Data quality in an e-Government perspective

Jim J. Yang, Norwegian Digitalisation Agency

Anne Karete Hvidsten, Norwegian Digitalisation Agency

Morten Borrebaek, Norwegian Mapping Authority

10th January 2020

Abstract

Data sharing and reuse is one of the key prerequisites for digitalization of public administration (e-Government). In order to reuse data, one needs to know which data already exist, the meaning of the data, whether it is open or restricted access to the data, and last but not least, the quality of the data. As a first step towards more data sharing and reuse across the public administration and with the private sector, we have established a national data catalog which gives an overview of the datasets that the public administration collects and produces.

In this paper we will present our approaches to cope with the major challenges that we met when establishing our national data catalog, regarding 1) making available standardized and machine-readable data quality descriptions and 2) ensuring unified understanding of the data quality descriptions across the public administration.

Introduction

Quality of data is becoming increasingly important, also accelerated by digitalization of public administration (e-Government).

Norway's *National geospatial strategy towards 2025* ([1]) states that "Society needs good, up-to-date data in private and public activities, within all the specialist areas and sectors. Data must be available in ways that meet the needs. The data must have known coverage and a quality adapted to the needs of the various actors, so that it can support their specific applications and be part of the relevant decision-making processes."

The Norwegian Government white paper *Digital agenda for Norway* ([2]) emphasizes a user-centric and efficient public administration. Both *Digital agenda for Norway* and the follow-up *Digitalization strategy for public sector 2019-2025* ([3]) also emphasize the importance of sharing and reusing data across the public administration and with the private sector. Using and reusing correct and updated information is crucial for the provision of seamless public services across the public sector and for the exercise of authority. Using correct information increases the quality of the public services and strengthens the rule of law for citizens. Public services can be improved and automated through access to quality-controlled information from all public authorities.

The quality of data may affect how suitable the data is for other uses than first intended. Documentation of data quality is therefore useful in the process of evaluating whether a dataset is fit for purpose, thereby increased ability for potential users to reuse the dataset. The Norwegian government *Guidance on sharing and reuse of public administration's data* ([4]) therefore requires that the quality of the data should be documented and known challenges should be explicitly described.

As a first step towards more data sharing and reuse across the public administration and with the private sector, we have established a National data catalog ([5]), which contains not only descriptions of open data but also descriptions of data with restricted access. The national data catalog is actually a portal of catalogs that are interlinked. It consists currently of a catalog of *datasets*, a catalog of *concepts*, a catalog of *APIs* and a catalog of *information models*. It gives an overview of the datasets that the public administration collects and produces (datasets), the meaning of the datasets (concepts), the distribution of the datasets (APIs) and how the datasets/concepts are modeled (information models). If needed, more catalogs may be included in the catalog portal in the future. In addition to the aspects as the purpose of the datasets, the meaning of the data elements in the datasets, the legal basis for non-disclosure or disclosure of the datasets, distributions of the datasets etc., the data catalog also contains descriptions of the quality of the datasets.

In this paper we will present the challenges that we met in achieving standardized and machine-readable data quality descriptions in our national data catalog, and our approaches and solutions to cope with those challenges.

Standardized and machine-readable descriptions of data quality

When we started to develop our national data catalog in early 2016 regarding the inclusion of descriptions of data quality into the data catalog, the first challenge that we met was the lack of suitable standards. Our national data catalog is based on a distributed architecture. The national data catalog should be able to automatically harvest data descriptions provided by various sectors and agencies. One crucial aspect is thus standardized and machine-readable descriptions.

The national data catalog is in compliance with the national *Standard for description of datasets and data catalogs* *DCAT-AP-NO* ([6]) which is based on *DCAT-AP* ([7]), a European application profile of the W3C recommendation *DCAT (Data Catalog Vocabulary)* ([8]). Using the same standard, the national data catalog automatically harvests from other sources, e.g. the national portal for metadata of geospatial data ([10]) which is in compliance with the INSPIRE legislation ([9]) (as for member states of the European Union).

However, except for a few data quality aspects, current versions of DCAT from W3C and DCAT-AP from the European Commission, do not yet specify or recommend specifically how to describe quality of data in a machine-readable way. As presented at the 2nd International Workshop on Spatial Data Quality by Borrebaek and Buskerud ([11]), a national working group got the mandate to establish suitable standards for machine-readable descriptions of data quality, based on the needs from the Norwegian public administration. The working group delivered a *Specification for description of quality of datasets* ([12]). The working group concluded to extend our Norwegian application profile *DCAT-AP-NO* with relevant parts of *DQV (Data Quality Vocabulary)* ([13]) from W3C. DQV provides a framework in which the quality of a dataset can be described, whether by the dataset publisher or by a broader community of users.

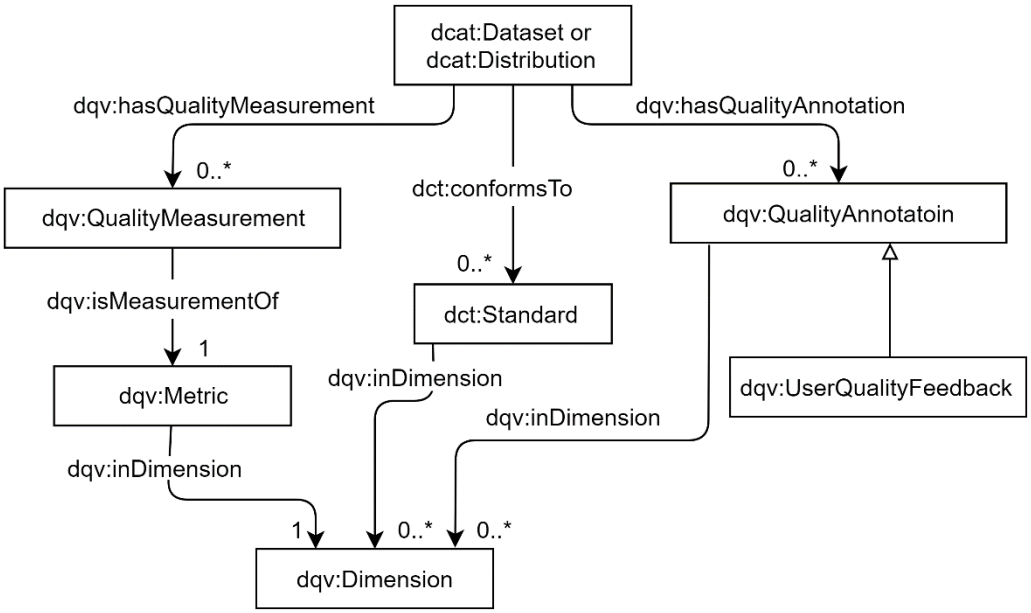

Figure 1: Simplified data model for extending DCAT-AP-NO with DQV for describing quality of datasets.

As shown in Figure 1, the working group suggested to start with the following types of quality descriptions based on the needs that were identified:

1. Description of quantitative data quality: One or more quantitative data quality measurements (dqv:QualityMeasurement) may be included in the description of a dataset (dcat:Dataset) using the property dqv:hasQualityMeasurement. Furthermore, using dqv:isMeasurementOf, one may specify which data quality metric (dqv:Metric) the data quality measurement is a measurement of, and using dqv:inDimension one may specify which quality dimension (dqv:Dimension) the data quality metric is within. E.g., "2%" as a measurement of the metric "rate of missing objects" in the quality dimension "completeness".
2. Description of data quality that conforms to given quality standards or specifications: Using the property dct:conformsTo one may specify that the quality of a dataset conforms to one or more given standards or specifications (dct:Standard). Similarly, using dqv:inDimension one may relate a standard/specification to one or more quality dimensions (dqv:Dimension).
3. Description of data quality in plain text: Using dqv:hasQualityAnnotation one may include one or more plain text descriptions of data quality in the description of a dataset, and relate the description to a quality dimension (dqv:Dimension) using dqv:inDimension. E.g. "2% missing objects" as a plain text description in the quality dimension "completeness".
4. Plain text user feedback on data quality: This is considered as a special case of plain text description mentioned above. The plain text description here is given by a user of the dataset, instead of the publisher of the dataset in the previous case.

The working group also identified the need to divide a quality dimension into "subdimensions", e.g. to divide the quality dimension "completeness" into "over-coverage" ("commission"), "under-coverage" ("omission") etc. "Subdimension" is not explicitly defined as a class in DQV but is possible to implement using DQV.

DQV is currently not yet a recommendation from W3C but a "Working Group Note". DQV is however the best specification that we found for machine-readable data quality descriptions covering the requirements from different domains. Nevertheless, based on the needs from Norway and several other European countries who are also using DQV, DQV is now indeed explicitly included in the upcoming European application profile of DCAT for base registries BRegDCAT-AP[1] ([14]).

Our national data catalog has already partially implemented DQV for describing quality of datasets.

Common definitions of data quality dimensions, quality subdimensions and quality metrics

The Data Quality Vocabulary (DQV) provides a generic framework, a vocabulary, for describing data quality. "The goal of the Data Quality Vocabulary is not to define a normative list of dimensions and metrics." ([13])

The second challenge that we met concerning data quality descriptions, was thus how to ensure that we have a unified understanding of the data quality descriptions in the data catalog, in order to achieve and increase semantic interoperability across the public administration.

The Norwegian geospatial community has assigned quality information to spatial datasets for several years, based upon *ISO 19157 Data Quality* ([15]) and *ISO 19115-1 Metadata* ([16]), the latter according to the European directive of *INSPIRE*. In the early days before we had international standards suitable for this purpose, we used a simple quality assignment in the form of measurement method (horizontal and vertical), positional accuracy (horizontal and vertical) and a rough statement on the visibility of the features from a photogrammetric point of view.

Other government agencies in Norway have been using standards such as *ISO/IEC 25012:2008 Data quality model* ([17], [18]) and *ISO/IEC 25024:2015 Measurement of data quality* ([19]). Some government agencies have similar quality elements specified in other specifications and regulations, such as *Eurostat's RAMON* ([20]), *Regulation (EC) No 223/2009* ([21]) of the European Union.

In 2019, we had a working group with the mandate to establish a set of common definitions based on ISO standards and other relevant standards and specifications, and to map the resulting definitions into the framework of DQV. The focus was standardized quality metrics. Since quality metrics should be related to quality dimensions, the working group had also the mandate to establish common definitions of the relevant quality dimensions and subdimensions.

The working group used the following criteria to decide what to define:

1. The mandate for the working group was to define metrics (dqv:Metric), i.e., only quantitative quality descriptions are included in the work.
2. Quality metrics that are only relevant for the data production phase are not included in the work, because it is about the quality of the datasets that are made available for reuse. E.g. metrics like "punctuality" are not included in the work.

[1] At the time of submitting this paper, BRegDCAT-AP is not yet finalized but a "stable draft".

3. Quality metrics that are already defined in existing standardized vocabularies are not included in the work. Examples of metrics that are already defined elsewhere and thus not included in the work are: “frequency at which dataset is published” (dcat:accrualPeriodicity) and “spatial/geographical coverage” (dct:spatial).
4. Sector specific quality metrics are not included in the work. Later in the process we became aware of that according to recommendations from ISA[2], geospatial should not be considered as sector specific, but generic.
5. Only inherent data quality metrics ([17], [18]) are included in the work. E.g. quality metrics like “accessibility” are not included in the work.

Table 1: Quality dimensions, quality subdimensions and quality metrics defined by the working group.

Quality dimension	**Quality subdimension**	**Quality metrics** (with data type)
completeness	under-coverage	missing objects (boolean)
		number of missing objects (integer)
		rate of missing objects (percentage)
		number of objects with missing value for a given property (integer)
		rate of objects with missing value for a given property (percentage)
	over-coverage	excess objects (boolean)
		number of excess objects (integer)
		rate of excess objects (percentage)
	imputation	number of objects with imputed value for a given property (integer)
		rate of objects with imputed value for a given property (percentage)
currentness	delay	overall time difference (xsd:duration)
consistency	consistency within the dataset	rate of objects with inconsistent properties (percentage)
		rate of objects with inconsistency between given properties (percentage)
accuracy	identifier correctness	number of objects with incorrect identifiers (integer)
		rate of objects with incorrect identifiers (percentage)
	classification correctness	number of incorrectly classified objects for a given property (integer)
		rate of incorrectly classified objects for a given property (percentage)

[2] Interoperability solutions for public administrations, businesses and citizens, https://ec.europa.eu/isa2/home_en

As shown in Table 1, the working group established a set of common definitions of quality metrics within the quality dimensions “completeness”, “currentness”, “consistency” and “accuracy”. The definitions of the mentioned quality dimensions, quality subdimensions and quality metrics, with examples, are listed in Appendix B of this paper. The definitions together with a guideline for how to use them, have been through a broad national commenting process.

Summary and future work

One of the key prerequisites for digitalization of public administration (e-Government) is data sharing and reuse. In order to reuse data, one needs to know which data already exist. Furthermore, quality of data is one of the aspects that is important for potential users of a dataset, to evaluate whether the dataset is reusable or not.

As a first step towards more data sharing and reuse, we have established a national data catalog which contains standardized and machine-readable descriptions of datasets that are collected and produced by the public administration. Among of the aspects that are described in our national data catalog is the quality of datasets.

As illustrated and summarized in Figure 2:

- In order to have standardized and machine-readable data quality descriptions in our national data catalog, we have chosen to incorporate Data Quality Vocabulary (DQV) into our national standard for description of datasets and data catalogs (DCAT-AP-NO).
- In order to ensure unified understanding of the quality descriptions across the public administration, we have chosen to establish common definitions of quality dimensions, quality subdimensions and quality metrics.

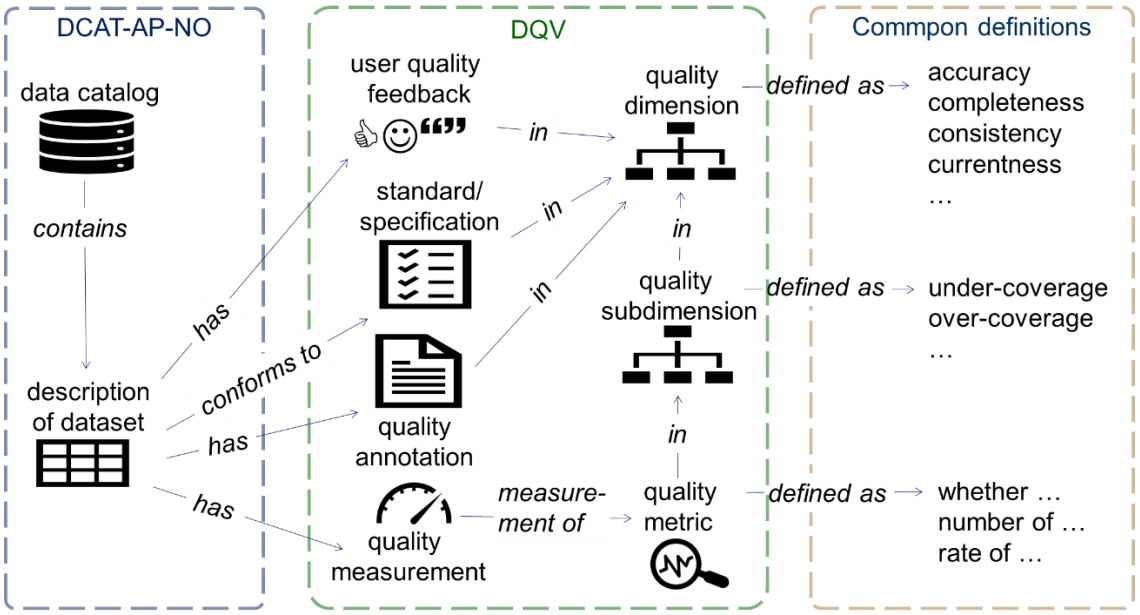

Figure 2: Incorporating DQV into DCAT-AP-NO for describing the quality of datasets, referring to common definitions of quality metrics, quality subdimensions and quality dimensions.

Future work:

- Our national standard DCAT-AP-NO will be revised (probably during spring 2020), with DQV explicitly incorporated, and aligned with DCAT-AP which was recently revised.
- The definitions from the working group will soon be published, with the preferred terms and definitions in both Norwegian and English, also in machine-readable formats (e.g. RDF).
- When and if needed, more definitions will be established and published bilingually and machine-readably. Geospatial quality is among the domains that will be prioritized.
- When and if needed, we will also establish a solution for making accessible and machine-readable sector specific metric definitions.

References and links to online resources

[1] *Everything happens somewhere - National geospatial strategy towards 2025*, https://www.regjeringen.no/contentassets/6e470654c95d411e8b1925849ec4918d/en-gb/pdfs/en_nasjonal_geodatastrategi.pdf

[2] *Digital agenda for Norway in brief*, https://www.regjeringen.no/en/dokumenter/digital-agenda-for-norway-in-brief/id2499897/

[3] (title freely translated from Norwegian) *One digital public sector: Digitalization strategy for public sector 2019-2025*, https://www.regjeringen.no/no/tema/statlig-forvaltning/ikt-politikk/digitaliseringsstrategi-for-offentlig-sektor/id2612415/ (in Norwegian).

[4] (title freely translated from Norwegian) *Guidance on sharing and reuse of public administration's data*, https://www.regjeringen.no/no/dokumenter/retningslinjer-ved-tilgjengeliggjoring-av-offentlige-data/id2536870/ (in Norwegian).

[5] The National Data Catalog, https://fellesdatakatalog.brreg.no/about

[6] (title freely translated from Norwegian) *Standard for description of datasets and data catalogs (DCAT-AP-NO)*, https://doc.difi.no/dcat-ap-no/ (in Norwegian).

[7] DCAT Application Profile for data portals in Europe (DCAT-AP), https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

[8] *Data Catalog Vocabulary (DCAT)*, https://www.w3.org/TR/vocab-dcat-2/

[9] INSPIRE Legislation, https://inspire.ec.europa.eu/inspire-legislation/26

[10] Map Catalogue, the national portal for metadata of geospatial data, https://www.geonorge.no/en/

[11] Morten Borrebaek and Magni Busterud, *From quality on spatial data to data quality vocabulary (DQV) for the Semantic Web – the Norwegian experience of aligning ISO 19157 Data Quality to DQV*, https://eurogeographics.org/wp-content/uploads/2018/06/11-ISO19157.pdf.

[12] (title freely translated from Norwegian) *Specification for description of quality of datasets*, https://doc.difi.no/data/kvalitet-pa-datasett/ (in Norwegian).

[13] *Data on the Web Best Practices: Data Quality Vocabulary* (DQV), https://www.w3.org/TR/vocab-dqv/

[14] *Specification of Registry of Registries*, https://joinup.ec.europa.eu/solution/abr-specification-registry-registries/news/stable-draft-bregdcat-ap.

[15] ISO 19157:2013 *Geographic information — Data quality*, https://www.iso.org/standard/32575.html

[16] ISO 19115-1:2014 *Geographic information — Metadata — Part 1: Fundamentals*, https://www.iso.org/standard/53798.html
[17] ISO/IEC 25012:2008 *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*, https://www.iso.org/standard/35736.html
[18] ISO 25012, ISO 25000 Portal, https://iso25000.com/index.php/en/iso-25000-standards/iso-25012
[19] ISO/IEC 25024:2015 *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*, https://www.iso.org/standard/35749.html
[20] EuroStat *RAMON - Reference And Management Of Nomenclatures*, https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC
[21] *Regulation (EC) No 223/2009 of the European Parliament and of the Council*, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009R0223
[22] BLUE-ETS, BLUE-Enterprise and Trade Statistics, http://www.pietdaas.nl/beta/pubs/pubs/BLUE-ETS_WP4_Del2.pdf
[23] (title freely translated from Norwegian) *Standards for Geographic Information – Geodata quality*, https://kartverket.no/globalassets/standard/bransjestandarder-utover-sosi/geodatakvalitet.pdf (in Norwegian).

Appendix A – Prefixes used in this paper

Table 1: Prefixes used in this paper.

Prefix	Namespace	Name of the vocabulary
dcat	http://www.w3.org/ns/dcat#	Data Catalog Vocabulary
dct	http://purl.org/dc/terms/	(Dublin Core) DCMI Metadata Terms
dqv	http://www.w3.org/ns/dqv#	Data Quality Vocabulary
oa	http://www.w3.org/ns/oa#	Web Annotation Ontology
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema

Appendix B – Quality metrics and the relevant quality subdimensions and quality dimensions that are defined

Note: At the time of submission of this paper, the definitions listed in this appendix are not yet publicly published. There might therefore be some minor changes in the final published version.

Table 2: Definitions of quality metrics and the relevant quality subdimensions and quality dimensions.

Quality dimension	Quality subdimension	Quality metric (with data type)
completeness the degree to which subject data	**under-coverage**	**missing objects** (boolean) whether objects are missing in the dataset (based on ISO 19157, [15])

Quality dimension	**Quality subdimension**	**Quality metric** (with data type)
associated with an entity has values for all expected attributes and related entity instances in a specific context of use (ISO 25012, [18])	data absent from a dataset (ISO 19157, [15]) *Alternative term: omission*	Example: “false” (the dataset contains all buildings)
		number of missing objects (integer) number of objects that are not present in the dataset but are expected to be (based on ISO 19157, [15]) Example: “2” (Two buildings are missing in the dataset)
		rate of missing objects (percentage) number of missing objects in relation to the number of objects that should be present in the dataset (based on ISO 19157, [15]) Example: “0.02%” (0.02% of buildings are missing in the dataset)
		number of objects with missing value for a given property (integer) number of objects in the dataset with missing value for a given property (our own definition) Example: “2” (Two buildings in the dataset do not have value for the property “usable area”)
		rate of objects with missing value for a given property (percentage) number of objects with missing value for a given property in relation to the number of objects in the dataset (our own definition) Example: “0.02%” (0.02% of buildings in the dataset do not have value for the property “usable area”)
	over-coverage excess data present in a dataset (ISO 19157, [15]) *Alternative term: commission*	**excess objects** (boolean) whether there are objects incorrectly present in the dataset (based on ISO 19157, [15]) Example: “true” (some buildings in the dataset are not supposed to be there)
		number of excess objects (integer) number of objects in the dataset that should not have been present (based on ISO 19157, [15]) Example: “3” (Three buildings in the dataset are not supposed to be there)

Quality dimension	**Quality subdimension**	**Quality metric** (with data type)
		rate of excess objects (percentage) number of excess objects in the dataset in relation to the number of objects that should have been present (based on ISO 19157, [15]) Example: "0.03%" (0.03% of the buildings in the dataset are not supposed to be there)
	imputation entering a value for a specific data item where the value is missing or unusable (EuroStat RAMON, [20])	**number of objects with imputed value for a given property** (integer) number of objects in the dataset with imputed value for a given property (our own definition) Example: "4" (Four buildings in the dataset have imputed value for the property "year of construction")
		rate of objects with imputed value for a given property (percentage) number of objects with imputed value for a given property in relation to the number of objects in the dataset (our own definition) Example: "0.04%" (0.04% of the buildings have imputed value for the property "year of construction")
currentness the degree to which data has attributes that are of the right age in a specific context of use (ISO 25012, [18])	**delay** age of the dataset described as the difference between two points in time (our own definition)	**overall time difference** (xsd:duration) length of time between data availability and the event or phenomenon they describe (EuroStat RAMON, [20]) Example: "24 days" (On average there will be 24 days from a building is completed or demolished, to it is included in or excluded from the dataset)
consistency the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data	**consistency within the dataset** the degree to which there is consistency between the properties in the dataset (our own definition)	**rate of objects with inconsistent properties** (percentage) number of objects with inconsistent properties in relation to the number of objects in the dataset (our own definition) Example: "0.03%" (0.03% of the buildings have inconsistency between some properties)
		rate of objects with inconsistency between given properties (percentage)

Quality dimension	**Quality subdimension**	**Quality metric** (with data type)
regarding one entity and across similar data for comparable entities. (ISO 25012, [18])		number of objects with inconsistency between given properties in relation to the number of objects in the dataset (our own definition) Example: "0.03%" (0.03% of the buildings in the dataset have "usable area" larger than "gross area")
accuracy the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use (ISO 25012, [18])	**identifier correctness** the degree to which the objects in the dataset have the correct identifiers (based on BLUE-ETS, [22])	**number of objects with incorrect identifiers** (integer) number of objects in the dataset with incorrect identifiers (our own definition) Example: "1" (One building in the dataset has wrong identifier)
		rate of objects with incorrect identifiers (percentage) number of objects with incorrect identifiers in relation to the number of objects in the dataset (our own definition) Example: "0.01%" (0.01% of the buildings in the dataset have wrong identifiers)
	classification correctness comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference data) (ISO 19157, [15])	**number of incorrectly classified objects for a given property** (integer) number of objects in the dataset that are incorrectly classified for a given property (based on ISO 19157, [15]) Example: "1" (One building in the dataset is classified with wrong occupancy code)
		rate of incorrectly classified objects for a given property (percentage) number of objects that are incorrectly classified for a given property in relation to the number of objects in the dataset (based on ISO 19157, [15]) Example: "0.01%" (0.01% of the buildings in the dataset are classified with wrong occupancy codes)