



data.
europa.
eu
academy 

Training session on Data and Metadata Quality

24 February 2022



data.europa.eu The official portal
for European data



Introduction



Esther Huyer
data.europa.eu
Project manager



Laura van Knippenberg
data.europa.eu
Support to EU Institutions
& Member States



Benjamin Dittwald
Fraunhofer FOKUS
Technical team

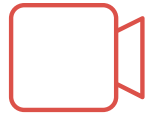


Lina Bruns
Fraunhofer FOKUS
Technical Team

Agenda of today

- Opening
- Introduction to data quality
- How metadata quality can be determined on data.europa.eu?
- Break
- How to improve data quality?
- Q&A and summary
- Feedback poll and closing

Rules of the game



The webinar will be recorded



Please mute yourselves during the webinar

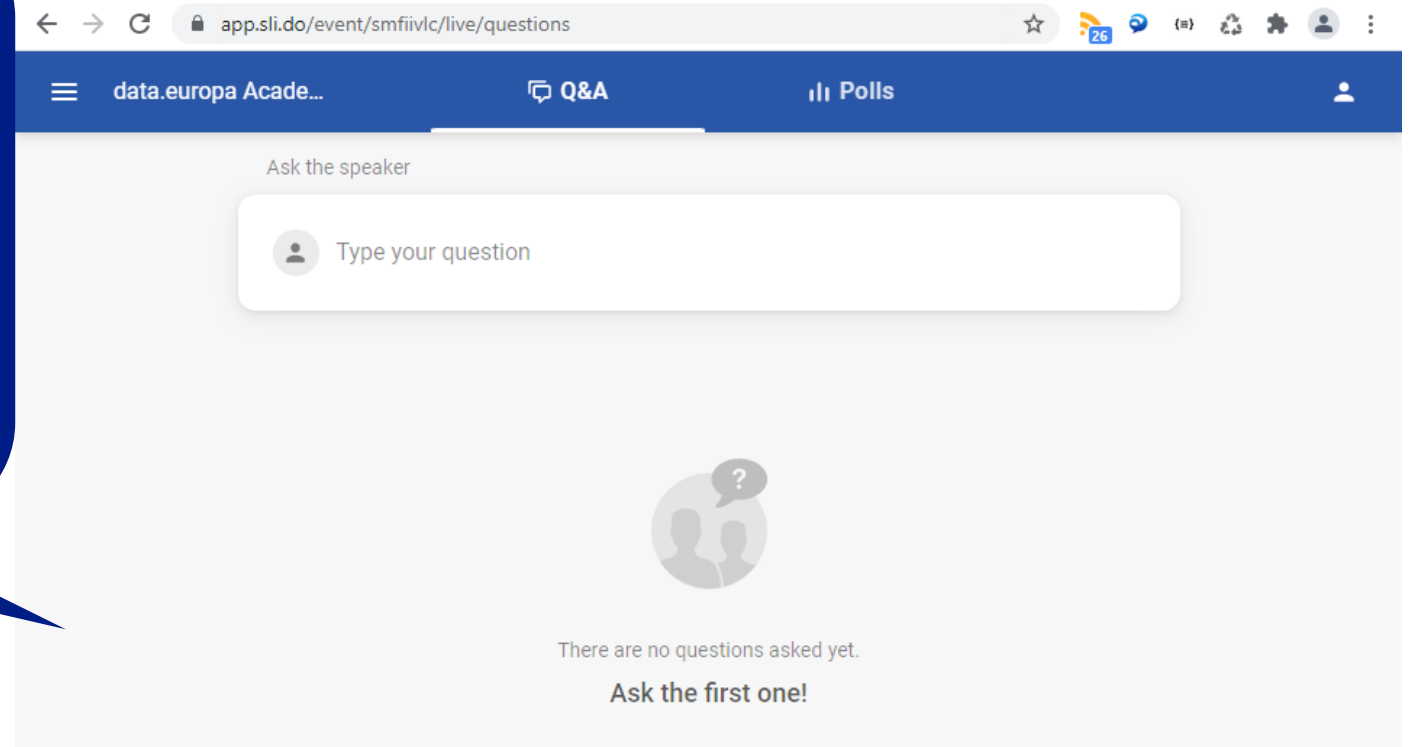
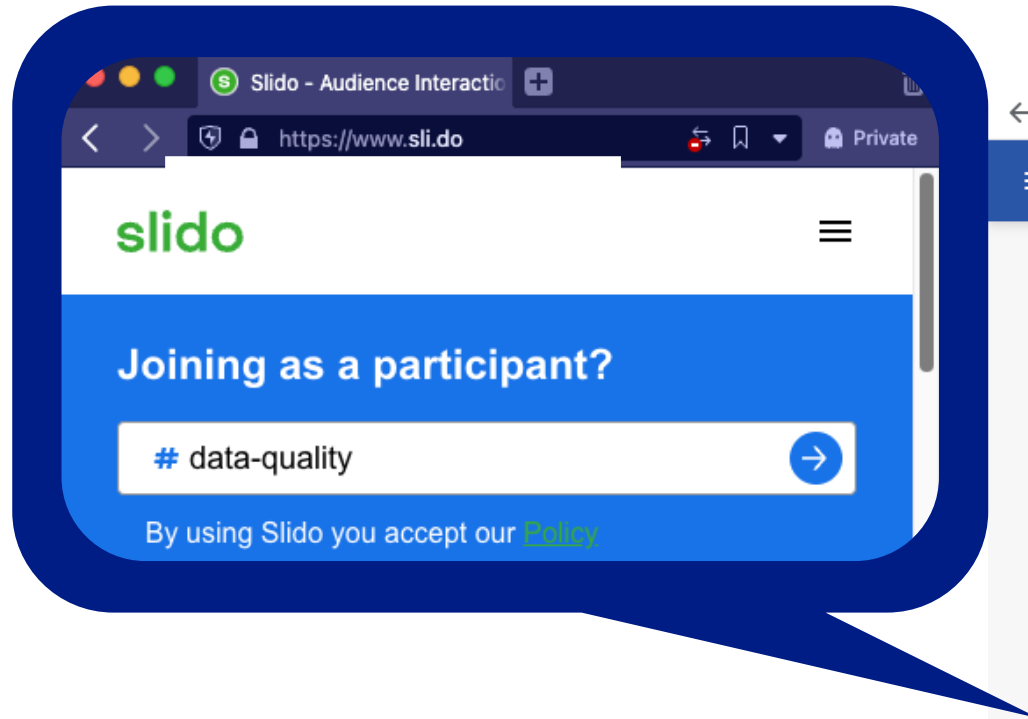


Please reserve 3 min after the webinar to help us improve by filling in our feedback form



For questions, please use [sli.do](#). Vote for the questions that are of most interest to you, those will be discussed later

Ask your questions in sli.do using the code *data-quality*



A close-up photograph of a hand with a diamond ring pointing to a hand-drawn map on a textured, light-colored surface. The map features dashed lines, a red spiral, and a red 'X'.

Where are you now?

Please go to sli.do
and enter *data-quality*

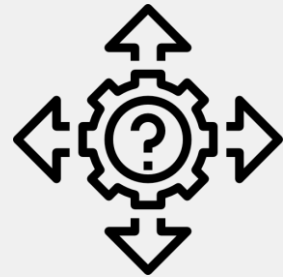


Introduction
to data quality

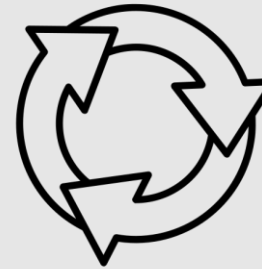


**Name three words you
associate with data quality**

Why data quality is important



**Informed
Decision-Making**



Reuse of data

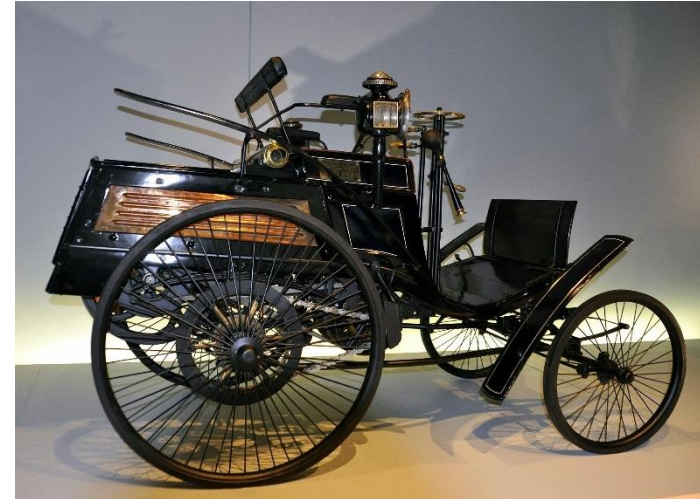
Quality dimensions and indicators

Findability	Accessibility	Interoperability	Reusability
Completeness	Accessibility / Availability	Conformity / Compliance	Timeliness
Findability		Openness	Consistency
		Machine Readability / Processability	Accuracy
			Relevance
			Understandability
			Credibility

Data quality is subjective



Public Transport Data



Historical Data

Quality of metadata and data

Metadata



Issues

- Title not understandable
- No description given
- Author unknown
- No publication year given
- Unusual format
- ...

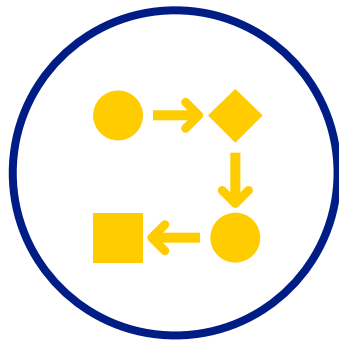
Data



Issues

- Does not open
- Missing pages or words
- Written in a foreign language
- Content is outdated
- Contains unknown letters and words
- ...

How to increase (meta)data quality



Introduce processes

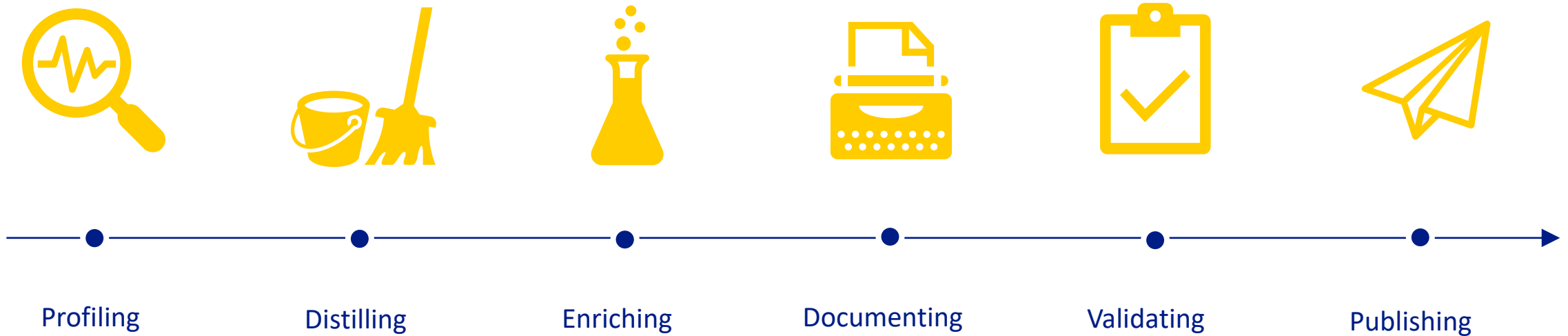


Make use of tooling



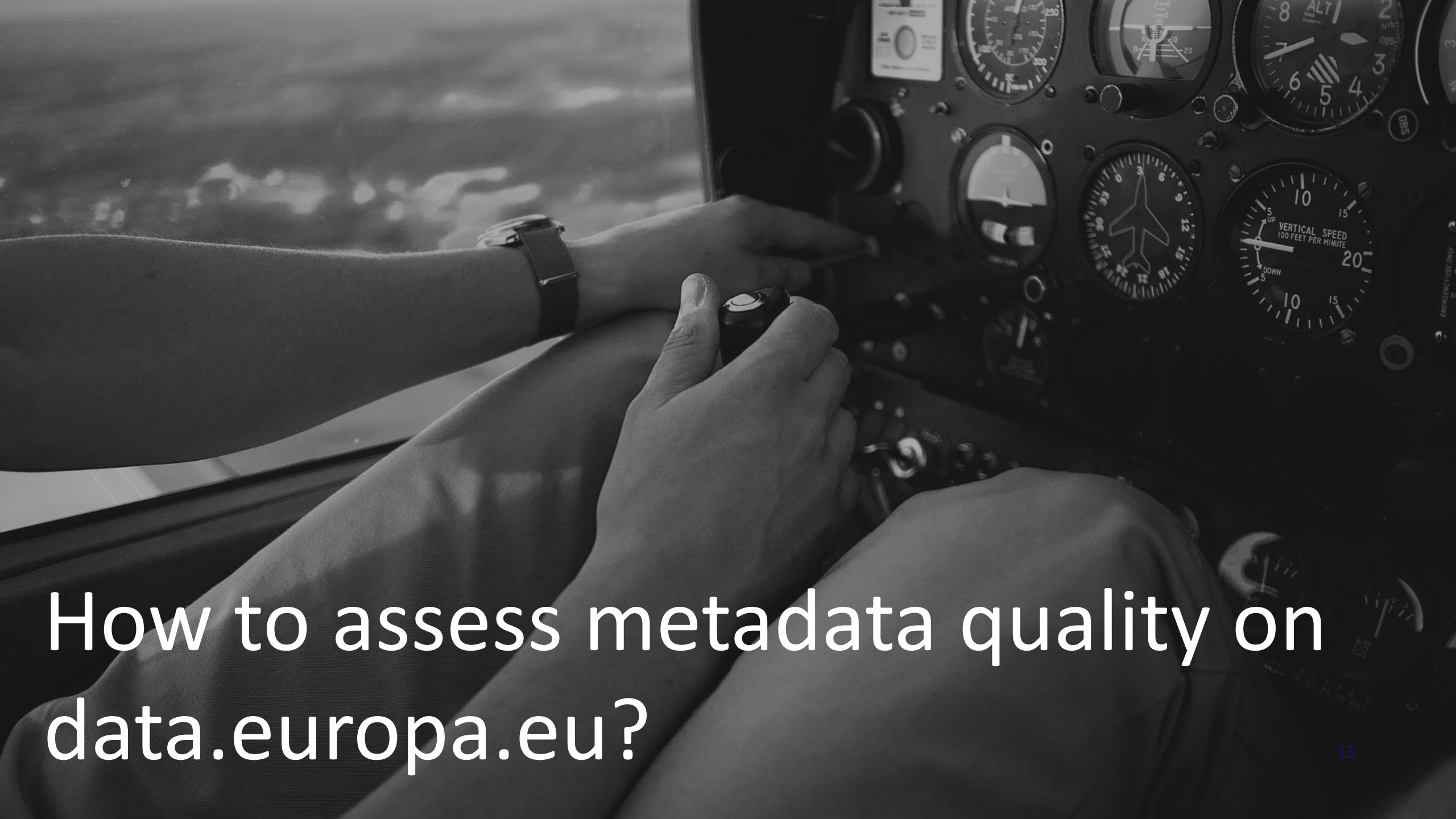
Acquire competencies

Data preparation process



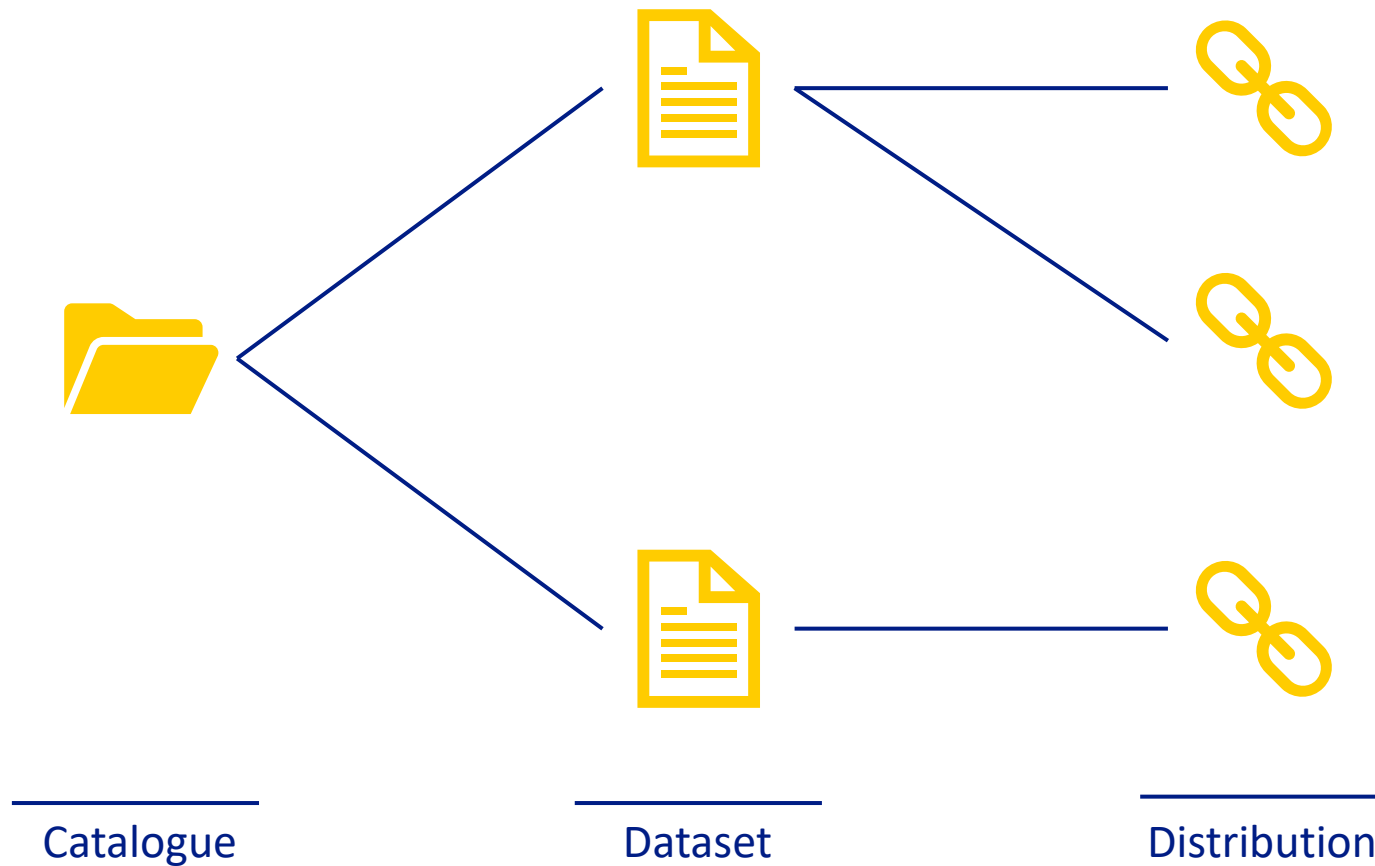


Questions?



How to assess metadata quality on data.europa.eu?

DCAT-AP introduction



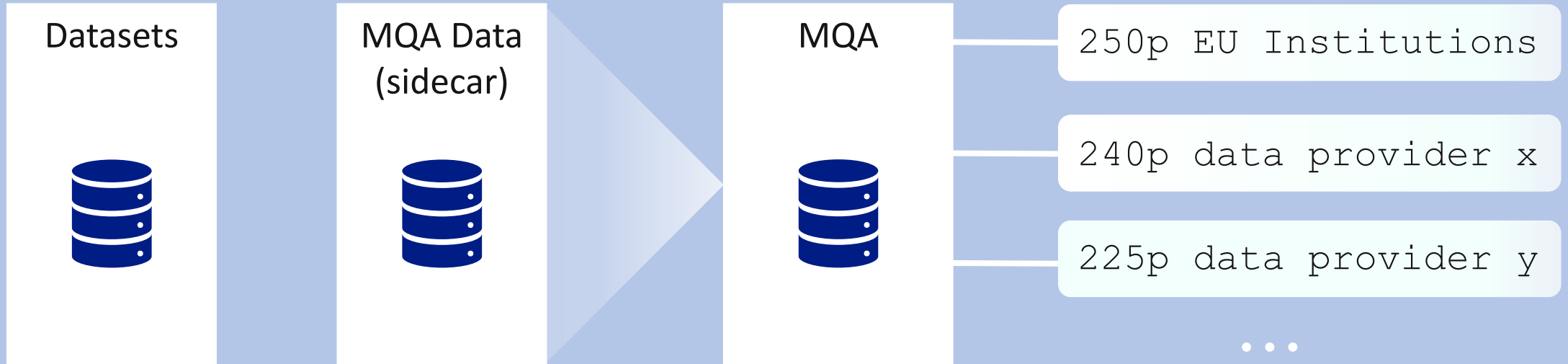


Dataset

MQA Data

How the scoring works

data.europa.eu



The “Metadata Quality Assurance” section

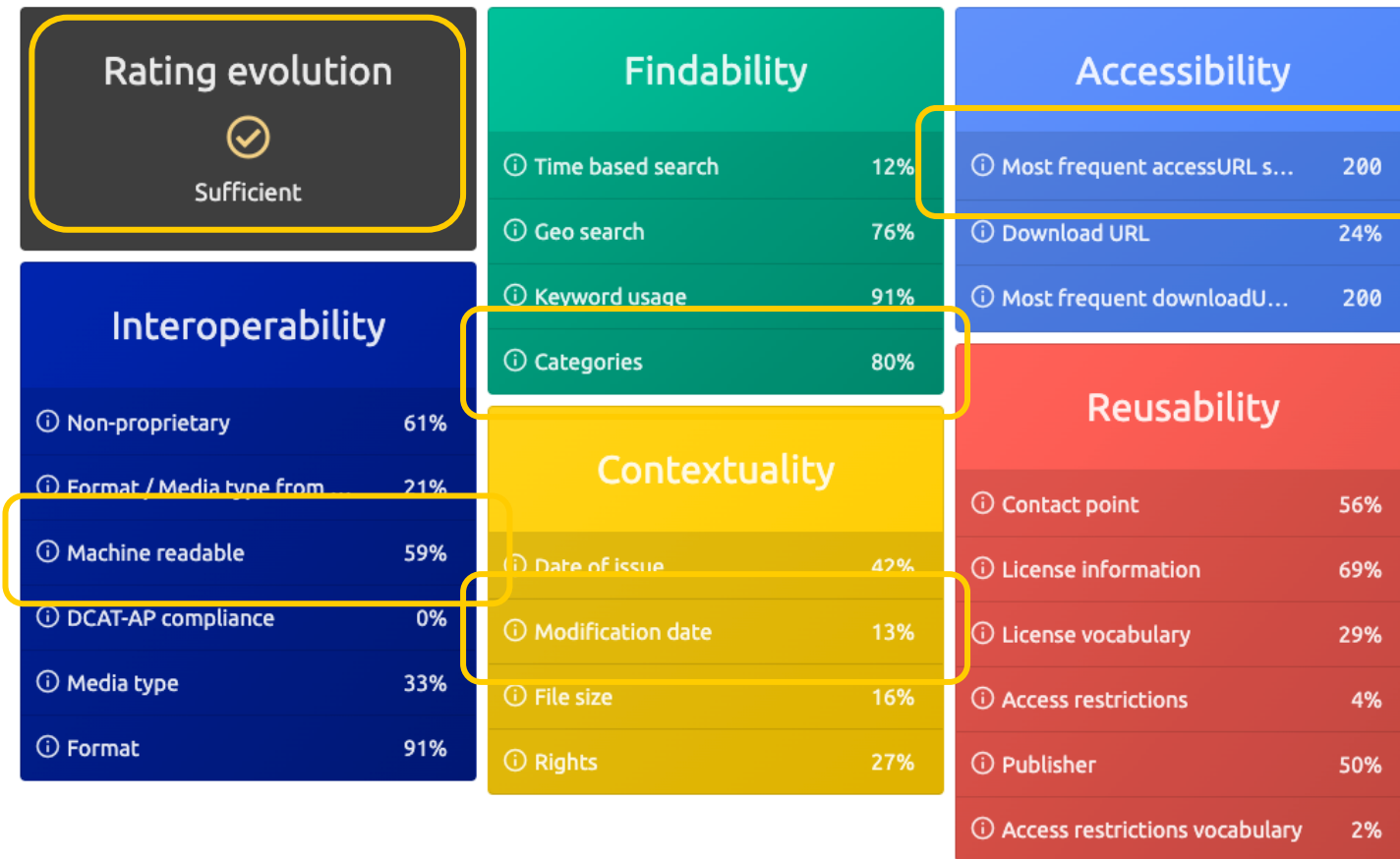
The screenshot shows the homepage of data.europa.eu. The navigation bar includes 'Data', 'Studies', 'data.europa academy', 'News', and 'Contact'. Below the navigation bar, there are links for 'Databases', 'SPARQL Search', 'Statistics', 'Metadata Quality', 'Persistent URIs', and 'EU Vocabularies'. A yellow arrow points to the 'Metadata Quality' link. The main content area features a search bar, a grid of category icons (Economy & Finance, Education, Culture & Sport, Energy, Environment, Government & Public Sector, Health, International Issues, Justice, Legal System & Public Safety, Population & Society, Regions & Cities, Science & Technology, Transport), and a 'Trending datasets' section. At the bottom, there are sections for 'Open data news' and 'Studies'.

The screenshot shows the 'Dimensions' page on data.europa.eu. The page title is 'Dimensions' with a sub-heading 'Overview'. The main content is a grid of metrics for Metadata Quality Assurance. A yellow box highlights the 'Dashboard', 'Catalogues', and 'Methodology' buttons at the top, and the 'Download as report' button. Another yellow box highlights the entire grid of metrics.

The Metrics Grid:

Rating evolution Sufficient	Findability <ul style="list-style-type: none"> Time based search: 12% Geo search: 76% Keyword usage: 91% Categories: 80% 	Accessibility <ul style="list-style-type: none"> Most frequent accessURL s...: 200 Download URL: 24% Most frequent downloadU...: 200
Interoperability <ul style="list-style-type: none"> Non-proprietary: 61% Format / Media type from ...: 21% Machine readable: 59% DCAT-AP compliance: 0% Media type: 33% 	Contextuality <ul style="list-style-type: none"> Date of issue: 42% Modification date: 13% File size: 16% 	Reusability <ul style="list-style-type: none"> Contact point: 56% License information: 69% License vocabulary: 29% Access restrictions: 4%

The “Metadata Quality” section



The “Metadata Quality” section

The screenshot shows the 'Catalogues' section of the data.europa.eu website. A search bar contains the text 'safety'. Below the search bar is a table with columns for Country, Name, and Description. The 'European Maritime Safety Agency (EUROPE)' entry is highlighted with a yellow box. The footer includes a newsletter subscription form and social media links.

Country	Name	Description
🇪🇺	Directorate-General for Health and Food Safety (EUROPE)	Directorate-General for Health and Food Safety
🇪🇺	European Agency for Safety and Health at Work (EUROPE)	European Agency for Safety and Health at Work
🇪🇺	European Food Safety Authority (EUROPE)	European Food Safety Authority
🇪🇺	European Maritime Safety Agency (EUROPE)	European Maritime Safety Agency
🇪🇺	European Union Aviation Safety Agency (EUROPE)	European Union Aviation Safety Agency

The screenshot shows the 'European Food Safety Authority' metadata quality dashboard. It features a search bar, a 'Download as report' button, and a table of metadata quality metrics. The 'European Food Safety Authority' entry is highlighted with a yellow box. The dashboard is divided into several sections: Rating evolution, Findability, Accessibility, Interoperability, Contextuality, and Reusability.

Section	Metric	Value	
Rating evolution	Good	Good	
	Interoperability	50%	
Findability	Time based search	37%	
	Geo search	26%	
	Keyword usage	100%	
	Categories	100%	
Accessibility	Most frequent accessURL s...	200	
	Download URL	98%	
	Most frequent downloadU...	200	
Contextuality	Date of issue	100%	
	Modification date	0%	
	File size	0%	
	Rights	0%	
	Reusability	Contact point	100%
		License information	100%
License vocabulary		100%	
Access restrictions		100%	
Publisher		100%	

Metadata quality of single datasets #1

The screenshot shows the 'data.europa.eu' website. The navigation bar includes 'Data', 'Studies', 'data.europa academy', 'News', and 'Contact'. Below the navigation bar, there are links for 'Dataset', 'Categories', 'Similar Datasets', and 'Quality'. A yellow arrow points to the 'Quality' link. The main content area displays the title 'Consolidated list of persons, groups and entities subject to EU financial sanctions' and includes a 'Quality' tab.

The screenshot shows the 'data.europa.eu' website with the 'Metadata Quality' section. The section is titled 'Metadata Quality' and includes a description: 'The Metadata Quality Assurance is intended to help data providers and data portals to check their metadata against various indicators. For information on which metrics we use for indicator measurements, please have a look at our methodology page (make a link to the methodology page)'. Below the description, there are five quality indicators: Accessibility, Contextuality, Reusability, Findability, and Interoperability, each with a table of metrics and values.

Accessibility		
Download URL		78 %
Most frequent accessURL status codes	200: 89 % 404: 11 %	
Most frequent downloadURL status codes	200: 86 % 404: 14 %	86 %

Contextuality	
File size	0 %
Rights	89 %
Datasets: Modification date	yes
Datasets: Date of issue	yes
Distributions: Modification date	11 %
Distributions: Date of issue	22 %

Reusability	
Access restrictions	yes
License information	100 %
Access restrictions vocabulary	yes
Contact point	yes
Publisher	yes

Findability	
Keyword usage	yes
Categories	yes
Geo search	no
Time based search	no

Interoperability	
DCAT-AP compliance	no

Metadata quality of single datasets #2

<input checked="" type="radio"/> Media type	89 %
<input checked="" type="radio"/> Format / Media type from vocabulary	89 %

Distribution Quality

The following lists the quality measurement of all distributions of the dataset. For information on which metrics we use for indicator measurements, please have a look at our methodology page (make a link to the methodology page).

Consolidated Financial Sanctions In PDF Format			
Consolidated Financial Sanctions File 1.0			
Consolidated Financial Sanctions File 1.1			
Consolidated Financial Sanctions File 1.1			
Financial Sanctions Files (FSF) website			
Accessibility			
Download URL	yes	DownloadURL status code	404
AccessURL status code	404		
Reusability			
License Information	yes		
Contextuality			
File size	no	Modification date	no
Rights	yes	Date of issue	no
Interoperability			
Format	yes	Format / Media type from vocabulary	yes
Media type	yes		
Sanctions List			



Questions?

A black and white photograph of a coffee machine dispensing coffee into two cups. The coffee is being poured from a spout into a cup in the foreground, with another cup visible behind it. The machine's handle and spout are in the upper left, and the coffee is being poured into the cups. The background is blurred, showing the rest of the machine and some indistinct shapes.

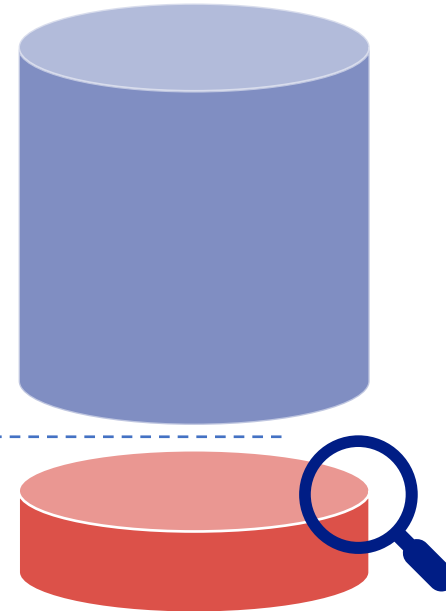
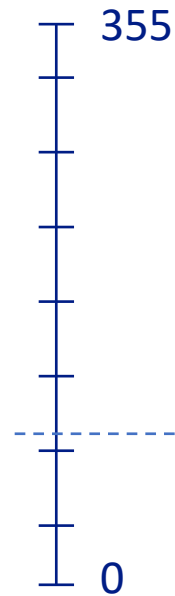
Break

Please be back
in 5 minutes!



How to improve
data quality?

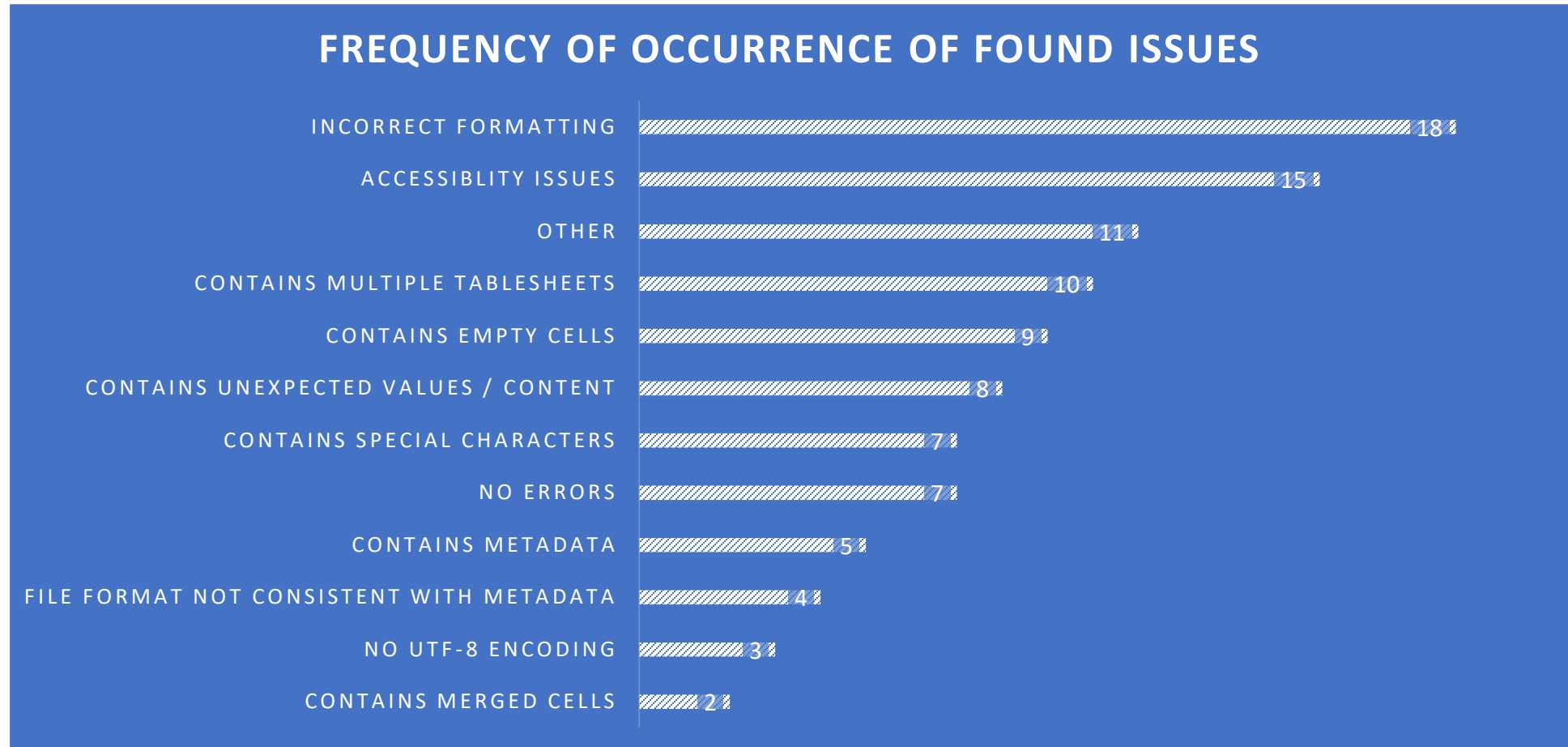
Identifying data for analysis



Automated metadata analysis

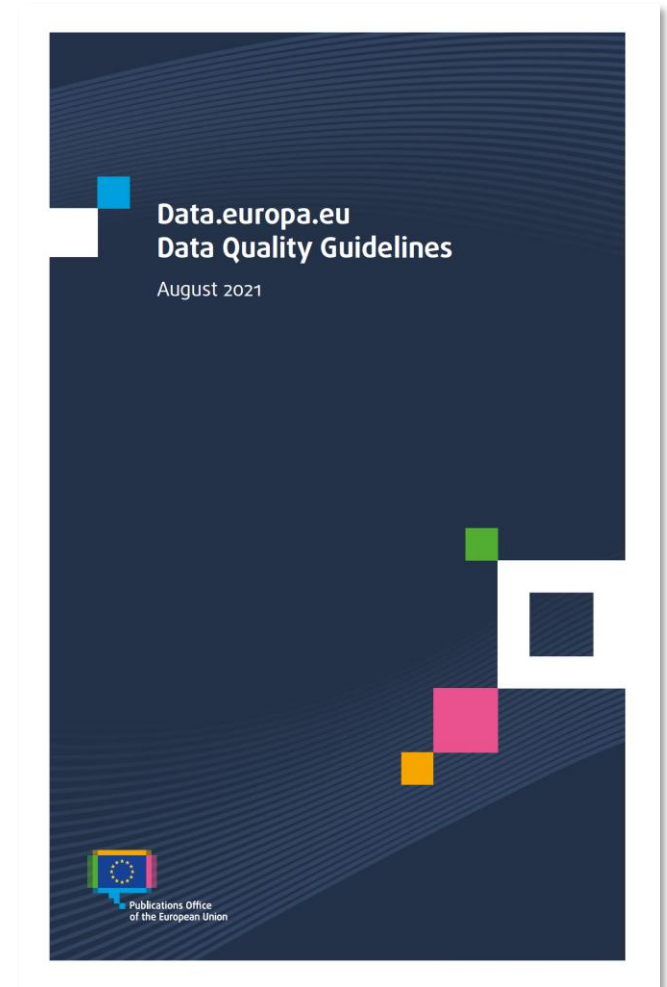
Manual data analysis

Frequency of occurrence of found issues



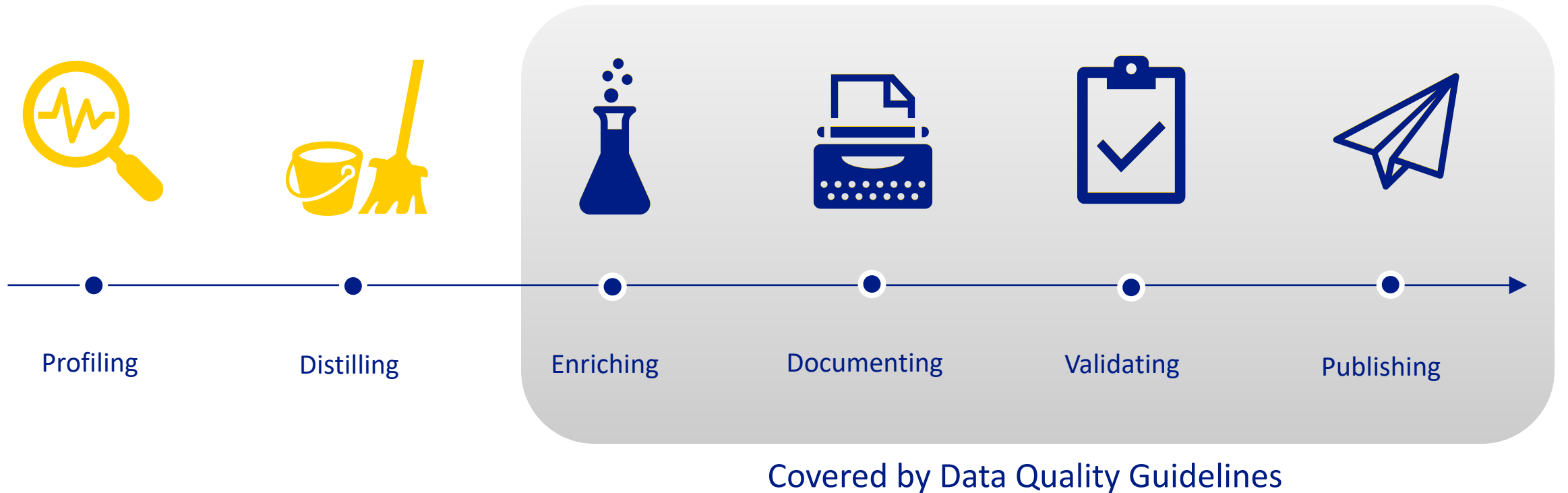
Data Quality Guidelines

- Project launched by the Publications Office of the European Union in 2019
- Aim: analysing major quality issues and providing a set of recommendations for data providers from the EU and its Member states.
- Consisted of three parts:
 - Data analysis for identifying most common quality issues
 - Identification of data quality dimensions, indicators and metrics
 - **Recommendations for delivering high-quality data**
- For any feedback regarding the publication please contact: OP-DATA-EUROPA-EU@publications.europa.eu



Download: <https://op.europa.eu/en/publication-detail/-/publication/023ce8e4-50c8-11ec-91ac-01aa75ed71a1/language-en>

Data preparation process



Findability



The findability of a data set depends on the metadata:

the better the data is described through metadata, the easier it is for users to find the data.

How to improve the quality:

- Choose precise title
- Assign keywords
- Assign categories
- Add temporal information
- Add spatial information
- Add description

Findability - Choose a precise title

The image shows three overlapping data cards from the Data.europa.eu portal, illustrating the importance of precise titles for findability. Each card is highlighted with a colored border and a circular icon indicating its status.


- Card 1 (Red border, red X icon):** Title: "CPI all households (elaborate)". Description: "Consumer price indices all households... Consumer goods 1996 - 2002; January... Changed on September 09 2003. F".
- Card 2 (Red border, red X icon):** Title: "Consumer price index". Description: "Consumer price index". Updated: 26.01.2022.
- Card 3 (Green border, green checkmark icon):** Title: "Consumer Price Index Hamburg June 2019". Description: "No description available". Download options: Excel XL, ZIP, PDF. Updated: 29.01.2022 08:46. Created: 18.07.2019 02:00. Source: GovData (Germany).

Findability - Provide detailed description

Consumer Price Index Hamburg June 2019

No description available


Updated: 29.01.2022 08:46 Created:



Consumer price index

Consumer price index


Updated: 26.01.2022




CPI all households (elaborate)

Consumer price indices all households. Index figures
Consumer goods 1996 - 2002; January 1996 - December 2002
Changed on September 09 2003. Frequency: Discontinued.

[JSON](#) [Atom Fee](#)

 Dataportaal van de Nederlandse overheid



Findability - assign categories

Categories ②	
Environment	262 701
Agriculture, fisheries, fores...	260 718
Justice, legal system and p...	183 495
Science and technology	72 837
Government and public sector	72 136
Economy and finance	63 879
Population and society	48 283

262 701 datasets found

Categories: **Environment** x

Air data Western Esplanaden

Continuous air measurements are carried out in central Umeå at Västra Esplanaden. The data set shown here shows the measurements available for nitrogen dioxide (NO₂) and Particles (PM₁₀) on Västra Esplanaden. New data is collected every fifteen minutes, but may be missing if the measuring station for some reason does not work for a period. What is shown is...

applicat text/csv

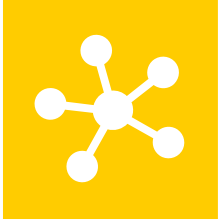
Updated: 08.02.2022 02:01 Created: 08.02.2022 02:01

 Sveriges dataportal

Findability - MQA Scoring

Indicator	Metric	Weight
Keyword usage	Check whether keywords are defined	30
Data themes	Check whether categories are assigned	30
Geo search	Check whether property is set or not	20
Time based search	Check whether property is set or not	20
Total		100

Interoperability



Data is interoperable if it can be easily retrieved, processed, re-used and operated by others.

How to improve the quality:

- Make data DCAT-AP compliant
- Consider standards e.g. for date, times and formats
- Consider conformity of licenses
- Use a machine readable and open format
- Use controlled vocabularies

Interoperability - consider standards I

Statistical data from ciclosurers from 1.1.2018

Statistical data from ciclosurers from 1.1.2010. CSV UNKNOWN ZIP

Cze

Tropical Cyclone Dorian in Bahamas (2019-09-01)

Activation time (UTC): 2019-09-01 08:12:00 Event time (UTC): 2019-09-01 18:00:00 Event type: Storm (Tropical cyclone, hurricane, typhoon) Activation reason: Tropical Cyclone Dorian is set to make landfall in the northern Bahamas with sustained winds of up to 240 km/h and a storm surge of over 5 metres...

Updated: 01.09.2019 Created: 01.09.2019 Esri Sha Joint Research Centre

Interoperability - consider standards II

Formats [Ⓢ]	
CSV	193 242
WMS	141 217
WFS	134 227
JSON	94 552
HTML	69 838
GML	63 848
Excel XLSX	60 392
Esri Shape	57 739
ZIP	53 955

PDF	34 423
Plain text	31 743
XML	25 393
KML	14 011
WMS	13 894
GeoJSON	12 049
TSV	11 584
WFS	9 960
TIFF	7 452

RDF XML	2 137
SERVICE	1 903
.xlsx	1 831
download	1 710
xlsx	1 566
view	1 502
web page	1 452
OWL	1 362
arcgis geoservices rest api	1 149

Interoperability - use open and machine readable formats



Austria's National Air Emission Projections 2021 for 2020, 2025 and 2030.

This report covers the results for projections of the air pollutants sulphur dioxide (SO₂), nitrogen dioxide (NO_x), non-methane volatile organic compounds (NMVOCs), ammonia (NH₃) and particulate matter (PM_{2.5}) under the scenarios "with existing measures" (WEM) and "with additional measures..."

PDF

Updated: 02.02.2022 08:56 Created: 21.09.2021 12:35



Cosmetic ingredient database (Cosing) - List of substances prohibited in cosmetic products


List of substances prohibited in cosmetic products from Annex II of the Regulation (EC) No 1223/2009 of the European Parliament and of the Council as amended. The list contains the substance identification (Chemical name/INN, CAS Number and EC number) of each prohibited substance. The...

Excel XL CSV HTML

PDF

Updated:
14.12.2018

Created:
13.01.2016

 Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs

Interoperability - MQA Scoring

Indicator	Metric	Weight
Format	Check whether property is set or not	20
Media type	Check whether property is set or not	10
Format / Media type from vocabulary	The media type is checked against a list of controlled vocabulary	10
Non-proprietary	The format is checked against a list of non-proprietary formats	20
Machine readable	The format is checked against a list of machine readable formats	20
DCAT-AP compliance	The metadata is validated against a set of SHACL shapes	30
Total		110



Questions?

Summary

- Introduction to data quality
 - Dimensions and indicators
 - Quality concerns data and metadata and is subjective
 - Data preparation process
- How metadata quality can be determined on data.europa.eu
 - How the scoring works
 - Where to find the metadata quality section
- How to improve data quality
 - Introduction of data quality guidelines

THANK YOU

**One word or line
on your
experience today**

on [sli.do](#)

Please
provide us
your
feedback!





data.
europa.
eu
academy 

Thank you very much!

OP-DATA-EUROPA-EU@publications.europa.eu



data.europa.eu The official portal
for European data

