www.data.europa.eu

Metadata Quality Assurance

27.04.2025

How data.europa.eu measures the quality of harvested metadata

The Metadata Quality Assessment (MQA) is a tool developed by the consortium of data.europa.eu to study the quality of metadata harvested by data.europa.eu. It is intended to help data providers and data portals to check their metadata quality and to receive suggestions for improvements. The results are presented via the MQA and are also available as download. In the following we describe the functionality of the MQA and the methodology it uses.

If this page still does not answer all your questions, please feel free to contact us via our feedback form at the end of the page.

Metadata Quality Assessment Methodology

Scope of research

With the MQA, we want to answer the following question:
What is the metadata quality for public sector data in the pan-European region and where are the biggest hurdles to achieving better quality?

Based on this, the MQA is currently investigating the following concrete questions:

- Complies with DCAT-AP and DCAT-AP derivatives
- Disclosure of information to which DCAT-AP is not obligated
- Accessibility of the data referenced in the metadata through the Access and Download URL
- Machine readability of the referenced data
- Use of licenses

Each question again results in individual investigations, which are described in detail below.

What do we not cover

The MQA is limited by the metadata it can examine. The investigation is limited exclusively to the metadata that data.europa.eu collects during the harvesting process. If there are errors in the source metadata, these can falsify the overall result. To limit this error potential, the MQA provides a validation service that can be used by data providers to validate their metadata for valid formats and compliant DCAT-AP before integrating it into the harvesting process.

DCAT-AP SHACL validation service web page

DCAT-AP SHACL validation service API

The MQA Process

With each harvesting, the metadata is also checked by the MQA. The MQA measures the quality of various indicators, each indicator is explained in the tables below. The results of the checks are stored as Data Quality Vocabulary (DQV). DQV is a specification of the W3C that is used to describe the quality of a dataset.

As accessibility can be volatile, repeated checks for the accessURL and downloadURL are necessary. For this reason, the MQA regularly checks the accessibility of all distributions. In contrast to the verification of the other indicators, this has a higher runtime, since the distributions are checked via HTTP and each requested URL may have a longer response time. The MQA uses a mechanism that takes into account that each URL is re-examined for accessibility within a few weeks of the last review.

Assumptions

The MQA is based on the following assumptions.

Use of non-obligatory fields

We believe that filling the DCAT-AP mandatory fields alone is not sufficient to provide high quality metadata. For this reason, the evaluation also checks fields that are not specified as mandatory according to DCAT-AP. The exact fields that are checked are listed below.

Identical content for multiple distributions

If a dataset contains more than one distribution, all distributions are identical in content, they differ only in the representation of the data. For example, a dataset can have two distributions, one offering the data as PDF and the other offering the identical data as machine-readable RDF/XML.

Dimensions

This section describes all dimensions that the MQA examines in order to determine the quality. The dimensions are derived from the FAIR principles .

Findability

The following table describes the metrics that help people and machines in finding datasets. A maximum of 100 points can be scored in this area.

Indicator	Description	Metrics	Weight
Keyword usage	Keywords directly support the search and thus increase the findability of the data dataset.	The system checks whether keywords are defined. The number of keywords has no impact to the score. Dataset dcat:keyword	30
Categories	Categories help users to explore datasets thematically.	It is checked whether one or more categories are assigned to the dataset. The number of assigned categories has no impact to the score. Dataset dcat:theme	30
Geo search	Usage of spatial information would enable users in order to find the dataset with a geo facetted search.	It is checked wether the property is set or not. Dataset dcat:spatial	20
Time based search	Usage of temporal information would enable users for a timely based facetted search.	It is checked wether the property is set or not.	20

Accessibility

The following table describes which metrics are used to determine whether access to the data referenced by the distributions is guaranteed. A maximum of 100 points can be scored in this area.

Indicator	Description	Metrics	Weight
AccessURL accessibility	The AccessURL is not necessarily a direct link to the data, but also may refers to a URL that gives access to the dataset or where more information about the dataset is available.	The specified URL is checked for accessibility via a HTTP HEAD request. If the responded status code is in the 200 or 300 range, the accessibility of the resource is evaluated positively. Distribution dcat:accessURL	50
DownloadURL	The downloadURL is a direct link to the referenced data.	It is checked wether the property is set or not. Distribution dcat:downloadURL	20
DownloadURL accessibility	If a downloadURL exists, the accessibility is checked.	The specified URL is checked for accessibility via a HTTP HEAD request. If the responded status code is in the 200 or 300 range, the accessibility of the resource is evaluated positively. Distribution dcat:downloadURL	30

Interoperability

The following table describes the metrics used to determine whether a distribution is considered interoperable. According to the assumption 'identical content with several distributions', only the distribution with the highest number of points is used to calculate the points. A maximum of 110 points can be scored in this area.

Indicator	Description	Metrics	Weight
Format	This field specifies the file format of the distribution.	It is checked wether the property is set or not. Distribution dct:format	50
Media type	This field specifies the media type of the distribution.	It is checked wether the property is set or not. Distribution dcat:mediaType	10
Format / Media type from vocabulary	Checks whether format and media type belong to a controlled vocabulary.	The format vocabulary can be found in the data.europa.eu GitLab repository. The media type is check against the IANA list Distribution dct:format dcat:mediaType	10
Non-proprietary	Checks if the format of the distribution is non-proprietary.	The distribution is considered as non-proprietary if the specified format and media type is contained in the corresponding data.europa.eu GitLab repository vocabulary. Distribution dct:format	20
Machine readable	Checks if the format of the distribution is machine-readable.	The distribution is considered as machine-readable if the specified format and media type is contained in the corresponding data.europa.eu GitLab repository vocabulary. Distribution dct:format	20
DCAT-AP compliance	DCAT-AP compliance is calculated across all sources and datasets available on a catalogue. This check is only performed if the metadata is originally harvested as DCAT-AP or as a valid derivate. DCAT-AP is a specification for describing linked public data in Europe. The data.europa.eu portal may also harvest metadata which does not fully comply to DCAT-AP. In order to increase conformity to DCAT-AP, the MQA checks each metadata for its DCAT-AP compliance.	The metadata is validated against a set of SHACL shapes. The metadata is not compliant, if the SHACL validation reports at least one issue. The MQA uses data.europa.eu's DCAT-AP SHACL validation service. SHACL is a recommendation from the W3C and is used for validating RDF graphs against a set of shapes.	30

Reusability

The following table describes which metrics are used to check the reusability of the data. A maximum of 75 points can be scored in this area.

Indicator	Description	Metrics	Weight
License information	A license is valuable information for the reuse of data.	It is checked wether the property is set or not. Distribution dct:license	20
License vocabulary	We would like to limit the indication of incorrect license information. For example, we encounter many CC licenses that lack versioning.	This section describes all dimensions that the MQA examines in order to determine the quality. The dimensions are derived from the FAIR principles. The MQA recommends and credits the usage of controlled vocabularies. The data.europa.eu portal publishes its controlled vocabularies in GitLab. The vocabularies are derived from the EU Vocabularies. Distribution dct:license	10
Access restrictions	This field indicates whether the access to the data is public or restricted.	It is checked wether the property is set or not. Dataset dct:accessRights dcat:mediaType	10
Access restrictions vocabulary	The use of a controlled vocabulary increases reusability.	It is checked whether the controlled vocabulary for access rights is used. Dataset dct:accessRights	5
Contact point	The contact point contains inform whom to address in case of questions regarding the data.	It is checked wether the property is set or not. Dataset dct:contactPoint	20
Publisher	The publisher is a person or organisation that has published the data.	It is checked wether the property is set or not. Dataset dct:publisher	10

Contextuality

The following table show some light weight properties, that provide more context to the user. A maximum of 20 points can be scored in this area.

Indicator	Description	Metrics	Weight
Rights	In some cases, a specific license cannot be applied to a dataset. The 'Rights' field can be used to specify a reference to a resource that will inform a user about the rights he has when using the dataset.	It is checked wether the property is set or not. Distribution dct:rights	5
File size	Specifies the size of the file in bytes.	It is checked wether the property is set or not. Distribution dct:byteSize	5
Date of issue	The date on which the dataset or distribution was released.	It is checked wether the property is set or not. Dataset and Distribution dct:issued	5
Modification date	The date on which the dataset or distribution was last changed.	It is checked wether the property is set or not. Dataset and Distribution dct:modified	5

Rating

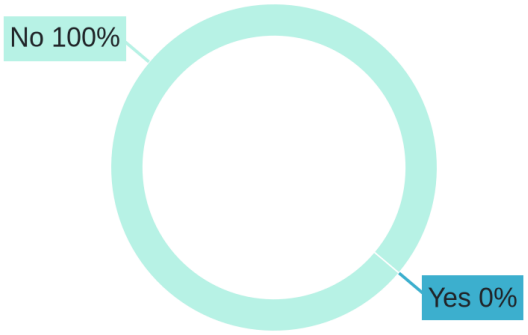
The final rating happens via four rating groups. The mapping of the points to the rating category is shown in the table below. The representation of the rating in the MQA is expressed exclusively via the rating categories. This enables providers to achieve the highest rating even with a slight deduction of points.

Dimension	Maximal points
Findability	100
Accessibility	100
Interoperability	110
Reusability	75
Contextuality	20
Sum	**405**

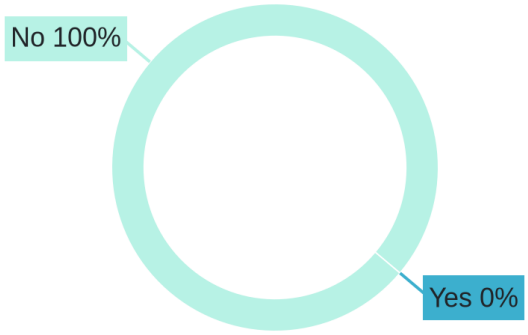
Rating	Range of points
Excellent	351 - 405
Good	221 - 350
Sufficient	121 - 220
Bad	0 - 120

Contextuality

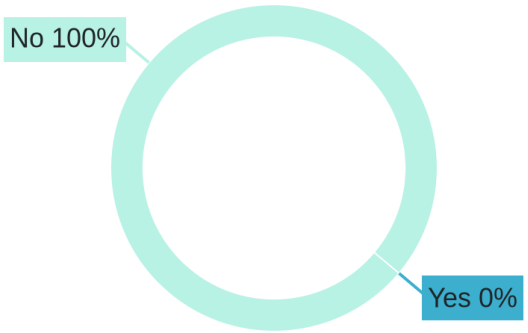
Rights



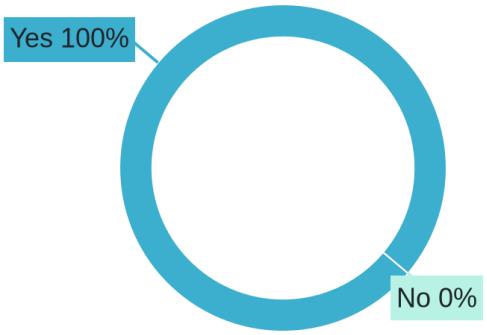
File size



Date of issue



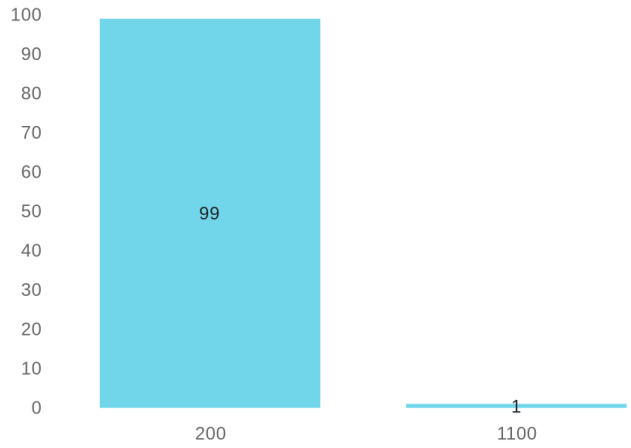
Modification date



European Central Bank

Accessibility

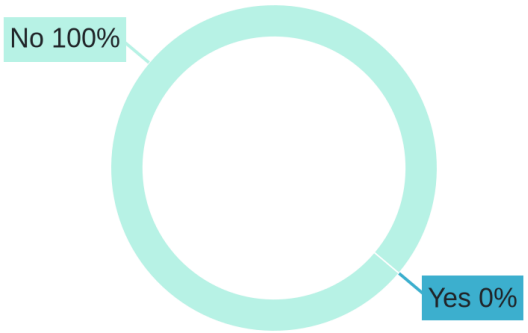
Most frequent accessURL status codes



Most frequent downloadURL status codes

No measurements available for this indicator

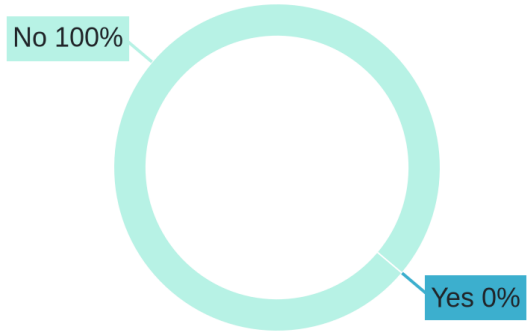
Download URL



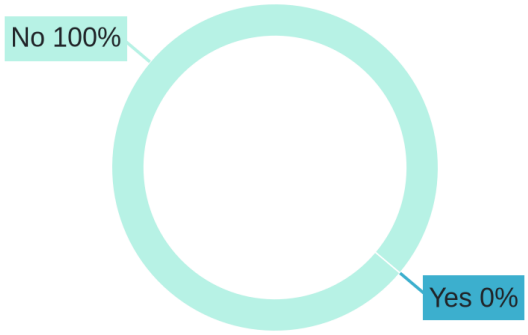
European Central Bank

Reusability

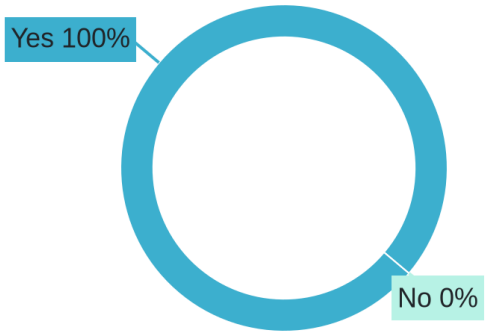
License information



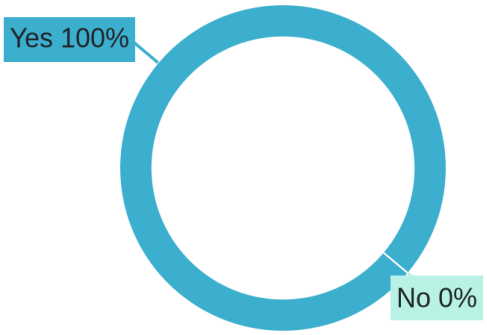
License vocabulary



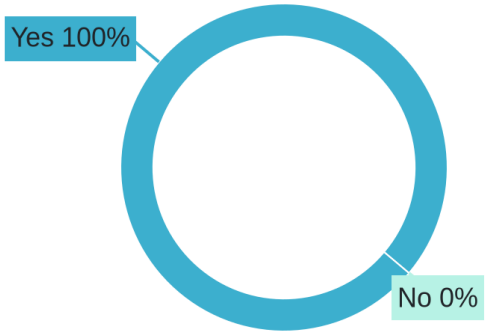
Access restrictions



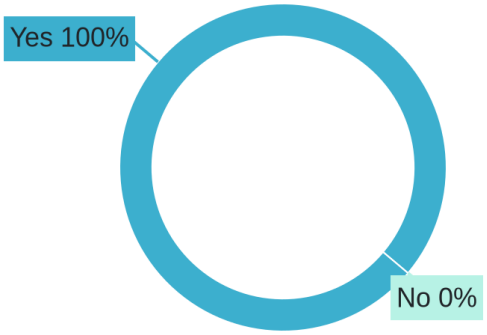
Access restrictions vocabulary



Contact point



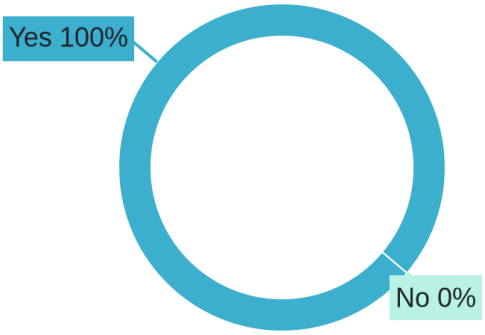
Publisher



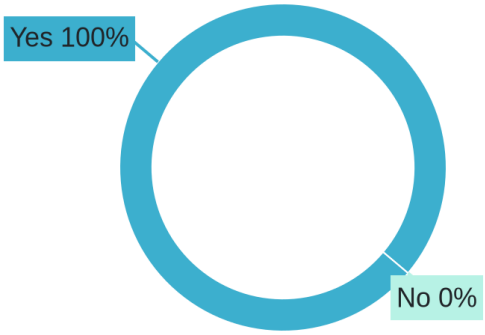
European Central Bank

Interoperability

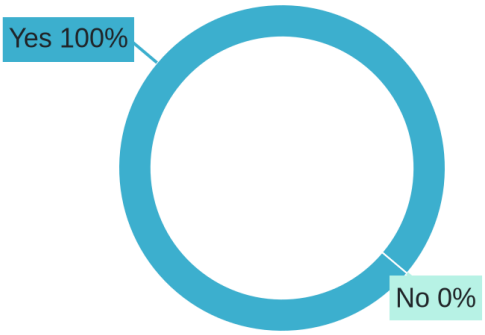
Format



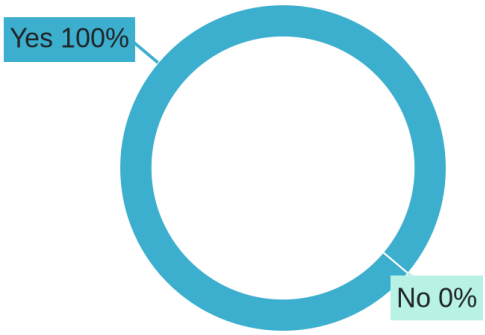
Media type



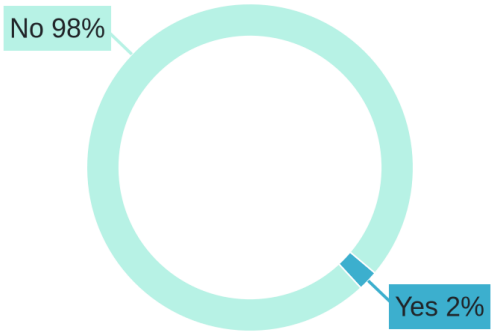
Format / Media type from vocabulary



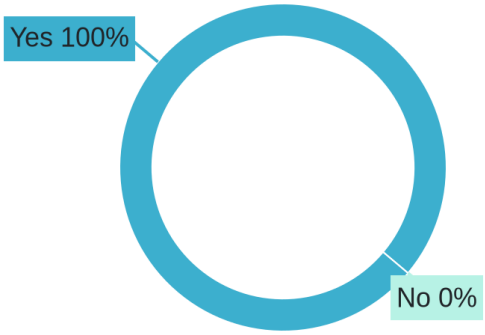
Non-proprietary



Machine readable



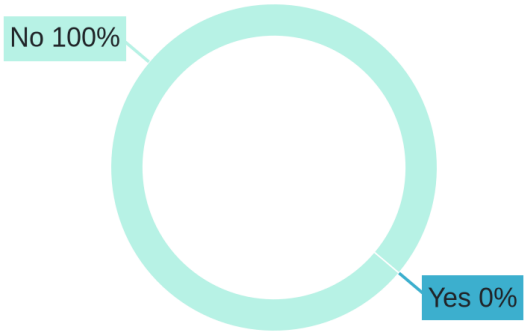
DCAT-AP compliance



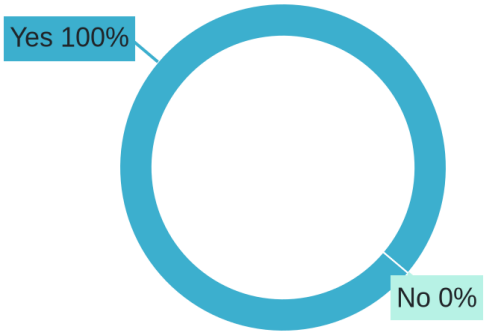
European Central Bank

Findability

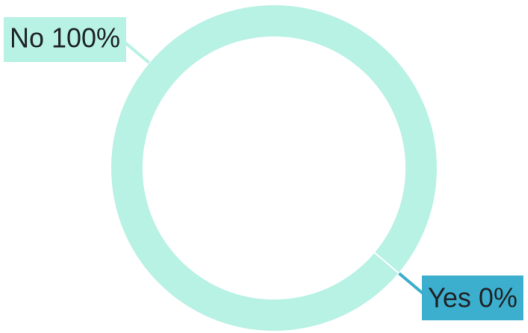
Keyword usage



Categories



Geo search



Time based search