	20	0	2	DQ_TheNQADisLic_CR	F
DQ_QualityOfFreeText	Dataset (title)	DQ_QFTDatTitO_QR	22,406	125	10	66	DQ_QFTDatTitO_CR	F
	Dataset (description)	DQ_QFTDatDesO_QR	22,406	125	10	6	DQ_QFTDatDesO_QR	T

TABLE 7. Results of quality elements associated with automated measures based on error rates.

		Quality element with quantitative result				Qual. el. with conformance result	
Quality element type	**Scope**	**Identifier**	**pop.**	**errors**	**error rate**	**Identifier**	**pass**
DQ_Completeness Commission	Dataset (identifier, modified, issued, accrualPeriodicity, license, valid, publisher)	DQ_ComComDat_QR	22,406	0	0.00%	DQ_ComComDat_CR	T
	Distribution (identifier, accessURL, mediaType, byteSize)	DQ_ComComDis_QR	112,874	0	0.00%	DQ_ComComDat_CR	T
DQ_Completeness Omission	Dataset (title, publisher, description, theme, distribution)	DQ_ComOmiDat_QR	22,406	1	0.04%	DQ_ComOmiDat_CR	T
	Distribution (accessURL, mediaType)	DQ_ComOmiDis_QR	112,874	112,874	100.00%	DQ_ComOmiDis_CR	F
DQ_TemporalConsistency	Dataset (issued, modified, valid)	DQ_TemDatIss_QR	22,406	1,731	7.73%	DQ_TemDatIss_CR	F
DQ_TemporalValidity	Dataset (*harvest date* vs issued, modified, valid)	DQ_TemDatHar_QR	22,406	226	1.01%	DQ_TemDatHar_CR	T
DQ_NonQuantitative AttributeCorrectness	Dataset (references)	DQ_TheNQADatRef_QR	5,473	306	5.59%	DQ_TemDatHar_CR	F
	Dataset (conformsTo)	DQ_TheNQADatCon_QR	176	26	14.77%	DQ_TheNQADatCon_CR	F
	Distribution (accessURL)	DQ_TheNQADisAcc_QR	104,119	8,546	8.21%	DQ_TheNQADisAcc_CR	F
	Distribution (license)	DQ_TheNQADisLic_QR	207	70	33.81%	DQ_TheNQADisLic_CR	F
DQ_PositionalCorrectness	Dataset (spatial)	DQ_PosCorrDatSpa_QR	15,758	634	4.02%	DQ_PosCorrDatSpa_CR	T

TABLE 8. Results of quality elements associated with automated measures based on rates of correct items.

		Quality element with quantitative result				Qual. el. with conformance result	
Quality element type	**Scope**	**Identifier**	**pop.**	**correct items**	**correct rate**	**Identifier**	**pass**
DQ_Conceptual Consistency	Dataset	DQ_LogConDat_QR	22,406	22,406	100.00%	DQ_LogConDat_CR	T
	Distribution	DQ_LogConDis_QR	112,874	0	0.00%	DQ_LogConDis_CR	F
	Distribution (format vs accessURL)	DQ_LogConDisFor_QR	112,874	59,387	52.61%	DQ_LogConDisFor_CR	F
DQ_DomainConsistency	Dataset.title	DQ_LogDomDatTit_QR	30,061	30,061	100.00%	DQ_LogDomDatTit_CR	T
	Dataset.description	DQ_LogDomDatDes_QR	22,889	22,889	100.00%	DQ_LogDomDatDes_CR	T
	Dataset.theme	DQ_LogDomDatThe_QR	22	22	100.00%	DQ_LogDomDatThe_CR	T
	Dataset.keyword	DQ_LogDomDatKey_QR	60,720	60,720	100.00%	DQ_LogDomDatKey_CR	T
	Dataset.identifier	DQ_LogDomDatIde_QR	18,263	16,562	90.69%	DQ_LogDomDatIde_CR	F
	Dataset.issued	DQ_LogDomDatIss_QR	10,441	10,441	100.00%	DQ_LogDomDatIss_CR	T
	Dataset.modified	DQ_LogDomDatMod_QR	3,384	3,384	100.00%	DQ_LogDomDatMod_CR	T
	Dataset.accrualPeriodicity	DQ_LogDomDatAcc_QR	6,284	6284	100.00%	DQ_LogDomDatAcc_CR	T
	Dataset.language	DQ_LogDomDatLan_QR	5	5	100.00%	DQ_LogDomDatLan_CR	T
	Dataset.publisher	DQ_LogDomDatPub_QR	133	0	0.00%	DQ_LogDomDatPub_CR	F
	Dataset.spatial	DQ_LogDomDatSpa_QR	72	52	72.22%	DQ_LogDomDatSpa_CR	F
	Dataset.temporal	DQ_LogDomDatTem_QR	4,670	4,670	100.00%	DQ_LogDomDatTem_CR	T
	Dataset.valid	DQ_LogDomDatValid_QR	63	63	100.00%	DQ_LogDomDatValid_CR	T
	Dataset.references	DQ_LogDomDatRef_QR	5,473	5,472	99.98%	DQ_LogDomDatRef_CR	T
	Dataset.conformsTo	DQ_LogDomDatCon_QR	176	176	100.00%	DQ_LogDomDatCon_CR	T
	Dataset.distribution	DQ_LogDomDatDis_QR	112874	112874	100.00%	DQ_LogDomDatDis_CR	T
	Distribution.identifier	DQ_LogDomDisIde_QR	64,922	64,524	99.39%	DQ_LogDomDisIde_CR	T
	Distribution.title	DQ_LogDomDisTit_QR	50,878	50,878	100.00%	DQ_LogDomDisTit_CR	T
	Distribution.accessURL	DQ_LogDomDisAcc_QR	104,119	104,114	99.99%	DQ_LogDomDisAcc_CR	T
	Distribution.format	DQ_LogDomDisFor_QR	68	40	58.82%	DQ_LogDomDisFor_CR	F
	Distribution.byteSize	DQ_LogDomDisByt_QR	8,835	8,701	98.48%	DQ_LogDomDisByt_CR	T
	Distribution.license	DQ_LogDomDisLic_QR	207	207	100%	DQ_LogDomDisLic_CR	T
DQ_QualityOfFreeText	Dataset (title vs language)	DQ_QFTDatTitR_QR	22,406	13,242	59.10%	DQ_QFTDatTitR_CR	F
	Dataset (description vs language)	DQ_QFTDatDesR_QR	22,406	12,783	57.05%	DQ_QFTDatDesR_CR	F

The results for the temporal consistency between the recorded dates for creation, modification and validity are shown in Table 7. We considered, because all time properties are optional, that each missing date represents a valid date as stated in section IV-B. Even with this assumption, the associated quantitative result has a high rate of errors (7.73%). It is interesting to note that the majority of these errors refer to a modified date which is previous to the corresponding issued date. In addition, Table 7 also presents the results of temporal validity, which refers to the cases where the date of metadata

harvesting is a date belonging to the interval that starts at the date of creation (*dct:issued*) or modification and ends at the date when validity ends (*dct:valid*). It is interesting to notice here that *dct:valid* is an optional property: only 260 *Datasets* have this property and 226 of them fail the test (86.92%). Because *dct:valid* is optional, we can consider that the absence of the *dct:valid* property means no end of validity date. Due to this, the catalog complies with the quality requirements because only 1% of the *Datasets* have a validity date before the harvesting date.

With respect to the automated controls on non-quantitative attribute correctness, also shown in Table 7, it must be noted that all the analyzed properties containing URIs fail the control because the percentage of non-reachable URIs surpass the AQL, especially the URIs linking to license resources.

Table 7 also includes the results of the evaluation of correctness of location resources referred in *dct:spatial* properties of *Datasets*. It only presents results for *Dataset* instances with non-empty *dcat:spatial* properties. In addition, it must be noted that because the proposed automatic method uses textual location references in *dct:title* and *dct:description* free-text properties, there are 5,121 of those records whose positional correctness could not be evaluated (32.50% of instances lack textual location references). As *dct:spatial* is not mandatory, we assumed that these *Datasets* are correct.

About the quality elements related to the readability of free text in Table 8, it must be noted first that previous to the computation of the readability indexes, some values of *dct:title* and *dct:description* were revised because they did not contain plain text; sometimes free text appears as HTML encoding (e.g. "&gt" instead of ">", or "públicos" instead of "públicos"). In addition, it must be noted that we have analyzed only *dct:title* and *dct:description* properties of *Dataset* entities that indicate Spanish in their *dct:language* property. We imposed this constraint because of the language limitations in the readability index proposed in section IV-B. With respect to the quantitative results, neither *dct:title* nor *dct:description* values reach a 60% of acceptable readability. In addition, it must be noted that the *dct:title* property of *Distributions* has not been used. Although this property fulfills with NTI rules (*"Brief title or name given to the distribution"*), these titles consist of a reduced set of words (e.g. "2016", "PRIM_RES_XCYS_P" or "Pc-Axis").

C. RESULTS WITH MQA

We have also evaluated the corpus of *datos.gob.es* according to MQA. Instead of retrieving the statistics currently shown at the European Data Portal,[5] we have developed our own implementation (in Python and accessing a SPARQL endpoint) of MQA methodology because we wanted to evaluate the exact contents of the corpus described in section V-A.

Table 9 shows the evaluation results of *datos.gob.es* for each quality dimension and indicator (*indicator* is the name given to metrics in MQA methodology). In this table, each indicator is described with the following information: the maximum points that can be assigned (*weight* column); the individual percentage of achievement (% column); and the weighted value of the indicator ($percentage \times weight$) contributing to the final rating of the catalog (*points* column). The final points of the catalog are 209.65. These points are translated by MQA into four possible rating categories: excellent (351 - 405 points), good (221 - 350 points), just enough (121 - 220 points) and bad (0 - 120 points). This means that *datos.gob.es* is rated as 'just enough'.

In addition, with respect to findability, it must be noted that the availability of categories is 100% because *dcat:theme* is a mandatory property in the NTI metadata model. With respect to accessibility, the no availability of download URL is explained by the NTI model, which does not include this property for distributions.

Regarding interoperability, it must be noted that *datos.gob.es* only uses *dct:format* to indicate the format of distributions. This explains the no availability of media type. Anyway, we have checked the correspondence of *dct:format* values with the vocabularies proposed by MQA.[6] In addition, DCAT-AP compliance fails for all *Dataset* instances. In the case of *Datasets*, this is because the *dct:publisher* property should have the *foaf:Agent* range. However, *datos.gob.es* defines publishers as instances of *skos:Concept*. Besides, from the perspective of DCAT-AP, some *Dataset* instances include the unexpected properties *dct:valid* and *dct:references*. In the case of distributions, some problems arise because some of the values of *dcat:byteSize* are not decimals as expected.

About reusability, it must be noted that *dct:accessRights* and *dcat:contactPoint* are not included in the NTI model for datasets. In addition, the 100% for availability of publishers is due to the fact that this property is mandatory for datasets. Last, as concerns contextuality, it must be noted that NTI does not take into account rights (*dct:rights*) associated with distributions, just the license. With respect to the 100% availability of issued date, this can be explained because of the fact that most CKAN servers generate an issued date by default every time a dataset is inserted.

VI. DISCUSSION

In section V, we presented the results of applying the ISO 19157 based method proposed in section IV and the MQA methodology. An important difference of the ISO 19157 based method with respect to MQA is that many ISO 19157 measures are computed using the distinct values of properties as the input population. This decision can derive in worse results than the correspondent MQA indicator because in many cases the faulty values are a minority with respect to the complete population of entities, but they may represent an important fraction when considering only distinct property values (e.g. 58% of correct items for *DQ_LogDomDisFor_QR* in

[5]https://www.europeandataportal.eu/mqa/catalogues/datos-gob-es/

[6]See controlled vocabularies for non-proprietary and machine-readable formats at https://gitlab.com/european-data-portal/edp-vocabularies/-/blob/master/Custom\%20Vocabularies/.

TABLE 9. Results of MQA evaluation. Notes: the indicator column includes the names of the checked entities and properties.

Dimension	Indicator	weight	%	points
Findability	Keywords available (Dataset/keyword)	30	83%	24.9
	Category available (Dataset/theme)	30	100%	30
	Spatial information available (Dataset/spatial)	20	70%	14
	Temporal information available (Dataset/temporal)	20	21%	4.2
Accessibility	Most frequent AccessURL status code=200 (Distribution/accessURL)	50	92%	46
	DownloadURL available (Distribution/downloadURL)	20	0%	0
	Most frequent DownloadURL status code=200 (Distribution/downloadURL)	30	0%	0
Interoperability	Format available (Distribution/format)	20	100%	20
	Media type available (Distribution/mediaType)	10	0%	0
	Format/ media type from vocabulary (Distribution/format or Distribution/mediaType)	10	90%	9
	Non-proprietary (Distribution/format or Distribution/mediaType)	20	55%	11
	Machine readable (Distribution/format or Distribution/mediaType)	20	53%	10.6
	DCAT-AP compliance (all entities and properties)	30	0%	0
Reusability	License available (Distribution/license)	20	100%	20
	License from vocabulary (Distribution/license)	10	10%	1
	Access rights available (Dataset/accessRights)	10	0%	0
	Access rights from vocabulary (Dataset/accessRights)	5	0%	0
	Contact point available (Dataset/contactPoint)	20	0%	0
	Publisher available (Dataset/publisher)	10	100%	10
Contextuality	Rights available (Distribution/rights)	5	0%	0
	File size available (Distribution/byteSize)	5	15%	0.75
	Issued date available (Dataset/issued or Distribution/issued)	5	100%	5
	Modified date available (Dataset/modified or Distribution/modified)	5	64%	3.2
Total		405		209.65

Table 8 vs 90% of *Format/media type from vocabulary* in Table 9).

Another important difference between MQA and ISO 19157 is that the first one is designed for a general DCAT-AP metadata model, while the last one is focused on NTI metadata, the original model used in the corpus. This explains why the values obtained for some MQA indicators against *datos.gob.es* are low. The NTI metadata model used in *datos.gob.es* is a subset of DCAT-AP and some properties checked by MQA cannot be found.

Despite these differences, we can state that both methods are complementary. MQA indicators checking the availability of some metadata properties not considered mandatory in DCAT-AP or NTI provide additional valuable information with respect to the completeness measures obtained with the ISO 19157 method, overall with properties related to interoperability (analysis of available formats) and reusability (analysis of licenses and rights). MQA DCAT-AP compliance overlaps conceptually the domain consistency measures of ISO 19157 method. However, each method allows to study the population from a different perspective: the entities, or different property values. In addition, the ISO 19157 based method pays much more attention on the accuracy of the content of metadata properties and the definition of special controls on spatial, temporal and free-text properties.

With respect to the global result obtained with both methods, the satisfaction is a bit higher in the case of the ISO 19157 method because 26 out of 45 proposed metrics were passed for *datos.gob.es*.[7] Just taking into account the number of passed metrics, this ranks *datos.gob.es* in second quartile. However, the global punctuation according to MQA is 'just enough', the third quartile of the possible ratings. Figures 12 and 13 show two dashboards with the results of ISO 19157 and MQA respectively. The aim of the dashboards is to present all the results in a single joint vision. The ISO 19157 dashboard presents three different parts: a left part with the legends and colors for each quality category; a central part dedicated to the automatic evaluation with full inspection that presents the quantitative results of each quality element in a circular bar chart (rates are normalized to use percentages of correct items); and a right part that presents the conformance results of the quality elements evaluated manually by sampling (according to the application of ISO 2859-2). The MQA dashboard is also structured in three parts: a left part with legends and colors for quality dimensions; a central part with a circular bar chart with the percentages of achievement for each indicator; and a right part with the rating in each dimension and the global rating.

One of the limitations, but also a strong point, of the method based on ISO 19157 is the control of aspects that cannot be automated. Nowadays, not all aspects of the quality of a metadata set can be automated, which means that if we do not perform manual evaluations, the quality of that metadata is not really known. These aspects are often overlooked since manual evaluation processes are expensive and, if quality assurance measures are not adopted, they can also be biased. We consider that the proposal made in this study, in which international quality control standards have been followed, is an adequate statistical approximation and aligned with industrial processes, which makes it very extensible to other similar cases. In addition, performing manual controls through a peer-review-agreement process reduces the risk of personal bias in this type of assessment.

Last, we must note also the volatility of the values obtained for some measures. Depending on the date of experiments, the non-quantitative attribute accuracy metrics of the ISO 19157 method that check the accessibility of URLs may

[7] Completeness: 3 out of 4. Logical consistency: 19 out of 25. Temporal quality: 1 out of 2. Thematic accuracy: 1 out of 9. Positional correctness: 1 out of 1. Quality of free text: 1 out of 4.

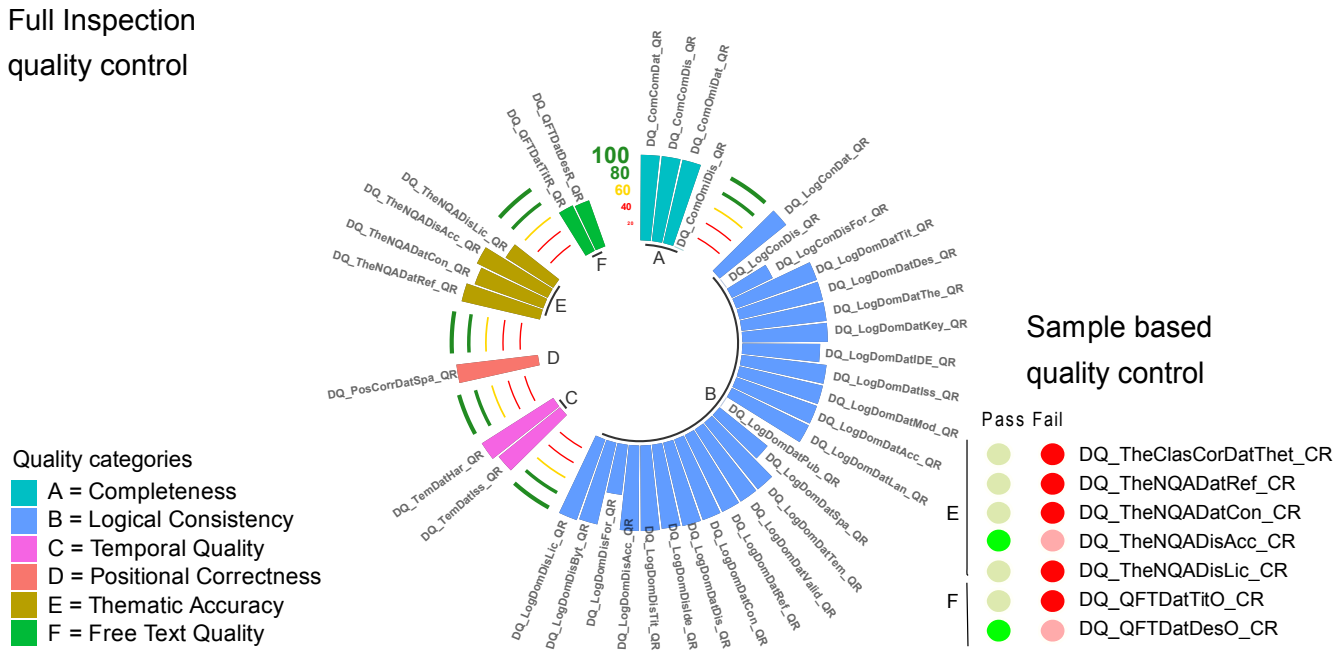

FIGURE 12. ISO 19157 dashboard summarizing the results shown in Tables 6, 7, and 8 (the labels correspond with the quality element identifiers in these tables)

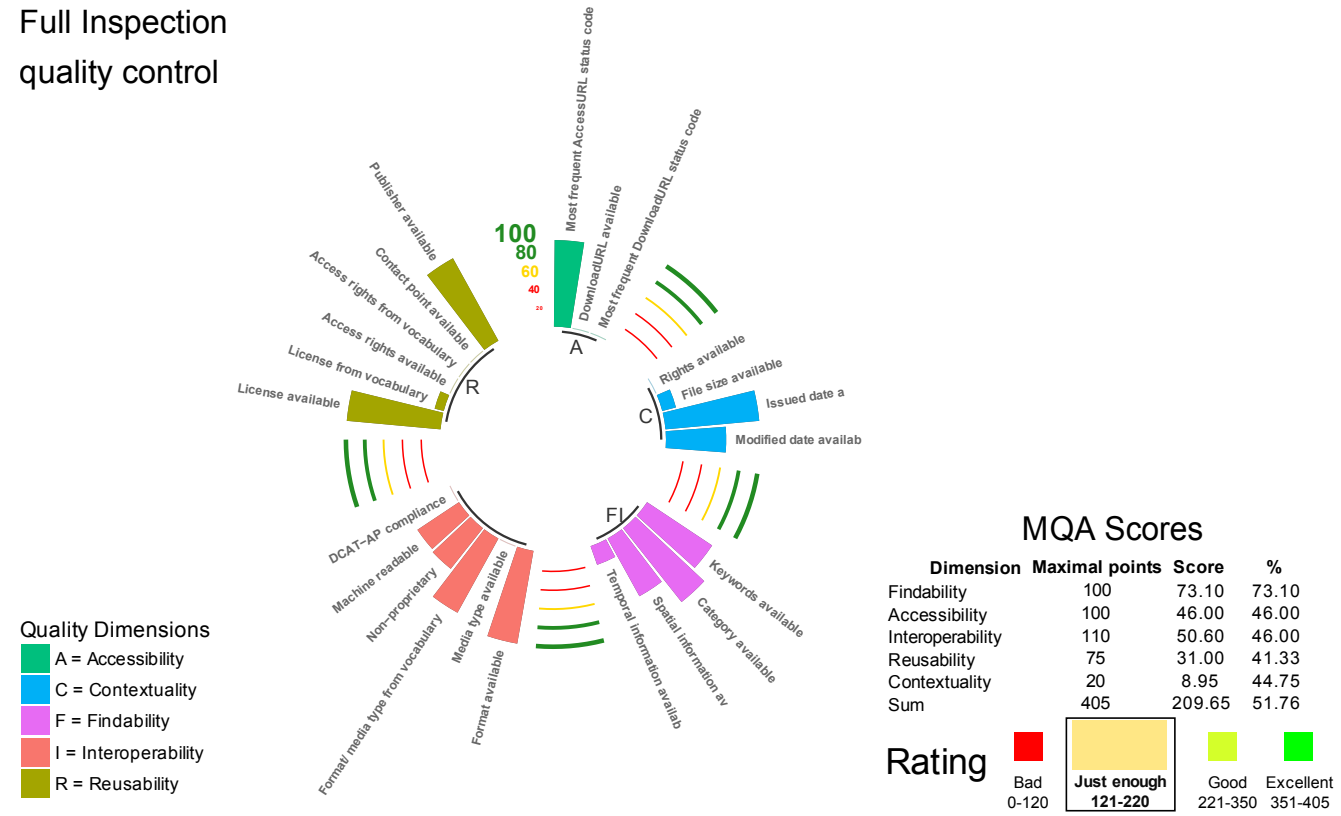

FIGURE 13. MQA dashboard summarizing the results shown in Table 9

vary. In the case of MQA, the European Data Portal verifies regularly the accessibility of access and download URLs. However, as the results shown in section V-C are obtained using our own implementation, the accessibility percentages are just a snapshot of a specific moment. In addition, it must be noted that MQA updates regularly the controlled vocab-

ularies for formats, licenses and access rights. The results obtained with MQA consider the controlled vocabularies used in April 2020.

VII. CONCLUSIONS

In this paper, we have proposed a new method for the evaluation of the quality of Open Data metadata based on ISO 19157. The benefits of applying the quality elements of ISO 19157 come from several lines. First of all, it is a quality model having almost twenty years of applied experience. This model includes more quality elements than other models (e.g., ISO 8000), giving more versatility to apply it. Furthermore, it offers numerous standardized measures that are directly applicable and, in addition, it is possible to define new measures according to new needs. Last, it is an international standard, which means that it has been generated in a high-level technical discussion process.

The original ISO 19157 based method for analyzing the quality of geographic metadata [12] has been customized in our proposed extension to the context of a DCAT-based metadata vocabulary. Although we have initially focused this extension of the method on the NTI metadata model, it could be easily applied to other vocabularies based on DCAT, i.e. any metadata schema defined in terms of different properties for Datasets and Distributions.

In addition, we have also demonstrated that the results of quality evaluation can be properly represented by means of the DQV vocabulary. Although there is not a 1:1 correspondence between the structure of quality elements and DQV concepts, it is a vocabulary that is being widely adopted in related methodological approaches like MQA.[8] Therefore, we have shown how ISO 19157 concepts can be expressed in the DQV vocabulary: the hierarchy of ISO 19157 quality categories and quality element types can be modeled as DQV categories and dimensions; ISO 19157 measures can be expressed as DQV metrics; and the instances of ISO 19157 quality elements and their associated results can be expressed as DQV measurements.

With respect to the feasibility of the proposed extended method, we have shown how to apply it to the metadata corpus of the Spanish Open Data catalog and we have compared the results with the ones obtained using the MQA methodology proposed by the European Data Portal. In general, we can conclude that the method based on ISO 19157 and MQA provide complementary perspectives thus being, in conjunction, one of the most appropriate approaches for the analysis of Open Data portals.

As future work, we plan to study the applicability of the proposed method for other DCAT-AP profiles. For instance, GeoDCAT-AP [36] is a metadata profile extending DCAT-AP, which was initially proposed by the European Commission in 2016 for the description of spatial data. The descriptors of this metadata schema have been designed to assure compatibility with the INSPIRE metadata regulation [37], and consequently it also covers the main elements of ISO 19115 metadata [38]. Its purpose is to give owners of geospatial metadata the possibility to achieve a wider audience by providing an additional RDF syntax binding, which can be more easily integrated in Open Data portals. Therefore, the evaluation of GeoDCAT-AP opens the door to verify if geospatial data published in Open Data portals is better described than geospatial data offered through traditional Spatial Data Infrastructures using ISO 19115 geographical metadata (the metadata schema for which the ISO 19157 based method was initially designed).

[8] See https://gitlab.com/european-data-portal/edp-vocabularies/-/blob/master/CustomVocabularies/edp-dqv-vocabulary.ttl.

ACKNOWLEDGMENT

This work has been partially supported by the Regional Government of Aragon (Spain) through the project T59_20R, and by the Regional Government of Andalusia (Spain) through the project PAIDI-TEP-164.

REFERENCES

[1] P. McDermott, "Building open government," *Government Information Quarterly*, vol. 27, pp. 401–413, 2010.
[2] P. Murray-Rust, "Open data in science," *Serials Review*, vol. 34, no. 1, pp. 52–64, 2008.
[3] E. Lakomaa and J. Kallberg, "Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs," *IEEE Access*, vol. 1, pp. 558–563, 2013.
[4] W3C, "Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation 04 February 2020," 2020. [Online]. Available: https://www.w3.org/TR/vocab-dcat/
[5] European Commission, "DCAT Application Profile for data portals in Europe, DCAT-AP v2.0.1," 2020. [Online]. Available: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0
[6] Spanish Ministry of Finance and Public Administration, *Technical Interoperability Standard for the Reuse of Informacion Resources*, 2013, ch. ANNEX III. Catalogue's document and information resource metadata. [Online]. Available: https://datos.gob.es/sites/default/files/doc/file/english_interoperability_agreement_for_the_reuse_of_information_resources.pdf
[7] Agenzia per l'Italia Digitale, "DCAT-AP_IT v1.0 - Italian profile of DCAT-AP," 2021. [Online]. Available: https://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0
[8] CKAN Association, "The CKAN website," 2021. [Online]. Available: https://ckan.org/
[9] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, and Y. L. Traon, "Comparison of metadata quality in open data portals using the analytic hierarchy process," *Government Information Quarterly*, vol. 35, no. 1, pp. 13 – 29, 2018.
[10] Publications Office of the European Union, "Metadata Quality Assessment Methodology. How EDP measures the quality of harvested metadata," 2020. [Online]. Available: https://www.europeandataportal.eu/mqa/methodology
[11] International Organization for Standardization (ISO), "ISO 19157:2013 Geographic information - Data quality," Geneva, CH, 2013.
[12] M. Ureña-Cámara, J. Nogueras-Iso, J. Lacasta, and F. Ariza-López, "A method for checking the quality of geographic metadata based on iso 19157," *International Journal of Geographical Information Science*, vol. 33, no. 1, pp. 1–27, 2019.
[13] W3C, "Data on the Web Best practices: Data Quality Vocabulary, W3C Working Group Note 15 December 2016. World Wide Web Consortium." 2016. [Online]. Available: https://www.w3.org/TR/vocab-dqv/
[14] S. Ma, C. Lu, X. Lin, and M. Galloway, "Evaluating the metadata quality of the IPL," in *Proceedings of the American Society for Information Science and Technology*, vol. 46, no. 1, 2009, p. 1–17.
[15] J.-R. Park, "Metadata quality in digital repositories: A survey of the current state of the art," *Cataloging & Classification Quarterly*, vol. 47, no. 3-4, p. 213–228, 2009.

[16] M. A. Gonçalves, B. L. Moreira, E. A. Fox, and L. T. Watson, "What is a good digital library?–A quality model for digital libraries," *Information processing & management*, vol. 43, no. 5, pp. 1416–1437, 2007.
[17] N. Veljković, S. Bogdanović-Dinić, and L. Stoimenov, "Benchmarking open government: An open data perspective," *Government Information Quarterly*, vol. 31, no. 2, pp. 278–290, 2014, dOI: 10.1016/j.giq.2013.10.011.
[18] S. Neumaier, J. Umbrich, and A. Polleres, "Automated quality assessment of metadata across open data portals," *Journal of Data and Information Quality (JDIQ)*, vol. 8, no. 1, Article 2, 2016.
[19] T. L. Saaty, *Decision Making for Leaders: the Analytic Hierarchy Process for Decisions in a Complex World*. RWS publications, 1999, vol. 2.
[20] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
[21] D. Nebert, Ed., *Developing Spatial Data Infrastructures: The SDI Cookbook*. Global Spatial Data Infrastructure (GSDI), 2004.
[22] International Organization for Standardization (ISO), "ISO 19115-1:2014. Geographic information - Metadata - Part 1: Fundamentals," Geneva, CH, 2014.
[23] ——, "ISO/IEC 25012:2008(en) Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model," Geneva, CH, 2008.
[24] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web*, vol. 7, pp. 63–93, 03 2015.
[25] International Organization for Standardization (ISO), "ISO 2859-2:1985 Sampling procedures for inspection by attributes – Part 2: Sampling plans indexed by limiting quality (LQ) for isolated lot inspection," Geneva, CH, 1985.
[26] J. Fernández-Huerta, "Medidas sencillas de lecturabilidad," *Consigna*, vol. 214, pp. 29–32, 1959.
[27] Z. Hernández-Figueroa, F. J. Carreras-Riudavets, and G. Rodríguez-Rodríguez, "Automatic syllabification for spanish using lemmatization and derivation to solve the prefix's prominence issue," *Expert Systems with Applications*, vol. 40, pp. 7122–7131, 2013.
[28] F. Szigriszt Pazos, "Sistemas Predictivos de Legibilidad del Mensaje Escrito: Fórmula de Perspicuidad [Readability predictive systems of the written message: Perspicuity formula]," Ph.D. dissertation, Universidad Complutense de Madrid, Ciudad Universitaria. Madrid (Spain), 1992.
[29] R. Flesch, *Marks of a Readable Style: A Style in Adult Education.Contributions to Education, No. 897*. Teachers College, Columbia University, New York: Bureau of Publications, 1942.
[30] W. Dubay, "Unlocking language: The classic readability studies," *Professional Communication, IEEE Transactions on*, vol. 51, pp. 416 – 417, 01 2009.
[31] M. Serban, "A readability analysis of French language online information on hearing related websites," University of Canterbury. New Zealand, 2018. [Online]. Available: https://ir.canterbury.ac.nz/bitstream/handle/10092/15642/Serban%2C%20Marius_MAUD%20Thesis.pdf?sequence=1
[32] C. Bocchetti, "Flesch Reading Ease e gli indici di leggibilità di un testo," 2018. [Online]. Available: https://www.asocialman.com/flesch_reading_ease-e-gli-indici_di_leggibilita-di-un-testo/
[33] International Organization for Standardization (ISO), "ISO/TS 19157-2:2016 Geographic information - Data quality - Part 2: XML schema implementation," Geneva, CH, 2016.
[34] J. Debattista, C. Lange, and S. Auer, "daQ, an Ontology for Dataset Quality Information," in *Proceedings of the Workshop on Linked Data on the Web (LDOW 2014), co-located with the 23rd International World Wide Web Conference (WWW 2014)*, 2014. [Online]. Available: http://ceur-ws.org/Vol-1184/ldow2014_paper_09.pdf
[35] J. Debattista, S. Auer, and C. Lange, "Luzzu—a methodology and framework for linked data quality assessment," *Journal of Data and Information Quality (JDIQ)*, vol. 8, no. 1, Article 4, 2016.
[36] European Commission, "GeoDCAT Application profile for data portals in Europe, GeoDCAT-AP v1.0.1," 2016. [Online]. Available: https://joinup.ec.europa.eu/release/geodcat-ap/101
[37] ——, "Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata," 2008. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2008/1205/oj
[38] INSPIRE Maintenance and Implementation Group (MIG), "Technical Guidelines for implementing dataset and service metadata based on ISO/TS 19139:2007. Version 2.0.1," 2017. [Online]. Available: http://inspire.ec.europa.eu/id/document/tg/metadata-iso19139

JAVIER NOGUERAS-ISO holds MS and PhD degrees in Computer Science from the University of Zaragoza. In 1998, he started his research at the Advanced Information Systems Laboratory of the University of Zaragoza. Currently, he is an Associate Professor of Computer Science at that University. Between 2011 and 2017 he was Director of Catedra Logisman on 'Technological Document Management', and between 2015 and 2019 he was Associate Director of the Aragon Institute of Engineering Research (I3A).

His research interests are focused on Information Retrieval and Semantic Web technologies applied to different domains, although with a special emphasis on Geographic Information Infrastructures.

JAVIER LACASTA holds a PhD in Computer Science since 2009, and he currently works as tenured Assistant Professor at the Computer Science and Systems Engineering Department of the University of Zaragoza (Spain).

His research work is focused in the field of Knowledge Management applied to Spatial Data, semantic web, information retrieval and data mining. Along the last years, he has co-authored numerous publications in books, journals or conference proceedings. He has also collaborated in several R+D projects in this field.

MANUEL ANTONIO UREÑA-CÁMARA is Associate Professor at the University of Jaén (Spain). He obtained his PhD in Geodesy and Cartographic Engineering in 2004 and a BSc in Computer Sciences.

He is member of the "Ingeniería Cartográfica" Research Group and his research lies on Geographic Information Systems, metadata, data modelling, quality control, generalization and digital photogrammetry.

FRANCISCO JAVIER ARIZA-LÓPEZ is Engineer (1991), and obtained his PhD from the University of Córdoba (1994).

Since 1989 his research has been devoted to working with spatial data. His expertise is in spatial data and data quality. He belongs to several standardization committees and has developed some standards on spatial data quality.

Apart from being Professor at the University of Jaén (Spain), he is currently the director of the Master's Degree of Science in Quality Assessment and Management of Geographical Information. He has numerous scientific publications in the field of spatial data and several manuals dedicated to the quality of spatial data. He has directed numerous doctoral theses and is an international consultant with a broad experience.