

Data.europa.eu and Citizen-generated Data

Identification of other types of citizen-generated data
sources

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:

Óscar Corcho

Javier Jiménez

Elena Simperl

Last update: 12 May 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, 2023

© European Union, 2023



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/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

ISBN: 978-92-78-43680-3

doi: 10.2830/95762

OA-09-23-329-EN-N

Table of Contents

EXECUTIVE SUMMARY	4
INTRODUCTION.....	6
CONTEXT AND PREVIOUS WORK	6
PURPOSE, MOTIVATION AND MAIN GOALS	8
METHODOLOGY	9
RESULTS AND DISCUSSION	12
CITIZEN-GENERATED DATA ALREADY INCLUDED IN OPEN GOVERNMENT DATA PORTALS.....	12
<i>Participants</i>	12
<i>Data and methods</i>	12
<i>Results</i>	12
CHALLENGES AND OPPORTUNITIES IN INTEGRATING CITIZEN-GENERATED DATA INTO OPEN GOVERNMENT DATA PORTALS.....	17
<i>Participants</i>	17
<i>Data and methods</i>	17
<i>Results</i>	17
ADDING MORE CITIZEN-GENERATED DATA TO OPEN GOVERNMENT DATA PORTALS	20
<i>Data and methods</i>	20
<i>Results</i>	20
DISCUSSION.....	23
CONCLUSIONS AND FUTURE WORK	25
REFERENCES	26
ANNEX A: INTERVIEW QUESTIONS	29
PRIVACY AGREEMENT	29
GENERAL QUESTIONS.....	29
<i>Status of the portal in what concerns CGD</i>	29
<i>Next steps in this context</i>	29
<i>Inclusion of CGD provided by different sources</i>	29
<i>Quality, impact, and integration of CGD with official data</i>	29
<i>Personal comments</i>	30
ANNEX B: LIST OF CITIZEN-SCIENCE PROJECTS.....	30
ANNEX C: CITIZEN-GENERATED DATA ANALYSED	32

Executive summary

This is the second and final report about citizen-generated data (CGD) and open governmental data portals (OGDPs). In the first report (Corcho, Jiménez, Morote, & Simperl, 2022) a literature review was undertaken to define CGD and its main attributes. National and local OGDPs that are in the scope of data.europa.eu were then analysed to see what CGD datasets they already publish. The analysis was summarised in 10 findings, accompanied by 10 recommendations.

The current report starts where the previous left off. To add context to the original findings and recommendations, the authors undertook a mixed methods study with three parts: (1) interviews with representatives of key stakeholder groups in the CGD space; (2) a system analysis of five OGDPs that came out on top in the previous report with respect to the number of CGD datasets published; and (3) a system analysis of national citizen-science portals (CSPs) to find additional CGD datasets that OGDPs could consider adding to their catalogues.

Among the interviewees was a representative of the French OGDP. The French OGDP is among the most advanced among all data.europa.eu portals when it comes to allowing data submissions from citizens. In addition, the authors also interviewed three European researchers active in citizen science, the main community concerned with CGD to date. As they were not able to recruit additional OGDPs to participate in the study, the authors complemented the interviews with a deep dive into CGD activities on OGDPs using the method of system analysis – in addition to France, which took part in the interview study, they analysed the top four portals from the 14 considered in the first report, based on the number of CGD datasets published: Czechia (30 CGD datasets), Helsinki (26), Madrid (14) and Zaragoza (12). Finally, they wanted to seek out existing CGD datasets that are not yet on the radar of national-, regional- and city-level OGDPs: all national CSPs of EU Member States were reviewed to identify citizen-science projects with CGD. The authors then analysed these CGD datasets to identify those which could add value to OGDPs, especially in high-value domains as per Commission Implementing Regulation (EU) 2023/138.

The main takeaways of this follow-up study are as follows.

1. There are four categories of CGD that should be considered for inclusion in OGDPs and data.europa.eu.
 - a. Data collected by public administrations from citizens (surveys, population statistics, etc.).
 - b. Data collected by citizens with the intention to influence policy or trigger government action. This data may be reused by administrations to inform administrative and policy decisions, such as crowdsourced reports of potholes and other problems that need the attention of a local authority. However, it should be subject to increased scrutiny, following best practices from data quality management and data justice.
 - c. Datasets that have been collected for scientific purposes. These may complement existing datasets published by government authorities, such as environmental monitoring datasets from citizen-science projects.

- d. Data collected through feedback mechanisms or automatically logged by portals. This data can provide insights into the information needs of portal users. There are established methodologies for publishing such data in an aggregated, privacy-minded way. Far from being a breach of citizens' privacy, this data can create opportunities to showcase how public authorities respond to user needs, and facilitate user behaviour analyses to improve user experience on the portal.
2. A methodology is needed to decide what CGD to include in OGDPs and how to present them to allow public authorities and others to trust them. While designing such a methodology was not in the scope of this report, the interviews hinted at several important dimensions, drawing on theory and practice from data quality management, data justice and data governance.
3. In areas such as environmental monitoring (e.g. pollution and biodiversity), which have been engaging citizens for many years in collecting and curating rich datasets, it is essential that OGDPs, should they wish to include these datasets in their catalogues, reuse and interoperate with existing data infrastructure in those areas rather than trying to define new formats or publishing practices.
4. As CGD datasets often complement existing datasets that are already listed on OGDPs, it was suggested to provide, alongside the data, tools that allow potential users to understand the commonalities and differences between CGD datasets and other datasets and decide which to use with confidence. Similarly, public authorities should encourage the publication of comparative analyses or commission them themselves for users to be able to take better decisions about which datasets to use and understand the ramifications of those decisions.

Introduction

Context and previous work

This report builds on (Corcho, Jiménez, Morote, & Simperl, 2022) **Error! Bookmark not defined.**, which provided a general analysis of citizen-generated data (CGD) in the context of open government data portals (OGDPs) in Europe. This is the second and final report in this series.

The first report surveyed extant literature to produce a glossary of common terms and a classification schema for CGD in OGDPs. Based on (Meijer & Potjer, 2018) the report defined CGD as ‘the data that individuals consciously generate and that are openly available for use in the public domain’. This definition highlights several attributes of CGD work, which are listed here because they have informed the design of the present study.

1. The data is consciously generated. This includes contexts where a citizen explicitly collects the data, for example by taking a picture of a broken street lamp and reporting it to the city council or uploading their cycling route to generate better open maps. In both cases the citizen submits the data, which is then published alongside other data points, possibly following a range of data processing, cleaning and aggregation steps.
2. In the authors’ view, the definition also includes datasets about citizens, such as survey responses and aggregated footfall datasets generated via smart sensors. The first case is fairly well understood: while the survey is not managed by the citizens, they are in control of submitting the responses and should be informed about the way their data is going to be processed and used. This report acknowledges that the second case is still an emerging area, where public authorities are piloting a range of mechanisms within the boundaries of the law to raise public awareness of the presence of such data infrastructure and the responsible downstream use of the data. While best practices are still under development (de Wijs, 2016), there is broad consensus that public buy-in is essential for smart city data-collection initiatives. In these cases, while citizens may not be directly responsible for recording and submitting the data themselves (like in surveys), they are aware to varying degrees that by using a public space their activities may contribute to the generation of datasets in an aggregated, anonymous form.
3. The definition touches on the use of the data ‘in the public domain’. This refers to licensing, but also to the use of the data ‘with a public purpose such as democratic debate or the development of solutions for public problems’ (Meijer & Potjer, 2018). Again, this study’s notion of CGD is more inclusive than the definition provided by (Meijer & Potjer, 2018). The first report distinguished between primary and secondary data to highlight that a CGD dataset could be the result of an activity that does not have to be explicit about or restrict how the data is used. However, that report also considered the expected policy or operational impact of CGD datasets, though this dimension proved challenging to operationalise in the report’s analysis. As the purpose of a dataset is often hard to reproduce without detailed documentation or access to those directly involved in scoping the data work, this report acknowledges the civic character of some CGD datasets, but does not exclude any CGD datasets for which that aspect could not be established with certainty.

Following a literature review, the first report defined a classification schema with 12 dimensions, divided into:

- portal dimensions (four dimensions: (1) the percentage of CGD datasets from the total number of datasets; as well as the availability of processes, methods, tools, guidance for CGD tasks related to (2) publishing (3) management and use, and (4) quality assurance); and
- dataset dimensions (eight dimensions: general information about area, format, license, whether it is a primary or secondary dataset, whether the data is generated by citizens or is merely about citizens, the roles different stakeholders play in the CGD life cycle, the presence of specific guidance and the expected policy or operational impact).

The report then applied the classification schema to 14 OGDs and discussed key findings and recommendations. These are reproduced here as they will be referred to in the present study.

Conclusions from the first report

- C1.** All OGDs had very few citizen-generated datasets, both in absolute terms and relative to the number of datasets hosted by the portals.
- C2.** In most cases analysed, citizens are mostly involved in generating or collecting the data, but the remaining work required to publish the data is driven by public administrations. The efforts are initiated by public administrations rather than bottom-up by citizens, who are also less involved in curating or maintaining the data.
- C3.** The most frequent areas of CGD published in OGDs are: 'questions and answers', 'surveys' and 'statistics'. This complements the domains commonly covered by citizen-science datasets (Ponti & Craglia, 2020).
- C4.** Primary CGD is more common than secondary CGD in open data portals.
- C5.** Most CGD that is made available on OGDs is shared with open licenses.
- C6.** Almost 30 % of the studied CGD datasets are available in open formats like JSON and XML. A much smaller percentage use proprietary formats, typically XLSX.
- C7.** None of the studied OGDs included documentation about how to contribute and use CGD datasets, nor about specific procedures to ensure data quality in this context. In fact, CGD is not explicitly identified as a data collection approach as such.
- C8.** Most portals do not offer tools to facilitate citizen contributions, either at the level of datasets (uploading their own data) or individual records (changing, curating or maintaining existing data).
- C9.** This report could find no evidence of participatory approaches to designing data pipelines or to collecting and implementing feedback from citizens on broader data strategy.
- C10.** No general guidelines on how to govern CGD datasets in OGDs are provided, which seems to limit the emergence of more of these types of datasets.

Recommendations from the first report

- R1.** Actively seek valuable CGD assets through open calls and partnerships with key citizen-science players such as the European Citizen Science Association (ECSA), national and regional offices in citizen science, and citizen-science projects.

- R2.** Facilitate the discovery of CGD in OGDs by tagging all CGD datasets with a specific tag such as 'CGD' or 'citizen-generated data' (e.g. the portal of the City of Dublin data.smartdublin.ie although the tagging should be made more specific).
- R3.** Include keywords/tags in official languages of the EU to facilitate comparative studies using multiple datasets.
- R4.** Establish procedures to capture CGD processes and data validation methods to increase trust of third-party data users.
- R5.** Extend data and metadata quality capabilities with metrics specific to CGD.
- R6.** Include CGD aspects in upcoming open government data (OGD) reports.
- R7.** Collect new and tag existing use cases from data.europa.eu to showcase the value of CGD datasets (e.g. the French national portal).
- R8.** Link use cases to applications and co-locate tools and documentation to encourage reuse by diverse audiences, including people with varying levels of data literacy.
- R9.** Create tools and applications that consume this type of data and allow citizens to contribute – via data collection or curation – to the original data sources.
- R10.** Allow citizens to contribute information within the portal, allowing not only the upload of complete datasets, but also the addition or maintenance of instances to existing records.

Purpose, motivation and main goals

This second report takes the analysis a step further. While the first report concluded with a series of recommendations for OGD publishers, in this report the authors wanted to understand specific challenges and opportunities relating to including CGD more systematically in European OGDs from the point of view of CGD stakeholders. They also wanted to identify sources of CGD for OGDs and provide more detailed guidance for the integration of CGD in OGDs and data.europa.eu for specific domains.

The study undertaken in this report consisted of three parts.

- First, a series of interviews with CGD stakeholders to highlight their experiences, priorities and recommendations concerning CGD datasets.
- Second, an analysis of those few OGDs which publish some CGD datasets, according to the first report.
- Third, following recommendation R1 from the previous report, seek out CGD datasets that are not yet included in OGDs that public authorities should consider.

The study's findings are for those organisations interested in discovering, reusing or publishing CGD on their portals; in particular, it presents challenges and opportunities associated with including CGD in existing OGD publishing pipelines, drawing on theory and practice from data quality management, data justice and data governance. The authors of this report believe discussing these challenges and opportunities is important to support an informed, balanced debate about the types of CGD datasets public administrations should aim to engage with more, and about the impact (economic, societal, political and ethical) that these data sources could have in European societies. The report provides a curated set of areas in which Europe has already invested in producing CGD datasets, either in the context of citizen science or in participatory policymaking. Some of these CGD are already used by

official agencies, but not all are indexed by OGD. In other cases, there are legitimate concerns around biases and other validity issues in data collection, which call for more research into robust quality assurance and sustainable, fair data governance.

Methodology

As stated in the first report (Corcho, Jiménez, Morote, & Simperl, 2022), the aim of this study is to gather additional insights from OGD publishers and from more established CGD stakeholders such as citizens, scientists either creating or researching CGD and civil society activists. The aim was also to seek out actual CGD datasets that have so far been missed by OGDs.

The first part of the study was for OGD managers. To interview them, a questionnaire was designed using EU Survey. The questions were divided into six categories:

- information about the participant (name, organisation);
- current status of the portal with respect to CGD datasets, including an estimate number of CGD datasets and their importance, details on the methods applied to produce and curate CGD datasets and document these activities, and details on data ownership and governance;
- next steps and planned activities, hinting at different roles citizens could take in the CGD life cycle, and the provision of user-centric tools;
- ways to source additional CGD datasets, which refers to active measures to encourage CGD processes, collaborating with other initiatives already producing CGD such as citizen-science projects, identifying new sources, etc.;
- quality, use and impact of CGD, with a focus on existing quality management protocols, ways in which CGD is or could be used, and links to impact creation in different areas;
- personal views on the role of CGD in OGDs, comments, suggestions, etc.

The questionnaire can be found in [Annex A](#) and at the following [link](#).

The authors reached out to representatives of the 14 data.europa.eu portals surveyed in (Corcho, Jiménez, Morote, & Simperl, 2022) and to leading researchers and practitioners from the CGD field, whose works informed the initial analysis.

To complement the responses, four additional portals were sampled, based on the number of CGD datasets published: Czechia (30 CGD), Helsinki (26), Madrid (14) and Zaragoza (12), and a system analysis (Bentley, 2007) was applied to complement the insights from the interviews.

The authors of this report also contacted the lead authors of seven papers selected in the first report (page 10) and followed up with other authors when these authors were not available. For this second group of interviews, they followed a similar protocol, but focused on questions 4 to 12, as the first three questions were only for participants who run OGDs. Interviews were analysed thematically for cross-cutting themes, with the aim to compile a list of challenges and opportunities and a list of additional sources of CGD which OGD initiatives might not be aware of.

The answers from the semi-structured interviews were analysed question by question, with emerging themes across questions guiding the subsequent system analysis of the additional portals to confirm findings or add new perspectives.

Finally, the authors looked for CGD datasets that could be added to OGDPs. They followed the following five steps:

1. define topics;
2. select portals with CGD datasets;
3. search for datasets on those portals;
4. analyse datasets with respect to license, publisher, publication dates and updates;
5. create a list of CGD datasets that have an open-data licence and a known publisher, and that are regularly maintained, and map them to categories from data.europa.eu.

The first two steps will now be explained in greater detail to allow others to replicate and reproduce this methodology. Steps 3–5 are the results of the analysis and are hence discussed in the next section.

Define topics

As the aim is to identify useful CGD datasets that complement data already available on OGDPs, the starting point was the topics for high-value datasets (HVD) established in December 2022 in Commission Implementing Regulation (EU) 2023/138. This document lays down a list of specific HVD alongside arrangements for their publication and reuse, and will enter into force 16 months after publication. The topics match the domains identified in Directive (EU) 2019/1024. The list consists of the following topics.

1. **Environment.** Air, climate, emissions, nature preservation and biodiversity, noise and waste.
2. **Meteorological.** Observation data measured by weather stations, climate data, validated observations, weather alerts, radar data, numerical weather prediction (NWP) model data.
3. **Companies.** Basic company information, company documents and accounts.
4. **Statistics.** Tourism flows in Europe (yearly and monthly), population, fertility, mortality, national accounts – GDP main aggregates (yearly and quarterly), national accounts – key indicators on corporations, national accounts – key indicators on households, government expenditure and revenue, consolidated government gross debt (yearly and quarterly), poverty rate, inequality rate, employment (yearly and quarterly), unemployment (yearly and quarterly) and potential labour force.
5. **Geographical.** Administrative units, geographical names, addresses, buildings, cadastral parcels, reference parcels and agricultural parcels.
6. **Mobility.** Transport networks as set out in Annex I to Directive 2007/2/EC and inland waterways.

From this list, only a subset of datasets are core to this report's analysis. Companies, statistics, geographical and transport network datasets tend to be published by government authorities. While there are some initiatives to enrich, update or even recreate some of these datasets by enlisting the help of citizen volunteers, in most cases, official publishers are aware of these initiatives and are often establishing their own bespoke processes to integrate the additional data that citizens provide into official data releases. For example, for geospatial data there are community projects such as OpenStreetMaps and OpenAddresses. For company data, OpenCorporates curates an open database

with information about more than 200 million companies from several countries. For statistics, there is a range of private data providers that use citizen sensing rather than citizen-science approaches and aggregate datasets on, for example, footfall or traffic from mobile devices, wireless network signals, etc. (Datarade.ai, 2023).

The authors' research on CGD datasets aims to highlight less-known datasets, which tend to be published on scientific data portals and are thus less known to OGD publishers. These tend to be in areas related to the environment and weather (Spasiano, Grimaldi, Braccini, & Nardi, 2021). For example, the Horizon 2020 project 'Action' produced several open environmental datasets related to regional pollution concerns (air, noise, waste, water and biodiversity) across Europe (ACTION, 2022). Given the high number of environmental citizen-science projects in the EU, this report will consider all relevant keywords to search for CGD: environment, air, climate, emissions, nature, preservation, nature preservation, biodiversity, noise, waste and water. Light was also added as a keyword, as there are many citizen initiatives in Europe addressing this form of pollution even if light is not mentioned as such in the list of environmental HVDs. An example for meteorological datasets is the initiative of the European Space Agency, which launched the app Camalot in the summer of 2022. Citizens use the app to record small variations in satellite signals, and the data is used to train machine-learning algorithms that analyse weather patterns. In this category of HVD, this study looked for the topics pertaining to observations and weather alerts, along with radar and satellite data and NWP model data.

Select portals

Now that a list of topics for CGD had been obtained, the next step was to put together a list of locations, i.e. portals, repositories and catalogues, where one could find such data. This is not a trivial task; both across Europe and internationally, there are several platforms where citizen-science data is published – the equivalent of data.europa.eu does not exist (yet). In addition, many citizen-science projects do not publish data, and if they do, project data is often made available as dashboards, visualisations, maps, etc., which are more accessible to diverse audiences than the typical formats found in OGDs. Projects that publish data often do not use portals, but make the data available on their individual websites. Where portals are used, they host a variety of datasets, commonly scientific datasets alongside citizen-generated or citizen-science datasets. This is because the citizen-science community overlaps with the open science one, and hence follows the same practices and uses the same tools as professional scientists.

To identify a list of popular, cross-initiative portals, this study started from the most well-known citizen-science hubs and from prior studies that produced lists of citizen-science projects, tools and technologies. It looked first at the [eu.citizen-science list of platforms and networks](#) and selected those entries which include links to relevant datasets or European citizen-science projects which may have published datasets on their own. These are [Scistarter](#), a globally acclaimed online citizen-science hub, and the national citizen-science portals (CSPs) in [Belgium and the Netherlands](#), [Czechia](#), [Germany](#), [Spain](#), [Austria](#), [Slovenia](#), and [Sweden](#). To complement the list of national CSPs, this study also surveyed a comprehensive list provided by the [Austrian Center for Citizen Science](#) and found an additional four Member State CSPs from [Denmark](#), [Ireland](#), [France](#) and [Italy](#). The list of portals hence contains 12 project catalogues, which link to various environmental and weather projects. Because of the large number of projects, the authors surveyed at most five projects per country. Not included in the analysis in the light of the resources available are some popular biodiversity portals, such as: the

[National Biodiversity Data Centre Ireland](#) (172 datasets); [eBird](#) (among the world's largest biodiversity-related science projects, focused on ornithology); and the [Global Biodiversity Information Facility](#) (the world's largest specified occurrence database, with around 2 000 CGD datasets). Also not included in the analysis are open-science portals such as Zenodo and OpenAire, as they do not provide easy means to identify datasets generated by citizens.

Results and discussion

Citizen-generated data already included in open government data portals

Participants

The study was carried out from June to September 2022. From the 14 portals reached out to for interviews, only one, the French OGD, contributed to the study. As noted earlier, the authors then decided to do a system analysis for the remaining four leading portals themselves, leading to a total number of five OGDs.

Data and methods

The analysis covers five OGDs. Information was obtained via:

- one interview with a representative of the French OGD;
- system analysis of four additional portals: Czechia, Helsinki, Madrid and Zaragoza, which were in the top five in the previous report with respect to the number of CGD datasets published (the fifth was France).

Results

As the first report concluded, citizen generation constitutes a recent interest in OGD (conclusion C1). This means that even the best-performing portals are still exploring the best ways to integrate CGD more systematically in their open data strategies and practices, and to produce specific guidance and tools (conclusions C7-C10). The presence of CGD datasets in the five portals suggests that they are considered as a potentially useful source of information (question 1). In all cases, there is evidence that CGD is part of a wider citizen engagement strategy (questions 2–3). When it comes to the French OGD, the team invests considerable effort and resources into fostering the creation and use of CGD, though the number of CGD datasets published compared to the total number of datasets remains low. There are ambitions to change this status quo in France (questions 6–8), including a new citizen initiative accelerator (see Figure 5).

The interviewee noted an increase in public awareness and their abilities to produce useful data, and in the use of CGD by companies (question 10). They also mentioned COVID-19 as an example of a situation in which the authorities reached out to the public to collect more and better data (questions 6 and 10). Other examples include: for Helsinki, surveys and sets of questions and answers, which are the dominant categories of CGD across the 14 portals analysed previously (conclusion C3); crowdsourced accessible maps on the Czech OGD; and the Aprende tu barrio project involving two

Madrilenian districts'. In the latter case (Figure 1), the geospatial dataset is entirely produced, curated and administrated by citizens and local communities to provide evidence for social policy (questions 2–7).

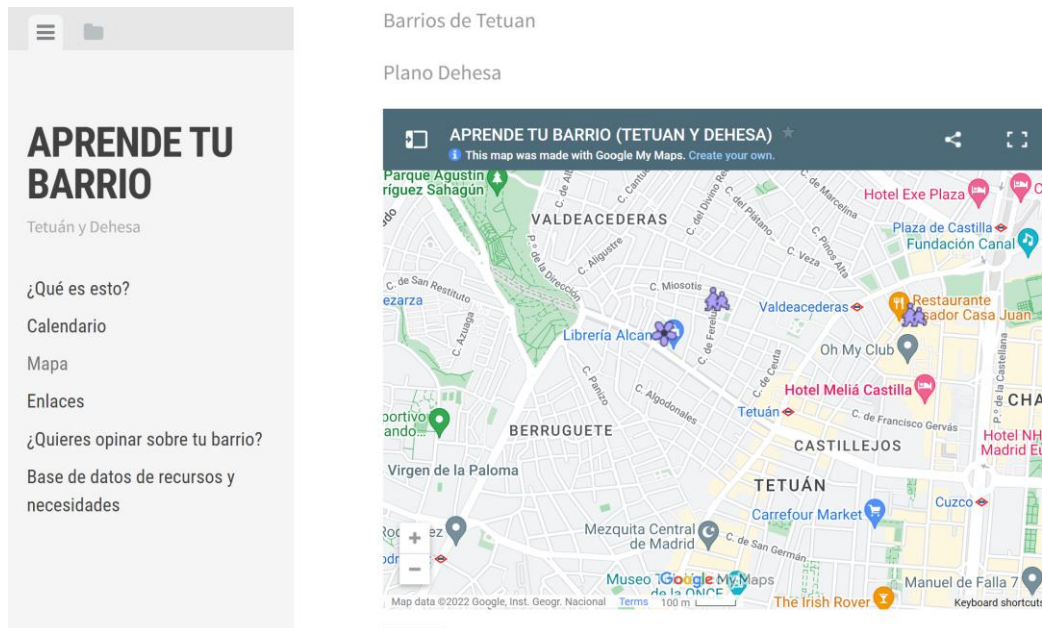


Figure 1: Example of a citizen initiative to produce and curate geospatial CGD in Madrid.

Best practices on how to encourage data contributions from the public are still under development. A portal should prioritise citizen empowerment and allow citizen data publishers to upload their own data without restrictions and decide on key questions around license and exploitation (Ada-Lovelace-Institute, 2021). It should support community building, including tools such as discussion forums and data requests (Kacprzak, 2019), along with transparent, accountable means to identify the best initiatives proposed by citizens and help with funds and promotion. Figure 2 shows the page on data requests of the Czech portal; its equivalent in Madrid is depicted in Figure 3 (question 5). Further on, Figure 4 provides a snapshot of open-data-related discussions on the French portal (question 5), while Figures 5 and 6 show the front pages of the French [Accélérateur d'Initiatives Citoyennes](#) initiative and its counterpart at the city level in Madrid (question 6).

Časová značka	Rádla byste	Jednoznačně identifikujte dataset, u který Vám jde.	Vyberte, ve kterých oblastech by se podle Vás měla kvalita datasetu zlepšit:	Pomůže nám, pokud požadavky na zlepšení dokážete blíže specifikovat:	Kdo je gestorem zadaného datasetu?	Je podle Vašeho názoru možné Vaše žádání splnit realizováním i bez přispění gestora?	Pomozte nám, prosím, odhadnout potřebnost tohoto datasetu a jeho cílovou skupinu.	žáky data
11/25/2016 14:43:10	získat dataset, který dosud není publikován	http://dataface.policie.cz/oblasti/1000000	přístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat), API, průhlednost vyhledávání, listing všech záznamů	Seznam v článku: https://www.tomas-dvorak.cz/po-dobrotena-data-kadobtena-wai-ai/	MPČR	Ano, například https://pcr.tiborak.net/out.com	Každý, kdo kupuje nebo pronajímá vozidla, pracuje s RZ, sdílí údaje se záměrem na vyhledávání spolek, webovou aplikaci na vyhledávání epoků, připojení ma a konkrétních vyhledávacím spolek, projekty vyhledující státní adresy dopravy aj.	Data + nevyř. požád.
11/25/2016 14:52:37	vyhledat existující dataset	Dataset jízdních řádů ČR - http://ftp.dopr.cz/	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat), průhlednost vyhledávání)	Podle legislativy by měly být všechna data ve stejné formě, ve skutečnosti data již mnohde transparentně dopravy jsou jen v PDF formátech.	Ministerstvo dopravy ČR	Ne		
11/25/2016 15:33:00	vyhledat existující dataset	http://www.mlsr.cz/oblast/	přístupnost datasetu (omezení přístupu z jedné IP adresy apod.)	Omezení na veřejná lozic neposkyt za dat, neexistence stamp	MPČR obira	Taovněky ano, ale MPČR by mělo číst sam	MEGA NEJVIC	
11/25/2016 15:36:21	získat dataset, který dosud není publikován	https://www.sof.co/cz/oblast/1000000	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Chybí IC, neexistuje export dat (jen screenshot), neexistují stará data za minulá léta	SZF	konkrétně ano, ale SZF by se měl číst sam	střední až nízká	Data + nevyř. požád.
11/25/2016 15:47:41	získat dataset, který dosud není publikován	endopdata.mlr.cz	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Rozhraní CEDR-III není pro běžnou veřejnost jednoduše ovladatelné	MPČR	Ano, za znalosti SPARQL a datových definic: CEDR-III	Vysoká	Data + nevyř. požád.
11/25/2016 15:56:23	získat dataset, který dosud není publikován	endopdata.mlr.cz	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Přístup a CEDR-III je náročná pro běžnou veřejnost	MPČR	Ano, se znalostí datového modelu a jazyka SPARQL	Vělnost, dostupnost nutná	Data + nevyř. požád.
11/25/2016 15:56:24	získat dataset, který dosud není publikován	endopdata.mlr.cz	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Přístup a CEDR-III je náročná pro běžnou veřejnost	MPČR	Ano, se znalostí datového modelu a jazyka SPARQL	Vělnost, dostupnost nutná	Data + nevyř. požád.
11/25/2016 15:56:26	získat dataset, který dosud není publikován	endopdata.mlr.cz	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Přístup a CEDR-III je náročná pro běžnou veřejnost	MPČR	Ano, se znalostí datového modelu a jazyka SPARQL	Vělnost, dostupnost nutná	Data + nevyř. požád.
11/25/2016 15:56:26	získat dataset, který dosud není publikován	endopdata.mlr.cz	kvalita dat (neprobíhá data, nepřístupnost datasetu (omezení přístupu z jedné IP adresy apod.), metadata ("popis" dat))	Přístup a CEDR-III je náročná pro běžnou veřejnost	MPČR	Ano, se znalostí datového modelu a jazyka SPARQL	Vělnost, dostupnost nutná	Data + nevyř. požád.

Figure 2: Data requests on the Czech OGD.

- Propuestas recibidas
- Proponer un conjunto de datos**
- Los conjuntos mejor valorados
- Aplicaciones
- Periodismo de datos
- Informar sobre una aplicación realizada
- Registro de reutilización

Proponer la publicación de un nuevo conjunto de datos

← Volver

Por favor, **utilice este formulario EXCLUSIVAMENTE** para proponer la publicación de **nuevos conjuntos de datos abiertos** relativos a servicios del Ayuntamiento de Madrid.

Su petición será revisada y publicada en el apartado de "Propuestas recibidas", si constituye un aporte de información nueva para este portal.

NO utilice este formulario para realizar **sugerencias o reclamaciones** sobre los servicios que presta el Ayuntamiento. Esas sugerencias o reclamaciones las puede realizar **aquí**.

Los campos marcados con **★** son obligatorios.

01 02

01. Formulario de envío

DATOS DE CONTACTO

Nombre y apellidos **★**

Correo electrónico **★**

Organización

Sexo (uso estadístico)

Hombre


Mujer

Figure 3: Data requests on the Madrid OGD. Requests are then voted on as in Figure 5.

Discussions (325) + Start a new discussion

Discussion between the organization and the community about this dataset. Discussion creation ▾

On se réveille Copy permalink ↗


 Stephane Gaillard
June 7, 2022

Bonjour ,
Vu le caractère particulier de cette épidémie et le fait qu'on pourra prendre des mesures contraignante le moment venu, il me parait normal de continuer à publier.
Aussi , nous sommes mardi matin et des données du vendredi précédent , ca commence à dater.
Je peux comprendre qu'elles ne soient pas publiées le WE mais il me serait agréable que les publication continuent (à votre rythme).

Autre remarque , les hopitaux savent quel patient à le covid , mais ils savent aussi quand le patient a été vacciné.
De façon étonnante aucune donnée concernant le status vaccinal (ex nombre de mois du dernier rappel) n'a jamais été publiée.
On dira , ca fait un gros fichier , si on y ajoute une tranche d'age...
Je ne pense pas être le premier à penser à ces notions.
Il n'est pas compliqué de comprendre que depuis fin 2021 avoir ces données présenterait un certain intérêt que je laisse deviner.

Reply >

Bug dans les comptages des entrées à l'hôpital par classe d'âge ? Copy permalink ↗

 Jean-Christophe DUTHOU
April 10, 2022

Bonjour,

Le fichier (1) covid-hosp-txad-age-fra ne me semble pas cohérent avec le fichier (2) covid-hosp-ad-age quand on compare les admissions à l'hôpital par tranche d'âge.
Dans le premier fichier, les plus de 90 ans sont sur-représentés (presque deux fois plus que dans le deuxième fichier).
Le deuxième fichier est par contre cohérent avec un troisième fichier DREES cette fois (3) <https://data.drees.solidarites-sante.gouv.fr/explore/dataset/covid-19-resultats-par-age-issus-des-appariements-entre-si-vic-si-dep-et-vac-si/information/>

J'ai loupé quelque chose ou bien il y a un bug dans (1) ?

Merci.

Reply >

Figure 4: Dataset discussions on the French OGDG.

 **L'Accélérateur d'Initiatives Citoyennes**
Le service public augmenté

Le programme d'accélération ▾ Les premières initiatives lauréates La boîte à outils Foire Aux Questions

 **La première promotion de l'Accélérateur est constituée !**
Sélectionnées par un panel citoyen et l'administration, les initiatives citoyennes retenues seront accompagnées pour quelques mois.

[Découvrez les initiatives lauréates](#) [Suivez nos actualités](#)

 Coaching et stratégie de croissance du projet	 Accès aux données, outils et briques techniques de l'État	 Mise en relation avec les bons contacts dans l'administration	 Conformité technique et juridique
 Distribution de la solution et terrains d'expérimentation	 Promotion et valorisation du projet	 Solutions de financement et aide au montage juridique	 Mise en réseau avec les lieux et communautés d'innovation

Figure 5: Encouraging citizen initiatives in the public sector in France.

The screenshot shows a web interface for citizen initiatives. At the top, there's a navigation menu with 'Propuestas recibidas' highlighted. Below it, a list of categories includes 'Proponer un conjunto de datos', 'Los conjuntos mejor valorados', 'Aplicaciones', 'Periodismo de datos', 'Informar sobre una aplicación realizada', and 'Registro de reutilización'. The main content area is titled 'Propuestas recibidas' and contains a sub-section 'Envíenos su propuesta' with a button 'Envíenos su propuesta'. Below this is a search bar 'Búsqueda de propuestas'. A pagination bar shows 'Total: 283' and 'Mostrados: 1-50'. The main proposal is titled 'Uso de aparcamientos Madrid (PARCIALMENTE PUBLICADA)' with '25 votos'. It includes a summary, reception date (08/10/2022), proposer (Solicitante), and status (PARCIALMENTE PUBLICADA). A 'Detalle:' link is also present.

Figure 6: Similar citizen initiative in Madrid.

In relation to the identification of new sources of CGD that might be useful for European OGDs (question 8), the interviewee suggested feedback from data users. Specifically, the French administration is starting to add feedback pop-ups at the end of all electronic procedures that can be carried out via their web pages, and also collecting information on dataset use by citizens. This way, they are creating a potentially useful database with the purpose of improving their online services, making them more efficient and user-friendly by means of analysing data produced in a real environment. This is in line with previous work by the European data portal, which has long argued for re-designing OGDs from a user-centric point of view (Simperl, 2017). There is a rich body of methods and studies in dataset retrieval, including work funded by data.europa.eu, which has shown how such data could inform the design of OGDs, including (Kacprzak, 2019), (Ibáñez, 2022). Data requests, as implemented by three of the four portals analysed (Czechia, France and Madrid), provide feedback from the public and other data reusers on topics and sources of new datasets, which could be produced and curated with citizen participation.

While all portals analysed have documented quality control methods (question 9), evidence of bespoke methods for CGD could not be found. However, the French OGD supports discussion around datasets, which sometimes touches on quality issues (see bottom example in Figure 4). Overall, this is a risk that emerged from the unstructured interviews in the next section.

It was difficult to answer questions 10 and 12 with the data and methods applied. The Compair project (ECSA, 2022) listed on the ECSA website could be a good example of how citizen science may produce useful data for public administrations (question 11). It engages citizens in studying weather conditions, particularly air quality, in real time at thousands of locations, complementing the coverage and granularity of data produced, for instance, by environmental agencies. A similar initiative is run in Valencia with the [VLC per l'aire](#) project, which aims to collect data to inform and influence environmental policy (Azorín Chico, González Galindo, & Raga i Domingo, 2018). In the final part of this study an analysis of citizen-science datasets was carried out starting from the ECSA website to suggest

a core set of topics of CGD that could complement official data, provided they reach a comparable level of quality and have clear protocols for maintenance, governance and use.

Challenges and opportunities in integrating citizen-generated data into open government data portals

Participants

The interviews were carried out in the summer of 2022. All authors listed on the seven papers which were analysed in (Corcho, Jiménez, Morote, & Simperl, 2022) were reached out to. Three authors agreed to be interviewed; these authors co-authored four of the seven papers.

Data and methods

The analysis is based on three semi-structured interviews (questions 4 to 12 in Appendix A). A list of challenges and opportunities is provided across the three interviews. Related to opportunities, a list of CGD projects which may prove useful for OGDs is provided. The interviewees are referred to as P1, P2, and P3 and no further information about them (e.g. affiliation, age or gender) is provided in order to maintain anonymity.

Results

During the interview with P1 several aspects of current CGD initiatives in Europe were discussed, such as the volume of CGD available and the potential impact that CGD could have on policy. However, the main part of the interview touched on an important characteristic of data work, which is particularly pertinent in CGD: its biases and the implications that specific choices in the data collection and publication process have in economic and social terms. While all datasets are affected by such choices (Gitelman, 2013), in CGD the processes followed are often ad hoc or follow emerging practices, and there is much less transparency, accountability and scrutiny of the results (Roman, 2021). As such, there are questions around the use of such CGD in decision-making if it is meant to serve, intentionally or otherwise, a particular agenda. While these concerns apply to all CGD, they are critical when CGD covers topics directly relevant to policymaking or which can influence elections.

This was a theme that emerged strongly from all three interviews. Misinformation and disinformation are some of the great challenges of our times. Publishing CGD alongside official data should ensure that standards of quality are maintained. A staged approach, starting with domains where the CGD can be cross-checked against complementary sources of data, potentially in domains that are not greatly debated publicly, could allow public authorities to gain an understanding of existing practices in citizen communities and co-design bespoke processes and tools with the appropriate level of transparency, accountability and quality (Kosmala, 2016). This view complements ongoing research in the area of data justice (Taylor, 2017), which may provide a useful framework of analysis to come up with a methodology to decide which CGD to include in OGDs and how to improve and maintain them. Data justice is generally concerned with how people are made visible, represented and treated as a result of their production of digital data. Surprisingly, none of the interviewees touched upon biases in datasets that may negatively impact the citizen subjects of some datasets. In addition to the fact that this reinforces existing inequalities, those who are missing from a CGD dataset are also invisible for any use and value generation, and hence once again disadvantaged (Milan, 2020).

The interview with P2 focused on how to develop new partnerships between citizens and public administrations, with data work being seen as a means to encourage different parts of the population to get involved in local policy. In this context, the purpose of data collection is to influence policy change. Citizen concerns become visible to decision-makers, in line with the first pillar of the data justice framework (Taylor, 2017). However, as P2 argued, sustained citizen engagement is only possible if the impact of citizen data work is clearly visible and communicated (Michels, 2010) and the benefits of data collection are shared equitably (Taylor, 2017); this is also related to the question of rewards and data ownership, which P3 also touched upon.

Just like P1, P2 saw the quality of CGD as a critical issue from several points of view: the need for data literacy training for citizens to empower them to generate better data, the perception of some official data publishers that CGD cannot be trusted, and the brittleness of the CGD data collection pipeline, which may be subject to adversarial attacks, trolling, etc. P2 and P3 both recommended involving researchers or practitioners to moderate CGD projects to ensure that the collected data fulfils quality standards (e.g. that the data is representative and valid) (Bird, 2014), without compromising citizens' autonomy in co-designing the initiative and its main activities. P2 mentioned bespoke online training as a means to improve citizen data literacy. These considerations fall into the second pillar of Taylor's data justice framework, which is concerned with engagement with technology (Taylor, 2017).

In addition to what was raised by P1 and P2, P3 commented on data provenance and infrastructure while providing many useful examples of CGD projects that are already used by established data publishers in those fields. Often, official datasets are created with the participation of citizens, but are not labelled as such, noted P3, which also resonates with two of the recommendations from the previous report (R2–3). Such opaque data flows impact the use of CGD in several ways: first, by virtue of the way they are produced and the number of participants involved, data flows in CGD datasets can be more complex, which makes it harder to consider accountability and fair use (Dencik, 2022). One of the advantages of joining together CGD and OGDs is a potential higher use of CGD, as suggested by P1, and flagship EU initiatives such as data.europa.eu could play an important role in promoting the datasets they index. However, citizens can be unaware of the use of the data they generated, which raises questions around the appropriate level of access and licenses for such datasets. If misuse cannot be prevented or policed, then the case that CGD should be made widely available, possibly under an open license (see conclusion C5 from the former report), does not universally hold for every type of data and domain currently present in the surveyed OGDs. P3 brought up areas such as the environment and biodiversity and emphasised that in those areas the priority should be to integrate the new CGD datasets with already available data, adopting the same quality assessment standards and governance formats. P3 also commented on the impact of CGD projects, which is unlikely to be accurately predicted from the beginning, so it is important to keep an open mind and continue to fund and oversee these projects in the medium term, as impacts start to materialise. This also resonates with concerns brought up by P1 and P2, who seemed disillusioned about the short-termism and unreasonable expectations they experienced in some of the CGD projects they took part in..

Table 1: Summary of challenges, opportunities and dataset domains.

Challenges	Opportunities
Data politics: data is collected to pursue a particular agenda (P1, P3).	Citizen empowerment: engage citizens in policymaking and local decisions (P1, P2, P3).
Ambitions vs reality: citizen initiatives are challenging to set up and sustain and often remain at the level of good intentions without clear sustainability and data governance plans (P1, P2, P3).	New data: publish data on issues that matter to people, adding relevance to OGD initiatives that have lost momentum (P1, P2, P3).
Sustained citizen participation: rewarding participation, engaging citizens in data stewardship, communicating the impact of citizen data work in policymaking and local government initiatives (P2, P3).	Better data quality: CGD can reach areas outside public spaces and provide better coverage and granularity (P3); collaboration with researchers is often useful to achieve and demonstrate quality (P2, P3).
Trust in data: demonstrating and documenting that CGD is useful and valid (P2, P3), identifying and mitigating biases (P1, P2, P3), designing robust data pipelines (P2), documenting data ownership (P3).	Higher use of existing CGD: OGD and data.europa.eu could make CGD initiatives widely available to enhance research and analyses (P1).
Data literacy: providing accessible ways to teach citizens the fundamentals of data work, including quality assurance (P2, P3).	Impact: CGD can drive political and societal change and bring about behavioural change (P2).
	Synergies with other data initiatives: established CGD initiatives have developed sophisticated data infrastructures in their fields, which could be reused for open government data; the European open science cloud and data spaces are areas to explore (P3).
CGD domains recommended by interviewees	
Environmental monitoring, including different forms of pollution (air, water, soil) and pollutants (noise, smell, pesticides).	
Biodiversity.	
Crowdsourced geospatial urban datasets, such as accessibility maps, public safety and use of public spaces.	

Surveys, such as household spending and energy consumption.

Adding more citizen-generated data to open government data portals

Data and methods

The analysis was done in February–March 2023. As noted earlier, finding CGD datasets is a challenging, tedious process because of the lack of data discovery infrastructure and the limited use of fair and open science practices among some citizen-science initiatives. Topic-wise the authors focused on environmental and weather datasets, which they sought to find on portals and websites of citizen-science projects listed on Scistarter and the national citizen-science hubs of 11 Member States – the 12th portal was not online at the time of the analysis. The keywords were those corresponding to the two HVD categories of environment and weather: air, climate, emissions, nature preservation and biodiversity, noise, waste (for environment) and observations, weather alerts, radar, satellite and NWP (for weather). The keywords were translated to match the languages of the citizen-science hubs of the Member States. For Scistarter, results were filtered by location (best filter option: Europe). Results across Scistarter and the national citizen-science hubs may include duplicates. Results for individual searches may include duplicates as well, as some projects may be matched to several keywords (e.g. environment and noise).

Results

Table 2 lists the number of citizen-science projects found on Scistarter and the Member State citizen-science hubs matching the chosen keywords. A total of 1 872 such projects were found. While some national hubs are richer than others in terms of projects listed, and there is no additional information on how complete the data is, a first observation is that a cluster of countries such as Germany, Spain and France are much more advanced in terms of citizen-science initiatives than the rest of the sample surveyed (Figure 7). Furthermore, as shown in Figure 8, most projects belong to the environment category, with topics such as nature, biodiversity and water pollution. Concerning weather data, the most popular topic seems to be related to observations.

Table 2: Number of citizen-science projects on the 12 sites surveyed

Keywords	Scistarter	BE/NL	CZ	DE	IE	ES	FR	IT	AT	SL	SE
Environment	89	5	4	0	0	127	95	0	15	13	1
Air	12	24	1	0	4	3	4	1	12	3	2
Climate	66	0	0	42	0	23	4	0	12	0	3
Emissions	4	0	0	0	0	0	1	0	1	0	0
Nature	75	10	0	89	0	39	83	0	13	0	0
Preservation	4	0	0	0	0	1	35	0	1	5	0
Nature preservation	0	0	0	0	0	0	99	0	13	1	0
Biodiversity	62	2	1	49	6	112	58	16	11	0	0
Noise	10	1	0	0	1	6	2	0	2	1	0
Waste	5	0	2	0	3	0	4	0	2	1	0

Water	75	24	2	26	4	36	22	11	3	3	0
Light	23	6	0	0	0	0	1	0	7	0	0
Observations	81	2	11	0	0	1	65	0	21	6	0
Weather alerts	0	0	0	5	0	0	0	0	3	0	0
(Radar) satellite	11	1	1	0	0	20	0	0	2	0	0
NWP	0	0	0	0	0	0	0	0	3	0	0

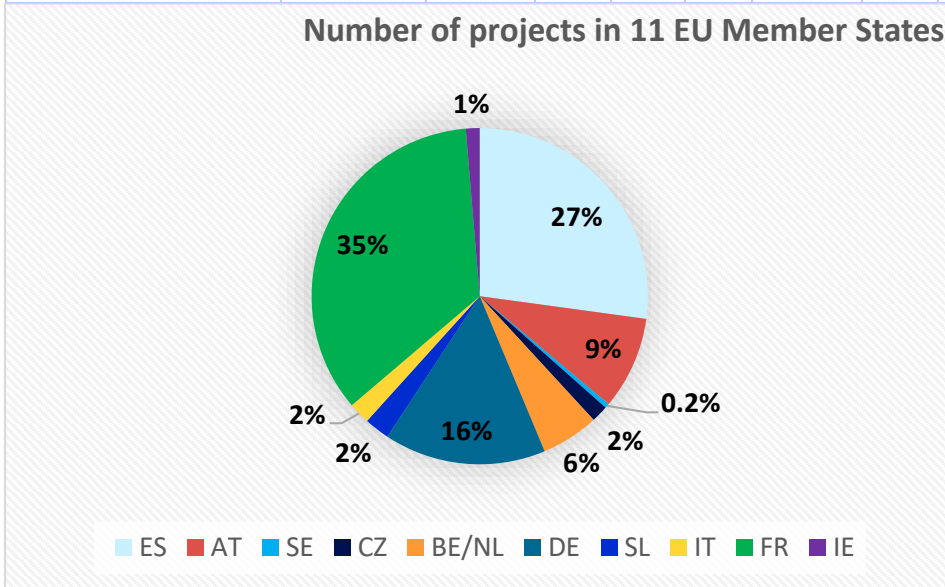


Figure 7: Citizen-science projects in 11 Member States

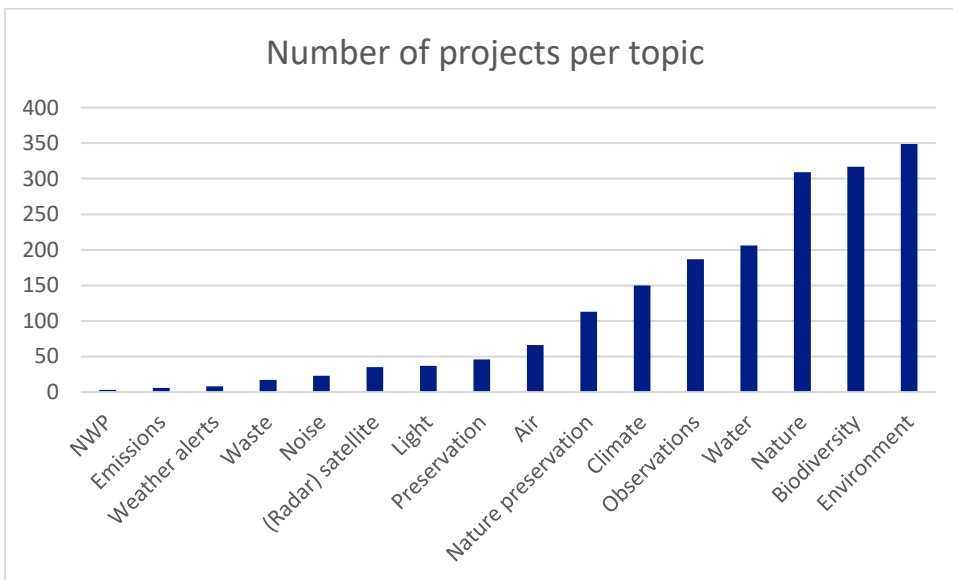


Figure 8: Topics across all sites surveyed

Across these 1 872 projects, the top five projects returned on each site per topic were then selected. Duplicates were removed, which were projects that matched more than one keyword in the search,

and the end result was the 53 projects from Annex B: List of citizen-science projects. Among them, 49 had produced at least one dataset, but in four cases the data could not be located. From the remaining ones, after removing the few instances where the dataset link did not work or the site was under maintenance, the result was a list of 47 CGD datasets. These are listed in Annex C: Citizen-generated data analysed.

14 of the 47 CGD datasets had an open license and 20 had a closed license. In the remaining 13 cases no license information could be found. From the 14 open CGD datasets, some had share-alike constraints, which are not compatible with the definition of open government data, which does not impose such constraints but encourages commercial exploitation.

It was impressive to see that 40 of the CGD datasets were up to date, and from the remaining seven only two provided no information about maintenance, with three publishers mentioning that the data will not be maintained in the future.

In terms of publishers, almost a quarter of the datasets were published by citizens, therefore entirely bottom-up, and in only two cases were the publishers difficult to ascertain. The remaining 33 CGD datasets were published by a range of organisations and initiatives, which can be contacted if the data were to be included in an OGD.

Mapping these 47 CGD datasets to the data.europa.eu dataset categories can be done in multiple ways. This report opted for an inclusive approach, taking all categories that applied rather than deciding on a ranking. Most datasets belong to the environment category; this was to be expected given the bias in environmental, rather than weather, citizen-science projects. However, within this category, some datasets belong to health and agriculture, fisheries, forestry and food, and pertain to policies at different regional or city levels.

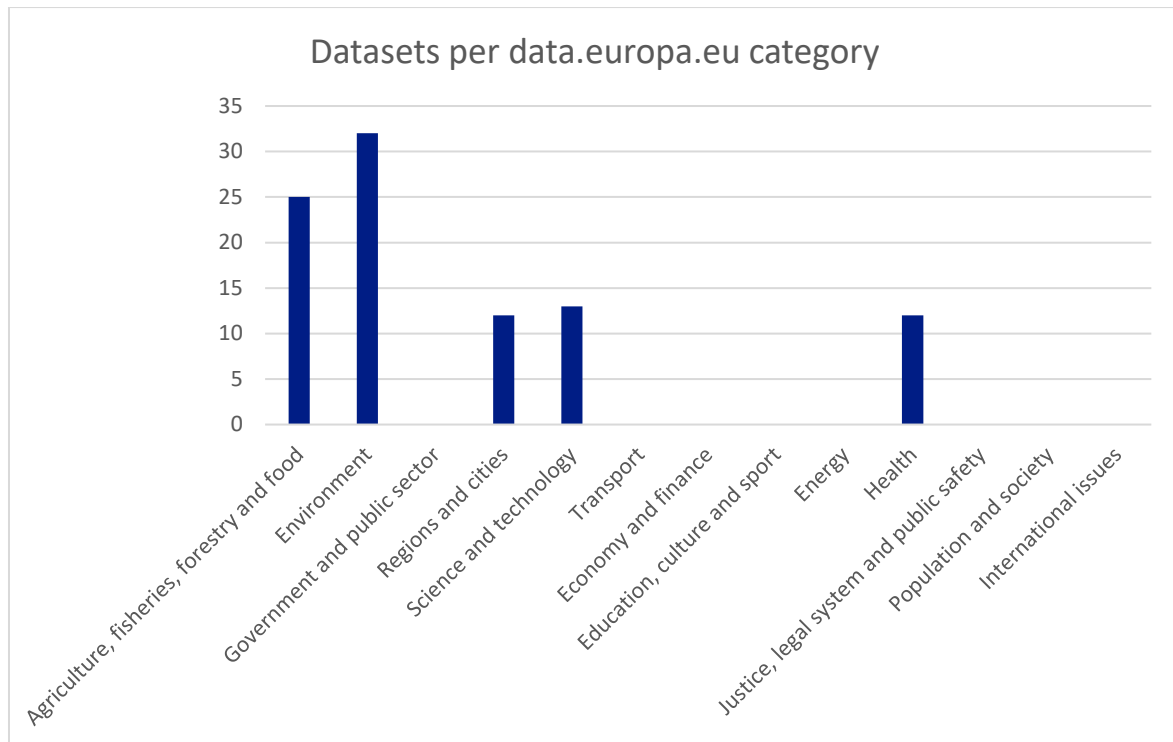


Figure 9: CGD datasets and the data.europa.eu categories they belong to, multiple categories possible

Discussion

This study combined insights about CGD activities linked to OGDs in the scope of data.europa.eu and opinions of CGD stakeholders directly involved in citizen projects in various domains with research into these projects

Even among the five portals analysed, it is clear that citizen engagement with data publishing can take many different forms. While some recommendations from the first report remain, there are promising best practices around data requests, discussion forums and using data collection as a tool to bootstrap citizen participation in policy decisions, which other OGDs could follow.

There were concerns around the purpose of CGD collection and the ways in which CGD initiatives could be hijacked to serve opaque agendas. The classification scheme from (Corcho, Jiménez, Morote, & Simperl, 2022) does not consider such aspects in depth. That report discussed primary and secondary data, along with data by or about citizens, but only considered data quality and biases at a high level. While data justice offers a useful framework to think about equitable data collection and use, the interview participants did not raise major concerns about equality of access, representation and participation, especially relating to already marginalised communities. There is a considerable body of work on data and algorithmic fairness (Garcia, 2016), which adds an extra dimension to the ongoing conversation around selection or coverage biases in CGD datasets and the robustness of CGD processes.

There was consensus about the benefits that CGD datasets may bring to European societies, given that they can be produced continuously and almost everywhere. However, it is important to remember

that this type of data still presents challenges with respect to the control of its bias, acceptability and trust. As with any other major source of data, it is challenging to manually control its results, but finding mechanisms to establish trust, for example by documenting uses, producing testimonials, commissioning quality assessments and audits, is essential. Data justice may provide a useful framework to consider these challenges in OGDs in the future.

Respondents seemed to suggest that at least for some categories of CGD – such as those created with the explicit purpose of driving policy change – there needs to be mechanisms to capture and assess the full context of data collection and ensure transparency, accountability, fairness and the rule of law. For other CGD – especially when it is already used or vetted by scientific organisations or public bodies – they felt that there are opportunities to establish more collaborations and explore synergies, including regarding the use and integration of different data infrastructures. Where information cannot be collected in the public space, public administrations are reliant on citizen participation (e.g. for surveys), and CGD projects in which citizens actively decide on goals and data tasks can lead to healthier public engagement with established institutions and stronger democratic societies in the framework of the European Union.

In terms of technologies, the interviewees suggested tools to support comparative analyses both between official datasets and complementary CGD datasets and across datasets from different regions and countries to spot commonalities and differences and link existing practices to data qualities. This is particularly pertinent as the final part of the study highlighted numerous datasets collecting similar or complementary (environmental) observations with the help of citizen volunteers.

One of the main findings of this final part of the study must be how tedious CGD discovery remains. The scope of the analysis was limited by the manual nature of the process, starting from finding citizen-science initiatives producing datasets and then analysing the datasets to understand their current status. The authors opted for a top-down methodology starting from national citizen-science hubs. An alternative could have been starting from open-science repositories like OpenAire and Zenodo, though those do not allow for the easy discovery of dataset artifacts created by or with the help of citizens.

The chosen approach to finding citizen-science projects was useful to an extent, as most projects identified had produced at least one dataset. However, among these CGD datasets, the majority would not be directly usable because of license restrictions or because the license is unknown. A follow-up study could aim to investigate why CGD datasets are licensed the way they are, and whether the possibility of including a CGD dataset on an official OGD may be incentive enough for publishers to reconsider the terms of use. However, with 25 % of the CGD datasets being published bottom-up by citizens, the question of participatory data stewardship (Ada-Lovelace-Institute, 2021) to ensure the sustainability of the CGD datasets, whether within an OGD or elsewhere, becomes paramount.

This analysis did not include an assessment of the quality of the CGD datasets – this is outside of the scope of the authors' work as in most cases it requires domain-specific methods and expertise. However, most datasets were maintained and up to date. This is encouraging and indicates that these datasets are used or considered useful by their publishers, as otherwise the effort to continue to maintain the datasets would hardly be justified.

Another area where methodological innovation is needed is finding similar datasets: when provided with a dataset, such as one of the 47 CGD datasets analysed in this study, what is the best way to find data.europa.eu datasets that are related in terms of location and topic? Keywords and facets provide a basic way to filter, but what is needed are algorithms that can compare two heterogeneous datasets to decide whether they aim to measure or represent the same things.

Conclusions and future work

The aim of this study was to revisit the conclusions and recommendations from the previous report through interviews with OGDs and other CGD stakeholders. Besides the domains listed in Table 1, potential new sources of CGD to be included in OGDs can be classified as follows.

1. Data collected by public administrations from citizens (surveys, population statistics, etc.).
2. Data collected by citizens with the purpose of influencing policy or triggering government action. This data may be reused by administrations to inform administrative and policy decisions, such as crowdsourced reports of potholes and other problems that need the attention of a local authority. However, it should be subject to increased scrutiny, following best practices from data quality management and data justice.
3. Datasets that have been collected for scientific purposes. These may complement existing datasets published by government authorities, such as environmental monitoring datasets from citizen-science projects.
4. Data collected through feedback mechanisms or automatically logged by portals. This data can provide insights into the information needs of portal users. There are established methodologies to publish such data in an aggregated, privacy-minded way (Navarro-Arribas, 2012), (Samavi, 2018), (Gotz, 2011). Far from being a breach of citizens' privacy, this data can create opportunities to showcase how public authorities respond to user needs, and facilitate user behaviour analyses to improve user experience on the portal.

Of the 47 datasets from Annex C: Citizen-generated data analysed, only 14 had open licenses. These datasets fall into categories 2 and 3 from the list above. Additionally, one could contact the publishers of the 13 datasets with an unknown license to understand whether they would consider releasing the data with a license compatible to the open data definition. To recommend specific datasets for inclusion in OGDs, the next step would be to understand their quality. This information is not documented in a standardised, domain-independent way, and requires expertise in the respective domains. To facilitate this, this report issues the following seven additional recommendations for OGD publishers, based on the earlier discussion.

- R11.** Co-design approaches to decide what CGD to include in OGDs and how to present it to allow public authorities and others to use them with confidence. For instance, the Horizon 2020 project 'WeObserve' developed a [toolkit](#) to set up citizen observatories in environmental monitoring, including training, resources and best practices in co-design methods, data collection, validation, analysis, evaluation, and advocacy with public sector stakeholders. Other examples can be found in the projects funded by the European Commission in a dedicated [Horizon Europe call](#) on the 'Uptake and validation of citizen observations to

complement authoritative measurement within the urban environment and boost related citizen engagement', with projects expected to start in 2023.

- R12.** Reuse and interoperate with existing data infrastructure in areas where CGD is an established source of official data. This will require a follow-up analysis of the infrastructures underlying the portals listed in Annex B to identify common software architectures and products that are in use in the high-value domains, along with closer ties to standardisation efforts in the citizen-science community towards metadata and exchange formats. Relevant stakeholders include the [working group](#) on projects, data, tools and technology at ECSA.
- R13.** Provide, alongside the data, tools that allow potential users to understand the commonalities and differences between CGD and other datasets and decide which to use with confidence. Such tools would compare a CGD dataset with an existing dataset of a similar scope, possibly published by official authorities, in terms of attributes and common data quality attributes such as level of granularity, geospatial coverage, noise in the data, etc. Depending on the type of data, comparisons could be tabular (e.g. listing attributes side by side), text based (e.g. providing contrastive descriptions of the datasets), or visual (e.g. showing the two datasets on a map or in a time series graph).
- R14.** Invest in scalable CGD discovery tools. Scalable means automated – such tools do not exist at the moment, hence this report's analysis has been performed manually and could only cover a limited number of projects (53) and datasets (47). Building such a discovery tool would be considerably harder than data.europa.eu or Google Dataset Search because CGD datasets do not use standard metadata schemas such as data catalog vocabulary and are published in a decentralised way, typically on the websites of the citizen-led initiatives that produced them.
- R15.** Increase awareness of and provide training on the importance of licenses for reuse when publishing CGD datasets.
- R16.** Work towards metadata standards, with co-located methods, to document the quality of CGD datasets. Such standards will build on established reporting mechanisms in the high-value domains analysed. For instance, for environmental data, the European Environment Agency specifies the metadata for reporting air quality data by its member countries. Furthermore, Commission Implementing Regulation (EU) 2023/138 lays down a list of HVD alongside arrangements for their publication and reuse – for each of those quality dimensions, one would need to invest in validating and documenting the quality of CGD datasets to understand how they compare with alternative datasets published by official authorities.
- R17.** Work on algorithms that, given a CGD dataset, provide a list of related datasets published by public authorities in the same or other jurisdictions.

References

Ada-Lovelace-Institute. (2021). *Participatory data stewardship: A framework for involving people in the use of data*. Ada Lovelace Institute.

Azorín Chico, F., González Galindo, I., & Raga i Domingo, E. (2018). *Citizen-Generated Data from the Valencian Context: Report on the Identification and Characterisation of CGD Initiatives*.

Valencia: Valencia Town Hall, Department of For Open Government, Transparency, Citizen Cooperation and Participation.

- Bentley, L. D. (2007). *Systems analysis and design for the global enterprise*. New York: McGraw-Hill/Irwin.
- Bird, T. J. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173, 144-154.
- Corcho, Ó., Jiménez, J., Morote, C., & Simperl, E. (2022). Opportunities and challenges associated to the inclusion of Citizen-Generated Data in data.europa.eu. *Publications Office of the European Union*.
- David, M. &. (2004). *Social research: The basics* (Vol. 74). Sage.
- de Wijs, L. a. (2016). How smart is smart? Theoretical and empirical considerations on implementing smart city objectives—a case study of Dutch railway station areas. *Innovation: The European Journal of Social Science Research*, 29(4), 424–441.
- Dencik, L. &-M. (2022). Data justice. *Internet Policy Review*, 11(1).
- ECSA. (2022, September 02). *COMPAIR | together for a better air*. Retrieved from ECSA: <https://ecsa.citizen-science.net/cases/compair-together-for-a-better-air/>
- Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, 33(4), 111-117.
- Gitelman, L. (. (2013). *Raw data is an oxymoron*. MIT Press.
- Ibáñez, L. D. (2022). A comparison of dataset search behaviour of internal versus search engine referred sessions. *ACM SIGIR Conference on Human Information Interaction and Retrieval*, (pp. 158-168).
- Kacprzak, E. K. (2019). Characterising dataset search—An analysis of search logs and data requests. *Journal of Web Semantics*, 55, 37-55.
- Kosmala, M. W. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10), 551-560.
- Meijer, A., & Potjer, S. (2018). Citizen-generated open data: An explorative analysis of 25 cases. *Government Information Quarterly*, Vol. 35, 613-621.
- Michels, A. &. (2010). Examining citizen participation: Local participatory policymaking and democracy. *Local Government Studies*, 36(4), 477-491.
- Milan, S. a. (2020). The Rise of the Data Poor: The COVID-19 Pandemic Seen From the Margins. *Social Media + Society*, 6(3).

Roman, D. R. (2021). An analysis of pollution Citizen Science projects from the perspective of Data Science and Open Science. *Data Technologies and Applications*.

Simperl, E. &. (2017). *The Future of Open Data Portals*. European Commission, European Data Portal.

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2).

Annex A: Interview questions

The questions used in the interview are listed below.

Privacy agreement

- These interviews are being recorded, transcribed and published in a written format as an annex to the report D2.1.5.2 of data.europa.eu.
 - Do you agree to be recorded during the interview?
 - Do you want your answers to be anonymized before publication?

General questions

Status of the portal in what concerns CGD

1. Which is the importance of CGD nowadays in your portal?
2. Is the method by which open data was produced/curated publicly explained?
3. Who owns these datasets?

Next steps in this context

4. Is it reasonable to allow citizens playing more roles? Concretely, those of Initiators, Curators, Keepers or Project Administrators.
 - a. If yes, what tools do you consider the most important ones to do it?
 - b. If not, why?
5. Would you be willing to include tools to ease citizen collaboration? Concretely: specific guidelines, cooperation forums, maps, visualization tools, query engines, etc.

Inclusion of CGD provided by different sources

6. Have you considered to actively include – primary, created by citizens – more CGD in the portal you manage? Why?
7. What could be the role of public administrations in citizen science projects? Should this projects – e. g., those available at ECSA – be allocated in open data portals?
8. Could you think of different CGD sources that might be interesting from the perspective of the portal you represent?

Quality, impact, and integration of CGD with official data

9. Is there any quality-control procedure in your portal? Does it specifically contemplate CGD aspects? Should it?
10. Is CGD re-usable? Have you considered to include a reutilizations register?
11. Is CGD used to complement official data?
 - a. If yes, may you offer an illustrative example? What are the outcomes?
 - b. If no, would you find it interesting?
 - c. In what areas is this articulation more convenient? Are those the same we found as the most common ones in open data portals?
12. In those projects with a previously defined impact, is it typically reached? E. g., participative budgets.

Personal comments

13. In your view, is CGD a priority in open data portals? Should it? Do you expect an improvement in the near future?
14. Do you have any comments, suggestions or thoughts you would like to share?
15. Do you have any question back?

Annex B: List of citizen-science projects

List of citizen-science projects (column 'NAME') listed on a citizen-science (CS) portal (column 'PORTAL'). For each project, it is mentioned if the project produced any CGD (column 'ASSOCIATED DATASET(S)') and whether the authors could find it (column 'AVAILABLE DATASET(S)').

ID_PROJECT	PORTAL	NAME	TOPIC	ASSOCIATED DATASET(S)	AVAILABLE DATASET(S)
SCI01	Scistarter	GLOBE Observer:clouds	Environment, satellite	Yes	Yes
SCI02	Scistarter	ISeeChange	Climate, environment	Yes	Yes
SCI03	Scistarter	AirCasting	Air, climate	Yes	Yes
SCI04	Scistarter	NaSuchu	Biodiversity, observations	Yes	Yes
SCI05	Scistarter	GlobeAtNight	Light, observations	Yes	Yes
SPA01	Spanish CS portal	Sinobas	Climate, weather alerts	Yes	Yes
SPA02	Spanish CS portal	OpenTEK	Climate, weather alerts	Yes	Yes
SPA03	Spanish CS portal	Coast Snap Cádiz	Water, biodiversity	Yes	Yes
SPA04	Spanish CS portal	Vigilantes del aire	Air, climate	Yes	Yes
SPA05	Spanish CS portal	Marnoba	Water, waste	Yes	Yes
AUS01	Austrian CS portal	Fire Database	Climate, preservation	Yes	Yes
AUS02	Austrian CS portal	CrowdWater	Water, weather alerts	Yes	Yes
AUS03	Austrian CS portal	Pollen Diary	Biodiversity, observations	Yes	Yes
AUS04	Austrian CS portal	CamalioT	Weather, observations	Yes	Yes
AUS05	Austrian CS portal	Ornitho	Biodiversity	Yes	Yes
SWE01	Swedish CS portal	Lufdata	Air, climate	Yes	Yes
SWE02	Swedish CS portal	Forkskarfredag	Divulagation	No	No
SWE03	Swedish CS portal	Artportalen	Biodiversity	Yes	Yes
CZE01	Czech CS portal	Proč se zabýváme zrovna petrkličí?	Biodiversity	Yes	No
CZE02	Czech CS portal	Lesodiverzita	Biodiversity	Yes	Yes
CZE03	Czech CS portal	NetLake	Water	Yes	Yes
CZE04	Czech CS portal	Fenofáze	Biodiversity	Yes	Yes
CZE05	Czech CS portal	Intersucho	Climate, water	Yes	Yes
BEL01	Belgium CS portal	Satellite Streak Watcher	Light, observations, satellite	Yes	Yes
BEL02	Belgium CS portal	Active Asteroids	Light, observations	Yes	Yes
BEL03	Belgium CS portal	Marsh Explorer	Biodiversity, water	Yes	Yes
BEL04	Belgium CS portal	InfluencAir	Air	Yes	Yes
BEL05	Belgium CS portal	Odour Collect	Air	Yes	Yes

GER01	German CS portal	OUCO Berlin	Climate, weather alerts	Yes	Yes
GER02	German CS portal	TreeChecker	Biodiversity, climate	Yes	Yes
GER03	German CS portal	Schweinswale – bitte melden!	Biodiversity	Yes	Yes
GER04	German CS portal	Mikroplastik auf der Spur	Water, waste	Yes	No
GER05	German CS portal	Die Apfelblütenaktion	Biodiversity, climate	Yes	Yes
SLO01	Slovenian CS portal	Monitoring Metuljev	Biodiversity, climate	Yes	Yes
SLO02	Slovenian CS portal	LifeBeaver	Biodiversity	Yes	No
SLO03	Slovenian CS portal	INCREASE	Biodiversity	Yes	No
SLO04	Slovenian CS portal	CianoSLO	Water	Yes	Yes
SLO05	Slovenian CS portal	Howling	Biodiversity	Yes	Yes
ITA01	Italian CS portal	NOSE	Air	Yes	Yes
ITA02	Italian CS portal	Mammalnet	Biodiversity	Yes	Yes
ITA03	Italian CS portal	i-Rosalia	Biodiversity	Yes	Yes
ITA04	Italian CS portal	Simile	Water	Yes	Yes
ITA05	Italian CS portal	Butterfly monitoring system	Biodiversity	Yes	Yes
FRA01	French CS portal	BioLit	Water, biodiversity	Yes	Yes
FRA02	French CS portal	Phenoclim	Climate, biodiversity	Yes	Yes
FRA03	French CS portal	Oiseaux des Jardins	Biodiversity	Yes	Yes
FRA04	French CS portal	GhostMed	Water, waste	Yes	Yes
FRA05	French CS portal	Faune France	Biodiversity	Yes	Yes
IRE01	Irish CS portal	Hushcity	Noise	Yes	Yes
IRE02	Irish CS portal	National Biodiversity Data Centre	Biodiversity	Yes	Yes
IRE03	Irish CS portal	Irish Whale and Dolphin Group	Biodiversity	Yes	Yes
IRE04	Irish CS portal	CleanAirTogether	Air	Yes	Yes
IRE05	Irish CS portal	River Obstacles	Water	Yes	Yes

Annex C: Citizen-generated data analysed

ID_DA TASET	ID_PR OJECT	LICENCE – OPEN	LICENCE DETAILS	PUBLISHER	UPDATED	MAINTAINED	DATA EU CATEGORY.
SCI01_01	SCI01	No (non-commercial)	https://www.globe.gov/about/policies/terms-of-use	US National Aeronautics and Space Administration	Up to date	Yes	5
SCI02_01	SCI02	No (license transferred to company)	https://www.iseechange.org/terms	Users	Up to date	Yes	1, 2
SCI03_01	SCI03	Yes	https://www.habitatmap.org/airbeam/faq	Users	Up to date	Yes	2, 5
SCI04_01	SCI04	No (non-commercial by default)	https://www.inaturalist.org/pages/terms	Users	Up to date	Yes	1, 2, 4
SCI05_01	SCI05	Yes (CC BY 4.0)	https://creativecommons.org/licenses/by/4.0/	Dennis L. Ward, © Copyright 2007 University Corporation for Atmospheric Research	2021	Unknown	5
SPA01_01	SPA01	No (data belongs to State Meteorological Agency)	https://www.aemet.es/en/nota_legal	Spanish State Meteorological Agency	Up to date	Yes	2, 5
SPA02_01	SPA02	Specific for each contribution, most of them are non-commercial	https://hackmd.io/@opentek/BjJZoZx?utm_source=preview-mode&utm_medium=rec	Local Indicators of Climate Change Impacts	Up to date	Yes	1, 2
SPA03_01	SPA03	It is assumed that Facebook owns the photos	–	Facebook users	Up to date	No	2, 4
SPA04_01	SPA04	Yes (CC BY SA)	–	Users	Up to date	Yes	2, 4, 10
SPA05_01	SPA05	No (non-commercial)	https://marnoba.vertidoscero.com/aviso-legal	Users	Up to date	Yes	1, 2, 10
AUS01_01	AUS01	Unknown	–	Unknown	Up to date	Yes	1, 2

AUS02_01	AUS02	Yes (CC0)	https://www.spotteron.net/terms-of-use	Users	Up to date	Yes	1, 2, 5
AUS03_01	AUS03	No (non-commercial)	https://www.pollenwarndienst.at/ueberuns/pollendaten.html?iframe=0ghjscjm %2527	Users and employees	Up to date	Yes	1, 10
AUS04_01	AUS04	No (user grants license)	https://www.camalio.org/en/terms/	Users	Up to date	Yes	5
AUS05_01	AUS05	Unknown	https://www.ornitho.at/index.php?m_id=42&item=23	Users	Up to date	Yes	1, 2
SWE01_01	SWE01	Yes (open content)	https://opendefinition.org/	Users	Up to date	Yes	2, 10
SWE02_01	SWE02	–	–	–	–	–	12
SWE03_01	SWE03	Unknown	Website under maintenance	https://www.artdatabanken.se/en/sok-art-och-miljodata/terms-and-conditions-for-account-holders-at-the-swedish-species-information-centre-artdatabanken/			1, 2
CZE01_01	CZE01	Unknown	Website under maintenance	https://nurmenukk.ee/sobre-primula/terminos-y-condiciones			1, 2
CZE02_01	CZE02	Seems open, but specifies scientific purposes; it is not completely defined	https://lesodiverzita.cz/mobilni-aplikace	Users and institutions	Up to date	Yes	1, 2, 4
CZE03_01	CZE03	Unknown	–	Botanickém ústavu	2017	Yes	1, 2, 10
CZE04_01	CZE04	Yes, but only for contributors	https://www.fenofaze.cz/cz/chci-byt-pozorovatelem-faq/	The Institute of Global Change Research of the Academy of Sciences of the Czech Republic and the Czech Hydrometeorological Institute	Up to date (yearly)	Yes	1, 2, 4
CZE05_01	CZE05	Yes, Similar to CC BY SA	https://www.intersucho.cz/cz/o-nas/licence-a-odpovednost/	Intersucho	Up to date	Yes	1, 2
BEL01_01	BEL01	Yes	https://www.anecdota.org/pages/terms	Users	February 2023	Yes	5

BEL02_01	BEL02	No	https://www.zooniverse.org/about/faq	Zooniverse	Not defined (they publish results from time to time)	Yes	5
BEL03_01	BEL03	No	https://www.zooniverse.org/about/faq	Zooniverse	Not defined (they publish results from time to time)	Yes	5
BEL04_01	BEL04	Yes	https://influencair.be/sensor-and-data/	Influencair	Up to date	Yes	2, 5, 10
BEL05_01	BEL05	Unknown	–	Ibercivis	Up to date	Yes	5, 10
GER01_01	GER01	No (CC BY-NC-ND 3.0 DE)	https://creativecommons.org/licenses/by-nc-nd/3.0/de/	Freie Universitaet Berlin	Up to date	Yes	1, 4
GER02_01	GER02	Unknown	–	Users	Up to date	Yes	1, 2, 4
GER03_01	GER03	Unknown	–	Schweinswale e.V.	2018	No	1, 2
GER04_01	GER04	–	–	–	–	–	–
GER05_01	GER05	Unknown	–	Research Group Earth Observation (Suedwestrundfunk - SWRR)	Up to date (daily)	Yes	1, 2
SLO01_01	SLO01	No (European Butterfly Monitoring Scheme (eBMS) license)	https://butterfly-monitoring.net/sites/default/files/eBMS%20Licence%20%3D%20Annex%20A%20v%202022%2002%2002.pdf	eBMS	Up to date (each 15 mins)	Yes	1, 2
SLO02_01	SLO02	–	–	–	–	–	–
SLO03_01	SLO03	–	–	–	–	–	–
SLO04_01	SLO04	No (photos are not published)	–	Inštitutu za biologijo	Not defined (they check photos 'as soon as possible')	Yes	1, 10
SLO05_01	SLO05	Unknown	–	Dinaricum	2021	No	1, 2, 4

ITA01_01	ITA01	No (login required)	https://nose-cnr.arpa.sicilia.it/	Nose	Up to date	Yes	4, 10
ITA02_01	ITA02	Yes (CC BY-SA 4.0)	https://creativecommons.org/licenses/by-sa/4.0/	Durham University	Up to date	Yes	2, 5
ITA03_01	ITA03	Unknown	–	Users	Up to date	Yes	1, 4
ITA04_01	ITA04	No (login required)	https://simile.como.polimi.it/SimileWebAdministration/faces/index.xhtml	Unknown	Unknown	Unknown	10
ITA05_01	ITA05	No (eBMS license)	https://butterfly-monitoring.net/sites/default/files/eBMS%20Licence%20%3D%20Annex%20A%20v%202022%2002%2002.pdf	eBMS	Up to date (every 15 minutes)	Yes	1, 2
FRA01_01	FRA01	Yes (but there is no specific license)	https://www.biolit.fr/vos-donnees-d-observation	BioLit (after validation)	Up to date	Yes	1, 2
FRA02_01	FRA02	Yes (CC BY 4.0)	https://phenoclim.org/politique-de-confidentialite-2/	Centre de Recherches sur les Écosystèmes d'Altitude Mont-Blanc	Up to date	Yes	1, 2
FRA03_01	FRA03	No (those who upload data own them)	https://www.oiseauxdesjardins.fr/index.php?m_id=36	Ligue par la Protection des Oiseaux and the Muséum national d'Histoire naturelle	Up to date	Yes	1, 2
FRA04_01	FRA04	Unknown	–	GhostMed	Up to date	Yes	1, 2
FRA05_01	FRA05	Unknown	https://www.faune-france.org/index.php?m_id=22&item=7	Ligue par la Protection des Oiseaux France	Up to date	Yes	1, 2
IRE01_01	IRE01	Unknown	–	Hush city	Up to date	Yes	4, 10
IRE02_01	IRE02	No (specific for each dataset, but most of them have restrictions)	https://maps.biodiversityireland.ie/Dataset/78	The Global Biodiversity Information Facility	Up to date	Yes	1, 2
IRE03_01	IRE03	No (only-use license)	https://iwdg.ie/terms-conditions/	Irish Whale and Dolphin Group	Up to date	Yes	1, 2
IRE04_01	IRE04	Unknown	–	CleanAirTogether	2022	No	4, 10
IRE05_01	IRE05	Yes (but there is no	https://www.river-obstacles.org.uk/about/	River Obstacles	Up to date	Yes	1, 2, 5

		specific license)					
--	--	----------------------	--	--	--	--	--

ISBN: 978-92-78-43680-3



■ Publications Office
of the European Union