Assessment of the value of data and other types of assets in data.europa.eu

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:
Oscar Corcho (https://orcid.org/0000-0002-9260-0753)
Ahmad Alobaid (https://orcid.org/0000-0001-8637-6313)
Elvira Amador (https://orcid.org/0000-0001-6838-1266)

Last update: 11 April 2023
Website: https://data.europa.eu/

Luxembourg: Publications Office of the European Union, 2023

ISBN: 978-92-78-43560-8 doi: 10.2830/192960 OA-03-23-107-EN-N

Contents

Executive summary

Previous research performed in the context of the former European Data Portal explored concepts, methods and architectures to make (open) data portals sustainable. They helped bring about a paradigm shift in the open data community, namely a shift away from understanding portals primarily as a means to publish and discover data and towards rethinking of portals as sources of added-value content and resources that facilitate data reuse and foster data communities.

This report is the first in a series. The series will build on this previous work and focus both on the value of the data and metadata that the data.europa.eu portal holds and on the value of other resources available on the portal (e.g. documentation in the form of reports and publications, tools and data stories). To provide such an analysis, we are developing methods (including computational methods, also known as software prototypes) to:

- assess the value of datasets and their related resources by studying their presence on other digital platforms, tools and applications that are commonly used by data communities;
- add value to datasets and resources by recommending other related datasets and resources and enriching their content to make it easier to use in downstream applications (e.g. in machine learning).

Throughout this report series, our work will result in methods as well as insights, produced through data analyses, literature reviews, user studies (including interviews and workshops) and meta-analysis.

In this, first, report, we focus in particular on methods that allow an understanding to be gained of the presence of data.europa.eu assets in external sources (e.g. GitHub, Stack Overflow and Reddit) that are commonly used by software developers. Our aim is to gain an understanding of the current presence of data.europa.eu content in those communities and to provide reproducible methods that can be used to gather comparable data in the future to facilitate the understanding of trends. Therefore, this report is mostly focused on data collection methods and provides only a preliminary analysis of the data acquired from these external sources.

Some of the conclusions from our initial work presented in this report are as follows:

- So far, the presence of data.europa.eu content on these sites where user communities (related to software, data, etc.) are created is very limited. As a recommendation for the future, to increase uptake it may be useful to reinforce messages for these communities (e.g. by providing code that can be used with a specific dataset together with another dataset or associated with a data story and allowing users to browse data.europa.eu content in those repositories, thereby promoting their use and presence).
- The data that can be obtained automatically from the application programming interfaces (APIs) of these sites are generally not sufficient to sufficiently understand the behaviour of users and the creation of data communities. Further work with additional techniques (literature reviews and user surveys) will be needed in the reports that will follow to better understand such usage.

In the future, these data from external sources may also be combined with the data that are being acquired using web analytics techniques on the use of data.europa.eu, as well as with additional data

about the quality of the metadata of data assets and other indications of the quality of data and other data.europa.eu resources.

1 Introduction

The work presented in this report stems from earlier work done in the context of former versions of the European Data Portal. Such studies (Ibáñez et al., 2020; Ibáñez and Simperl, 2021) focused on the analysis of the behaviour and needs of European Data Portal users when searching for datasets, as well as on the characterisation of the demand of datasets by users (e.g. highly demanded datasets and their categories, the combined usage of datasets and the frequency of downloads). Different types of methods were proposed to provide answers to research questions around these topics, by analysing the data obtained through search and interaction logs collected with the Matomo web analytics platform, which was in place for the European Data Portal.

These analyses helped bring about a paradigm shift in the open data community, namely a shift away from understanding portals primarily a as means to publish and discover data and towards rethinking of portals as sources of added-value content and resources that facilitate data reuse and foster data communities.

In this series of reports, we will go a step further by considering more than just the data that can be obtained through the existing search and interaction logs available through the current web analytics platform (Piwik, now Matomo). Our aim is to take our analysis further and consider the use of data.europa.eu assets (data, documents, tools and stories) on external sites (e.g. GitHub, Stack Overflow and Reddit), that are commonly used by data users and software developers. To do so, we first perform an exploratory study on external sites, analysing how many of the portal contents are referenced there, the context in which they are referenced and the use of its different types of resources. The analysis shown in this report allows the first of our set of research subquestions to be answered (Q1.1 – To what extent is data.europa.eu content (datasets and other related resources) present on external sites?). The complete list of research questions to be addressed in this report series is described in Section 2.

The report is structured as follows. Section 2 presents the whole set of research questions that will be addressed in this series of reports. Section 3 presents the materials and methods, understood as the data collection processes used for the external sites that are considered at this stage, and some preliminary exploratory data analysis to better understand the types of data that have been obtained from those sites. Finally, in Section 4, we provide some conclusions from this initial work and recommendations for future work that will be addressed in the upcoming reports. The annex discusses the methodology that will be used in the upcoming periods to address the rest of the proposed research questions.

2 A comprehensive set of research questions

In this section, we present first the full list of research questions that we will be addressing in this series of reports. These research questions have been derived from the results and conclusions from the aforementioned reports (Ibáñez et al., 2020; Ibáñez and Simperl, 2021), as well as through interactions with the data.europa.eu team. They are divided into four main groups:

- **Q1. Metadata and engagement.** In previous work (Ibáñez, Kacprzak, Koesten, & Simperl, 2020; Ibáñez & Simperl, Analytical Report 19: Understanding Supply and Demand in Dataset Search on the European Data Portal, 2021), questions related to user engagement were addressed using two sources of data: the European Data Portal itself, in the form of analytics collected using Matomo and Google Analytics, and GitHub, as an example of what future-proof, user-community-centric alternatives to current portal architectures may look like.
  In this work, we will go beyond the state of the art by analysing not only these sources (or equivalent sources, such as the current Piwik/Matomo logs), but also other external sources of engagement data. This means that, rather than looking at data collected by the owners of a data portal or intermediaries, we will look at digital platforms, tools and applications in which the data or other related resources may actually be used or referred to (e.g. Reddit, Twitter, GitHub and scientific papers). We will use statistical analysis to understand the relationship between various variables, some related to the datasets themselves and others to engagement, similar to the methods that we used in Koesten et al. (2020). We will need to collect such data from different sources, which means that we will provide several iterations of our analysis throughout the years. The research subquestions will be:
  - Q1.1. To what extent is data.europa.eu content (datasets and other related resources) present on external sites? Presence is understood in a generic manner, including the reuse and republication of datasets on other external sites and the existence of references to datasets on those external sites.
  - Q1.2. To what extent does the introduction of descriptive statistics (average values of columns, the standard deviation, the number of values for a specific type, etc.) on the dataset description page or as part of the metadata affect the presence of datasets on external sites?
  - Q1.3. To what extent does having data stories (or other types of similar content) associated with the datasets affect their presence on external sites?
  - Q1.4. Is there a significant difference between the reuse of old datasets and the use of new datasets? In other words, are old datasets more or less present on external sites than new datasets? Is there a period (e.g. between the last 2 and 5 years) during which the data reuse peaked?
- **Q2. Access and findability.** In this series of reports, we will also update the previous analyses already referenced above (Ibáñez et al., 2020; Ibáñez and Simperl, 2021), as well as those in Kacprzak et al. (2019), with the data that can be obtained from the analysis of social media reporting, web clipping and sentiment analysis, web analytics and content evolution, which is also

being done in the context of data.europa.eu. We will also carry out qualitative user studies to add context to some of the key findings of these prior analyses, including a consideration of the results from the user surveys that will be run in data.europa.eu. We will design a task-based think-aloud study in which we will observe users while searching for data. The research subquestions will be:

  - Q2.1. One of the main insights was that there are probably two groups of distinct users, namely those associated with native searches (which start on data.europa.eu and use the built-in search affordances) and those associated with external searches (which use, for example, a search engine or a conversational platform and are likely to be more interested in having a question answered than in downloading a dataset). What are the typical tasks of these two groups of users? How do these tasks relate to the characterisation that will be done with the results from the user survey?
  - Q2.2. Following on from Q2.1, how could the data.europa.eu site help users who are not looking for a particular dataset? To what extent do data users utilise data.europa.eu as a stepping stone to find other external datasets or sources (e.g. the uniform resource locator (URL) of the official website of an organisation or dataset host)? Are the users who are accessing data.europa.eu interested only in the content hosted on data.europa.eu or do they also use the portal to find other external datasets and sources, based on the findings of Piccardi et al. (2021)?
  - Q2.3. How do users find datasets and other related resources on data.europa.eu (e.g. through search engines, conversational platforms, forums, social media or research papers)?
  - Q2.4. How quickly do users find the dataset or resources that they are looking for on data.europa.eu? How easy it is to find relevant data? Do users adjust their search strategies depending on the results they receive? This is relevant, as previous studies of ours (Kacprzak et al., 2019) have suggested that people expect dataset searches to produce results of very limited precision, with users deliberately expressing their data needs in very broad, abstract terms to make sure that they do not miss any datasets in the results.
  - Q2.5. One of the limitations of the analyses of the data.europa.eu logs carried out so far is that we could not look at the data needs across individual search sessions. From the literature, we know that users often need more than one dataset to meet a data need (Koesten et al., 2017), which translates into multiple, related searches. How do users decide what to look for in subsequent searches? Do they combine dataset searches with searches of other related resources? How could data.europa.eu support complex query needs that span multiple searches and sets of interconnected results (Li, Schijvenaars and de Rijke, 2017)?

- **Q3. Topical analysis and external referrals.** There is a lot of research into the value of open resources and online ecosystems, including resources such as OpenStreetMap and Wikipedia, and of user-generated data platforms. First, we will analyse the existing literature to see whether the conclusions reached for those resources are also applicable to data.europa.eu content. Using scrapping and data application programming interfaces (APIs), we will create a dataset of topics, datasets and external platforms in which they are mentioned and link these data, where possible, to engagement or quality metrics, such as those that are being captured in service 4. Just like for

Q1, we will need to produce iterations of the analysis as we collect more data. We will answer, among others, the following subquestions, whose answers will be presented as guidance to data publishers on how to prioritise future publishing efforts:

  - Q3.1. Which topics, datasets and resources are most commonly shared in research papers and on social media platforms?
  - Q3.2. What is the impact of the quality or reach of the social media post on the reuse on the abovementioned datasets?
  - Q3.3. What is the impact of research papers that reference data.europa.eu datasets (e.g. impact factor/quantile) on the reuse (traffic) of the abovementioned datasets?
  - Q3.4. Which organisations generate the most traffic to data.europa.eu?
- **Q4. Data services and architecture.** Building on our previous work on alternative architectures of open data portals, we will look at specific add-on features that could increase the usefulness of existing dataset retrieval search algorithms implemented on data.europa.eu. We will survey existing recommender algorithms from related areas to see how they apply to the context of datasets, for which we have very limited information about prior dataset searches by the same users. We will consider mostly content-based methods, which are applicable to sparse data, to implement and evaluate dataset recommenders. This answers the following subquestion:
  - Q4.1. Does the offering of suggestions to other relevant datasets or other resources on data.europa.eu increase the reuse of data.europa.eu?

The work presented in this report specifically addresses the first research subquestion (Q1.1). More specifically, we aim to provide insights into the context in which data.europa.eu content is referenced in third-party sources. The rest of the research questions will be addressed in the following reports, as reported in Table 1.

Table 1. Research questions, methodology and expected outputs.

Research question	Methodology	Output	Associated report
Q1: Metadata and engagement	Data.europa.eu data collection from third-party sources and subsequent statistical analysis	Insights regarding the use of data.europa.eu resources outside its homonym platform Software developed	Report 1 (Q1.1) Report 2 (Q1.2, Q1.3 and Q1.4)
Q2: Access and findability	User study regarding access to and the findability of the resources on the data.europa.eu platform	Insights regarding the findability of the data.europa.eu resources	Report 2

Q3: Topical analysis and external referrals	Literature review, data collection and further analysis on the most relevant topics addressed in research papers and on social media related to data.europa.eu	Insights regarding (1) the topics in which data.europa.eu is mentioned and (2) the types of users and organisations actively using data.europa.eu Software developed	Report 3
Q4: Data services and architecture	Literature review on previously existing data portals. Development of a data portal that supports dataset recommendation	A recommender system for datasets, alongside other user-oriented assets. Insights on the benefits that the inclusion of these assets has on the overall usability of the platform	Report 3

This report, therefore, aims to address research question Q1 (metadata and engagement), focusing specifically on Q1.1 (To what extent is data.europa.eu content (datasets and other related resources) present on external sites?). In this context, different third-party sites with a prominent presence of software developers (GitHub, Reddit and Stack Overflow) are analysed. A fine-grained analysis of the presence of the different data.europa.eu assets (data stories and datasets) is conducted, giving an overview of the context in which these resources are employed.

3 Materials and methods

In this study, we analysed how the data.europa.eu resources are present (referenced) on external sites to understand their impact beyond the direct usage of data.europa.eu. This study comprised two general steps: data collection and data analysis. In the first step, data about data.europa.eu references were collected from a set of forums that were considered relevant for this purpose. In the second step, these data were then analysed to extract contextual information on their usage on these sites. This allowed some initial answers to be provided to the Q1 research questions.

All of the source code that was used for data collection and analysis is available on GitHub (Alobaid, Amador-Domínguez and Corcho, n.d.) so that it can be replicated in the future and the results obtained in this report can be updated.

3.1 Data collection

A group of three representative sites for the software development and data reuse communities were selected for this study, with the aim of providing a general overview of how data.europa.eu content is being referenced beyond the limits of the portal. These sites are:

- GitHub (https://github.com/). This site was already considered in one of the previous analytical reports. It provides support to software developers when producing and maintaining software. As of June 2022, GitHub reported having over 83 million developers and more than 200 million repositories, including at least 28 million public repositories. Other alternatives or additional sites that could be analysed in the future include Gitlab (https://about.gitlab.com/) and Software Heritage (https://www.softwareheritage.org/).
- Stack Overflow (https://stackoverflow.com/). This site provides support for collaboratively creating questions and providing answers in a multitude of domains, including software development and data science. As of March 2021, it had over 14 million registered users and had received over 21 million questions and 31 million answers.
- Reddit (https://www.reddit.com/). This site is advertised as a social news website and forum where content is socially curated and promoted by site members through voting. It covers a wider set of domains than the previous two sites. It has 430 million monthly active users (this has been a stable number since 2020) and 52 million daily active users (as reported in 2020).

We also conducted a superficial analysis on the mentions of data.europa.eu on Twitter. Nonetheless, no further information besides the number of mentions of data.europa.eu could be retrieved, as the Twitter API cannot be publicly accessed. Subsequently, Twitter was excluded from further analysis.

Table 2 summarises the data sources that we considered, as well as their availability and the explicit mentions of data.europa.eu content. It must be noted that the data collection process may have been non-exhaustive, as our data collection process may have omitted shortened URLs that did not contain the string 'data.europa.eu' but still referred to it (e.g. bit.ly links). This is a well-known challenge in the collection of references to web content, which may be dealt with more carefully in future reports. Another aspect to take into consideration is that, in our analysis, we also removed those mentions that were related to uniform resource identifiers (URIs) supported by the data.europa.eu domain but that did not refer to the data.europa.eu portal (using the list available at https://data.europa.eu/URI.html).

Table 2. Mentions of data.europe.eu on the different platforms

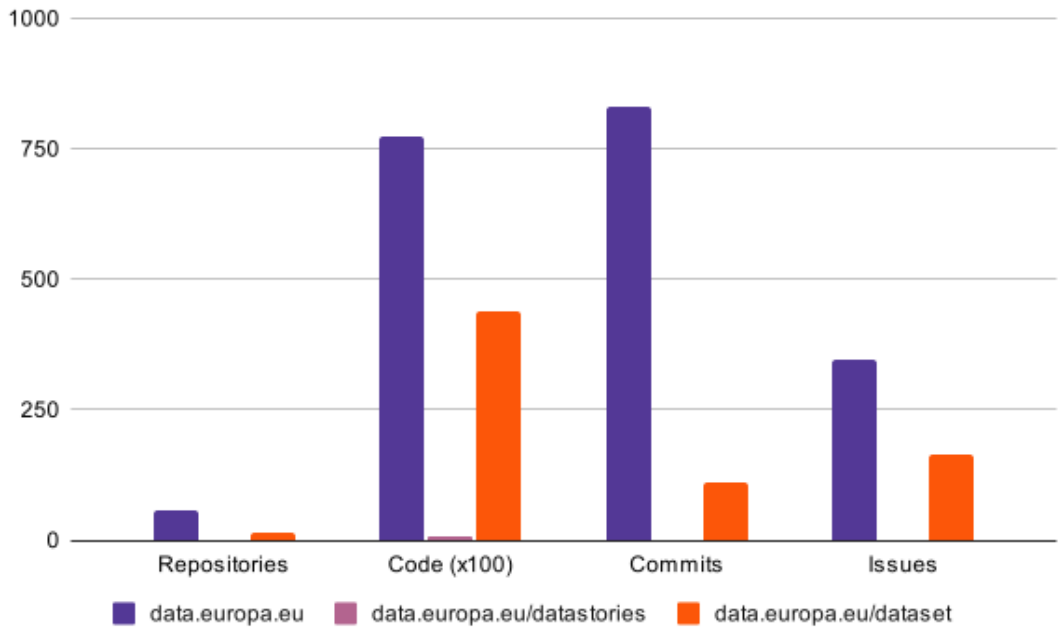
Source	Explicit mentions of data.europa.eu	Data source availability
GitHub	12 repositories 69 000 code files 35 commits	Publicly available Data collection was limited to 1 000 results per hour

Stack Overflow	28 posts	Publicly available Only queries of questions were allowed; therefore, mentions of data.europa.eu in answers may not have been identified
Reddit	—	Publicly available Search was limited to subreddits, not global
Twitter	+1 000 retweets +2 000 favourites	Not publicly available unless a commercial licence is acquired or an academic application is approved

3.1.1 Data collection from GitHub

The list of repositories, code files, commits and issues containing references to data.europa.eu was collected using GitHub's public API. The number of sources containing the different types of assets available on data.europa.eu (datasets and data stories) is summarised in Figure 1. The full list of sources is available at https://github.com/oeg-upm/dataeuropa-analysis/tree/main/data/github. As GitHub allows a limited number of queries per hour, data were harvested in an incremental fashion. Partial results were saved locally, grouped by type (repositories, commits, code files and issues), alongside the reference to the last retrieved element. On each execution of the analysis code, it was checked if there were remaining or new elements since the last retrieved element. If there were, they were first collected and included in the corresponding saved file, then the analysis was performed.

Figure 1. Distribution of data.europe.eu resources on GitHub.



For each of these sources (repositories, code files, commits and issues), the following parameters/attributes could be obtained:

- Repositories. The GitHub API allows repositories mentioning data.europa.eu to be obtained. This search returned only those repositories that included data.europa.eu in their description.
- Commits. The API allows GitHub commits to be retrieved that explicitly mention data.europa.eu. Most of these commits were related to dataset resources. Some examples of GitHub commits containing data.europa.eu were:
    - 'The excel file containing the data analysed. It was obtained from https://data.europa.eu/data/datasets/s2190_90_1_478_eng?locale=en. The data titled 'Link to ebs_478_volume_A_xls.zip'.'
    - 'Top 20 Dataset Platform you can use for practise and Research Purpose
      ---
      1. Google Dataset Search: https://lnkd.in/geXp7AZd

      2. Kaggle: https://lnkd.in/gBGjsRZi

      3. Awesome Public Dataset: https://lnkd.in/gpkzzBcd

      4. Global Open Data Initiative: https://data.europa.eu/en

      5. Google Scholar: https://lnkd.in/g27ZEJrc

      6. Amazon Web Services (AWS) Open Data: https://lnkd.in/gK9V9gB7

      7. UC Irvine Machine Learning Repository: https://lnkd.in/gZKWHgdx

      8. Microsoft Research Open Data: https://msropendata.com/

      9. Academic Torrents: https://lnkd.in/gRqrSBsQ

      10. FiveThirtyEight: https://lnkd.in/gwGPg5js

      11. Stocks data: Alpaca: https://alpaca.markets/

      12. Socrata: https://lnkd.in/g5iw3pmS

      13. Quandl: https://lnkd.in/gsxM3YWN

      14. Data is plural: https://lnkd.in/gtEE6eFi

      15. The World Bank: https://lnkd.in/gwTrWJXd

      16. World Health Organization: https://lnkd.in/grQBCXEG

      17. U.S. Census Bureau: https://lnkd.in/gHFxmGgU

18. UNICEF Data: https://data.unicef.org/

19. Google Cloud Public Dataset: https://lnkd.in/gZxtGfFc

20. NOAA: National Oceanic & Atmospheric Administration
Data: https://www.noaa.gov/'.

- Issues. Similarly, the API enables the retrieval of issues mentioning data.europa.eu. Issues include not only references that contain the string data.europa.eu inside repositories, but also suggestions of use in non-related repositories. The following are two examples of retrieved issues:
    - 'The data.europa.eu sparql endpoint only contains DCAT-AP instances from whole Europe. Therefore the result is empty, as expected.'
    - '… E.g. data.europa.eu creates a new URI for each dataset it harvests.
      So if you query the union of data.europa.eu and datavindplaats together you get your data RDF wise **twice**.'

- Code files. Code files were the GitHub source with the highest number of references to data.europa.eu. Code files include not only pure programming files (e.g. Java scripts and Python scripts), but also data files and configuration files:
    - config/install/field.storage.node.field_dsj_digital_technology.yml:
      'settings:
      allowed_values:
      -
      value: "http://data.europa.eu/uxp/3030"
      label: "Artificial Intelligence"
      -
      value: "http://data.europa.eu/uxp/c_04ae3ba8"
      label: Cybersecurity'
    - tests/test_data/package_F03_demo/transformation/resources/cpvsuppl.json
      '
      "type": "uri",
      "value": "http://data.europa.eu/cpv/cpvsuppl/AA"
      "conceptURI": {
      "type": "uri",
      "value": "http://data.europa.eu/cpv/cpvsuppl/AA26"'

3.1.2 Data collection from Stack Overflow

The Stack Overflow public API was used to retrieve those questions that made references to data.europa.eu content. Stack Overflow provides two different types of searches: the first type searches only in the body of the question, while the second uses other parts of the question and possibly answers, using an undisclosed algorithm (as per the official website). Using the body-only search, only seven results were returned, while the second search returned a total of 20 results (including the seven posts from the body-only search). We found 102 links to data.europa.eu content from the 20 posts obtained. Of these posts, only four had at least one reference to datasets and none referred to data stories.

For each question, 50 attributes can be obtained, including the:

- question ID
- body (the text of the question/post)
- creation date
- tags
- title
- owner.

As discussed above, the API does not allow answers to be queried directly and therefore we may have omitted references to data.europa.eu content in answers that were not included in the original questions. If we want to be more comprehensive in our analysis, new methods will have to be developed in the future to also obtain data in those cases and therefore expand the data collection process that we have presented here.

3.1.3 Data collection from Reddit

The search API of Reddit is limited, only allowing searches within subreddits. A subreddit is a discussion forum where users post about a specific topic. Therefore, we decided to first perform a manual search of the platform for the string data.europa.eu, which allowed several subreddits to be identified in which this type of content was available. The most prominent subreddits identified were 'Europe', 'datasets' and 'EuroStatistics'. Then, the Reddit API was used to search for the appearance of data.europa.eu within each of these subreddits. Similarly to what we did with GitHub, the results returned were categorised according to the type of resource on data.europa.eu. Figure 2 summarises these results.

Figure 2. Distribution of data.europa.eu resources on Reddit.

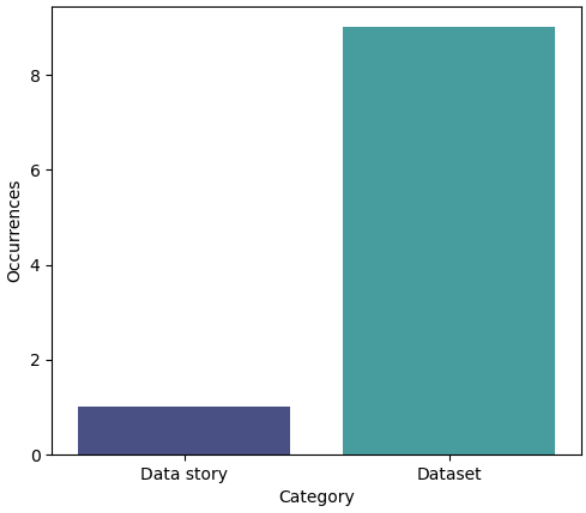
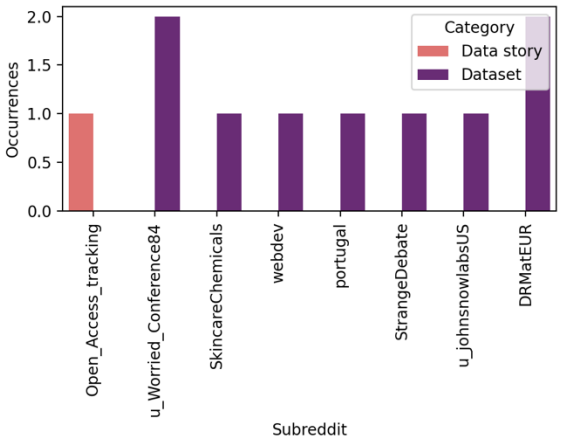


Figure 3 outlines the distribution of the different data.europa.eu resources detected across the main subreddits identified and the resource category.

Figure 3. Distribution of data.europe.eu resources across the main subreddits identified.



For each post, more than 100 attributes can be obtained, including the:

- ID
- author
- number of comments
- subreddit
- title
- ups (liked/preferred)
- downs (not liked)
- selftext (text body).

3.2 Preliminary exploratory data analysis on the selected external sites

For each of the data sources identified, we carried out a preliminary data analysis that aimed to further understand the types of resources that were being mentioned in these data sources. More context was obtained by using simple superficial means such as wordclouds or tag analysis on the related materials in which data.europa.eu was mentioned and applying simple classification techniques (e.g. classification of texts according to the DBpedia ontology). Some small variations were made for each data source, considering the characteristics of the data that can be retrieved from them and the types of resources that they contain (e.g. software code versus text). It is important to note that this analysis

was only **a first step towards better understanding the context in which data.europa.eu content is mentioned**. This will need to be further assessed in the future by a more comprehensive analysis of the resources obtained, via an evaluation framework and by means of other non-computational techniques (e.g. user surveys), to understand the context in which these resources are used, why they are selected, how they are found, etc.

As already discussed earlier, all of the software code and intermediate results that correspond to the computational methods developed are available at https://github.com/oeg-upm/dataeuropa-analysis. This means that the methods can be easily reproduced in the future by us and by others, as long as there are no major changes in the APIs used. Table 3 provides a summary of the data retrieved from each source, which served as a basis for the analysis. All of the data are available in the repository and under a CC-BY 4.0 licence. The code is available with an Apache v2 licence.

Table 3. Data retrieved from each source

Source	Data collected
GitHub	Repository descriptions, content of issues, commit descriptions of repositories, code files, issues and commits mentioning data.europa.eu, data.europa.eu/datasets and data.europa.eu/datastories
Stack Overflow	Questions containing references to data.europa.eu, data.europa.eu/datasets and data.europa.eu/datastories
Reddit	Reddit posts, and the subreddits they were posted to, containing references to data.europa.eu, data.europa.eu/datasets and data.europa.eu/datastories

3.2.1 Preliminary exploratory data analysis of mentions of data.europa.eu on GitHub

This analysis of the references to data.europa.eu resources on GitHub showed a predominance of the use of dataset resources, as may have been expected given the type of users and uses of GitHub. Other resources such as data stories were only occasionally featured among code files and issues.

A contextual analysis was conducted on the GitHub resources studied, removing those that did not apply to data.europa.eu portal content and taking into consideration the description of the GitHub repositories retrieved in the previous search. These descriptions were analysed using a word frequency approach and plotted using a wordcloud graph (the larger terms are those that were featured most

prominently across the different repository descriptions). Figure 4 plots the wordclouds for GitHub's code, repositories and commits.

Figure 4. Wordclouds of the terms contained in the code, commits and repository descriptions retrieved from GitHub code files.

This analysis provides a first insight into the context in which data.europa.eu assets are used. As denoted by the wordclouds (Figure 5), the projects in which data.europa.eu was referenced related mostly to 'services'. It is worth noting that more specific terms related to semantics were also prominently featured, such as 'ontology' and 'vocabulary'.

In addition, the descriptions of the GitHub repositories previously retrieved were categorised to further analyse the context of use. The only textual information available for the GitHub resources is the description of the repositories in which the resources are identified. Therefore, a URL categorisation into data.europa.eu categories could not be applied. Subsequently, a text classification model was employed to try and aggregate the resources identified into categories. The NLP4Types tool (Santana and Rico-Almodóvar, 2018) was used to categorise GitHub repository descriptions. This model categorises textual input according to the classes of the DBpedia ontology. Figures 5, 6 and 7 depict the results of this analysis.

Figure 5. Categorisation of the descriptions of the repositories retrieved from GitHub code searches.

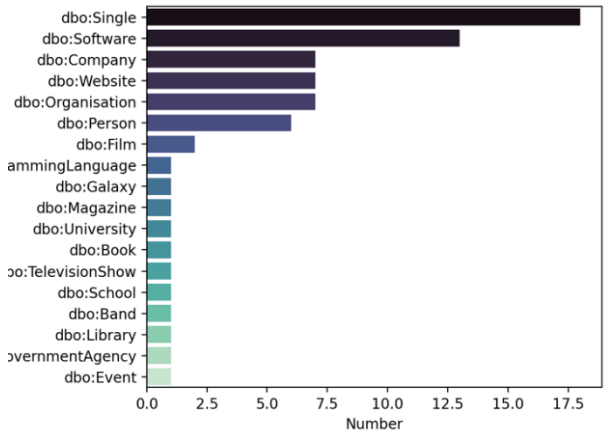


Figure 6. Categorisation of the descriptions of the repositories retrieved from GitHub commit searches.

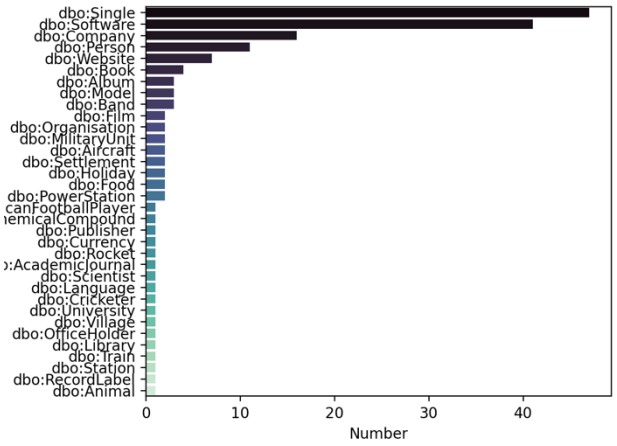
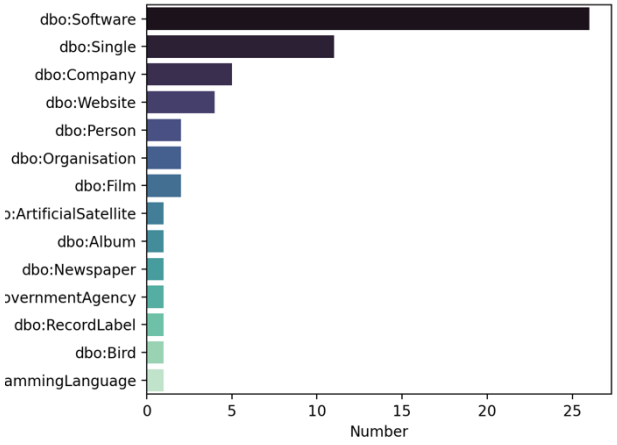


Figure 7. Categorisation of the descriptions of the repositories retrieved from GitHub repository searches.

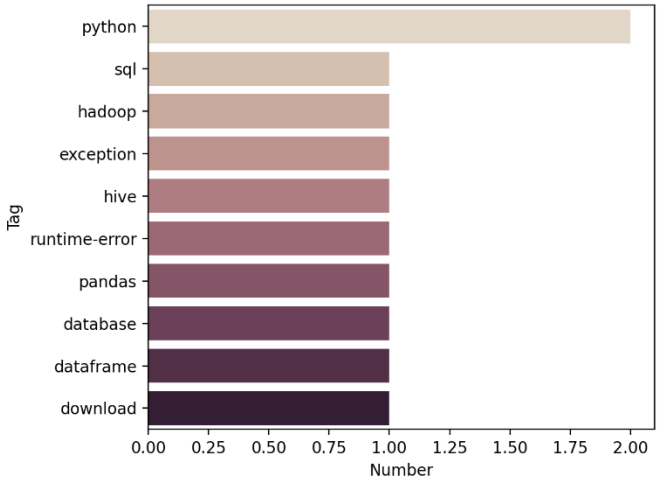


As shown in Figures 5–7, the most prominent categories were ‘single’ and ‘software’. Although the descriptions categorised as ‘software’ were accurate, this does not provide further insight into the context in which the resources were used. The category ‘single’, which relates to musical works, is also not descriptive of the resources in this context. This may be due to the limited length of GitHub descriptions, which may not be expressive enough for an accurate categorisation. Replacing repository descriptions with README.md files may be a potential solution for this issue, and will be explored in the future.

3.2.2 Preliminary exploratory data analysis of mentions of data.europa.eu on Stack Overflow

All of the results retrieved from Stack Overflow were related to the general topics of software and data science, which was expected given the nature of this forum. A more in-depth analysis of the specific techniques referenced alongside the data.europa.eu references is provided in this section, based on the tags of the posts in which data.europa.eu references appear. Figure 8 plots the frequency of the tags of the posts in which data.europa.eu was referenced.

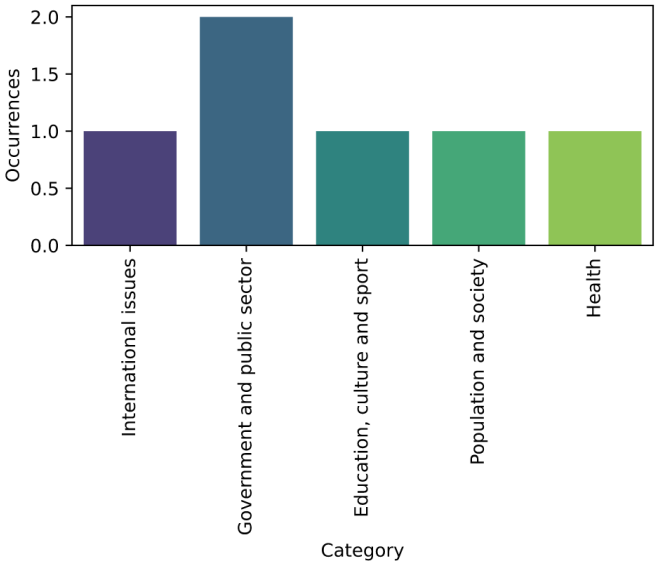
Figure 8. Tags referenced alongside data.europa.eu on Stack Overflow.



From a developer perspective, Python is the main language used to treat and use data, which is further evidenced by the frequency of the terms ‘dataframe’ and ‘pandas’, which refer to the specific Python library for data processing. Another trend was found within the tags referring to Semantic Web related terms, such as ‘rdf’ and ‘sparql’.

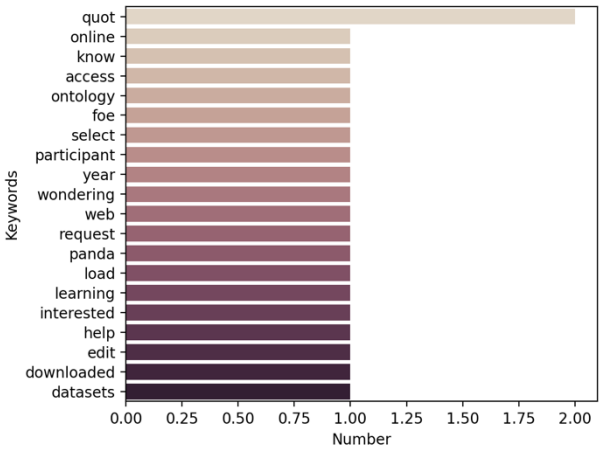
Among those posts specifically referencing datasets, a deeper analysis was performed to categorise the datasets identified into the data.europa.eu specified categories. Figure 9 plots the results of this analysis. As the number of available resources was fairly limited, no reliable conclusions can be drawn from this analysis.

Figure 9. Categorisation of the datasets identified from data.europa.eu on Stack Overflow.



Finally, similarly to the analysis conducted for GitHub and shown in the previous section, the most frequent terms featured in the resources retrieved were identified to detect trends (Figure 10). While in the case of GitHub, the kinds of terms featured in the descriptions were rather heterogeneous, in the case of Stack Overflow these terms were largely related to software (with general terms such as 'quot', 'url' and 'prefix' most prominent).

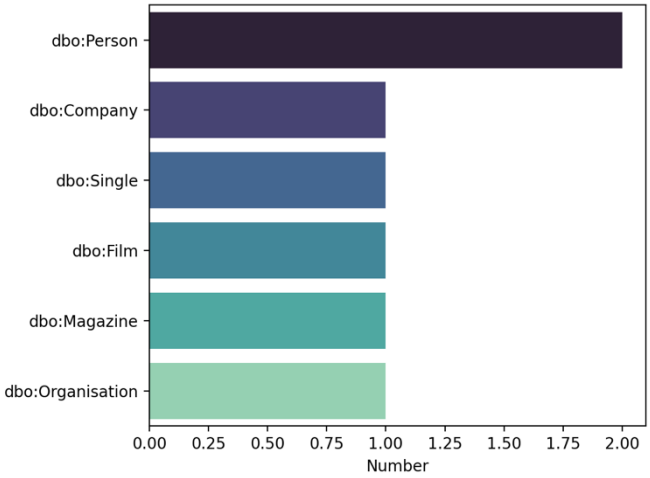
Figure 10. Term frequency of the data.europa.eu resources extracted from Stack Overflow.



3.2.3 Preliminary exploratory data analysis of mentions of data.europa.eu on Reddit

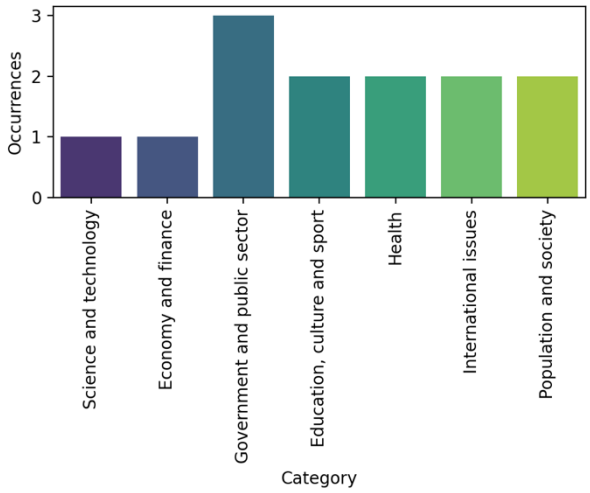
As in the case of GitHub, an initial classification of the posts mentioning data.europa.eu across the different subreddits was performed. Figure 11 showcases the result of this categorisation. While for GitHub, the predominant class was ‘single’, which is a concept related to the music industry and was the result of a wrong classification, in the case of Reddit the predominant class was ‘person’, which demonstrates a change in the context in which data.europa.eu is mentioned on this platform. The second most featured topic was, like for GitHub, ‘software’, which serves as an indicator of the type of users commenting on the subject, that is, mostly developers.

Figure 11. Categorisation of the posts retrieved from Reddit.



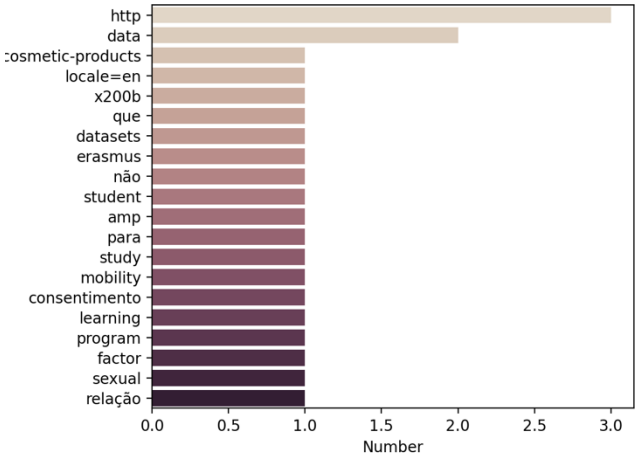
The same categorisation procedure used for the datasets detected in Stack Overflow posts was also applied to the datasets referenced in Reddit. Figure 12 plots the results of this analysis. As in the previous case, the number of results was insufficient to draw significant conclusions. However, it is worth noting that, as in the case of Stack Overflow (Figure 9), the most predominant category was ‘government and public sector’, maybe indicating a preference by the users for these types of datasets.

Figure 12. Categorisation of the datasets identified from data.europa.eu in Reddit.



Finally, a contextual analysis was also performed to extract the most frequent terms from the posts retrieved. Figure 13 plots the results of this analysis. Similarly to the previous scenario, the number of results was not sufficient to extract patterns from the terms detected. Nonetheless, there was a clear predominance of software-related terms ('http', 'account', 'post' and 'string'), which further strengthens the idea that most of the users of data.europa.eu are developers.

Figure 13. Term frequency of the most common terms detected in Reddit posts.



4 Conclusions and future work

The work presented in this report mostly focused on creating the necessary infrastructure and data collection and processing pipelines to enable the collection of as much data as possible from a selected set of external sites (GitHub, Stack Overflow and Reddit), on which references to data.europa.eu content can be identified. We expect these data to be useful for providing partial or complete answers to the research questions set out in Section 2. Furthermore, given that these data will be able to be retrieved in the future for further analysis, we will be able to track the presence of data.europa.eu content on those sites over time.

The initial exploratory analysis performed and discussed in Section 3.2 shows that a large percentage of the discussion related to data.europa.eu on these sites relates to technical aspects, mainly coming from software developers. This was expected, given the types of sites that were selected and explored.

4.1 Findings

As discussed in the introduction, we have mostly focused our work in this report on the first research subquestion (Q1.1), which aims to determine how much data.europa.eu content is present on those sites that are commonly used by communities of data scientists and software developers. Given the results of our data collection process and the initial exploratory analysis that we undertook, we can conclude that:

- data.europa.eu content (mostly datasets) is mostly referred to from a technical perspective. This is evidenced by the frequency of terms such as ‘pandas’ and ‘python’ alongside references to data.europa.eu, as shown in Figures 5–8, 10, 11 and 13.
- As the number of data points that we obtained for all of the external sources was extremely limited, we cannot make strong conclusions, as there is no statistical significance. However, as shown in Figures 9 and 12, it should be mentioned that categories such as ‘government and public sector’ are the most relevant, as was also expected.
- The data.europa.eu content that appears most often is datasets and data stories. Figures 1–3 summarise the distribution of the different resources across the external sources studied. Datasets were, by a significant margin, the most demanded resource. It is worth noting that data stories were significantly less referenced, which may be because these resources were only included more recently and are still generally unknown.

4.2 Limitations

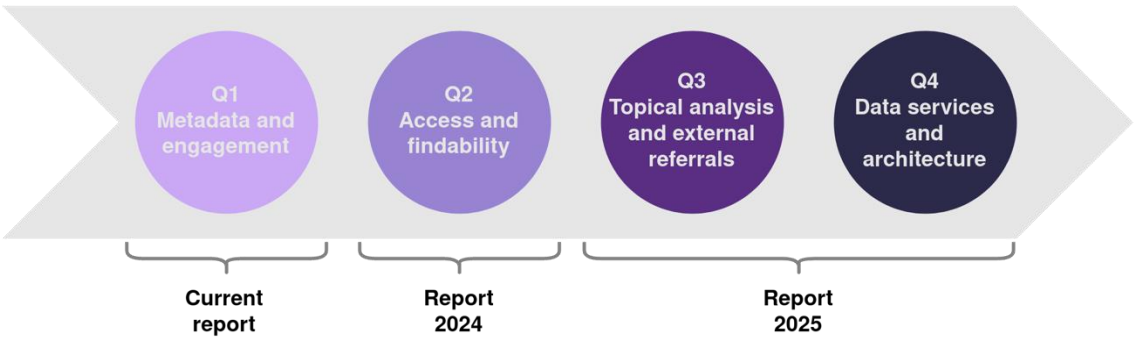
The results obtained so far in terms of data collection have also demonstrated that the presence of data.europa.eu content on these sites is still very limited, which influences the amount of statistical data analysis that can be performed and the conclusions that can be obtained from such analysis. Further types of methods and techniques will need to be used, specifically for deepening the analysis

and providing greater insights, including literature reviews and user studies. Through such methods, we can obtain additional qualitative data about the use of data.europa.eu content on these sites, which can also help us to better understand how these communities work and evolve.

There is also a need to refine the data collection methods for identifying mentions of and references to data.europa.eu content that may be partially hidden because of the use of URL shortener services, for instance. While this would be very relevant for sites that shorten URLs by default (e.g. Twitter and LinkedIn), it may not necessarily be a problem for the sites that we have identified, as they do not have those restrictions. However, attention will be paid to this problem in the future.

Future reports will focus on the other subquestions identified for Q1, as well as on the other three questions that were identified in Section 2 (Q2, Q3 and Q4; Figure 14). A description of the research methods that will be used to address these questions is provided in the annex.

Figure 14. Research questions addressed in each report.



5 References

Alobaid, A., Amador-Domínguez, E. and Corcho, O. (n.d.), 'Repository for the analysis of data.europa.eu done in task 3.4', https://github.com/oeg-upm/dataeuropa-analysis

Ibáñez, L. D. and Simperl, E. (2021), *Analytical Report 19: Understanding supply and demand in dataset search on the European Data Portal.* Publications Office of the European Union, Luxembourg

Ibáñez, L.-D., Kacprzak, E., Koesten, L. and Simperl, E. (2020), *Analytical Report 18: Characterising dataset search on the European Data Portal.* Publications Office of the European Union, Luxembourg

Kacprzak, E., Koesten, L., Ibáñez, L.-D., Blount, T., Tennison, J. and Simperl, E. (2019), 'Characterising dataset search: An analysis of search logs and data requests', *Journal of Web Semantics*, Vol. 55, pp. 37–55.

Koesten, L., Kacprzak, E., Tennison, J. and Simperl, E. (2017), 'The trials and tribulations of working with structured data: A study on information seeking behaviour', in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, Association for Computing Machinery, New York, NY, pp. 1277–1289.

Koesten, L., Vougiouklis, P., Simperl, E. and Groth, P. (2020), 'Dataset re-use: Toward translating principles to practice', *Patterns*, Vol. 1, No. 8, 100136.

Li, X., Schijvenaars, B. and de Rijke, M. (2017), 'Investigating queries and search failures in academic search', *Information Processing and Management*, Vol. 53, No. 3, pp. 666–683.

Piccardi T., Redi, M., Colavizza, G. and West, R. (2021), 'On the value of Wikipedia as a gateway to the web', in *Proceedings of the Web Conference 2021 (WWW '21)*, Association for Computing Machinery, New York, NY, pp. 249–260.

Santana, I. and Rico-Almodóvar (2018), 'NLP4Types: Predicting types using NLP', in *Proceedings of the EKAW 2018 Posters and Demonstrations Session*, CEUR Workshop Proceedings, Nancy, France, pp. 57–60.

6 Annex. New data sources and methodology for further research questions

In this annex, we describe the methodology that will be used in future reports to address the complete list of research questions that have been proposed in this report (Section 2). We begin with a general consideration of the data sources that can be used for this purpose.

Data sources

The results presented in Section 3 on the data collected by using the APIs of the three relevant sites (GitHub, Stack Overflow and Reddit) have demonstrated that the data that can be obtained from these sites are rather limited. This is not necessarily a problem, as it indicates that the data communities using data.europa.eu are not very active on those sites, but it makes further analyses associated with all of the research questions difficult, given the lack of statistical significance of the data obtained.

While we will continue using this API-based approach, both with the sites already used and with a new set of sites (e.g. Gitlab and Software Heritage for software-related practices), it may also be useful to explore a crawling-based approach on web content aimed at identifying other potential relevant sites and content. For this purpose, instead of performing selective crawls on these sites, we will explore the use of already crawled content such as Common Crawl (https://commoncrawl.org/), which may provide additional insights into the presence of data.europa.eu content on the web and allow other communities to be identified that might not initially have been considered. Owing to the large volume of this crawled content, the processing will take a long time and will require relevant computational resources associated with the crawled content.

As with the sites that have been used for this report, the challenge of shortened URLs will need to be addressed. This will require retrieving those URLs that are identified as shortened URLs and looking for the redirected URL, to identify whether it refers to data.europa.eu content or not.

Q1. Metadata and engagement

- **Q1.1. To what extent is data.europa.eu content (datasets and other related resources) present on external sites?** Presence is understood in a generic manner, including the reuse and republication of datasets on other external sites and the existence of references to datasets on those external sites.
    - We will use the same methodology that has been presented in this report, applying it to the same set of sources (GitHub, Stack Overflow and Reddit), as long as they continue to be publicly available and do not implement major changes in the capabilities of their APIs, and adding other sources such as GitLab or Software Heritage, or any other sources that may emerge in the future.

  - In addition, we will also perform this analysis at a global scale on the latest version that is available at each time of the Common Crawl. This will require developing new computational methods to process this content and filter the references to data.europa.eu content, and a manual analysis on the initially filtered references to determine to what extent they are relevant for deriving conclusions for this research question.
- **Q1.2. To what extent does the introduction of descriptive statistics on the dataset description page or as part of the metadata affect the presence of datasets on external sites? Q1.3. To what extent does having use cases/user stories (or other types of similar content) associated with the datasets affect their presence on external sites?**
  - We will select data.europa.eu datasets (note that this applies only to datasets and not to the other types of content available on data.europa.eu) that contain descriptive statistics associated with their content (Q1.2) or that have associated data stories (Q1.3).
  - Considering the data already retrieved for Q1.1, we will analyse whether there is a correlation between the presence of such statistics and an increase in the presence on external sites. It must be noted that, for this analysis to be representative, more appearances need to be identified on external sites than were identified on those that have been considered so far. The Common Crawl source will be used for this purpose, as it will go beyond sites on which data communities are active. If an insufficient number of mentions is finally obtained, this research question will be considered to remain unanswered, but the content processing and analysis pipelines will be ready for the future.
- **Q1.4. Is there a significant difference between the reuse of old datasets and the use of new datasets?** In other words, are old datasets more or less present on external sites than new datasets? Is there a period (e.g. between the last 2 and 5 years) during which the data reuse peaked?
  - The data gathered from Q1.1 (when expanded to other sources) will be used to compare and cross-reference the release and last update information related to specific content that mentions data.europa.eu content with the dataset publication dates.
  - Once these data are available, we will perform a quantitative analysis to see whether there is any relevant difference between the group of datasets that are mentioned and those that are not.

Q2. Access and findability

All of these questions will be addressed through a user study, as described in the following. The research questions are provided here for convenience.

- **Q2.1.** ... there are probably two groups of distinct users, namely those associated with native searches (which start on data.europa.eu and use the built-in search affordances) and those associated with external searches (which use, for example, a search engine and are likely to be more interested in having a question answered than in downloading a dataset). **What are the typical tasks of these two groups of users? How do these tasks relate to the characterisation that will be done with the results from the user survey?**

- **Q2.2. How could the data.europa.eu site help users who are not looking for a particular dataset? To what extent do data users utilise data.europa.eu as a stepping stone to find other external datasets or sources** (e.g. the URL of the official website of an organisation or dataset host)? Are the users who are accessing data.europa.eu interested only in the content hosted on data.europa.eu or do they also use the portal to find other external datasets and sources (based on (Piccardi, Redi, Colavizza, & West, 2021))?
- **Q2.3. How do users find datasets and other related resources on data.europa.eu** (e.g. through search engines, forums, social media or research papers)?
    - This has also been addressed partially in a previous report.
- **Q2.4. How quickly do users find the dataset or resources that they are looking for on data.europa.eu? How easy it is to find relevant data? Do users adjust their search strategies depending on the results they receive?**
- **Q2.5.** One of the limitations of the analyses of the data.europa.eu logs carried out so far is that we could not look at the data needs across individual search sessions ... **How do users decide what to look for in subsequent searches? Do they combine dataset searches with searches of other related resources? How could data.europa.eu support complex query needs that span multiple searches and sets of interconnected results** (Li, Schijvenaars, & Rijke, 2017; Ibáñez & Simperl, Analytical Report 19: Understanding Supply and Demand in Dataset Search on the European Data Portal, 2021)?
    - A set of controlled experiments will be performed with a group of participants who are familiar with open data and have experience in the usage of data.
    - The first set of experiments will divide users into two groups of participants with similar backgrounds. A set of coarse-grained and fine-grained content discovery tasks on data.europa.eu content will be proposed to all of them. One group will start from data.europa.eu and its search affordances and the group will have the freedom to use external search engines. Then they will be required to look for additional external content.
    - The interactions during content searches, as well as the time required to complete each task and the quality of the results obtained, will be recorded by the research team, and additional observations will be performed during the execution of the tasks.
    - All of these results will be summarised to provide insights on how to improve, if needed, the content search capabilities of the data.europa.eu portal.

Q3. Topical analysis and external referrals

- **Q3.1. Which topics, datasets and resources are most commonly shared in research papers and on social media platforms?**
    - The methodology to be followed will be similar to that proposed for those subquestions under Q1. The main difference will be the data sources that will be used for this purpose, which will address two areas that were not considered in Q1, such as research papers and social media platforms, although they may already be available from the analysis of Common Crawl that will be performed.
    - For research papers, the analysis will be based on the content available in the OpenAIRE research graph, which contains a large open set of references to research papers and their

content. This open data source will be processed similarly to what is proposed in Q1, and insights will be derived from the appearance of datasets in research papers. Datasets may appear directly referenced as scientific references, as footnotes or as part of the mentions of larger resources such as GitHub repositories or Zenodo bundles. All of these will be analysed to increase the comprehensiveness of the analysis.

  - For social media, the content produced as part of the work done in the analysis of website usage by other data.europa.eu teams will be used as a starting point.

- **Q3.2. What is the impact of the quality or reach of the social media post on the reuse on the abovementioned datasets?**
  - For those cases in which the mentions of data.europa.eu content appear on social media, descriptive statistics will be provided of some of the typical indicators used in social media posts (the reach, number of retweets/reposts and amount of discussion generated).
- **Q3.3. What is the impact of research papers that reference data.europa.eu datasets (e.g. impact factor/quantile) on the reuse (traffic) of the abovementioned datasets?**
  - For those cases in which the mentions of data.europa.eu content appear in scientific papers, descriptive statistics will be provided of some of the typical indicators used in these papers (the number of citations and impact factor).
- **Q3.4. Which organisations generate the most traffic to data.europa.eu?**
  - The data.europa.eu logs provided by Piwik/Matomo will be analysed to determine which organisations generate the most referrals. Descriptive statistics will be provided for this purpose.

Q4. Data services and architecture

- **Q4.1. Does the offering of suggestions to other relevant datasets or other resources on data.europa.eu increase the reuse of data.europa.eu?**
  - A recommender system for datasets or other types of content will be created, using content-based methods related to the content of data and to the metadata associated with each of the datasets or the content. This can be compiled upfront and offered as an additional metadata item in the portal, if such a change is possible in cooperation with Q1. These recommendations will be clearly marked as automatically suggested recommendations by an automated system, and will be applied to half of the datasets available on data.europa.eu.
  - An analysis will be undertaken, once this deployment has been running for a sufficiently long period of time, on whether the presence of these datasets in external sources has increased or not in comparison with the other datasets for which no suggestions are provided. The data collected for Q1 will be used for this analysis.
  - An additional analysis will be undertaken on whether the links to the suggested datasets have been used or not by users. Traffic data collected by Piwik/Matomo will be used for this purpose.

Publications Office
of the European Union

ISBN 978-92-78-43560-8