Improving data publishing by open data portal managers and owners

data.europa.eu
The official portal for European data

This study has been prepared as part of data.europa.eu. Data.europa.eu is an initiative of the European Commission. The Publications Office of the European Union is responsible for contract management of data.europa.eu.

For more information about this paper, please contact:

European Commission
Directorate-General for Communications Networks, Content and Technology
Unit G.1 Data Policy and Innovation
Email: CNECT-G1@ec.europa.eu

data.europa.eu
Email: info@data.europa.eu

Written by:
Dr. Michael Stemmer (Fraunhofer FOKUS)

Supported by:
Lina Bruns (Fraunhofer FOKUS)
Benjamin Dittwald (Fraunhofer FOKUS)
Merle Heider (Fraunhofer FOKUS)
Leonard Mack (Fraunhofer FOKUS)

Last update: 20 July 2023
https://data.europa.eu/

Disclaimer
By the European Commission, Directorate-General of Communications Networks, Content and Technology. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission’s behalf may be held responsible for the use which may be made of the information contained therein.

Luxembourg: Publications Office of the European Union, 2022

ISBN: 978-92-78-43683-4 doi: 10.2830/626320 OA-07-23-291-EN-N

Improving data publishing by open data portal managers and owners

CONTENTS

1. Introduction

Publishing high-quality data in a simple way is crucial for open data to create added value. The design of the data-publishing process is particularly important. On the one hand, this refers to the internal data publication processes within data-providing organisations (data provider). On the other hand, managers and owners [1] of open data portals can also positively influence the data-publishing process through various mechanisms, thus contributing to better data quality, easier data publication for data providers and a wider visibility of published data. This report is intended to address the latter perspective. We assume that open data portal managers and owners can make a significant contribution to the use and dissemination of open data within the following three areas of data publishing.

1. **Data and metadata quality.** Open data is much more useful if it is of high quality. High quality is important for metadata and data. High-quality metadata ensures that the data can be found. The better the quality of the data, the easier it will be for users to reuse the data. Open data portal managers and owners can offer tools, such as the metadata quality assurance service (MQA) from data.europa.eu, to reflect the quality of the (meta)data published directly back to the providers within the publishing process and thus provide valuable information for optimising quality.
2. **Methods, targets and processes.** The design of the publishing process offered by the portal influences how easily and efficiently data providers can publish their data. For example, the publishing process becomes more comfortable for data providers if they can publish their metadata automatically via application programming interfaces (APIs). Furthermore, open data portal managers and owners can also push the visibility of the published data if they have processes to republish the data on different targets.
3. **Feedback and improvement process.** Open data portal managers and owners can also contribute to improving the use of open data by continuously improving the data-publication process (e.g. by collecting feedback from data providers and data users and responding to their demand).

data.europa.eu currently harvests data from 176 catalogues, including national and regional open data portals and geoportals from the EU Member States and beyond. We assume that the respective publication process differs between these catalogues and that there are different mechanisms on how open data portal managers and owners address the beforementioned aspects.

The aim of this report is to investigate how data is made available in the different portals, what pain points exist in their publication processes, and what suggestions for improvement portal managers and owners can offer. Based on these findings, specific and harmonised recommendations for the general publication process have been formulated. Accordingly, the report thus shares knowledge and lessons learned that are rooted in the everyday practical experiences of open data portals.

The structure of the report is rather simple: after a short introduction, the methodological approach will be explained. In the following section, the key analysis results for each of the three areas will be introduced and improvement potential for the data-preparation process will be presented. After a short conclusion the report ends with an annex, containing in-depth results from the survey and the introduction and questions of the online survey.

[1] By ‘managers and owners of open data portals’ we mean anyone who collects and publishes data in a systematic way.

2. Methodology

The starting point for the report is the examination of the status quo of the data-publishing process in the various portals. For the collection of this data, two different methods were considered, both with their own advantages and disadvantages: an online survey and conducting interviews. The following table compares both methods.

	Online survey	Interviews
Target group	All open data portal managers and owners, which are harvested by data.europa.eu.	Selected open data portal managers and owners.
Analysis	Quantitative analysis of the results.	Qualitative analysis of the results.
Advantage	Less time-consuming and produces a representative database as all open data portal managers and owners can be invited to take part in the survey.	Deep dive and ad hoc follow-up questions possible.
Disadvantage	No deep dive and follow-up questions possible.	Time-consuming method, only a limited number of open data portal managers and owners can be interviewed, no representative database.
Risk	Online surveys always come with the risk of low response rates.	The findings are based on a non-representative database and might not mirror the reality in an adequate way.

We chose to combine both approaches. The online survey [2] was supplemented by a few additional interviews with open data portal managers and owners that gave interesting answers in the survey to gain more knowledge from them. For this purpose, we gave open data portal managers and owners the opportunity to give us the URL of their portal and their contact data with the survey. With this they had the possibility to choose between answering the survey anonymously or potentially being contacted by us for an additional expert interview.

We invited the portal managers and owners of all open data portals being harvested by data.europa.eu to take part in the survey. It was open from 17 November to 16 December 2022, and during this time frame we received 20 participants. From these, 12 stemmed from national portals, 7 from regional portals and 1 from a project-specific open data portal. 13 participants answered the survey in part and 7 completed it. From the 7 complete participations we selected 3 for additional interviews. The interviews took 20–30 minutes each and were conducted online in January 2023.

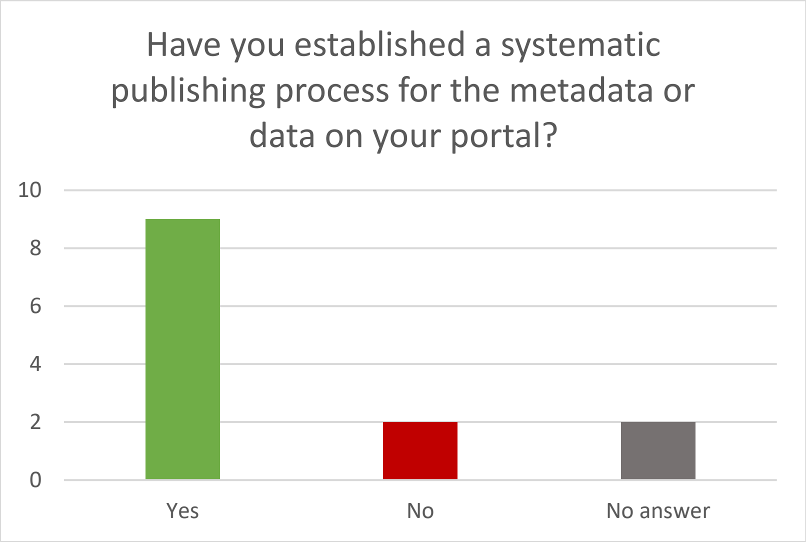
[2] Introduction and questions of the online survey can be found in Annex B.

3. Survey results

3.1. Overview of the publishing process and publishing mechanisms

The design of the publishing process offered by a portal influences how easily and efficiently data providers can publish their data. Thus, we subsequently look at the publishing processes, methods and targets of open data portals.

The following diagram shows that many portal managers and owners have established a systematic publishing process [3].



Publishing process and mechanisms

The process of publishing data varies with the different open data portal managers and owners. Some of them have no formalised publication process in place, others have established technical guidelines and/or make the publication process part of a quality management system.

The following steps show a typical example from one of the participants of our survey.

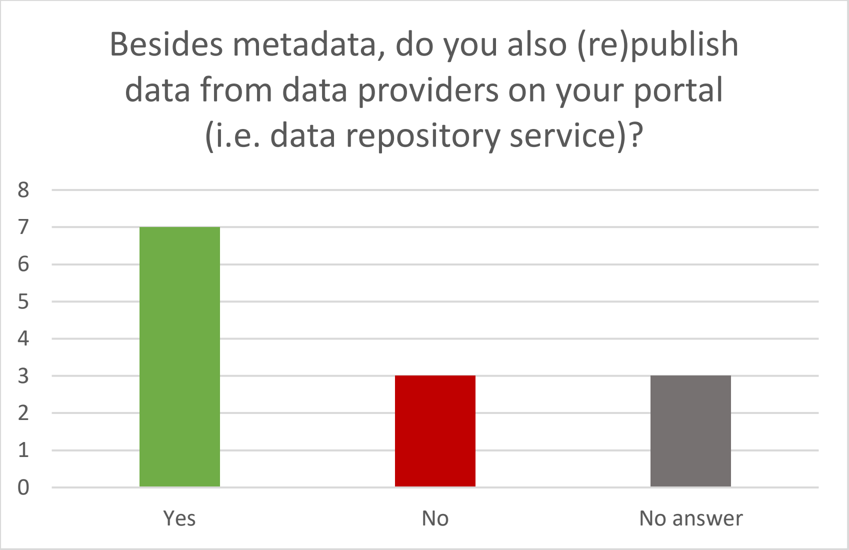
- Fill in the metadata (according to the Data Catalog Vocabulary application profile (DCAT-AP); main fields must be filled in).
- Choose the licence.
- Add the dataset (for example using an API, adding the file or a link).
- Publish the dataset.

Within the publication process, different mechanisms of publishing open data are used. The following were reported in our survey from free text answers:

- harvesting
- manually via Web UI
- push via API
- XML (Extensible Markup Language) (automated import).

[3] Out of all the valid participations in our survey (n=20), only seven completed every question. In some cases, participants were not shown certain questions because of their previous answers. Therefore, in this (n=13) and all following diagrams 'n' might be lower than '20' due to certain participants only completing some parts of the questions.

Regarding (re)publishing data and metadata, we received the following additional answers, as shown in the diagram below.



3.2. Overview of the pain points and room for improvement

Top quality issues

In our survey we asked about the most important quality issues relating to data publishing and metadata publishing. The detailed answers can be found in Annex A. Below we have clustered these answers to extract the main topics.

Regarding data publishing, the following pain points were identified from summarising and clustering the free text answers from the survey.

- Insufficient data quality:
  - data not up to date;
  - data not machine-readable;
  - no time series in the data sources;
  - no time stamps available;
  - no unique identifiers in the datasets;
  - broken links.
- Lack of standardisation and interoperability:
  - lack of data standardisation;
  - transition between International Organization for Standardization (ISO) 19139:2007 and DCAT;
  - putting different logic in datasets with same standard;
  - different title for same content data;
  - different formats and/or service types;
  - regional v national data;
  - limited formats for accessibility.

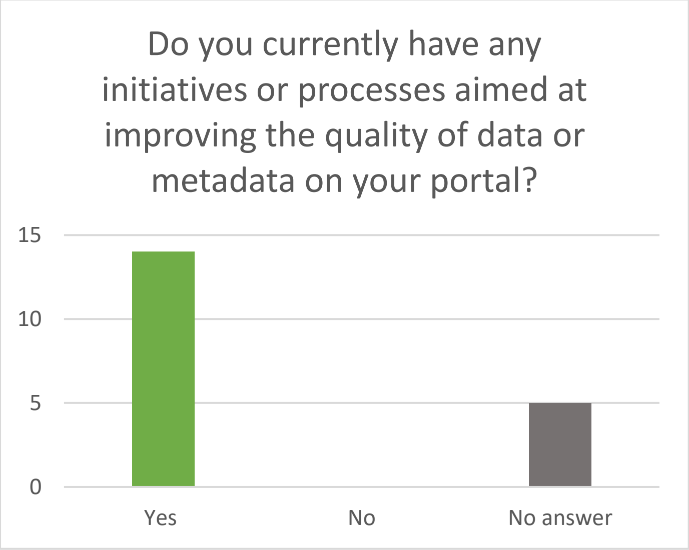
In addition to this, complex licensing and the fulfilment of different EU regulations in an 'Infrastructure for spatial information in Europe' (INSPIRE) context were identified as further pain points.

Regarding metadata publishing, the following pain points were identified from summarising and clustering the free text answers from the survey.

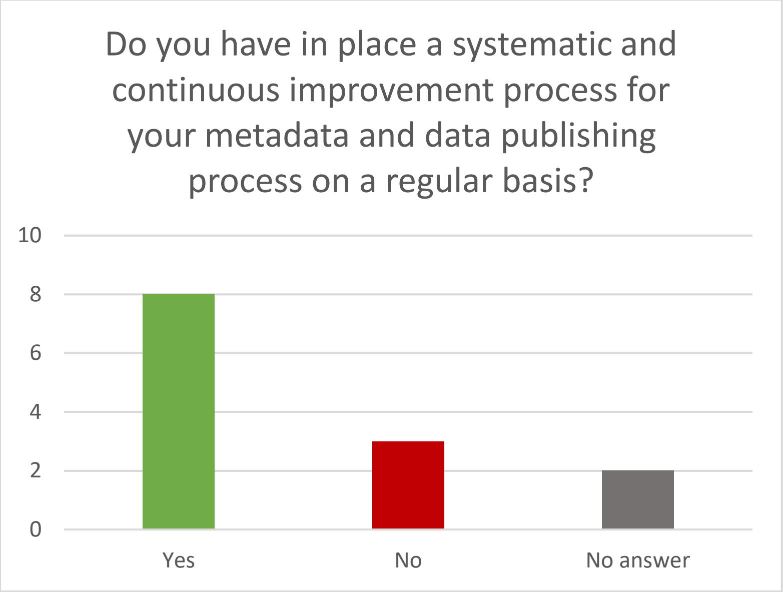
- Lack of standardisation:
  - no common standard for metadata;
  - use of very specific terminology;
  - not following the DCAT-AP controlled vocabulary.
- Lack of metadata quality:
  - not always comprehensive / precise / up to date;
  - misuse of description fields to provide narrative;
  - free text keywords;
  - missing, poor or incorrect licence choices and/or information;
  - missing feature type description;
  - quality of metadata harvesting from geoportal to national open data portal.
  -

Improving quality of data and metadata

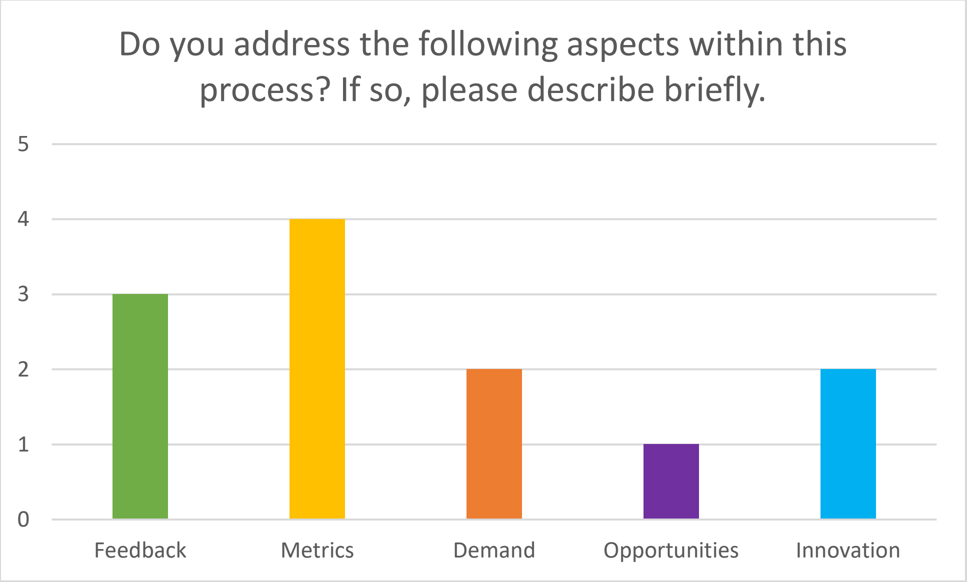
To overcome the stated quality issues and pain points regarding data and metadata publishing, we asked the portal managers and owners whether they had corresponding improving initiatives or processes. As the following diagram shows, the majority had such initiatives or processes in place.



As the next diagram shows, most of the portal managers and owners who have an improvement initiative or process in place state that this is implemented in form of a systematic and continuous improvement process.



Within the improvement processes, the following diagram shows which aspects are addressed.



A list of examples, we identified from the free text answers, can be found in Annex A. In the following section we give some details on what survey participants considered as improvement initiatives.

Best practices

Regarding the different aspects of improvement within the survey, the following examples of best practices were reported in more detail.

- Best practice example 1 (feedback).

  Metadata and data are monitored by the administrator and relevant data providers. In addition, information from portal users is used, transmitted via the ‘Send feedback’ button. The portal allows users to report their needs via the ‘Describe the data you have not found on the portal’ section. A questionnaire on the open data portal is conducted to help improve its content for public needs and increase the number of datasets that will be attractive for reuse. All responses and other information on this issue received from users are monitored and analysed by the administrator and open data team, and feedback is then passed onto data providers. An extensive analysis of the questionnaires received was provided in the diagnostic section of the 2021–2027 open data programme.

- Best practice example 2 (metrics).

  Statistics of the open data portal are analysed. Based on statistics, cooperation with open data officers is carried out to increase the level of data openness.

- Best practice example 3 (demand).

  The portal allows users to report their needs via the ‘Describe the data you have not found on portal’ section.

- Best practice example 4 (innovation).

  It is possible to submit a showcase example of an application using open data via a form. A section of the open data portal is dedicated to presenting submitted showcases.

Another general and detailed best practice example (best practice example 5), together with a list of tools for improvement identified by free text answers from the survey, can be found in Annex A.

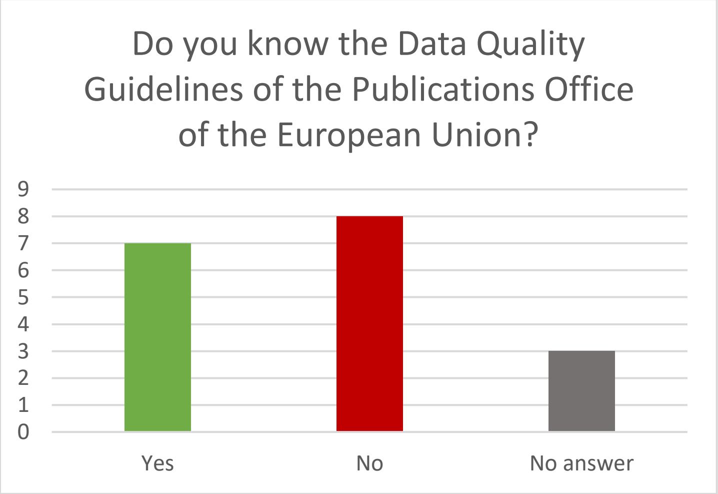
Experience with data quality guidelines and metadata quality assurance service

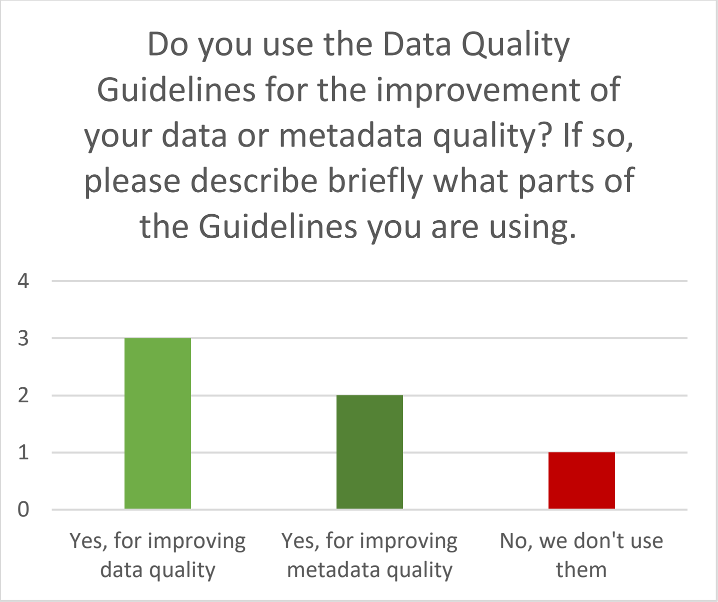
The following tools, provided by the Publications Office of the European Union and data.europa.eu, aiming at improving the quality of data and metadata publishing are already available to use:

- Data Quality Guidelines (DQG);
- Metadata Quality Assurance service (MQA).

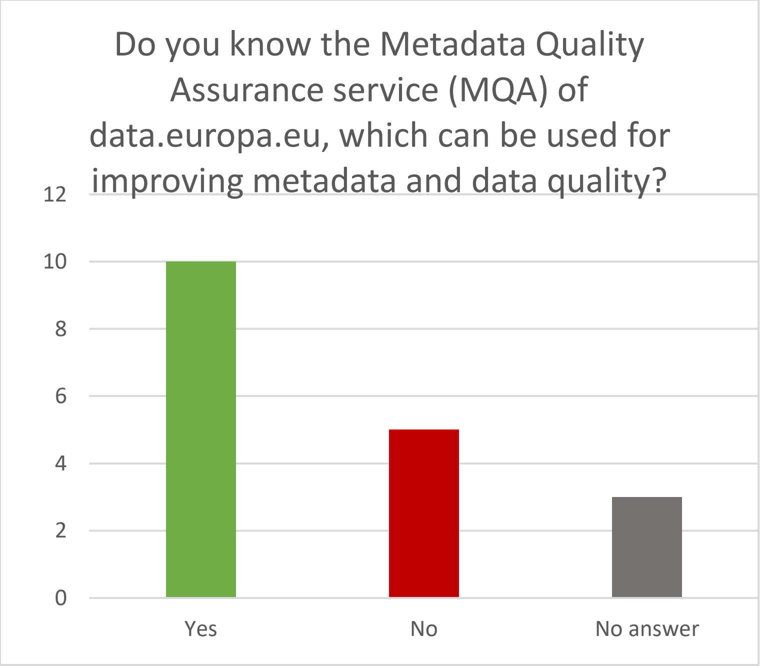
We asked if these were known, used and what experience users had with them.

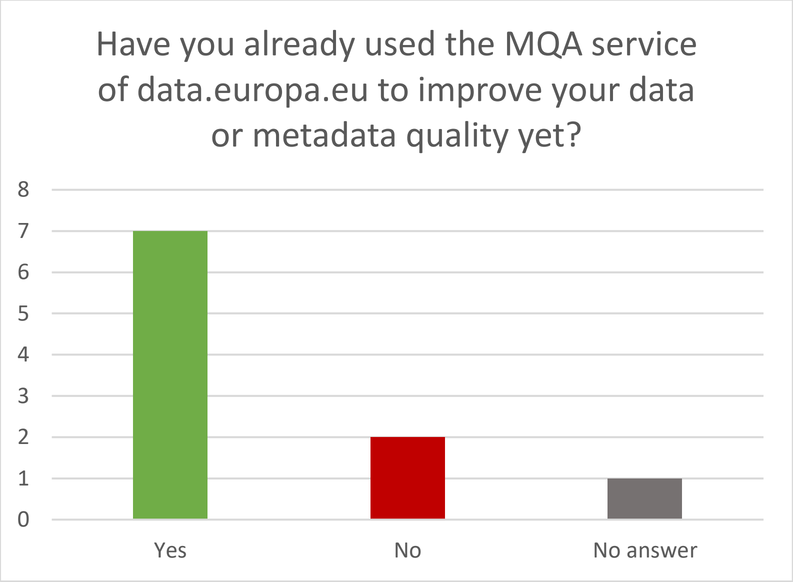
As the following two diagrams show, the data quality guidelines are known to nearly half of the answering participants, and if they are known, they are mostly used for improving data and metadata quality.





The MQA is known to most of the answering participants, and has, in most cases, also been used by those familiar with it to improve the quality of metadata.





Regarding experience with the MQA, detailed feedback for improvement reported in free text answers from the survey can be found in Annex A.

Obstacles in improving data and metadata quality

Regarding improvement initiatives and processes, we also asked about the biggest obstacles for improving the data and metadata quality of datasets. The detailed answers can be found in Annex A. Below we have clustered the answers to extract the main topics.

As one might expect, the biggest data and metadata quality issues, already identified above, turned out to be the same primary obstacles for improvement. Further additional pain points were identified from summarising and clustering the free text answers from the survey.

- Lack of information and knowledge:
  - very different level of knowledge on data;
  - insufficient knowledge and skills of data providers;
  - insufficient information from data providers;
  - lack of awareness of importance of metadata;
  - lack of technical skills;
  - misinterpretations;
  - poor guidelines (INSPIRE, high-value datasets (HVD)).
- Lack of resources:
  - not enough resources for data management / data governance;
  - staff cost of data providers.

In addition to this, complex schema, federalism and software limitations were also identified as further obstacles.

Suggestions for improving data and metadata quality

Finally, we asked for suggestions on how to improve data and metadata quality. The detailed answers can be found in Annex A. Below we have clustered the answers to extract the main topics.

The following suggestions were identified from summarising and clustering free text answers from the survey and answers from the additional interviews.

- Management and methods:
    - get wholistic view on open data / overall data strategy;
    - use of linked data;
    - find and eliminate duplicates from different national portals;
    - report errors to data providers within quality management process
    - concentrate on valuable and primary datasets;
    - check metadata for correctness;
    - use a common, widely accepted schema;
    - give better feedback for debugging the process;
    - appoint openness officers responsible for data/metadata;
    - make legal regulations specifying the quality of data/metadata.
- Knowledge and information:
    - increase knowledge (e.g. through communication, workshops, meetings);
    - establish a strategic level of improvement (e.g. through informing, events, training);
    - conduct training for data providers;
    - prepare video guidelines, online courses;
    - establish a national community.
- Harmonisation and standardisation:
    - follow standards and ontologies;
    - use common formats and services;
    - follow vocabularies such as DCAT or DCAT-AP;
    - reject data and metadata that is not compliant to quality standards;
    - prepare standards and guidelines;
    - introduce a common metadata standard;
    - foster harmonisation in an open government data (OGD) / INSPIRE context.
- Automations and tools:
    - use stable, well-supported software;
    - use common software for metadata management;
    - provide tools and mechanisms to ensure data/metadata quality verification;
    - use improved schema translations;
    - improve the MQA tool;
    - provide resources;
    - provide reference tools (open source);
    - automate process (such as data.europa.eu).

3.3. Specific and harmonised recommendations

With the answers, current situation, pain points, best practice examples and suggestions given in the survey, we recommend open data portal managers and owners above all to focus on the following areas to improve data publishing.

Metadata management

Metadata is crucial but still a pain point for data portal managers and owners. We recommend setting up a (total) quality management approach, if not already in place. This can help improve the quality of metadata along the whole value chain, eventually increasing the findability, accessibility, interoperability, reusability and contextuality of the data provided. A quality management approach should consist of

- strategic measures, such as quality objectives, feedback mechanisms, continuous improvement and a systematic process;
- tactical measures, such as like monitoring and analysing the quality of metadata; and
- operational measures under the form of regularly quality checks and tests, from validation and verification up to the rejection of non-compliant datasets.

In addition to metadata, data is also of interest, but data quality lies more in the hands of data providers than data portal managers and owners. Nevertheless, appropriate feedback mechanisms for data quality could be useful.

Specific recommendations include the following.

- Focus on primary datasets that have a high potential to add value when being reused.
- Appoint an open data officer responsible for data/metadata.
- Consider using the Semantic Web and linked data.
- Set up an internal data management pipeline to ensure compliance before publication.
- Be as close to the target data schema as possible.

Knowledge and information

Knowledge is essential, especially when relating to metadata. We recommend setting up initiatives and/or measures addressing information, awareness and learning for data providers and users.

Specific recommendations include the following.

- Establish a national or regional community (depending on the scope of the portal).
- Prepare and offer workshops, training sessions, discussions and events.
- Prepare and offer video guidelines and online courses.
- Participate actively in related working groups.
- Use and test other open data portals.

Harmonisation and standardisation

One root cause for the often poor quality of metadata is the lack of maturity of harmonisation and standardisation. Insufficient harmonisation and standardisation can have a detrimental effect on the interoperability, the accessibility and the reusability of data. On the strategic level, we recommend fostering harmonisation and standardisation and seeking cooperation with corresponding communities and authorities to push harmonisation and standardisation forward. On the tactical and operational levels, we recommend fostering compliance with standards by using constructive means, such as increasing knowledge, and analytical means, such as quality control'. A key success factor for harmonisation and standardisation is maintaining a balance between covering all aspects and keeping the standards simple and applicable.

Specific recommendations include the following.

- Use open and well-established formats and services.
- Follow specifications like DCAT-AP as closely as possible.
- Do not reinvent standards, reuse and combine existing standards instead.
- Implement procedures to check and, if applicable, reject data and metadata that are not compliant with standards; such mechanisms can in principle be automated, semi-automated or function manually.

Automation and tool support

Automation and tool support are powerful means to improve data publishing. We especially recommend automating the data-publishing process by providing and using defined transmission protocols and APIs. As the survey shows, these are now rarely adopted. Furthermore, there are various quality tools available, such as the MQA or validation tools. We recommend providing appropriate tool support and fostering its usage.

Specific recommendations include the following.

- Establish tools and mechanisms to ensure data/metadata quality verification.
- Implement processes, initiatives and campaigns to ensure data-quality-verification tools are used.
- Use common software for metadata management.
- Automate the data-publishing process.
- Share your solutions with the community.

4. Conclusion

Within this report we investigated how data is made available in the different portals harvested by data.europa.eu, and what pain points exist in the publication process. From these findings, we identified potential improvements for the publication process, and derived specific and harmonised recommendations. Open data portal managers and owners can therefore use the report as a primary source of information for the improvement of their own data-publishing processes. Additionally, the findings from the results can serve as a basis for the creation of webinars and e-learning nuggets.

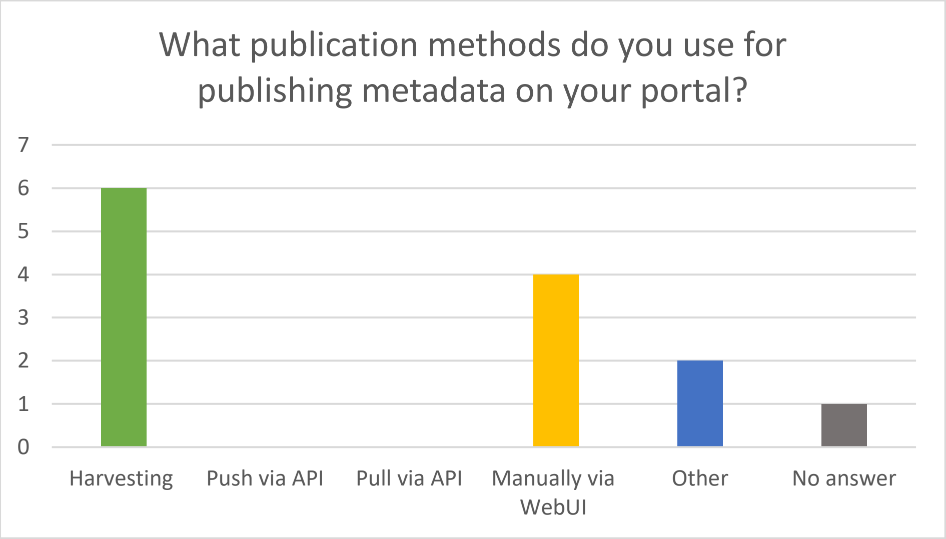
Annex

The annex contains further, in-depth and raw results from the survey and interviews (Annex A), along with the introduction texts and the full questions of the online survey (Annex B).

A. Further results from survey and interviews

Publishing process and mechanisms

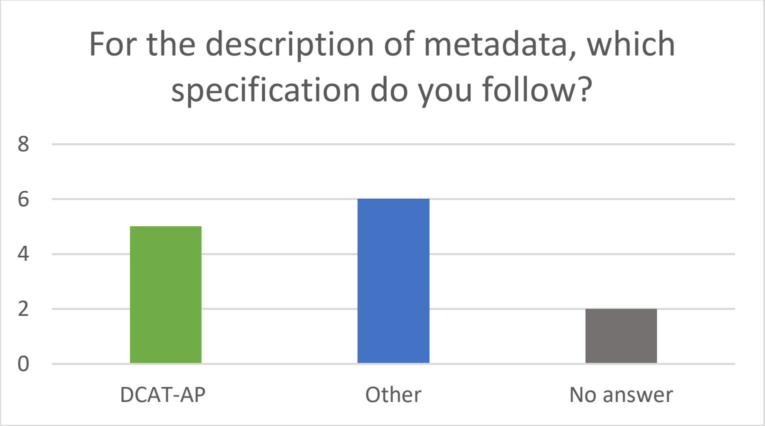
The following diagram shows the quantitative results relating to publishing metadata in the survey.



One participant, a national open data programme user, told us that twice a year they prepare open data schedules, which are published on their portal. Furthermore, open data officers monitor the timely publication of public data on the portal.

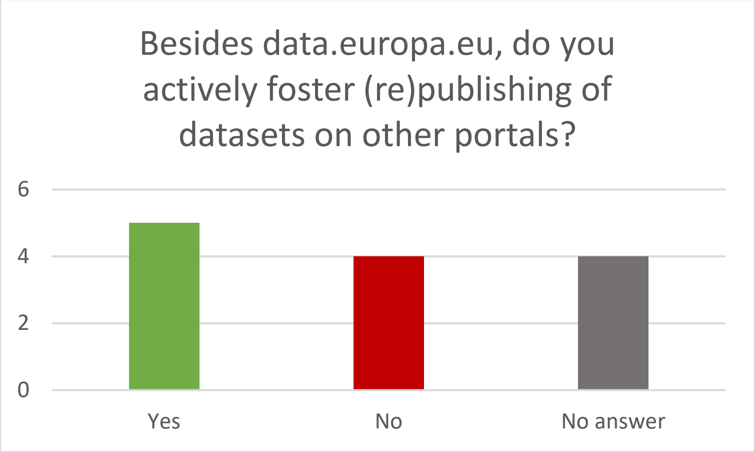
One participant told us that they were using the comprehensive knowledge archive network (CKAN) API and the Fiware Draco Generic Enablers processors specific for the publication of linked open data in CKAN, or harvesting of different Catalogue Service for the Web (CSW) services from data providers with GeoNetwork metadata software.

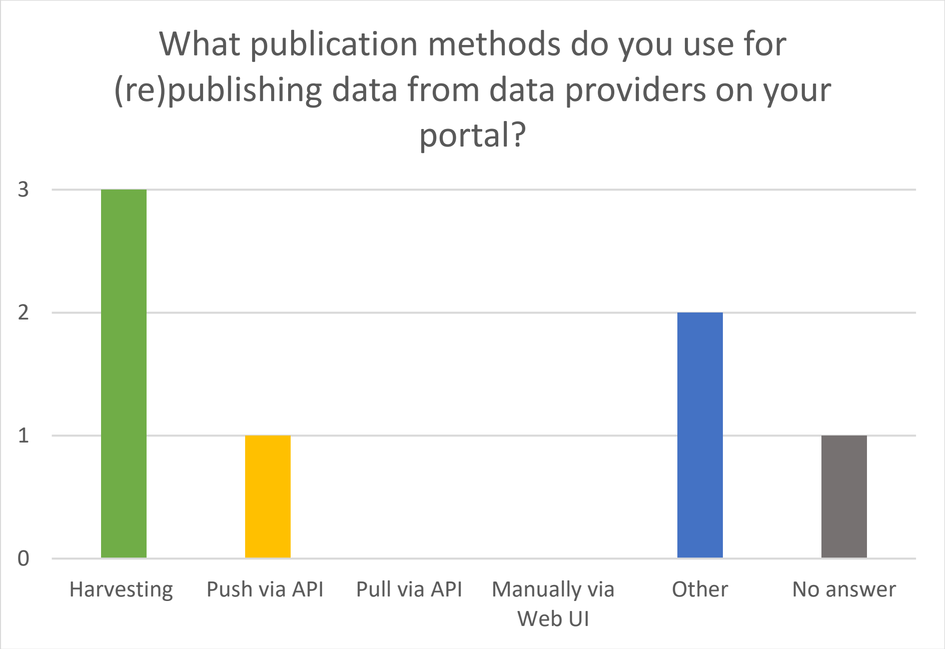
Regarding metadata, the following diagram shows the relevant specifications.

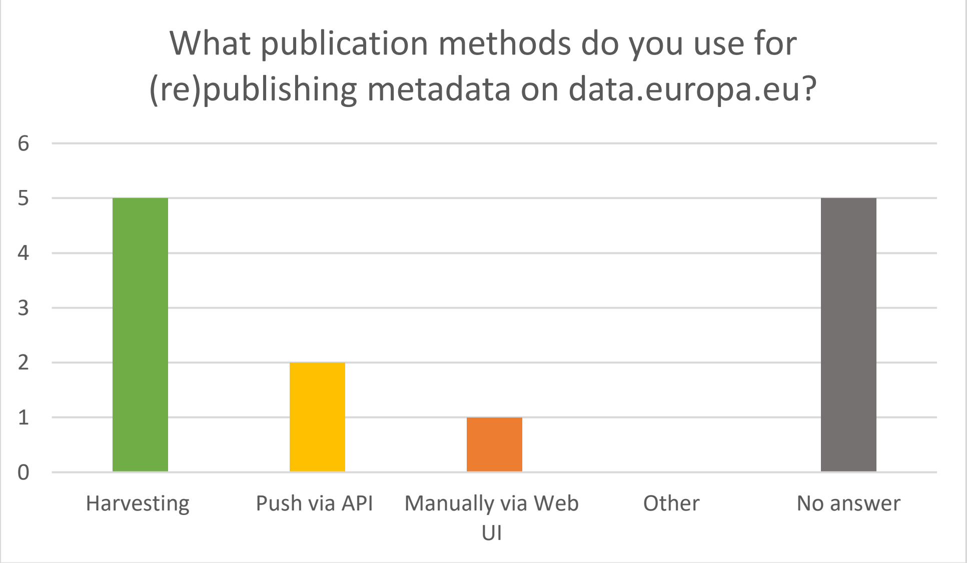


In addition to DCAT-AP as a general metadata standard, INSPIRE(ISO 19115/19139) could be identified as the most relevant standard for geodata.

Regarding the (re)publishing of data and metadata, we received the following additional answers (as shown in the diagrams below).







Top quality issues

Regarding data publishing, the following pain points were identified from the free text answers.

- Data not up to date / failure of publishers to update their data / no timely update / irregular update.
- Data not typically machine-readable.
- Lack of data standardisation / compliance with data openness standards.
- Using standards but nevertheless putting different logic in datasets.
- Lack of harmonisation / interoperability.
- Insufficient quality of data.
- No time series in the data sources.
- No time stamps available.
- Sometimes no unique identifiers in the datasets.
- Different title for same content data.
- Different formats and/or service types.
- Transition between ISO 19139 and DCAT.
- Limited formats for accessibility.
- Regional v national data.
- Complex licensing.
- Broken links.
- Fulfilment of different EU regulations in an INSPIRE context.

Regarding metadata publishing, the following pain points were identified from the free text answers.

- No common standard for metadata describing, such as DCAT-AP.
- Use of very specific terminology, which may not be understood by others.
- Not following the DCAT-AP controlled vocabulary.
- Metadata is not always comprehensive/precise and requires manual checks.
- To short or unclear data descriptions.
- Not providing optional metadata (e.g. data location).
- Authorities try to do only the bare minimum regarding metadata.
- Improper marking of metadata by providers (e.g. update frequency).
- Misuse of description fields to provide narrative.
- Free text keywords.
- Incorrect author and ownership information.
- Missing, poor or incorrect licence choices and/or information.
- Missing feature type description.
- Providing the size of the data because they are updating automatically.
- Catalogue maintenance.
- Timely update / outdated info / metadata not up to date.
- Temporal references.
- Metadata not always very user-friendly at the graphical user interface level.
- INSPIRE compliance.
- Passing the Shapes Constraint Language (SHACL).
- Poor (technical) implementation.
- Quality of metadata harvesting from geoportal to national open data portal.
- MQA calculation.
- MQA index not scoring correctly for geospatial data.

Improving quality of data and metadata

Examples we identified from the free text answers include the following.

- Ongoing project to harmonise data at the regional level and automate data submission.
- Constant development of the national data portal based on the users' feedback and business needs.
- Improvement process as part of a quality management system.
- Cooperation with statistics authority to help other authorities describe datasets in accordance with strict standards.
- National committee to discuss common actions.
- National assistance board to help data providers with technical issues.
- Use of quality measures typically using automated testing.
- Open Data Institute open data certification and validation.
- Weekly reports on quality of metadata.
- Quality checks in the local data catalogue.
- Manual inspection of records for scheme violations, URL issues and INSPIRE compliance.
- Open data portal's support team is in control of dataset descriptions being up to date.
- Mandatory metadata fields of the national portal, which cannot be left unfilled.
- Rebuild of data store in progress to promote better quality of data and metadata.
- Developing processors based on the FIWARE Draco GE to automate and improve the quality of the data published.
- Developing a python tool to measure the metadata quality before being harvested by data.europa.eu.
- Developing CKAN extensions to improve the metadata quality.
- Feedback field on the portal; regular surveys.
- Collect and analyse the statistics regarding the portal and make a decision based on them.
- Field/section on the portal, where users can leave data requests and feedback.
- Financial and time opportunities play a key role.
- Strive for innovation and try to follow best practices.
- Considering training and other interventions to encourage publishers.
- Multimedia training in open formats.
- Workshops aiming at increasing overall knowledge on data and metadata management.
- Training for nominated data experts.
- Improving the quality of data as one of the goals of open data programmes.

Best practices

- Best practice example 5 (general).
  - Systematic training and tailored workshops addressed to data providers are provided. The open data portal includes multimedia training on open data and preparation of data in open formats, in particular the following tutorials.
    - How to make data available on the portal?
    - How to add datasets/resources?
  - When adding a file to the open data portal or saving an external link, the validation process is carried out, which consists of three following steps.
    - Link validation. Consists of checking whether the resource can be downloaded (for a file) or can make a connection (for API or web pages).
    - File validation. Consists of determining the file format.
    - Data validation. Consists of checking whether the structure and quality of data allow a tabular view to be created (e.g. whether cells are merged or the date format is correct).

    Additional data quality validators in the file have been implemented with the selected technical standard rules (e.g. compliance of telephone number, PNA code, KRS number). In the automatic data validation process, the completeness of the metadata is checked when data is made available on the open data portal. The set of mandatory resource metadata elements made available in the open data portal was defined in the Regulation of the Council of Ministers on the Data Portal.

- Open data portal implemented the functionality of automatic notifications for editors about the need to update data.
- Open data portal has been adapted to the DCAT-AP standard.
- The evaluation of the technical and API standard of data openness has been published.
- Open Data Academy.
- Open data portal functionality enables automatic conversion of well-prepared data of xls/xlsx files to csv and json-ld.
- The *Guide for Data Providers* is regularly updated.
- The *Open Data Laboratory* publishes reports that support data providers in the legal, technical and security areas of opening public sector information for reuse.

Tools for improvement

Within the improvement process certain tools can be useful. The following could be identified from the survey:

- MQA
- Github.com/YourOpenData/mqa-scoring or mqa-scoring-api
- standards such as DCAT, DCAT-AP, Next Generation Service Interface – Linked Data (NGSI-LD)
- data validation tool
- validator for INSPIRE datasets
- metadata validators
- INSPIRE ETF validator
- INSPIRE linkage checker
- JSON Schema or schema.org ontologies
- various internal quality control tools
- manual or batch updates
- administration system for the state information system in Estonia (RIHA/RIHAKE): automated dataflow.

Experience with data quality guidelines and metadata quality assurance service

Regarding the MQA, the following experiences and feedback for improvement were reported in free text answers.

- 'The tool MQA is very useful in checking the quality of metadata.'
- 'The SHACL validation tool has problems. For example, to specify the theme and pass the SHACL you need to include Science and Technology.'
- 'MQA index not scoring correctly for geospatial data.'
  - 'The tool (MQA) is not working properly for our catalogue. Accessibility, interoperability, reusability and contextuality are not properly calculated. The quality report does not give enough technical information to be used to improve each individual record. I am guessing most issues are because our catalogue is a geonetwork node with records based on 19139.'
  - 'We identified reasons why MQA ratings are not achieving higher standards and contacted the "data.europa.eu" helpdesk in June 2022 for advice. For example, we noticed that for some areas of the MQA outputs, that a score has not been issued where metadata exists. E.g., the MQA dashboard for the HELCOM Metadata Catalogue does not award a scoring to "Download URL" under Accessibility. However, where applicable, these URLs exist. This is also the case for "time-based search", "date-of-issue", "format", and others. Further correspondence with the "data.europa.eu" helpdesk in September 2022 revealed that the MQA tool is not completely compatible for rating catalogues that feature ISO/INSPIRE metadata records, such as ours. We understand that efforts are underway to

ensure MQA ratings are not penalized if records are based on ISO/INSPIRE mapping. Compliance with the MQA tool is linked to milestones for our Baltic Data flows project with the deadline of 31 August 2023. We hope this issue will be addressed before this deadline to help improve our MQA ratings.'
  - 'We prioritize INSPIRE guidelines (e.g., guidelines for interoperability and reusability, e.g., guidelines for findability).'

Problems with the MQA and its scoring were therefore reported, especially relating to geospatial metadata.

Obstacles in improving data and metadata quality

Regarding improvement initiatives and processes, we also asked about the biggest obstacles to improving the data and metadata quality of datasets. The following pain points were identified from the free text answers.

- Data standardisation.
- Homogenisation of different data sources.
- Passing the SHACL.
- Major part of data is provided by other organisations.
- Priorities given to data volume v quality.
- Insufficient information from data provider.
- Lack of awareness of importance of metadata.
- Undetected changes.
- Complexity of schema.
- Datasets are not always well described by the schema as we deviate from spatial datasets.
- Poor dataset descriptions: very different level of knowledge on data.
- Still insufficient knowledge and skills of data providers.
- Lack of knowledge.
- Lack of technical skills.
- Not a common practice for the authorities.
- Lack of resources.
- Not enough resources for data management / data governance (no buzzwords like digitalisation, artificial intelligence, big data, blockchain).
- Staff cost of data providers.
- Federalism.
- Technical know-how.
- Software limitations.
- Metadata often not important enough for data providers.
- Misinterpretations.
- Poor guidelines (INSPIRE, high-value datasets (HVD)).
- Poor commitment to INSPIRE harmonised datasets (not very useful).
- Differences between open data – metadata and INSPIRE – metadata (DCAT/CKAN – ISO/CSW).

Suggestions for improving data and metadata quality

Finally, we asked about suggestions for improving the quality of data and metadata. The following points were identified by the free text answers.

- Improve the MQA tool.
- Follow standards and ontologies.
- Use of linked data.
- Follow vocabularies such as DCAT or DCAT-AP.
- Reject data and metadata that is not compliant to quality standards.
- Concentrate on valuable and primary datasets; check metadata for correctness (attribution etc.).

- Use a common, widely accepted schema and a stable, well-supported software for it.
- Improved schema translations and better feedback for debugging the process.
- Training for data providers.
- Preparation of standards and guidelines.
- Appointing openness officers responsible for data/metadata quality in institutions.
- Data portals should have tools and mechanisms to ensure data/metadata quality verification.
- Legal regulations should be implemented specifying the quality of data/metadata.
- Increase overall knowledge about the data field and have explicit guidelines, which can be implemented in real life.
- Use common formats and services.
- Use common software for metadata management.
- Find and eliminate duplicates from different national portals.
- Provide resources.
- Introduction of a common metadata standard (open data, INSPIRE, Dataspaces).
- Reference tools (open source).

From the interviews, we gained the following additional suggestions.

- Error handling report errors to data providers within quality management process
- Strategic level of improvement (e.g. through informing, events, training).
- Valuable datasets (quality over quantity):
    - direct link to INSPIRE data;
    - increase knowledge (e.g. through communication, workshops, meetings);
    - automate process (like data.europa.eu);
    - video guidelines, online courses;
    - existing national community (the Data Working Group); happy to invite representatives from the EU (e.g. open data maturity, talk directly to national community);
    - whole view on open data / overall data strategy;
    - harmonisation in open government data (OGD) / INSPIRE context.

B. Introduction texts and questions of online survey

[Title of survey]

Improving Data Publishing

[Description of survey]

Publishing high-quality data in a simple way is crucial for creating most added value. The design of the data-publishing process is particularly important. On the one hand, this refers to the internal data publication processes within the data-providing organisations (data provider). On the other hand, open data portal managers and owners can also positively influence the data publishing process in their portal through various mechanisms, thus contributing to better data quality, easier data publication for data providers and a wider visibility of published data. This survey is intended to address the latter perspective. We assume that open data portal managers and owners can make a significant contribution to the use and dissemination of open data. The aim is to investigate how data is made available in the different portals and what pain points exist in the publication process. These findings will be used to identify potential improvements for the general publication process.

[Welcome text]

Thank you very much for your willingness to take part in our survey. The survey should take about 5 - 15 minutes of your time and will help to analyse key aspects of data publishing by open data portal managers and owners.

[Question group I – General]

[G04Q21] What type of data portal do you work for? [mandatory question]

- ○ EU open data portal
- ○ Supra-national open data portal
- ○ National open data portal
- ○ Regional open data portal
- ○ Project-specific open data portal
- ○ Other: __________

[G02Q24] May we contact you for an additional short expert interview?

- ○ Yes [then next question is G01Q26]
- ○ No [then next question is G01Q01]
- ○ No answer [then next question is G01Q01]

[G01Q26] Please indicate the following contact data.

URL of the portal: _________

Contact information: _________

[Question group II – Data and metadata quality]

[Introduction text]

Open Data is only useful if it is of high quality. High quality is important for both, metadata and data. Only high-quality metadata can ensure that the data can be found. The better the quality of the data, the easier it will be for users to reuse the data. Open data portal managers and owners can offer tools for measuring data and/or metadata quality of published datasets that are then shared with the data providers.

Please be aware that in some of the following questions we make an explicit distinction between data and metadata.

[G01Q01] What are the 3 top-most quality issues you currently face relating to the data published on your portal?

Top 1: __________

Top 2: __________

Top 3: __________

[G01Q02] What are the 3 top-most quality issues you currently face relating to the metadata published on your portal?

Top 1: __________

Top 2: __________

Top 3: __________

[G01Q03] Do you currently have any initiatives or processes aimed at improving the quality of data or metadata on your portal?

- ○ Yes [then next question is G01Q04]
- ○ No [then next question is G01Q05]
- ○ No answer [then next question is G01Q05]

[G01Q04] Please tell us more about the current quality improvement initiatives.

[Free text field]

[G01Q05] Do you know the Data Quality Guidelines of the Publications Office of the European Union?

- ○ Yes [then next question is G01Q06 followed by G01Q07]
- ○ No [then next question is G01Q24]
- ○ No answer [then next question is G01Q24]

[G01Q06] Do you use the Data Quality Guidelines for the improvement of your data or metadata quality? If so, please describe briefly what parts of the Guidelines you are using.

- ○ Yes, for improving data quality: _____________

- Yes, for improving metadata quality: ____________
- No, we don’t use them

[G01Q07] If you have any other feedback on the Data Quality Guidelines, please let us know.

[Free text field]

[G02Q24] Do you know the Metadata Quality Assurance service (MQA) of data.europa.eu, which can be used for improving metadata and data quality?

- Yes [then next question is G01Q08]
- No [then next question is G01Q10]
- No answer [then next question is G01Q10]

[G01Q08] Have you already used the MQA service of data.europa.eu to improve your data or metadata quality yet?

- Yes [then next question is G01Q09]
- No [then next question is G01Q10]
- No answer [then next question is G01Q10]

[G01Q09] Please tell us about your experience with this tool. If you have any feedback for improvement, please let us also know.

[Free text field]

[G01Q10] Do you use other methods and/or tools for checking and improving your metadata quality or data quality?

- Yes [then next question is G02Q25]
- No [then next question is G01Q12]
- No answer [then next question is G01Q12]

[G02Q25] Please tell us, which tools you are using for checking or improving your data or metadata quality.

- Checking data quality: _____________
- Improving data quality: _____________
- Checking metadata quality: _____________
- Improving metadata quality: _____________
- Other: _____________

[G01Q12] What are the 3 top-most obstacles to improving the data quality of your datasets?

Top 1: __________

Top 2: __________

Top 3: __________

[G01Q25] What are the 3 top-most obstacles to improving the metadata quality of your datasets?

Top 1: __________

Top 2: __________

Top 3: __________

[G01Q13] What are your main general suggestions for improving data and metadata quality from your experience?

- Improving data quality: _____________
- Improving metadata quality: _____________

[Question group III – Methods and processes]

[Introduction text]

The design of the publishing process offered by the portal influences how easily and efficiently data providers can publish their data. Thus, we subsequently would like to know more about the publishing processes, methods and targets you are using in the portal you provide.

[G02Q14] What publication methods do you use for publishing metadata on your portal?

- Harvesting
- Push via API
- Pull via API
- Manually via Web UI
- Other____________
- No answer

[G03Q26] For the description of metadata, which specification do you follow?

- DCAT-AP
- Other____________
- No answer

[G03Q27] Besides metadata, do you also (re)publish data from data providers on your portal (i.e. data repository service)?

- Yes [then next question is G03Q28]
- No [then next question is G03Q29]
- No answer [then next question is G03Q29]

[G03Q28] What publication methods do you use for (re)publishing data from data providers on your portal?

- Harvesting
- Push via API

- Pull via API
- Manually via Web UI
- Other____________
- No answer

[G03Q29] What publication methods do you use for (re)publishing metadata on data.europa.eu?

- Harvesting
- Push via API
- Manually via Web UI
- Other____________
- No answer

[G03Q30] If you have any suggestions on how to improve the (re)publishing of metadata on data.europa.eu, please let us know.

[Free text field]

[G02Q15] Besides data.europa.eu, do you actively foster (re)publishing of datasets on other portals?

- Yes [then next question is G03Q31]
- No [then next question is G02Q16]
- No [then next question is G02Q16]

[G03Q31] What methods do you use to address other portals besides data.europa.eu to (re)publish metadata?

- Harvesting
- Push via API
- Pull via API
- Manually via Web UI
- Other____________
- No answer

[G02Q16] Have you established a systematic publishing process for the metadata or data on your portal?

- Yes [then next question is G02Q17]
- No [then next question is G03Q18]
- No answer [then next question is G03Q18]

[G02Q17] Please describe briefly the publishing process you have established.

[Free text field]

[Question group IV – Feedback and improvement process]

[Introduction text]

Open data portal managers and owners can promote the use of open data by continuously improving the data-publication process, e.g. by collecting feedback from data providers and data users and responding to their demands. In the following section we like to know more about your feedback and improvement process.

[G03Q18] Do you have in place a systematic and continuous improvement process for your metadata and data publishing on a regular basis?

- ○ Yes [then next question is G03Q19 followed by G03Q20]
- ○ No [then end of survey]
- ○ No answer [then end of survey]

[G03Q19] Please describe the process you have established briefly.

[Free text field]

[G03Q20] Do you address the following aspects within this process? If so, please describe briefly.

- ○ Feedback? _______________
- ○ Metrics? _______________
- ○ Demand? _______________
- ○ Opportunities? _______________
- ○ Innovation? _______________

[End message]

Thank you very much for taking part in this survey. If you have left your contact information, we might contact you for an additional interview. Your answers to the survey have been recorded and will be evaluated in preparing a report on ‘Improving Data Publishing’ in the context of data.europa.eu. The results will be published presumably in early 2023.

Publications Office
of the European Union

ISBN 978-92-78-43683-4