Principles and recommendations to make data.europa.eu data more reusable

A strategy-mapping report

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 the management of data.europa.eu.

For more information about this paper, please contact the following.

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

Authors
Oscar Corcho
Esteban González
Edna Ruckhaus
Daniel Garijo
Manuel Castillo
Elena Simperl

Last update: 22-04-2022
https://data.europa.eu/

DISCLAIMER
By the European Commission's Directorate-General for 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 that may be made of the information contained therein.

Luxembourg: Publications Office of the European Union, 2022
© European Union, 2022

OA-09-22-160-EN-N ISBN: 978-92-78-42938-6 doi: 10.2830/9342

The reuse policy of European Commission documents is implemented by Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (https://creativecommons.org/licences/by/4.0/). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

Contents

This is the first of a series of reports that focus on advancing the discussion on the mid- and long-term sustainability of (open) data portal infrastructures. This series of reports will analyse the role of data intermediaries in the broader data economy, will make several proposals for the evolution of open data portals towards mid- and long-term self-sustainability, and will design software tools and platforms to support data intermediaries.

This first report focuses on providing recommendations for open data providers and data intermediaries on how to make open data available so as to promote its reuse. It stems from the work previously carried out by the data.europa.eu team and our own research in open data management and human-data interaction. Taking the conclusions of this previous research into account – along with our experience in the reuse of datasets and the results of a set of workshops, training courses and other stakeholder engagements – we have created a list of 10 recommendations that we propose would increase the reusability of the data to be made available in the next generation of open data portals.

The following reports, which will be produced annually, will design and implement methods to assess the value of key data assets aggregated by data.europa.eu, and understand how users search for datasets and how to design dataset recommendation systems for them.

Executive summary

This report summarises the work done so far on thought leadership regarding the mid- and long-term sustainability of (open) data portal infrastructures. The main research questions we are addressing in this work, which spans from 2021 to 2025, are the following.

- What is **the role** of data intermediaries (such as data.europa.eu) in the broader data economy?
- How can open data portals evolve towards mid- and long-term self-**sustainability**?
- How can we design software **tools** and platforms to support data intermediaries?

We answer these questions in two ways.

- We provide thought leadership to advance our collective understanding of sustainable data portals and ecosystems by convening webinars with experts from different stakeholder groups. The webinars will feature representatives of open governmental data portals across Europe, technology developers, scientists, open data advocates and policymakers. They will be organised in 2022 and 2023 as part of the data.europa.eu academy.
- We issue an initial list of recommendations for open data portals and intermediaries, such as data.europa.eu, on how to make their data available so that its reuse can be improved. This list of recommendations has been drawn up using a mix of desk research of existing studies and approaches to improve data use and usability (carried out from May 2021 to September 2021), complemented by workshops, with diverse groups of open data users, undertaken from October 2021 to March 2022. The recommendations will evolve to accommodate new insights from the webinars, along with a second series of workshops with data users.

The following is a summary of the recommendations extracted from the results of our work.

- Clearly describe the column headers or attribute names of the published dataset distributions (recommendation 1).
- Provide statistical information about the dataset (recommendation 2) and generate visual previews of the data so that it can be better understood at a simple glance (recommendation 3).
- Add as much contextual documentation as possible to the datasets, describing the whole lifecycle of the data and including any relevant information about data quality assurance methods and assessments (recommendation 4), and ethics and data anonymisation techniques applied to the dataset (recommendation 5).
- Provide examples of data usage, including use cases where the dataset has been or can be used (recommendation 6), and link datasets to other digital artefacts that make use of them (recommendation 7), including software, technical reports, publications, etc.
- Improve the way in which data can be found, beyond current metadata models (search by variables, geospatial and temporal filters, etc.) (recommendation 8) and enable programmatic access to datasets (recommendation 9).
- Generate persistent identifiers for the datasets to allow them to be unambiguously identified and track their usage via citation tracking mechanisms (recommendation 10) to understand their use and impact.

1. Introduction

Making data available (e.g. as open data) does not necessarily mean that it will be easy to use and actually used or reused by data consumers. This is one of the main conclusions drawn from the discussion paper presented by Koesten et al. (2021) on dataset reuse. In this paper, the authors provide their insight based on their participation in many data-related projects over several years of research. This paper sets out several questions (such as: how comfortable are we with reading, interpreting and working with other people's data?) that are a good starting point for the work presented in this report. A clear message from that work is that using datasets beyond the context for which they were originally created remains challenging, even in those cases where the datasets were published according to existing guidelines and best practices.

The paper draws on interviews with data practitioners and user studies in which participants were asked to describe datasets to make it easier for other people to use them. From this research, the authors claim that it is crucial to consider the requests and experiences of those people who are trying to understand and work with datasets so as to make sure they have a meaningful user experience. They identify several points to be considered, such as better communication on how data is captured and curated, with clear textual descriptions (as they point out, there is a positive correlation between having textual descriptions of the data and incrementing the usage of the data), and considering the different tasks associated with the usage of the data, including quality assurance procedures, statistical information and ethical considerations.

Some of these recommendations echo prior work undertaken by the European Data Portal (now data.europa.eu) team in 2017, published in a report on the future of open data portals (Simperl, E. and Walker, J., 2020). The authors define ten dimensions of user-centric open data portal design, which include the following.

- Promote the use of open data through impact stories and examples.
- Be discoverable using metadata description standards.
- Increment the metadata considering other dimensions such as relevance (do I need this?), usability (can I use it?) and quality (how good is it?) to improve the publication of the metadata and increment the reusability of the data.
- Co-locate context-sensitive documentation within the published data, by including additional links and information.
- Design and use portal indicators that cover both the publisher perspective (usage) and the user perspective (quality).
- Co-locate tools for data manipulation to make the data more accessible to casual data users.
- Be accessible through open and machine-readable formats.

There are also recommendations for specific sectors or use cases, which build on these general-purpose insights. For example, the research carried out by Soylu et al. (2022) on data quality barriers for transparency in public procurement provides a set of recommendations for publishing better, more usable open procurement data. These recommendations include the following.

- Make your public procurement data available in a structured format and according to existing standards.
- Include identifiers of all the tenderers that participate in a contracting process.
- Include identifiers of the departments and suborganisations that act as tenderers.
- Include identifiers of the participating organisations in joint ventures' data.

- All notices and steps associated with a contracting process should be linked with the same identifier.
- Link invoices (and results) to the public procurement process to which they belong.
- The text of all documents used in a contracting process should be available for further processing and linked to their corresponding contracting process.
- Provide commonly agreed visualisations of public contracting data.
- Provide answers to the most common questions made by citizens and organisations.
- Use your own public procurement data internally (e.g. as a data back-end in your transparency portal).

Such recommendations could be co-created with relevant stakeholders in other high-stakes domains identified by the European Commission's data strategy ([1]), including the verticals of common data spaces and of so called high-value datasets from the open data strategy.

Other relevant sources that discuss the importance of making open data available and easier to use are developed for local public administrations in the context of the Ciudades Abiertas (Open Cities) project ([2]). Several data quality dimensions are identified as relevant not only for data providers, but also for potential data users. These dimensions include, among others: (i) completeness, where the dataset contains all the elements; (ii) unicity, dataset with no duplicate records; and (ii) accuracy, where the data is similar to the reality.

In this report we take these analyses further, following a methodology focused on understanding, at first hand, how real users feel about different forms of data publishing by open data portals. The results and corresponding recommendations, which build on all these prior works, are presented in Section 3.

2. Methodology

This section describes the methodology followed to derive the recommendations that are presented in Section 3. Our research (spanning from May 2021 to March 2022) has been done following the next steps.

- First, we started with the analysis of some of the previous results obtained by our research teams at Universidad Politécnica de Madrid and King's College London on how to make open data usable in different contexts: open data portals in general (Koesten, L. et al., 2021; Simperl, E. and Walker, J., 2020), public procurement data (Soylu, A. et al., 2022), open data published by local authorities (Corcho, O. and De Pablo, V., 2022) and open research data (Corcho, O. et al., 2021; Corcho, O., González, E. and Garijo, D., 2021). While such an approach may seem too narrow, as it seems to only be considering our own literature sources, it is based on several years of research – carried out by our teams – on dealing with data and exploring the main challenges in the usage of open data published by public administrations. This analysis has allowed us to identify the main groups of stakeholders involved in the usage of open data, and some of their main challenges when aiming for data reuse. A summary of the main findings and recommendations from these sources has already been provided in the introduction section of this report.
- Second, we ran four sessions between October 2021 and March 2022, with different groups of data reusers, to verify whether the aforementioned challenges are actual challenges for

([1]) https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

([2]) This project, led by the city of Zaragoza in Spain, focused on the creation and deployment of a reusable and interoperable technology platform to support Open Government processes, including open data, transparency and public participation (www.ciudadesabiertas.es)

them and to identify potential new needs and opportunities in open data publishing based on their experiences in using open data. The following sessions took place.

- A 4-hour brainstorming and challenge-mapping workshop – covering a variety of domains such as agriculture, transport and mobility, public procurement and earth sciences – with seven EU researchers from five countries (Belgium, Ireland, Greece, Spain and Poland) who have expressed their wishes on how they would like open data to be published so as to facilitate its use for their research purposes.
- A 30-minute brainstorming session with open data publishers and users in a municipality in Spain, following a survey sent to data reusers in that same municipality (mostly companies and freelancers), for which 31 responses were collected.
- A 2-month hands-on project on data identification, transformation and usage for data science purposes, carried out by an international group of approximately 40 data science master's students, with a background in computer science and data science.
- A 1-hour training session and a 1-week hands-on project on data cataloguing for public services, carried out by an international group of artificial intelligence master's students – with a diverse background, ranging from law, international relations, politics and public management to computer science – focused on the specific topic of public procurement data.

3. Principles and recommendations

One of the key objectives of publishing datasets in open data portals is to make them reusable by third parties (and by the public administrations that are publishing them), fostering the creation of new business opportunities, advancing science and improving public services.

With this objective in mind, and as a result of the methodology described in Section 2, we have compiled a list of principles and recommendations that can be used to shape the design of and improve current and future open data portals so they become more oriented towards the users' needs and experiences, rather than being an accumulation of datasets generated by public administrations.

3.1. Improve dataset descriptions to facilitate reuse

Recommendation 1: clearly describe the column headers or attribute names of the published dataset distributions

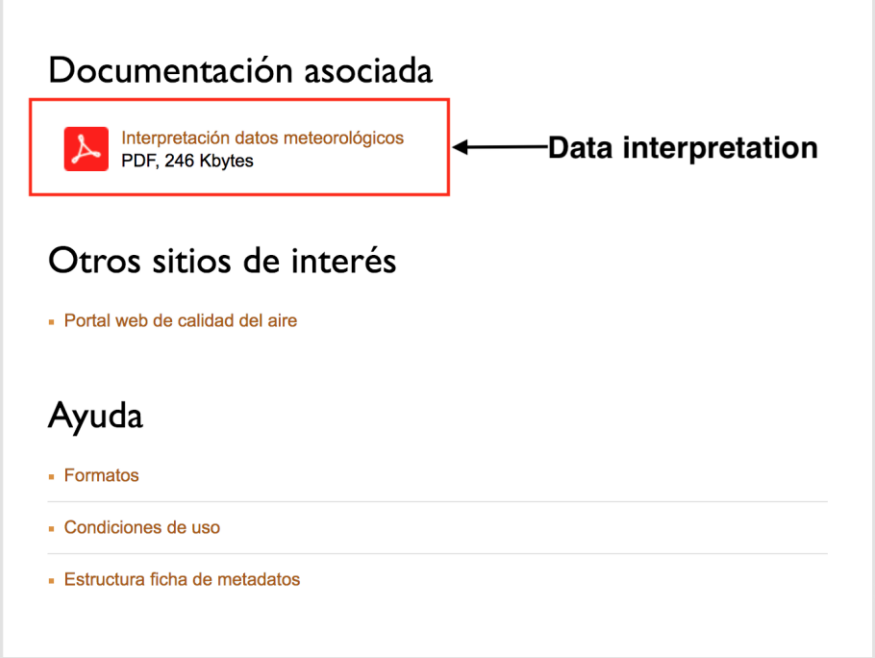
One of the problems data consumers face when trying to reuse a dataset they have found via the search features of an open data portal is related to understanding the meaning of the data inside the datasets they have initially identified. Sometimes, the attributes or variables of the dataset distribution (the column headers of the tables in the case of a tabular distribution such as CSV and XSLX, or the property names in the case of formats such as XML or JSON) have labels that are incomplete or difficult to understand.

These labels would be easier to understand if they referred to terms that are already defined in standard or widely used vocabularies and ontologies. Ideally, the column header would have a label such as 'streetAddress', which would then lead to a URL (https://schema.org/streetAddress), and

would therefore make it easier for data reusers to examine the dataset. However, these labels are not always aligned with existing vocabularies or ontologies, and when they are aligned, these vocabularies or ontologies may not be published according to best practices when making their definitions available on the internet. This then makes them difficult to find by people who are not experienced in using semantic technologies.

We recommend **clearly describing the column headers or attribute names of the published dataset**, by aligning the column headers or attribute names of the datasets to terms defined in existing vocabularies or ontologies – or at least by adding a complementary document to the dataset where data providers describe the intended meaning of these column headers or attribute names.

Examples of this practice can be found in the meteorological observations dataset of the Open Data Portal of the city of Madrid ([3]). Figure 1 depicts the box where data consumers can find this additional information ([4]) as a PDF document. However, this document is not available on the data portals that aggregate this dataset, such as datos.gob.es ([5]) or data.europa.eu ([6]), which means that this type of additional documentation should also be inserted in the dataset metadata and be aggregated so that it can also be shown in those data portals.



([3]) https://datos.madrid.es/FWProjects/egob/Catalogo/MedioAmbiente/DatosMeteorologicos/Ficheros/Interpretaci%C3%B3n_datos_meteorologicos.pdf

([4]) https://datos.madrid.es/sites/v/index.jsp?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

([5]) https://datos.gob.es/es/catalogo/l01280796-datos-meteorologicos-datos-horarios-desde-20191

([6]) https://data.europa.eu/data/datasets/https-datos-madrid-es-egob-catalogo-300352-0-meteorologicos-horarios

Figure 1: Data interpretation (based on the headers of the dataset).

Another interesting case is the one concerning the Open Data Portal of the city of Los Angeles. In this case, data variables are described in a table ([7]) that compiles the name, the description and the datatype, as shown in Figure 2.

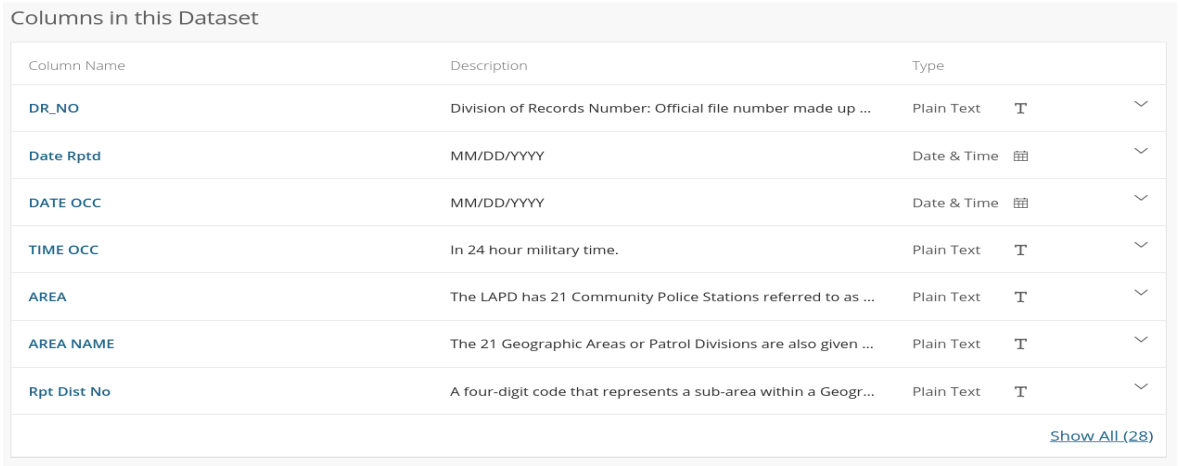


Figure 2: Column header interpretation in the Open Data Portal of the city of Los Angeles.

Recommendation 2: provide statistical information about the dataset

Open data portals commonly show some data quality indicators associated with each dataset. However, in general, these indicators are based on the dataset metadata and not on the data itself. Basic statistical information about the dataset – such as number of records, averages, means, modes, quartiles, outliers and empty fields – can be relevant in the case of numerical data, as shown in Figure 3. In the case of qualitative data, additional indicators such as the number of categories or number of samples per category can be relevant too. This information can be precalculated and offered as part of the metadata of the dataset.

We recommend **providing statistical information** for each dataset and showing this information together with the data. This type of statistical information may be visualised by means of charts such as scatter plots (to see the data distribution), histograms and box-and-whisker plots (to see quartiles and outliers).

([7]) https://data.lacity.org/Public-Safety/Domestic-Violence-Calls-from-2020-to-Present/qq59-f26t

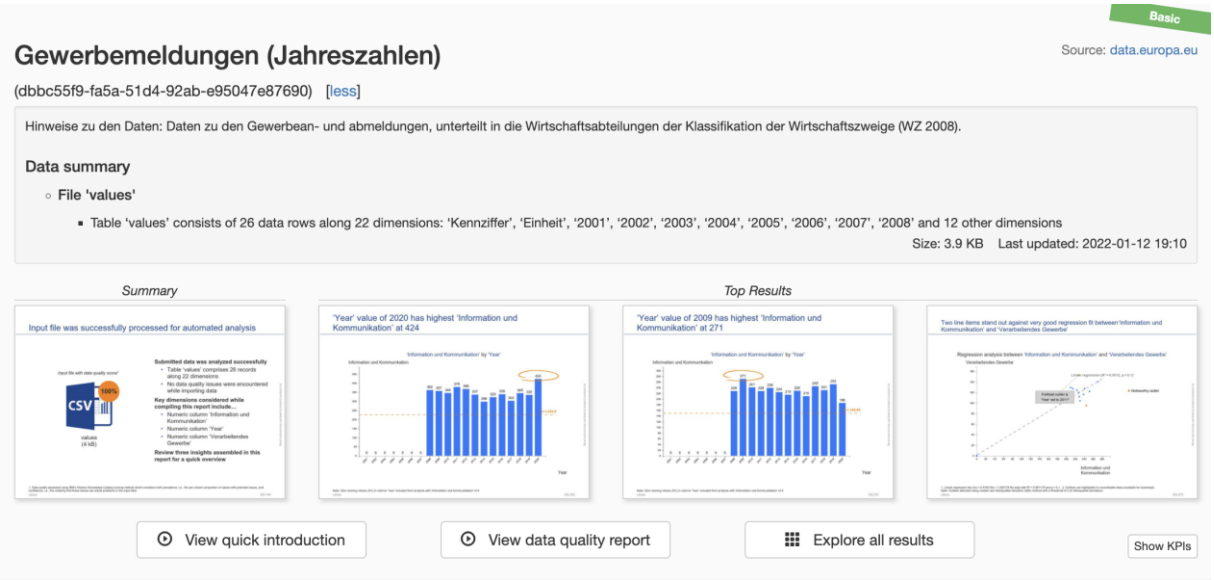


Figure 3: Statistical information associated with a dataset (*source:* https://analyst-2.ai/).

Recommendation 3: offer visual previews of the data

A common practice for data consumers is to download the data from the open data portal to explore it locally and analyse its main characteristics. This can be an annoying and tedious process in terms of time (for downloading and browsing) and space (in terms of the size of the dataset), especially for very large datasets (e.g. Excel files that cannot be opened in most personal computers due to their sizes). Some data repositories, frequently used in the research domain, such as Zenodo ([8]) and open data portals in general, provide simple data previsualisations (see Figure 4 for the dataset referenced above) in tabular formats, for instance.

([8]) https://zenodo.org/record/3378310

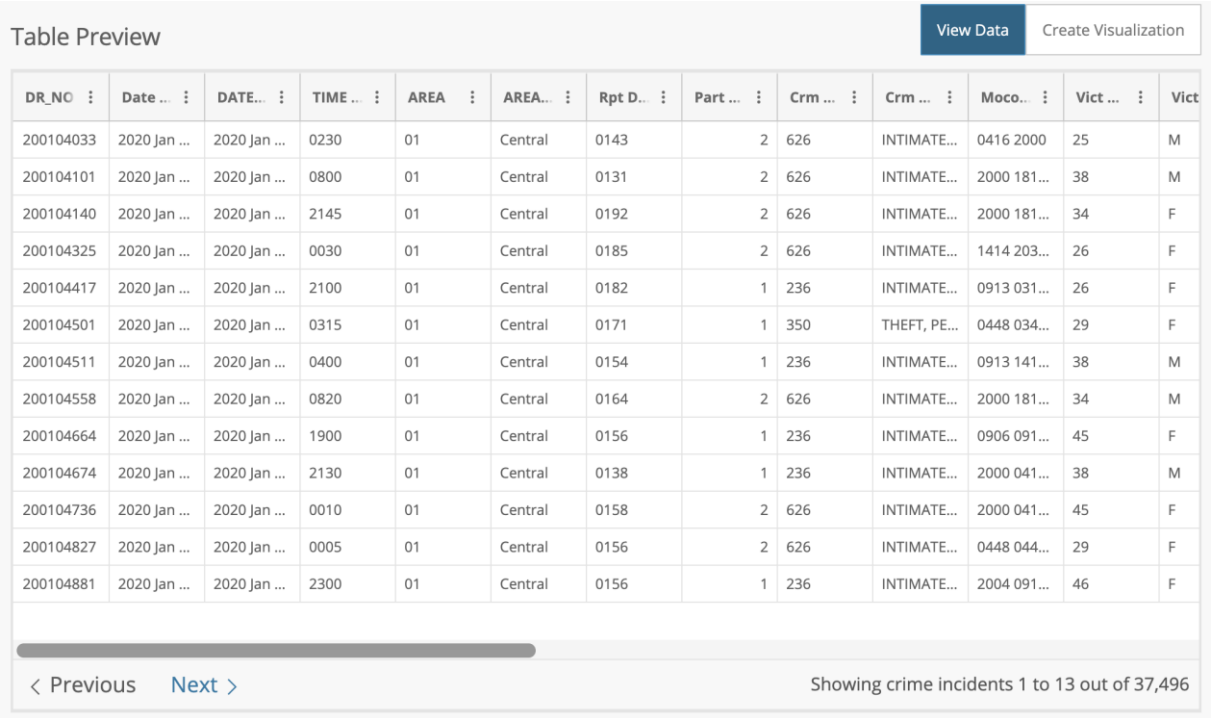


Figure 4: Data previsualisation in tables.

Data visualisations are also an excellent approach to allow potential data consumers to look into the data. The Open Data Portal offered by the STARS4ALL foundation allows sensor data observation to be explored using tables and plots ([9]) (see Figure 5).

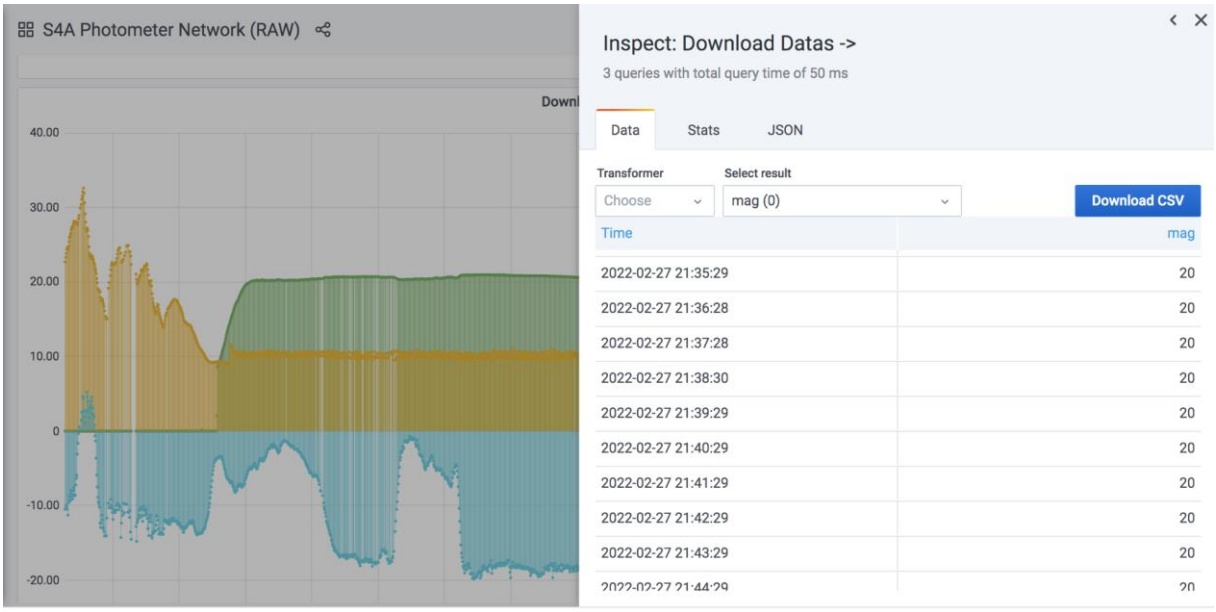


Figure 5: Data previsualisation in tables and plots.

([9]) https://tess.dashboards.stars4all.eu/

We recommend **adding previsualisations for each dataset**, whenever this is possible, using tabular formats, specialised charts or a combination of them.

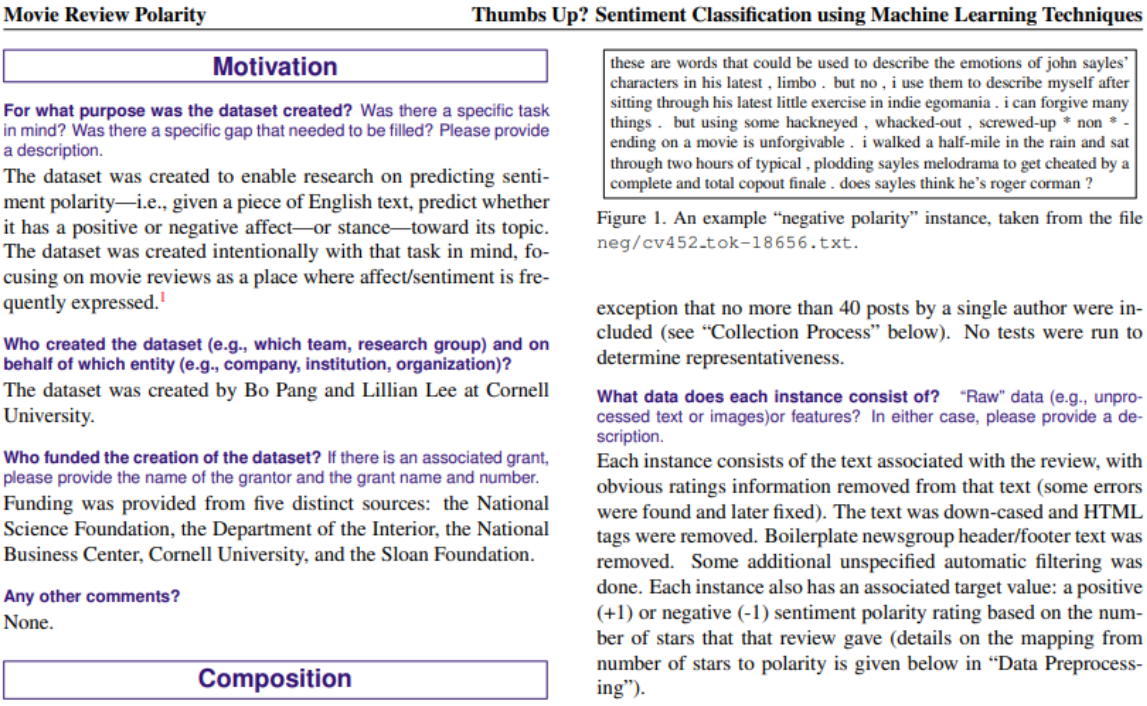
3.2. Add contextual information about the datasets

Recommendation 4: provide details about the dataset provenance

Understanding how a dataset has been generated, and how it is being maintained and managed by the data owner, is relevant for potential data consumers in order to understand how the data may be reused.

A good source of this type of information is what is commonly known as a **data management plan** (DMP), a document that is frequently used in research projects and initiatives, and which describes the lifecycle of the data that is used and/or produced in the context of scientific research (see Thuermer, G., 2020, for example). A DMP includes information on how the data has been collected, processed and published, how data quality is assured and how personal data is treated (see more in recommendation 5). In the different sections of a DMP, dataset owners describe any treatment, preparation, filtering or transformation done in the data, or any procedure used to obtain the samples.

In the context of artificial intelligence, another relevant example of dataset documentation is datasheet for datasets (Gebru, T. et al., 2021), proposed with the intention of reducing the bias in machine-learning models and incrementing their reproducibility. The information about the motivation, composition and collection process associated with a dataset is represented in a datasheet (see example in Figure 6).



Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Figure 6: Example extracted from the publication.

We recommend **including enough information about the provenance of the dataset**, potentially including a DMP or a completed datasheet.

Recommendation 5: add ethical information, including how data protection is handled

When a data provider decides to publish a dataset, those fields containing personal data may need to be dealt with properly. That part of the data may need to be removed before publishing it or be (pseudo-)anonymised or obfuscated (e.g. to hide locations of endangered species so as to protect them).

When publishing datasets that contain personal data (e.g. the full name of a person), a consent form may also be needed. Having all this information associated with the dataset provides sufficient details to potential data reusers on how the data has been treated, generating more confidence in its reuse.

Examples of the type of information that may be included are identified in the ethical canvas ([10]), proposed by the Open Data Institute, as shown in Figure 7. They created this canvas ([11]) to help users identify ethical issues in your data. This information could be included as an attached file similar to a DMP.

([10]) https://theodi.org/article/the-data-ethics-canvas-2021/
([11]) https://docs.google.com/document/d/10NXybDF6oFAP1dMNehqEbIeOIX_JLEP8PAejs3oOZ-M

Figure 7: Data ethics canvas from the Open Data Institute.

We recommend adding enough **information** to the dataset **to describe the process that has been followed in order to ensure that ethical aspects are considered, including the protection of personal data**. This may include the templates of the consent forms used to collect the data, the description of the (pseudo-)anonymisation or obfuscation process applied before publishing the data and a filled-in ethical canvas like the one proposed by the Open Data Institute.

3.3. Provide examples of data usage and link datasets to other digital artefacts

Recommendation 6: describe existing or potential use cases for the dataset and examples of data consumption

Datasets published in open data portals already provide basic textual descriptions and keywords. However, this is normally not enough for data consumers to know the situations in which this data has been or can be used/applied, which may be relevant to better understand the potential of the dataset. A good example of this recommendation can be found in the UK Office for National Statistics ([12]), where we can find stories and data linked to stories (as shown in Figure 8).

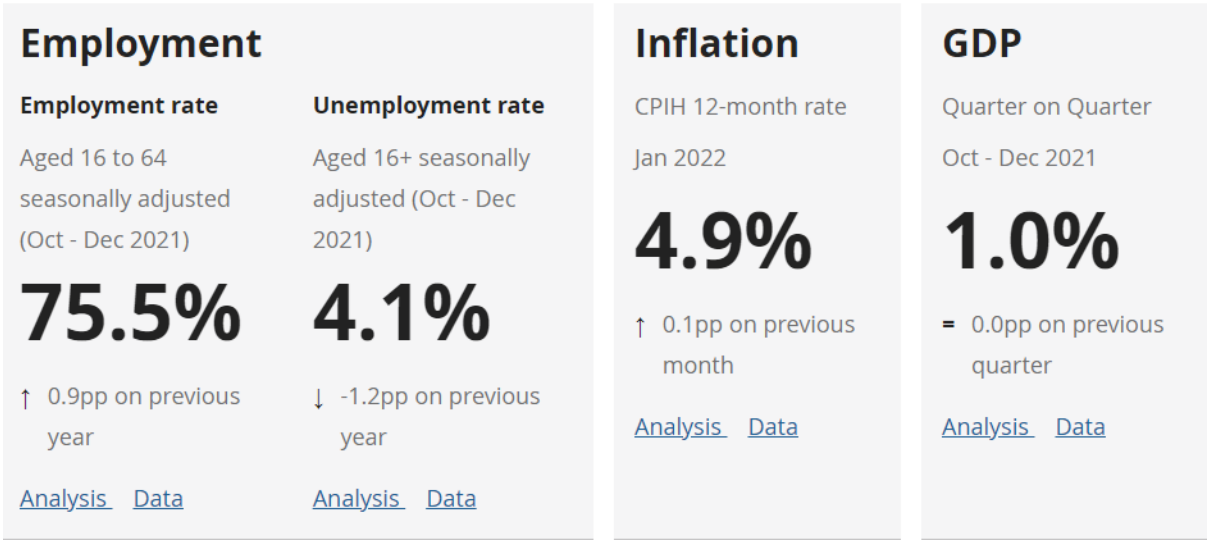


Figure 8: Data stories associated with one or several datasets.

We recommend **providing a list of examples, use cases and data stories** where the data has been used and, when available, provide simple snippets that illustrate how other consumers make use of the datasets (e.g. in Kaggle ([13])). This action will help users discover the potential of the data and better understand whether it can be used in their projects.

([12]) https://www.ons.gov.uk/
([13]) https://www.kaggle.com/

Recommendation 7: link datasets to other digital artefacts that make use of them

In the scientific context, it is common to see how scientific papers (the ones that are published by researchers in conferences and journals) are now being increasingly accompanied by the datasets, software and any other materials that have been used in the context of the research, with the aim of facilitating reproducibility of the experiments. For example, journals such as PLOS One allow users to upload data related to publications ([14]).

Initiatives such as Papers with Code ([15]) are also very relevant in this context. In this case, entries in this online database – which is so far mostly focused on machine learning and artificial intelligence advances – are not only pointing to the datasets and software used in some experiment that is reported in a scientific paper, but they also make the effort to describe such experiments in simpler terms so that non-scientists can replicate the experiments if they wish.

In the European Open Science Cloud, there are initiatives that are trying to systematise the way in which all these interrelated artefacts (code, papers, diagrams, charts, etc.) are being published and made available together. For instance, a well-known initiative is the research objects ([16]), which are understood to be aggregations of all the digital objects associated with a scientific experiment, such as papers, software, datasets and presentations. Some formal specifications to represent these aggregations and platforms (e.g. ROHub ([17])) are appearing to provide support to this new form of publication of research results – with the hypothesis that publishing all these results in this manner will facilitate reproducibility and increase reuse, thus improving impact.

Therefore, if we look at this from a dataset perspective, and from the point of view of an open data publisher such as data.europa.eu, we recommend **collecting the works (scientific or not) where a dataset has been used**, with clear **links to all the artefacts** that make use of such data, and providing all this information in association with the dataset. If possible, these links should be maintained automatically, something that would be facilitated by following what is proposed in recommendation 10.

3.4. Facilitate finding and accessing datasets

Recommendation 8: improve the way in which data can be found, beyond current metadata models (search by variables, apply temporal filters, etc.)

It is not always easy for data consumers to find datasets in a portal that contains a very large set of them, as is the case with intermediaries like data.europa.eu. Open data portals allow different types of searches to take place, which take elements of the used metadata model (e.g. data catalogue

([14]) https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231266
([15]) https://paperswithcode.com/
([16]) https://www.researchobject.org/
([17]) http://www.rohub.org/

vocabulary (DCAT)) into account. New dimensions, identified as relevant in our brainstorming and hands-on sessions, are the variables of the data (e.g. the data columns we referred to in recommendation 1) so that requests such as 'give me the datasets that contain the variable temperature' would be possible. Another relevant search feature would be to consider the temporal dimension on the temporal coverage of the dataset, and the spatial dimensions on the spatial coverage. These dimensions arecommonly used, for instance, in earth observation projects and applications, as shown in portals such as the Global Earth Observation System of Systems ([18]) (see Figure 9).

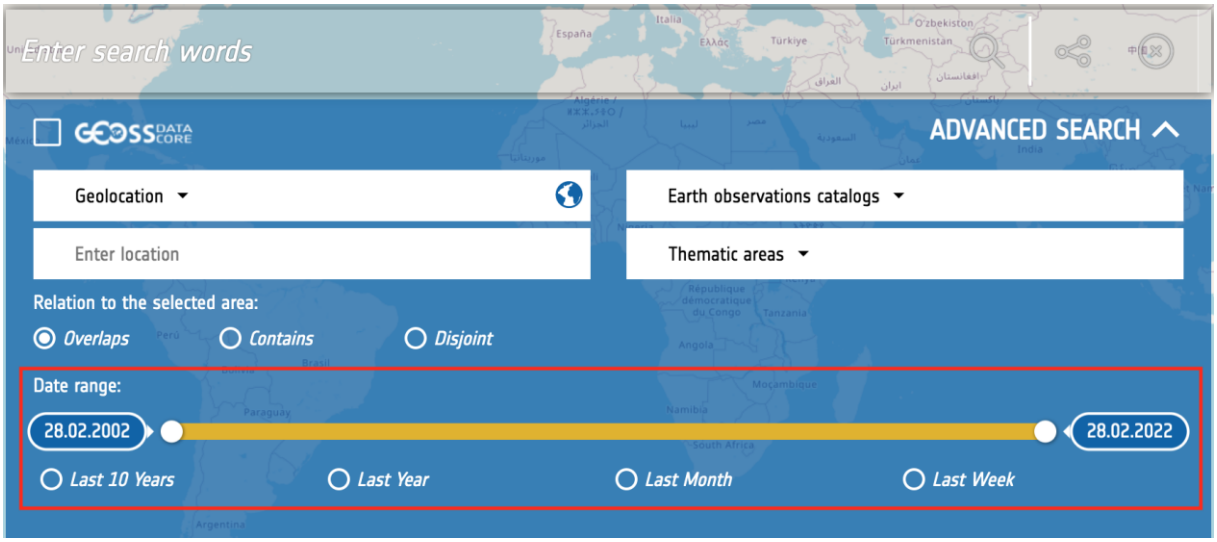


Figure 9: Temporal search dimension in the Global Earth Observation System of Systems platform.

We recommend including search functionalities that allow users to **search by the variables** present in a dataset, and offering search functionalities associated with the geospatial and temporal coverage of a dataset.

Recommendation 9: enable programmatic access to datasets

As discussed in a previous recommendation, a common practice for a potential data consumer, when interested in a dataset, is to download the full dataset and open it locally on their computer. This is not a major problem when the consumer has to deal with small datasets. However, for large datasets this is less manageable. Mechanisms such as application programming interfaces, which have been identified in the latest open data directive as a relevant type of data distribution for high-value datasets, facilitate the use of data by introducing features like pagination or filtering. Examples can be found in the Advanced Geospatial Data Management platform ([19]), where users can access satellite data, or in platforms such as Barcelona's Open Data Portal ([20]), where users can access the data via an

([18]) https://www.geoportal.org/
([19]) https://adamplatform.eu/
([20]) https://opendata-ajuntament.barcelona.cat/en

application programming interface and, in some cases, even execute structured query language queries over the datasets ([21]).

We recommend **providing an application programming interface to access large datasets**, therefore transforming the data catalogue into a large (possibly interconnected) database where datasets are not considered in isolation. This facilitates the development of new applications that consume public data (saving money for data consumers) and – with good access control, such as tokens – introduce new ways of deriving indicators about data usage (e.g. number of accesses, user profiles, etc.).

3.5. Facilitate data citation to improve data sharing and tracking

Recommendation 10: generate persistent identifiers for the datasets

A persistent identifier (PID) is a digital identifier that can be used to provide a long-lasting reference to a document, a file or a web page, for example. PIDs allow for the consistency and unambiguous identification of resources and facilitate their citation (Rueda, L., Fenner, M. and Cruse, P., 2017). For example, data.europa.eu generates persistent URIs for all datasets and provides a function that generates the references to such datasets in different formats (EU data citation, the American Psychological Association style, Harvard and Vancouver). Figure 10 is an example of one of the datasets that have been referenced in this report.

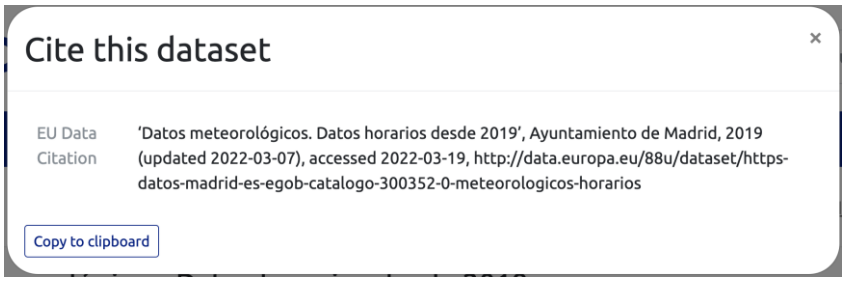


Figure 10: Example of dataset citation in data.europa.eu.

In addition to the use of persistent URIs, there is another type of PID that is commonly used in academic, professional and government contexts so as to facilitate the citation of digital resources such as scientific articles, reports, official publications and datasets. It is called the digital object identifier (DOI), as shown in Figure 11 for a research dataset archived in Zenodo, a service commonly used for the archival of digital resources associated with research in Europe.

([21]) https://opendata-ajuntament.barcelona.cat/en/desenvolupadors

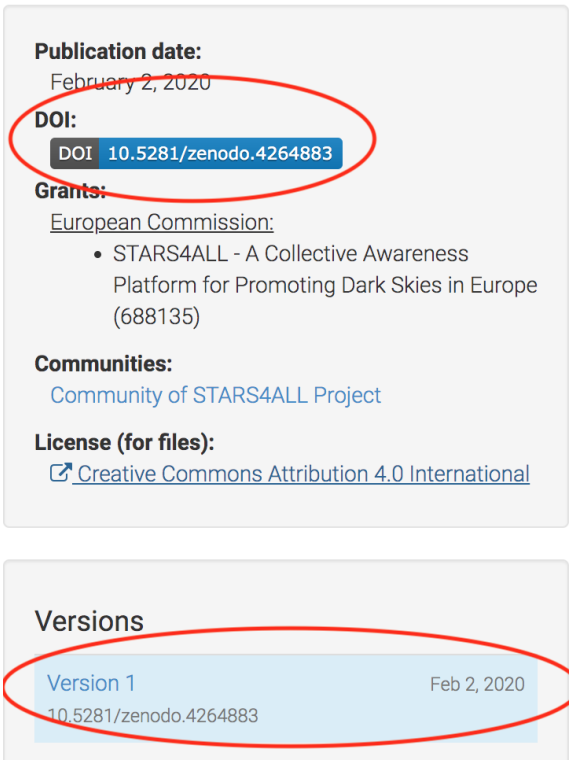


Figure 11: Generation of DOIs in Zenodo.

There are several registration agencies that can produce these references. For example, the Publications Office of the European Union has the rights to produce DOIs for resources from EU institutions, which come in either of the following forms: https://data.europa.eu/doi/<<identifier>> or https://doi.org/10.2906/<<identifier>>. The data.europa.eu portal can generate DOIs on request for the datasets of EU institutions. The following are some examples.

- http://data.europa.eu/88u/dataset/inventory-of-recognised-producer-organisations
  Directorate-General for Agriculture and Rural Development, 'Inventory of recognised producer organisations in the EU's agricultural sector', Publications Office, 2019, (https://doi.org/10.2906/097103114105/1).
- http:/data.europa.eu/88u/dataset/database-of-the-european-energy-storage-technologies-and-facilities
  Directorate-General for Energy, 'Database of the European energy storage technologies and facilities', Publications Office, 2020 (https://doi.org/10.2906/101110101114/1).

This not only facilitates data sharing but it also provides a homogeneous citation, making it easier to track where a dataset has been used and maintaining the same identifier across different versions of the datasets.

We recommend **generating PIDs** for those resources (datasets) that do not already have one and providing the possibility to generate DOIs on request for those that are going to be reused and need to be referenced in scientific publications or other types of reports.

4. Conclusions

The first generation of open data portals were mainly designed to act as repositories where publishers archived their data, together with the corresponding metadata, giving users the possibility to download datasets from them. In fact, most efforts were focused on data providers to push for the provision of good metadata, they were not yet focused on how to encourage the reusability of data by data consumers.

The new generation of open data portals should be more oriented towards data consumers, where they should be able to find additional information about the quality of the datasets, recommendations and how-tos, for example. Nevertheless, we believe that there is much room for improvement in this regard.

We started this report by summarising insights and recommendations from prior works, including work carried out by the data.europa.eu team and our own research in open data management and human–data interaction. Taking these conclusions into account, and our experience in the reuse of datasets and the results of a set of workshops, training courses and other stakeholder engagements, we have created a list of 10 recommendations that would increase the reusability of data to be made available in the next generation of open data portals.

5. References

Corcho, O. and De Pablo, V. (2022), 'Adaptación de estructuras de conjuntos de datos para asegurar su calidad y anonimización – Informe técnico del proyecto Ciudades Abiertas', Zenodo, February 2022. https://doi.org/10.5281/zenodo.5942552

Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M., Choirat, C., Van de Sanden, M. and Coppens, F. (2021), *EOSC interoperability framework – Report from the EOSC executive board working groups FAIR and Architecture*, Publications Office of the European Union. https://data.europa.eu/doi/10.2777/620649

Corcho, O., González, E. and Garijo, D. (2021), 'Deliverable D4.1 data cube metadata model', *Reliance*, Zenodo, June 2021. https://doi.org/10.5281/zenodo.5024537

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H. and Crawford, K. (2021), 'Datasheets for datasets', arXiv, December 2021. https://arxiv.org/abs/1803.09010v8

Koesten, L., Vougiouklis, P., Simperl, E. and Groth, P. (2020), 'Dataset reuse: Toward translating principles to practice', *Patterns*, Vol. 1, No 8, Cell Press, November 2020. https://doi.org/10.1016/j.patter.2020.100136

Rueda, L., Fenner, M. and Cruse, P. (2017), 'DataCite: Lessons learned on persistent identifiers for research data', *International Journal of Digital Curation*, Vol. 11, No 2, the University of Edinburgh, pp. 39–47. https://doi.org/10.2218/ijdc.v11i2.421

Simperl, E. and Walker, J. (2020), *The Future of Open Data Portals*, Publications Office of the European Union. https://data.europa.eu/doi/10.2830/879461

Soylu A.; Corcho, Ó., Elvesæter, B., Badenes-Olmedo, C., Yedro-Martínez, F., Kovacic, M., Posinkovic, M., Medvešček, M., Makgill, I., Taggart, C., Simperl, E., Lech, T. C. and Roman, D. (2022), 'Data quality barriers for transparency in public procurement', *Information*, Vol. 13, No 2, Multidisciplinary Digital Publishing Institute, February 2022. https://doi.org/10.3390/info13020099

Thuermer, G. (2020), 'D1.1 – Data management plan (v2)', *Action*, Zenodo, May 2020. https://doi.org/10.5281/zenodo.3885248