

# Analytical Report 18



Analytical Report 18

**CHARACTERISING DATASET SEARCH ON THE EUROPEAN DATA PORTAL: AN ANALYSIS OF SEARCH LOGS**

This study has been prepared by the University of Southampton as part of the European Data Portal. The European Data Portal is an initiative of the European Commission, implemented with the support of a consortium led by Capgemini Invent, including Intrasoft International, Fraunhofer Fokus, con terra, Sogeti, 52North, Time.lex, the Lisbon Council asbl and the University of Southampton. The Publications Office of the European Union is responsible for contract management of the European Data Portal. 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  
Daniele Rizzi – Policy Officer  
Email: [daniele.rizzi@ec.europa.eu](mailto:daniele.rizzi@ec.europa.eu)

### European Data Portal

Gianfranco Cecconi, European Data Portal Lead  
Email: [gianfranco.cecconi@capgemini.com](mailto:gianfranco.cecconi@capgemini.com)

### Written by:

Luis-Daniel Ibáñez

[L.D.Ibanez@soton.ac.uk](mailto:L.D.Ibanez@soton.ac.uk)

Emilia Kacprzak

[E.Kacprzak@soton.ac.uk](mailto:E.Kacprzak@soton.ac.uk)

Laura Koesten

[laura.koesten@kcl.ac.uk](mailto:laura.koesten@kcl.ac.uk)

Elena Simperl

[elena.simperl@kcl.ac.uk](mailto:elena.simperl@kcl.ac.uk)

### Reviewed by:

Eline N. Lincklaen Arriëns

[Eline.lincklaen.arriens@capgemini.com](mailto:Eline.lincklaen.arriens@capgemini.com)

Last update: 23.09.2020

www: <https://europeandataportal.eu/>

email: [info@europeandataportal.eu](mailto:info@europeandataportal.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, 2020

© European Union, 2020



OA-02-20-767-EN-N

ISBN: 978-92-78-42299-8

ISSN: 2600-0601

doi: 10.2830/776263



The reuse policy of European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Except 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.

## Table of Contents

Executive summary .....	4
1 Introduction.....	5
2 Background.....	7
2.1 Overview.....	7
2.2 The search process .....	8
2.2.1 Querying .....	8
2.2.2 Query handling .....	8
2.2.3 Data handling .....	8
2.2.4 Results presentation.....	8
2.3 What we know about dataset search and interaction .....	9
2.4 The EDP dataset search interface.....	10
3 Quantitative analysis .....	12
3.1 Key terms.....	12
3.2 The search and interaction logs corpus.....	13
3.3 Results .....	15
3.3.1 Dataset search in the context of the EDP.....	15
3.3.2 Dataset search strategies and search query characteristics .....	18
3.3.3 EDP vs. web search engines in dataset search .....	25
3.3.4 Success in dataset search .....	32
4 Discussion.....	37
4.1 Search experience on EDP.....	37
4.2 Search through web search engines .....	38
4.3 Curated datasets with context .....	39
4.4 The role of SEO .....	39
5 Summary and conclusions.....	40
6 References.....	42

## Executive summary

This report illustrates a quantitative study on dataset search through more than two years of European Data Portal (EDP) search and interaction logs. Understanding data search behaviour is key to developing better search algorithms and improving the search experience. In this study, we present current findings from key literature in dataset search and cover four key aspects in our analysis:

1. Dataset search in the context of the EDP,
2. Dataset search strategies and search query characteristics,
3. EDP versus web search engines, and
4. Assessing success in dataset search.

## 1 Introduction

This report describes a quantitative study of two years of European Data Portal (EDP) search and interaction logs to provide directions for further development informed by user search behaviour. In the EDP's Analytical Report 8 on "[The Future of Open Data Portals](#)", we highlighted the importance of the re-user when thinking about data discoverability, metadata, and co-location of documentation. Other EDP reports have looked at the [maturity of open data in Europe](#), suggesting it is undergoing a consolidation phase, following earlier intensive efforts to make as much data as possible accessible for everyone. As a critical mass of datasets has been published openly, the aim has gradually shifted towards ensuring that the available data is of value to users and has broad impact. According to prior EDP work, the impact remains the least mature open data dimension, hence calling for sustained efforts to continue to monitor and measure it in multiple ways.

Previous studies of dataset search have shown that making datasets available does not ultimately equate to accessibility or usefulness for users. It is therefore important for data portals to consider other ways to add value on top of the raw published data. This includes capabilities to find, download and make sense of datasets, potentially in new, unforeseen contexts.

The provision of datasets as well as dataset search functionalities is a key section of the EDP. We aim for EDP users to be supported in both the discovery and re-use of datasets. The more we understand about data search behaviour of EDP users, the better we will be able to develop capabilities and experiences that support them. This report is a first step in this direction. We provide some background on dataset search as an emerging area of research, detailing different subtopics from the literature that feed into the development of the dataset research agenda, as laid out in our previous work (Chapman et al., 2019). We detail what we have learned about dataset search in prior studies, analysing the search logs of four open data portals. This allows us to contrast and to extend the findings of this report with existing initial research on data search behaviour. We then describe a search and interaction log analysis of the EDP spanning three portal development cycles.

Our analysis spans the following questions, which we answer by analysing the logs and framing the results in the context of related literature and previous studies of ours on other portals:

### 1) Dataset search in the context of the EDP

- a) How is the dataset search section of the EDP used? Are there variations across portal versions?
- b) How does dataset search compare to other sections of the portal? Do users visit several sections of the portal in the same session? Are there variations of the above across portal versions?

### 2) Dataset search strategies and search query characteristics

- a) How do people search for datasets on the EDP? Can we identify particular search strategies that are more popular than the others?
- b) What are the most popular facet filters?
- c) What are the most popular combinations of facets?
- d) Is there any difference in the use of facets when the user issues queries via the dataset search box?
- e) What characteristics do the queries that are issued via the dataset search box have?
- f) How do they compare to previous studies on national open data portals?

### 3) EDP versus web search engines in dataset search

- a) What is the EDP's role in users' dataset search journeys? Do users prefer to search datasets using EDP search functionality or do they rely on a (generic) web search engine?
- b) Is there a difference in the characteristics of dataset search queries made to the EDP dataset search box (internal) with respect to those made on a web search engine (external)?

### 4) Success in dataset search

- a) Are users successful when they search for datasets on the EDP? Does success change across portal versions?
- b) Is there any difference in success between internal search (EDP's dataset search box) and external queries (from web search engines)?
- c) Is there one search strategy (search box only, facets only, mixed) more successful than others?

To address these questions, we conducted a quantitative analysis of 844,343 EDP user session logs from April 2018 to June 2020 (see Section 3 Quantitative Analysis). Search log analysis is routinely used to understand the behaviour of users and evaluate search on the web and elsewhere. A user session log includes, among other things, which website referred the visitor to the portal, the pages visited during a session, the queries issued to the dataset search section of the portal, and which datasets were accessed from the portal.

For answering **question (2)** we compute a similar set of indicators as in our previous research (Kacprzak et al. 2019) and highlight commonalities and differences. For **question (3)**, we compare two different types of sessions: those that started a dataset search after landing on the portal from a web search engine or website and those that started a dataset search from the portal's dataset search box. For **question (4)**, as no specific studies for satisfaction in dataset search exist yet, we rely on state-of-the-art techniques for measuring and inferring user satisfaction on document search engines and discuss how they transfer to dataset search.

This report concludes with a discussion of the main findings of the search and interaction logs analysis, including recommendations to emphasise and expand the tracking of user interactions in dataset search to allow for more detailed follow-up studies to inform search and UX design. It is vital for portals to understand the needs of their users, especially when thinking about which functionalities to prioritise for future development for success in user uptake. Our analysis suggests that while many EDP users land on the dataset section from searching with web search engines, there are alternative ways to add significant value to a user's dataset search journey. If we do not see dataset search on the EDP merely as an information retrieval problem, but consider richer types of contexts, such as additional material and re-use support, around datasets as a priority. This means portal designers could focus less on improving the accuracy of search results and instead understand the implications of people using external search tools on their user journeys. User experience can be improved by supporting users in finding value in the content published on the EDP and if the EDP facilitates re-use through search and by delivering related resources that general-purpose web search engines are less concerned about (e.g. visualisations, stories, comments). In time, this could bootstrap a real open government data community that can use the EDP as a learning hub that supports open data re-use.

## 2 Background

### 2.1 Overview

As a growing amount of data becomes available on the web, data searching becomes an increasingly important topic. Data search and discovery is researched in a range of complementary disciplines. However, despite advances in information retrieval (search, navigation, and query), the semantic web and data management, data search is not as advanced as related areas, such as searching for documents, both technologically (Cafarella et al., 2011) and from a user experience point of view (Gregory et al., 2017; Koesten et al., 2017). Prior studies investigating search strategies for datasets amongst users have shown that personal recommendation (Koesten et al., 2017), as well as links from literature to datasets still play a big role (Gregory et al. 2020).

The dataset search problem can be addressed at various levels. Services such as Google Dataset Search (Noy et al., 2019) and DataMed (Sansone et al., 2017) crawl across the web and facilitate a global search across distributed resources. Despite efforts in data search development, the limited amount of user interaction with the data restricts how these search tools are developed (Noy et al., 2019).

Existing data search approaches use tags found in metadata mark-up, expressed in vocabulary terms from schema.org<sup>1</sup> or DCAT<sup>2</sup>, to structure and identify the metadata considered important for datasets. The same approach is used in data portals, including open government data portals such as data.gov.uk<sup>3</sup>, organisational data lakes (Reynolds 2014), scientific repositories such as Elsevier's<sup>4</sup> and data markets (e.g. Grubenmann et al. 2018). EDP as a meta-portal sits in between these two levels – it aims to facilitate global search and discovery across multiple publication sites but has a focus on specific types of data and portals (open government data, public administrations in specific countries).

Dataset search and web retrieval is a relatively unexplored area compared to document search and retrieval. The literature suggests that dataset search has unique characteristics that result in requirements for users and on infrastructures: complex information needs leading to difficulty on expressing queries as keywords or questions, need for complex filter conditions, and larger inspection times to confirm if a dataset is relevant or not. Currently, there is a disconnect between what datasets are available, what datasets a user needs, and what datasets a user can find, trust, and is able to use. This is where data portals can add value and guidance for the user if their needs are addressed accordingly. This is not a question of doing information retrieval correctly, but of making sure the user is able to use the data effectively once they have found a set of datasets to choose from.

In a dataset search context, approaches need to consider aspects such as data provenance, annotations, quality, the granularity of content, and schema to effectively allow users to evaluate a dataset's fitness for a particular use (Chapman et al. 2019). The user does not have the ability to introspect over large amounts of data and they need guidance and support tools. Users are attempting to discover and assess datasets for a particular purpose. Supporting them requires frameworks, methods, and tools that specifically target data as its input form and consider the specific information needs of data professionals. This means thinking about how results are presented and co-located with other resources.

---

<sup>1</sup> <https://schema.org/Dataset>

<sup>2</sup> <https://www.w3.org/TR/vocab-dcat/#class-dataset>

<sup>3</sup> <https://data.gov.uk/>

<sup>4</sup> <https://datasearch.elsevier.com>

## 2.2 The search process

The dataset search process (Figure 1) can be described in four steps: 1) *querying*, 2) *query handling*, 3) *data handling* and 4) *results presentation*. Each of these steps are an area that specific research efforts can focus on (based on Chapman et al. 2019):



Figure 1: Main steps of the dataset search process

### 2.2.1 Querying

Querying occurs commonly in the form of keywords where filters can be applied. For instance, on the EDP users can filter by countries, catalogues, formats, licenses etc. In Section 3 on Quantitative analysis, we analyse queries issued on the EDP as well as queries issued to search engines leading to the EDP dataset section.

Users can be supported in issuing queries in several ways. One is to consider tasks associated with the search session and tailor filter functionalities to different task types, for example. In prior work we have identified two categories of data-centric tasks:

- process-oriented tasks in which data is used for something transformative, to do something with the data (e.g. using it to build a tool); and
- goal-oriented tasks in which data is, e.g., used to answer a question (Koesten et al. 2017).

These search tasks come with different information needs around the dataset. By considering them as distinct scenarios with different requirements, we can better understand what people do when searching for and engaging with datasets and support the decisions they make during data discovery.

### 2.2.2 Query handling

Most dataset search algorithms operate over the dataset's metadata. Results are produced based on how similar the metadata is to the search terms. Unfortunately, low metadata quality (or missing metadata) affects both the discovery and the consumption of the datasets within open data portals (Umbrich et al, 2015). The success of the search functionality depends on the publisher's knowledge of the dataset and the quality of the descriptions they provide.

### 2.2.3 Data handling

Publishers populate metadata using vocabularies such as DCAT, schema.org or CSV on the Web. This step is mostly manual and is resource-intensive, which means that dataset descriptions are often incomplete or do not contain enough detail, as has been shown by Koesten et al. in a study on dataset summaries (2020). This limits the capabilities of query handling methods, which attempt to match search terms to the descriptions and to achieve quality and entity resolution.

### 2.2.4 Results presentation

Choosing a dataset greatly depends on the information provided alongside it. Most Search Engine's Results Pages (SERPs) for dataset search currently follow a traditional 10 blue links paradigm. Clicking on a search result takes the user to a preview page that contains metadata, a free-text summary and sometimes a preview or visualisation of the data. Google Dataset Search also follows the traditional result presentation as a list, but they display a split interface. This presents a large number of search



results for scrolling on the left side and a reduced version of a dataset preview page with links to one (or multiple) repositories that hold the respective dataset on the right side.

*Ranking datasets* is a research problem on its own. The traditional web-based ranking is difficult due to limited links between datasets (Noy et al. 2019). The applicability of IR models built mainly for document retrieval is questionable (Carevic et al. 2020). Datasets might require different approaches to ranking due to their unique characteristics both in terms of their structure as well as concerning the types of search tasks users engage in (Chapman et al., 2019). At present, there are limited corpora available to train learning to rank algorithms for dataset search. The EDP could consider playing a role thereby releasing search and interaction logs to the research community to build such algorithms and publish them open source.

*Interactions with the search results page.* Interactive query interfaces facilitate ad-hoc data analysis and exploration, which needs to be informed by user behaviour in dataset discovery. This could include the ability to compare, contrast or even combine several datasets. Below we discuss insights from current literature that can be used to inform such functionalities; moreover, the analysis we present in this report can further add to the specifics of how users search for data on the EDP. Interaction does pose different requirements on the supporting in terms of computational resources and performance (Jiang et al., 2018) and has yet to be realised for larger data portals.

### 2.3 What we know about dataset search and interaction

In prior work (Kacprzak et al. 2019) we analysed search logs of four national open data portals to understand the characteristics of queries for datasets, how they differ from general web search and how users request data in a non-constrained form (as free-text data requests issues to an open data portal rather than keywords in a search box). We describe key findings from this work here and point to related findings from other literature:

- 1) **Dataset search is a work-related activity.** We found that most queries issued directly on the portals (i.e., the internal queries) were related to datasets in the area of business and economy. By contrast, external queries were topically more diverse, with topics such as society and towns and cities appearing regularly. We also noticed differences in the ratio of question queries - a larger percentage of external queries included question queries.
- 2) **Dataset queries are short.** Carevic et al.(2020) also found (comparing dataset search to publication search) that on average, the length of a dataset search query is shorter which is in line with our findings. In an interview study with data users in 2017, we saw that many users do not expect that the search functionality will be able to provide relevant data for longer and more specific queries and therefore issue short queries (Koesten 2017). The observation of short dataset queries conflicts with other work on the characteristics of dataset search queries (Kacprzak et al., 2018) but is likely due to the differences in study and the portal context. This underlines the importance of conducting a portal specific log analysis for the EDP as presented in this report.
- 3) **Data search queries on data portals are different from those issues on a general web search.** There is a difference in topics, length and structure between dataset queries issued directly to data portals and dataset queries issued to web search engines. For instance, a larger percentage of external queries included question queries, which might be due to the increasing ability of large search engines to support natural language type queries.

- 4) **Data requests describe the data by using boundaries and restrictions about location, temporal information, specific data type and/or specific granularity (e.g., year/month/day).** Geospatial and temporal search has shown to be more prevalent in dataset search in this study and an approach to enable this has been described by Neumaier et al. (2019). One key finding by Carevic et al. (2020) was that dataset queries contained significantly more numerical digits, which can be explained with the nature of periodic records in research data. In that sense, this confirms prior findings on the importance of time-based searches in dataset search (Kacprzak et al. 2019).
- 5) **Common properties to describe datasets are temporal and geospatial coverage, with varying levels of granularity.** Queries including some indication of the time were almost five times more frequent than in web search (Nunes et al., 2008), suggesting that datasets have a stronger relationship to time than documents. This can include the time frame the data represents (data about a particular year) or the creation time of a dataset (the time the data was collected and published, or the frequency of updates). DCAT already includes properties for the temporal and geospatial description of datasets, and our findings suggest that providing fine-grained descriptions of these properties could improve the search experience.
- 6) **Users have dataset-specific selection criteria.** Looking specifically at dataset search amongst researchers in a large scale survey Gregory et al. (2020) have found that for almost 90% data collection conditions and methodology was important or extremely important in their decisions, which was also considered the key fact to establish trust in the data. This was followed by information about data processing and handling as well as topical relevance. The ease of accessing data was also considered very important. Most of these results are mirrored in a mixed-methods study by Koesten et al. (2020) looking at selection criteria for datasets, different aspects of relevance, quality and usability.

We use these findings to inform the EDP log analysis presented in the following sections.

## 2.4 The EDP dataset search interface

EDP's dataset search interface follows a traditional structure in the style popularised by digital marketplaces, and in use by most national data portals across the world. Figure 2 shows the relevant components:

- (1) Dataset search box: Where users type their queries
- (2) Order by selector: Allows re-ordering results according to the following criteria:
  - (a) Relevance to the query keywords
  - (b) Descending date of modification
  - (c) Descending date of creation
  - (d) Ascending alphabetically by dataset name
  - (e) Descending alphabetically by dataset name

The leftmost column lists the available "Facets". Facets are filters that can be combined by users to narrow down the number of obtained results. Figure 2 shows three of the facets available in the EDP (3) location, (4) operator, and (5) country.

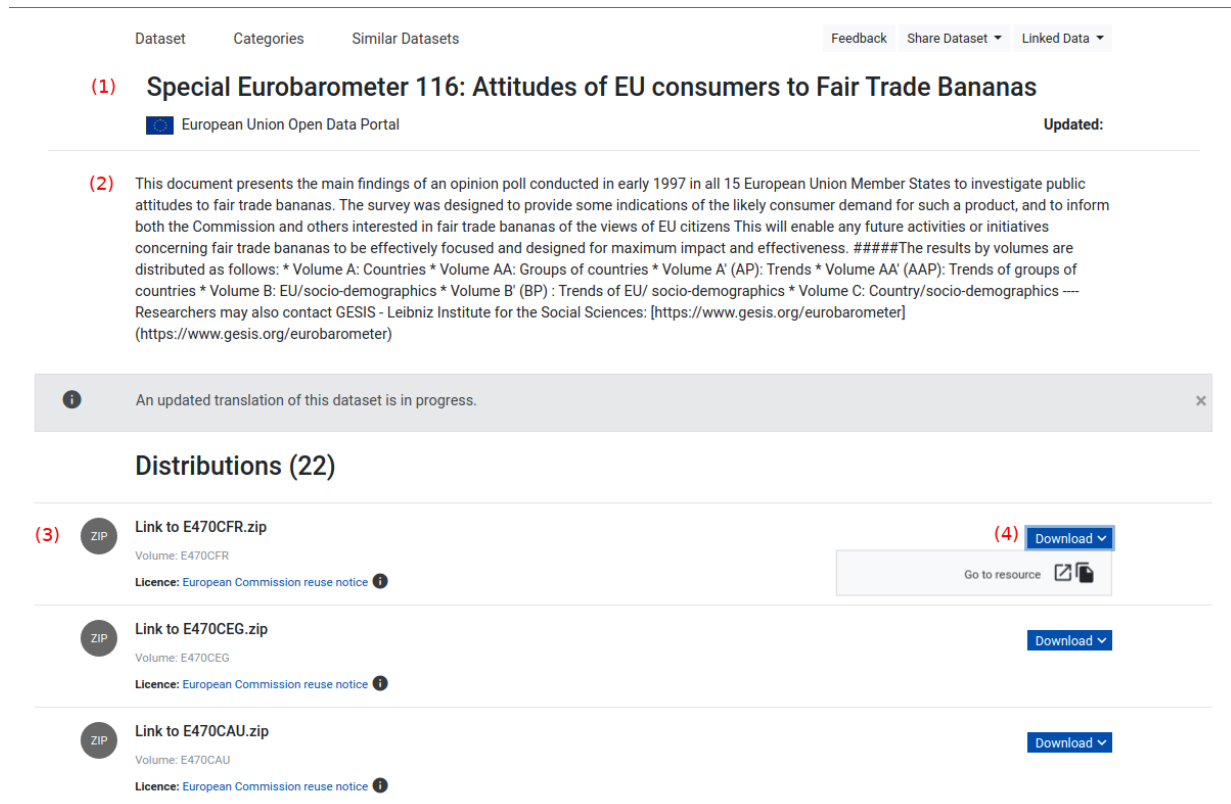
Figure 2: EDP's dataset search interface

The facets available on the EDP are:

1. Operator: Sets how multiple facets should be combined. Options: logical AND (show results that match all facets), logical OR (show results that match any of the facets)
2. Countries: Provenance of the dataset. Options: EU countries + EU Institutions.
3. Catalogues: Data catalogue from which the dataset was harvested. Options: All catalogues linked to at least one dataset in the result set.
4. Keywords: Dataset keywords according to the DCAT-AP description of the dataset. Options: All keywords detected in the datasets in the result set.
5. Licences: Licence(s) of distribution of the dataset. Options: All licences detected in the current result set.
6. Formats: Format(s) on which dataset distributions are available. Options: All format types detected in the current result set.
7. Location: Approximate geographic area covered or referred by the dataset: Control: Minimap where user can draw a rectangular area to approximate the location of interest, or name of the location (city, region, country).

After a query is issued, a ranked list of dataset summaries that are relevant to the query is shown in the centre of the screen. A dataset summary (6) is comprised of the dataset title and description as they appear on the metadata harvested by the EDP, the formats of the available distributions of the datasets, dates of creation and update, and the catalogue from where the dataset was harvested (In Figure 1, both datasets come from the EU Open Data Portal).

When a user clicks on a summary, they are redirected to the page of the corresponding dataset (Figure 3), with the full details about the dataset: (1) title, (2) description (3) distributions and (4) for each distribution, a link to download or go to the page of the resource in the catalogue of origin.



The screenshot shows the dataset page for 'Special Eurobarometer 116: Attitudes of EU consumers to Fair Trade Bananas'. At the top, there are navigation tabs for 'Dataset', 'Categories', and 'Similar Datasets', along with utility links for 'Feedback', 'Share Dataset', and 'Linked Data'. The main title is '(1) Special Eurobarometer 116: Attitudes of EU consumers to Fair Trade Bananas', with a sub-label 'European Union Open Data Portal' and an 'Updated:' field. Below the title is a detailed description '(2)' stating that the document presents findings from an opinion poll conducted in early 1997 across 15 EU member states. A grey notification bar indicates 'An updated translation of this dataset is in progress.' The 'Distributions (22)' section lists three items: '(3) Link to E470CFR.zip', '(3) Link to E470CEG.zip', and '(3) Link to E470CAU.zip'. Each item includes its volume ID and a 'Licence: European Commission reuse notice' link. To the right of each item is a '(4) Download' button and a 'Go to resource' link with an external link icon.

Figure 3: Dataset page

Having reviewed what we know about dataset search and retrieval from existing literature, including previous studies of ours, we now turn to the analysis of three years' worth of search and interaction logs of several iterations of the EDP platform.

### 3 Quantitative analysis

#### 3.1 Key terms

*User:* An online visitor of the European Data Portal.

*Query:* A set of keywords describing user needs.

*Actions and interactions* on the portal: For the web analytics package we use, actions are a superset of interactions. An interaction is a page view or a site search, while an action also includes downloads, outlinks, and other events.

*Search session:* A logged session that includes one or more queries.

*Search sessions vs. visits:* We differentiate between *search sessions* on the portal and *visits*, the latter being triggered by external queries leading to dataset pages.

*Issuing a query:* The activity of writing keywords in a search box.

### 3.2 The search and interaction logs corpus

The European Data Portal uses the [Matomo](#) Web Analytics suite to log the *actions* of users of the portal for each of their visits. The EDP uses the logs to create aggregate and anonymous statistics of user behaviour on the site, with multiple objectives:

- improving the site and ensuring its correct functioning;
- informing further design decisions;
- measuring its impact;
- customising the information or e-services of interest; and
- detecting and addressing any abuses or security issues.

A description of the capabilities of Matomo is available on this [link](#). In the following we provide a summary of those data attributes which feature in our analysis:

- **ID**: A unique identifier of the session.
- **Duration**: Duration of the session in seconds.
- **lastActionTimestamp**: timestamp of the last action of the visit, in UNIX time.
- **firstActionTimestamp**: timestamp of the last action of the session, measured in UNIX time.
- **actionDetails**: List of actions performed by the user. An action has the following fields:
  - **type**: Action type, can be one of:
    - **page URL**: an EDP page was loaded in the user browser.
    - **Click outlink**: User clicked on a link on the EDP that redirects to a non-EDP page.
    - **Download file**: User downloaded a file hosted in the portal.
    - **Search dataset**: User asked a query on the dataset search box.
  - **pageTitle**: If type = pageURL, the title of the page, else, blank.
  - **subtitle**: If type = pageURL, the subtitle of the page, else, blank.
  - **url**: For all action types except **search dataset**: URL clicked by the user in this action. Blank otherwise.
  - **siteSearchKeyword**: for actions of type **search dataset** and in the presence of user consent, contains the keywords typed on the dataset search box. If the action is not to **search dataset**, or the user did not give consent, this field is blank.
  - **Timestamp**: Timestamp of the action in UNIX time.
  - **TimeSpent**: Time spent on this action (in seconds).
- **referrerName**: Name of referrer website. A referrer website is a website from which the user clicked a link to land on an EDP page.
- **referrerUrl**, URL of referrer website or social network.
- **referrerTypeName**: Type of referrer, which can be a **search engine**, **website** or **social network**. When a referrer cannot be identified, this attribute is set to **direct entry**, that is, the user typed the landing URL directly onto the browser.
- **referrerSearchEngineUrl**: ULR of the search engine, if applicable.
- **referrerKeyword**: for search engine referrals, and if the referrer search engine makes them available, search keywords the user issued to the engine before getting to the EDP page. Unfortunately, in most cases, referrer search engines do not make this information available for privacy reasons.

We furthermore report three parameters of the EDP's Matomo configuration that affect data collection:

1. **Session timeout (in minutes):** This parameter refers to how long a web analytics package should wait after the last recorded action to consider the session finished. If a user returns to the portal within this time, their subsequent actions will be recorded as part of the same session; otherwise, the activities will be recorded as a new session. In EDP this value set as 30 minutes – this means that if no activity has been recorded for 30 minutes, any subsequent activity will be considered to belong to a new session. Setting the right value for this parameter should be the subject of future studies to better reflect the realities of dataset search and allow for more accurate follow-up session analytics.
2. **Exclusion of bots:** *Bots* are automated agents that crawl websites. As we are interested in understanding the behaviour of people on the portal, visits from bots should be excluded. EDP's web server proxy configuration provides a first line of defence against malicious bots. Matomo itself, with its default configuration, can filter out most bots that make it to the portal. Periodic analysis of this dataset for internal EDP reporting found no indication of skewing due to bot traffic.
3. **Time spent measurement:** Matomo's EDP kept a default configuration that does not allow the measurement of time spent on the last page of a session. This means that the available duration (in seconds) of a session is a lower bound of the real-time spent by the user. We consider this a limitation of this study. We suggest that data portals configure their web analytics packages for maximum accuracy. In Matomo, this can be achieved following these [instructions](#).

Since March 2018, the EDP has delivered **three major releases of the portal**. We describe the changes affecting the dataset search analysis below:

- EDPv1: From the 1st April 2018 to the 2nd April 2019 EDP's native **dataset search engine was based on CKAN**.
- EDPv2: From the 2nd April 2019 to the 6th March 2020, the **dataset section and search engine migrated** to a [solution developed in-house](#) by the EDP team. The URL scheme of the dataset section was changed, leading to a period where web search engines had to re-index those pages.
- EDPv3: From the 7th March 2020 onwards, **improvements were introduced on the dataset search engine following the evolution of the DCAT-AP standard**. This more or less coincided with the introduction of lockdowns in many European countries, which, as the logs show, affected traffic on the EDP.

In our analysis, **we split sessions on three disjoint datasets corresponding to each of the three releases of the portal**. Table 1 summarises the characteristics, date range and several sessions in each category. Note that v3 covers a shorter period than v1 and v2 (4 vs 11 months), meaning that we need to take care when comparing absolute values of statistics. We compare averages and medians, except when the highlight of an increase in absolute values is meaningful (e.g. increases of absolute visits to a section of the portal). One can also notice right away a decrease in the number of visits from v2 to v1, that have comparable lengths. We analyse the reasons for this decrease in our study.

**Table 1: Three corpora used to analyse sessions, corresponding to the three versions of the EDP in our dataset**

Portal version	Description	Date range (time in GMT time zone)	Number of sessions

<b>v1</b>	Dataset section and search engine based on CKAN	01 April 2018 00:00 to 01 April 2019 23:59	430,815
<b>v2</b>	Dataset section and search engine based on EDP's code base. Change of URL scheme on dataset section.	02 April 2019 00:00 to 06 March 2020 23:59	283,470
<b>v3</b>	Approximate start of COVID-19 outbreak in Europe. COVID-19 section on EDP Improvements on dataset indexation following DCAT-AP evolution	07 March 2020 00:00 to 30th June 2020 23:59	130,058

### 3.3 Results

In this section, we analyse the search and interaction logs to answer the four sets of questions introduced in Section 1.

#### 3.3.1 Dataset search in the context of the EDP

This first theme has been broken down into the sub-questions shown in Figure 4.

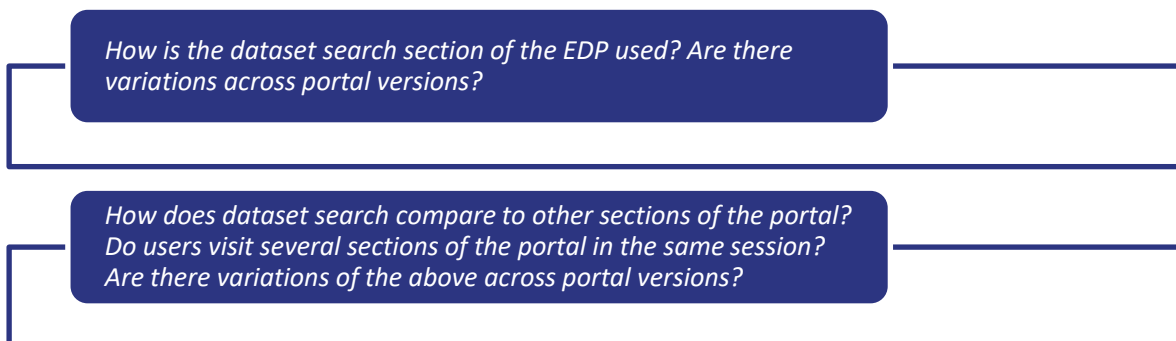


Figure 4: Questions of the dataset search in the context of the EDP theme

To answer these questions, **we had to find a way to classify sessions further depending on the main activities carried out by the user**. We did this for the three corpora corresponding to the three versions of the EDP covered by the logs. The classification looks as follows:

- 1) We call a session a **dataset search session** if one of the following applies:
  - a) The session includes actions related to entering a query into the search box one or more times.
  - b) The session includes actions related to filtering results with a facet one or more times.
- 2) We call a session a **dataset page session** if that session includes a visit to at least one dataset page (see Figure 3) and if the session is not a dataset search session. In other words, these are the sessions where the user landed straight on a dataset page without using the search box of the EDP. This happens, for instance, if the user was referred to that dataset page via an external website or by a web search engine.

- 3) We call a session a **homepage bounce** when the user visits the EDP front page or the dataset search the main page, but then does not follow up to any other parts of the portals and does not trigger any dataset searches.
- 4) We call a session a **section session** when the user visits at least one of the other main sections of the EDP site, including: News, Events & Highlights; Training & E-Learning; Reports & Studies; and COVID-19. The COVID-19 section is also divided into sub-sections similar to the ones of the main portal, including a dataset section featuring datasets manually curated by the EDP's editorial team. We consider the sessions that contain a visit to a page in the COVID-19 dataset section separately (we refer to that sub-corpus **COVID-Data** in our analysis, see Table 2), and those that visit one of the other COVID-19 sections, but do not visit any COVID-19 dataset section page (which we refer to as **COVID-Other**, see Table 2).

Figure 5 compares the percentage of sessions that visit each section, for each portal version. Absolute numbers are detailed in Table 2.

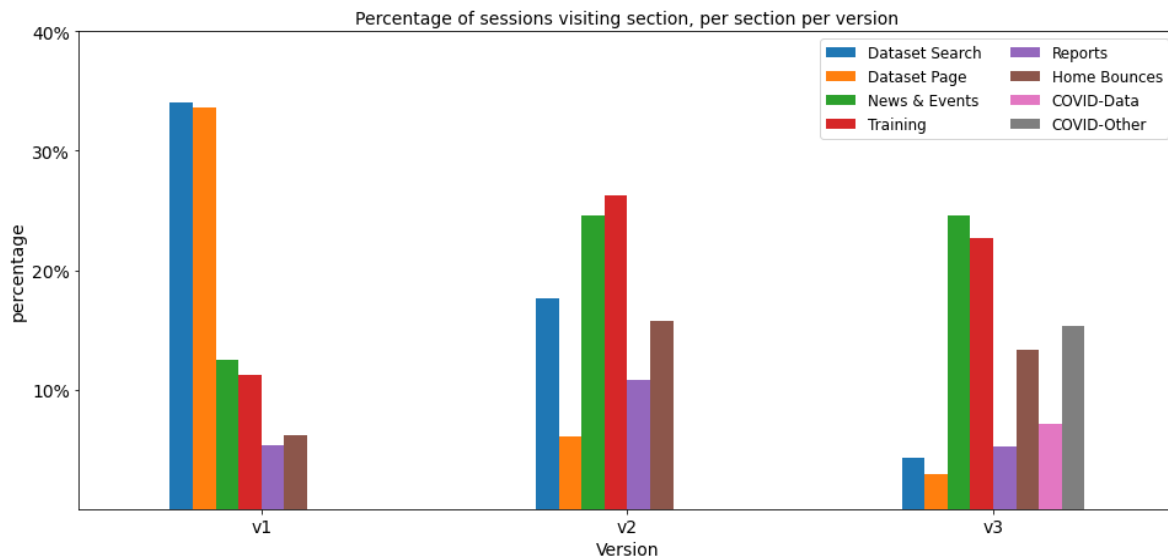


Figure 5: Percentage of sessions that visit each portal section

Table 2: Number of sessions per category per version of the EDP

Portal version	Total sessions	Dataset search	Dataset pages	Homepage bounces	News, events, highlights	Training, eLearning	Reports & Studies	COVID-Data	COVID-Other
v1	430815	146498 (34.0%)	144898 (33.63%)	26634 (6.18%)	53929 (12.52%)	48231 (11.2%)	23195 (5.38%)	N/A	N/A
v2	283470	49919 (17.61%)	17295 (6.1%)	44709 (15.77%)	69802 (24.62%)	90084 (31.78%)	30745 (10.85%)	N/A	N/A
v3	130058	5510 (4.24%)	3787 (2.91%)	17334 (13.3%)	32032 (24.63%)	29500 (22.68%)	6863 (5.28%)	9272 (7.13%)	19910 (15.31%)

During EDPv1, more than 67% of the sessions included a dataset search or a dataset page visit. The next most popular section was news, events & highlights with 12.5%.



For EDPv2 we observe a **sharp decrease in the number of overall visits to the portal**. The number of sessions including a dataset search or a dataset page visit decreased to 23%. All other categories experienced an increase in popularity, both in percentage and number of sessions, with training & E-learning being the most popular (with over 30% of sessions). We hypothesise that the reason for the decrease is caused by **web search engines not re-indexing a large number of pages of this section after the URL scheme was updated. We put this hypothesis to the test and further analyse the impact of web search engines on EDP's dataset search below.**

In EDPv3 **the percentage of sessions to the dataset section decreased to 7% of the total**. There is approximately the same number of sessions to the traditional dataset section than to the COVID-19 dataset section, making the overall percentage of dataset related visits increase to 14.3%, still lower than what was measured on EDPv2. Overall, all sections except for News & highlights experienced a percentage decrease in favour of the COVID-19 section, which was the second most popular during this time period with 22.45% of the sessions. Among the sessions related to COVID-19, **31% of users visited its dataset subsection.**

**To compute the number of cross-section sessions – that is sessions where the user visits multiple sections of the EDP site, we counted for each section the number of sessions that only visit that section and divided it by the number of sessions that visit that section or others.** Results are shown in Figure 6. We observe that across all versions of the portal, **more than 80% of visits to the dataset section do not crossover to other sections. There is a similar trend among the other sections except for the Reports section.** This suggests that we need more explicit links from dataset pages to related resources, as users will not be willing (or able, with the current designs) to find those added-value resources themselves.

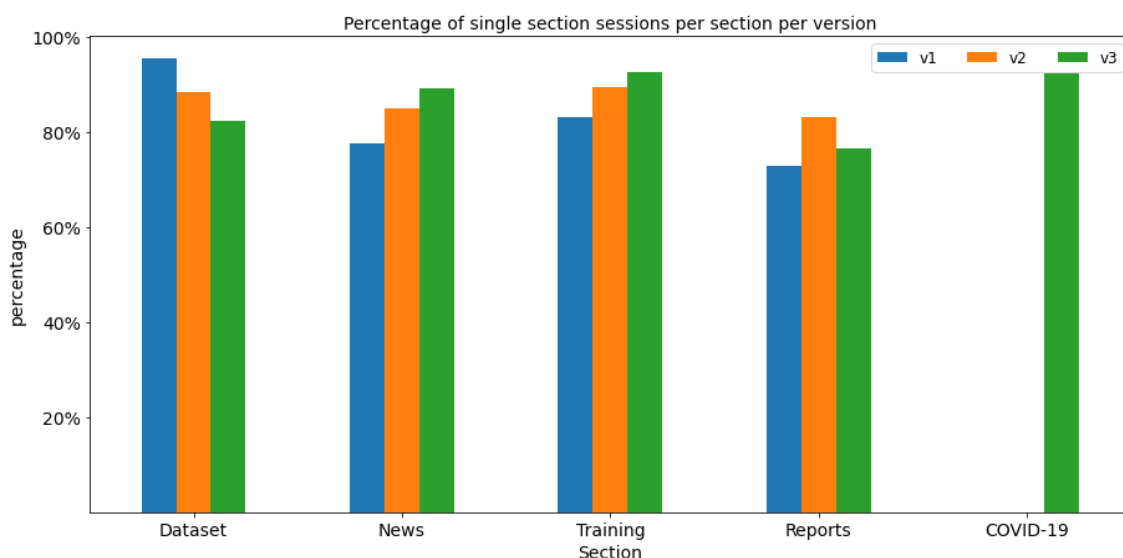


Figure 6: Percentage of single-section sessions

**We took a closer look at the crossovers between the dataset section and the others, as shown in Table 3.** Crossovers between sections in v1 were minimal, but increased for v2 and v3, in particular with the news, events & highlights sections. We believe this effect comes from the decrease in the number of visits to datasets (because of the change in URLs, for instance) combined with an increased use of highlights to announce new datasets. This showcases the added value of auxiliary content such as news, stories etc that facilitate re-use.

Table 3: Crossover between the dataset section and other sections

	Dataset section	Dataset only	Dataset + News	Dataset + Training	Dataset + Reports	Dataset + COVID-Data	Dataset + COVID-Others
v1	291395	277999 (95.4%)	8162 (2.8%)	4828 (1.66%)	3428 (1.18%)	N/A	N/A
v2	67214	59331 (88.27%)	5128 (7.63%)	2963 (4.41%)	1798 (2.68%)	N/A	N/A
v3	9297	7651 (82.3%)	859 (9.24%)	366 (3.94%)	265 (2.85%)	413 (4.44%)	116 (1.25%)

### 3.3.1.1 Summary of findings

*How is the dataset search section of the EDP used? Are there variations across portal versions?*

- Usage of the native search capability varied greatly among the three releases. Following v2, we saw a drop from 67% to 23% in share of sessions. In v1 the logs confirm the use of the EDP as a means to find data, but the changes in v2 meant that those numbers will need time to get back to their original levels, as external search engines re-index the resources. During v3 (COVID-19 outbreak), the number of visits to dataset sections further decreased to 7%, however we saw an additional 7% of sessions visiting the COVID-19 datasets section. In general, COVID-19 related content was well received, becoming the second most visited section of the portal (after news).

*How does dataset search compare to other sections of the portal? Do users visit several sections of the portal in the same session? Are there variations of the above across portal versions?*

- A large majority of users visit only one section of the portal at a time. For the dataset sites (search, dataset descriptions) the number of crossovers with other sections was less than 5% in v1, but increased up to 18% in v3, mostly due to more links to and from news, events & highlights section. The trend is opposite for the news and training sections. In summary, there are significant variations between different release versions of the portal. We recommend making links between datasets and other content more explicit, in both directions, and investing in automatic tools for creating these links at scale, similar to the work on interlinking datasets. We also believe the EDP should run subsequent analyses like ours to capture the effects of referrals and indexing by external search engines. Such analyses should not be a one-off activity, but be undertaken regularly.

Having gained a high-level understanding of the search and interaction sessions across the three versions of the EDP, we now deep dive into the search behaviour in terms of search strategies and query characteristics.

### 3.3.2 Dataset search strategies and search query characteristics

Users may follow three different strategies to search for datasets on the EDP. For instance, they could:

- type keywords into the search box;
- apply facets to filter datasets; and
- a combination of the above.

In this section, we aim to describe such strategies and associated queries. We pursue the following questions, depicted in Figure 7.

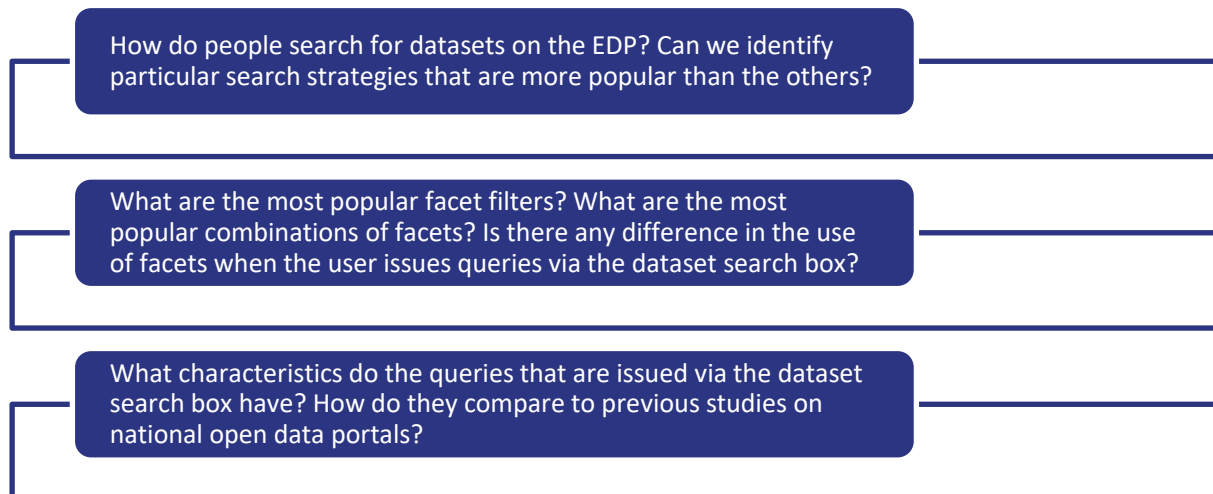


Figure 7: Questions addressed in the search strategies and query characteristics theme

### 3.3.2.1 Logs we analysed

Not all search and interaction logs are relevant to understand search strategies and search query characteristics. To restrict the analysis only to the relevant sessions, we had to remove, say, sessions where a user has visited the EDP to access news, learning materials etc rather than search for data. In this section, we explain how we put together the subset of logs that are relevant to dataset search sessions. We undertook several steps.

**Our starting point is the dataset search sessions dataset.** As explained earlier, it is made of **sessions that include at least a keyword query or the application of a filter via one or more facets** (see Figure 2).

Furthermore, we **distinguish in our analysis between two scenarios for users to reach EDP's dataset search pages:**

- 1) A user could land on an EDP page that is not dataset search related (e.g. the homepage). They could then move to the dataset search interface, enter a query, or apply filters. We refer to these sessions as **internal dataset search** sessions.
- 2) Alternatively, a user could reach straight a dataset search result page by following a link from another website or a result returned by an external search tool. Web search engines crawl and index result pages (e.g. <https://www.europeandataportal.eu/data/datasets?keywords=karte>). We refer to these as **external dataset search** sessions.

We wanted to focus our attention on **dataset search actions performed by users using EDP controls.** To achieve that, we first **divided the external dataset search sessions into two parts:**

- 1) External search sessions that **do not include any further queries or use of faceted search.** In other words, the user has issued a query outside of the EDP and clicked a link that sent them to EDP pages. We call this set of sessions in our corpus **external landing without further search.**
- 2) External sessions that **include further use of keywords or facets past arrival on an EDP page.** In the case, the user landed on some EDP content following an external link, then went ahead and searched on the EDP using native tools. We call this **external landing with a further search.**

The analysis hence covers:

- 1) internal dataset search sessions (where the user started the session on EDP and searched there) and
- 2) external landing with further search sessions (where the user came from somewhere else but went on to search natively on the EDP).

For the sake of simplicity, we refer to both these categories as **internal dataset search** for the remainder of this section.

Table 4 shows the number of sessions of each of the datasets described above

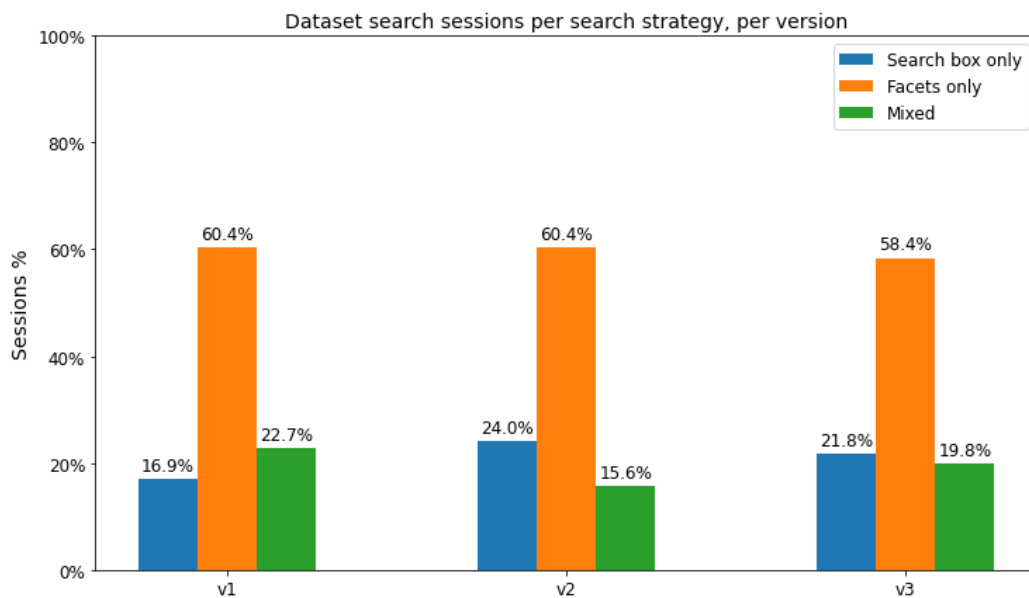
**Table 4: Breakdown of dataset search sessions by starting point and search continuation**

	All dataset search	Internal dataset search (A)	External landing with further search (B)	External landing without further search (C)	Logs we analysed (A) + (B)
v1	<b>146498</b>	49162	33279	64057	<b>82441</b>
v2	<b>49919</b>	25964	5847	18108	<b>31811</b>
v3	<b>5510</b>	3432	566	1512	<b>3998</b>

Now that we have described the log dataset which we used for the analysis, we will deep dive into two aspects: the use of search affordances on the EDP, and the types of queries people write.

### 3.3.2.2 Use of search box and facets

Earlier in this section, we noted that users can follow different strategies to look for the data they need, including queries, faceted search or both. Figure 8 shows the percentage of sessions per search strategy for the three versions of the EDP. Using only facets to search is significantly more popular than the other two throughout the evolution of the portal.



**Figure 8: Dataset search sessions per search strategy, per version**

Next, we calculated the share of sessions that use each of the available facets at least once. Figure 9 shows the results for sessions that used only (left) and those relying on keywords and facets (right).

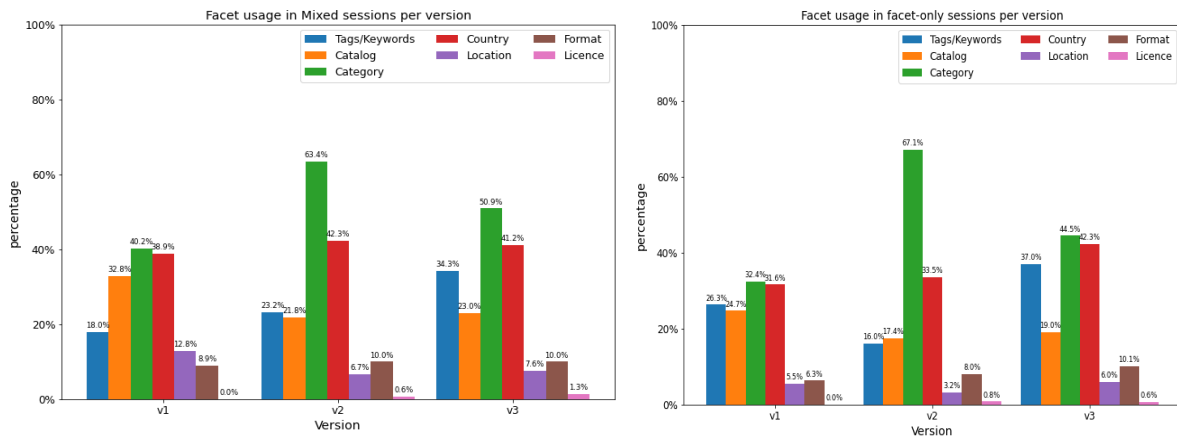


Figure 9: Use of facets per version for sessions that used facets (left) vs search box + facets (right)

For those cases where the users relied only on faceted search, we note a slightly more even distribution of facet types in EDPv1 than EDPv2 and EDPv3. V1 was also the corpus for which the majority of recorded user activity was dedicated to search. In V1, people used tags/keywords, catalogue, category and country extensively. For EDPv2 the category facet is by far the most popular; we hypothesise that this is a function of the homepage redesign, which includes a panel where users can start exploring the datasets along categories. For v3, the difference between a category and the others does not stand out as much. Overall, we observe that facets related to the format or license of the dataset, as well as its geolocation, are under-used. This is in stark contrast to some of the insights we gained in previous studies by analysing the queries people write or interviewing data practitioners about their search strategies (Koesten et al. 2017), which suggested that these are three key attributes that decide if a dataset will be eventually re-used.

In the sessions that included both search box and faceted search activities, we noticed similar distributions, suggesting that the use of the search box does not affect how facets are used. The category facet remains in high demand, even in EDPv1, compared to the sessions that did not use any queries. We expected this, as a query via the search box is conceptually similar to the tag/keyword facet – strictly speaking, the difference in the EDP implementation is that the facet suggests keywords to the user, whereas in the search box there is no auto-completion or other support.

Finally, we calculated the most popular facet combinations for the two types of sessions from Figure 9. This helps us understand the type of information needs people have and the attributes of the data that matter to them while looking for data. The table shows the combinations found in at least 5% of the sessions. Across all versions, the most common combination involves countries and categories. The popularity of this combination greatly increases from v2 onwards, possibly in relation to the new design of the site which promoted dataset exploration by category. We also note that during v1, there was a lesser use of combined facets, despite a higher share of search sessions overall.

Table 5: Percentage of sessions with more than one facet per portal version. We only show combinations that reached more than 5%. Pairs are not ordered.

Version	Only facets	Query + facets
EDPv1	(Country, Category) -> 5.4%	(Country, Category) -> 9.9% (Catalog, Category) -> 8.3% (Country, Location) -> 8.0%

		(Catalog, Country) -> 7.6% (Tags, Country) -> 6.6% (Country, Format) -> 5.6%
<b>EDPv2</b>	(Country, Category) -> 17.42% (Country, Keywords) -> 6.8% (Catalog, Category) -> 6.2% (Catalog, Country) -> 6.2% (Category, Keywords) -> 5.5%	(Country, Category) -> 23.1% (Country, Keywords) -> 10.7% (Catalog, Category) -> 10.3% (Category, Keywords) -> 10.0% (Catalog, Country) -> 9.0% (Country, Category, Keywords) -> 6.8% (Country, Format) -> 6.6% (Category, Format) -> 6.2% (Catalog, Keywords) -> 6.1% (Catalog, Category, Country) -> 5.7%
<b>EDPv3</b>	(Country, Category) -> 17.4% (Country, Keywords) -> 9.8% (Catalog, Country) -> 7.9% (Catalog, Category) -> 6.5% (Country, Format) -> 6.38% (Category, Keywords) -> 6.0%	(Country, Category) -> 18.6% (Country, Keywords) -> 11.7% (Category, Keywords) -> 9.9% (Catalog, Country) -> 9.9% (Catalog, Category) -> 8.4% (Country, Format) -> 6.0% (Country, Category, Keywords) -> 5.2% (Catalog, Category, Country) -> 5.1% (Catalog, Keywords) -> 5.1% (Category, Format) -> 5.0%

### 3.3.2.3 Search box query characteristics

In the previous section, we looked at people use the EDP search box and facets to find the data they need. Here, we analyse the queries they pose in the search box. As a reminder, we do not have access to the queries people use when starting their sessions elsewhere for privacy reasons. This is why we focus only on internal search logs described in Section 3.3.2.1.

For the analysis of the queries, we build on top of a previous study we have carried out (Kacprzak et al., 2019). It looked at search logs from two UK open government data portals, establishing benchmarks on **query length** and **types of keywords** used. The former is an indicator of style. The second could inform the design of facets, as well as metadata schemas. In addition to these two dimensions, we considered the **language used to write the queries**, accounting for the European character of the EDP. We wanted to know whether English is the main language of choice of EDP users to express their data needs or whether other languages are used as well.

The corpus we used consists of all queries made via the EDP search box as EDP's Matomo. Figure 10 shows the number of all queries compared to the number of unique queries – unique queries are queries resulting from eliminating duplicates from the list of all queries. As we discussed earlier, EDPv2 and v3 saw a fall in the share of dataset search sessions. This is consistent with the drop in the number of queries and unique queries.

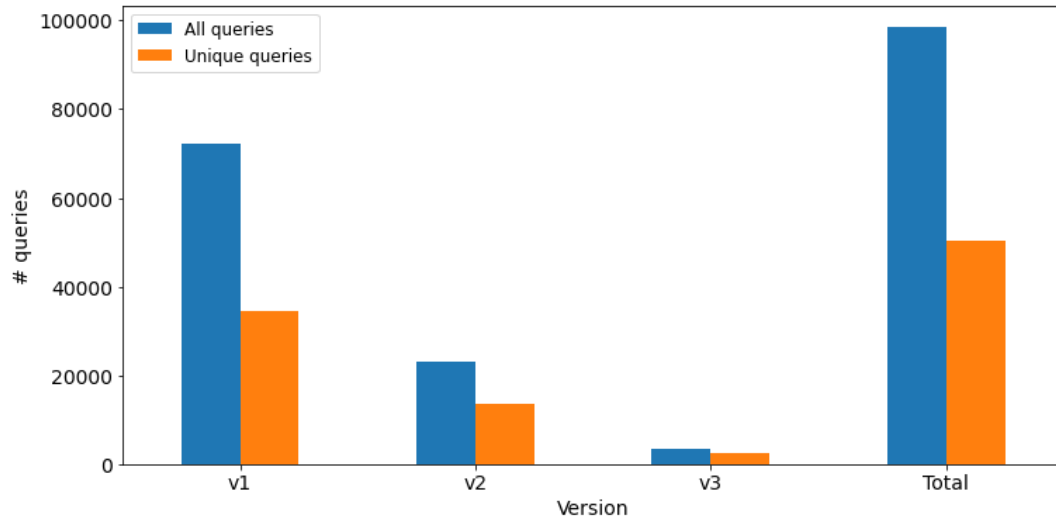


Figure 10: Number of queries and unique queries issued to EDP's dataset search box, per EDP version

### 3.3.2.3.1 Query length

Figure 11 shows the mean, median, and mode of all and unique queries. For all three releases of the EDP, the mode is 1. The median is also 1 for all combinations, except unique queries in v1 and the total. The mean also shows similar values among versions, close to 1.5 for all queries and close to 2 for unique queries. Due to the small variation of these statistics between versions we consider for the rest of our analysis a single dataset comprising the aggregation of all queries.

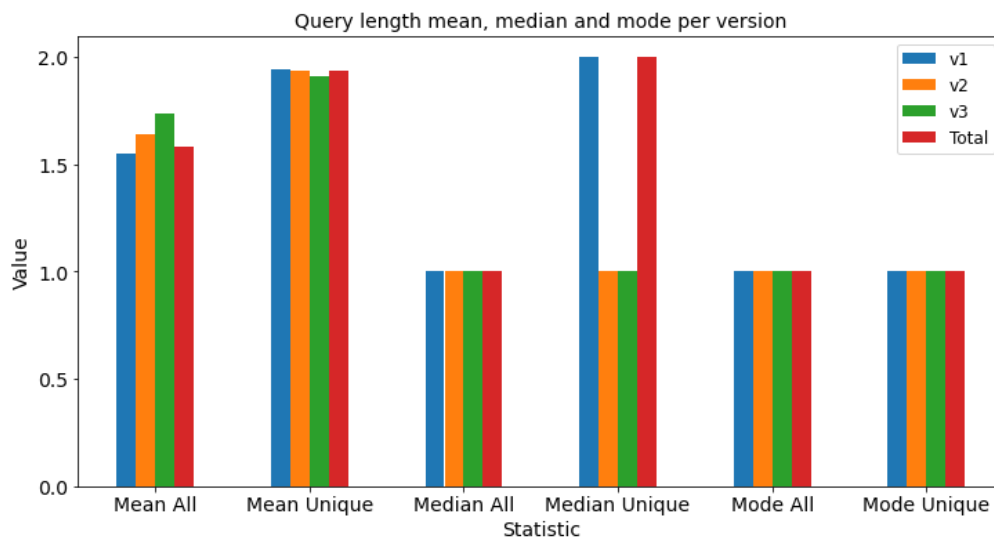


Figure 11: Query length mean, median and mode per EDP version

Figure 12 shows the query length distribution for all queries (left) and unique queries (right). The large proportion of single-word queries is consistent with results previously reported for UK open government data portals. This suggests that users may be using the search box in a similar way to a facet, that is, to sift through datasets rather than writing longer, more complex queries or ask questions, as it is often the case for informational queries on the web. It may also indicate users' perception that the search capabilities are limited – hence, queries are held more general and results are filtered manually or through filters.

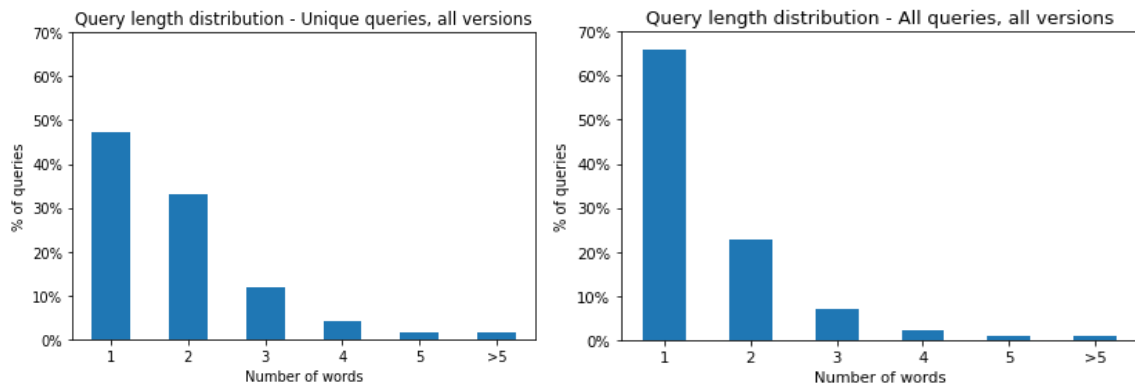


Figure 12: Query length distribution for all queries (left) and unique queries (right), for the 3 EDP versions

### 3.3.2.3.2 Query language

As automated language detectors are not accurate for very short snippets of text like the queries in our dataset, we manually inspected a subsample of the queries to get a sense of the languages in use. We sampled from the queries that have more than one word, which was asked in at least 5 sessions. We ended up with a total of 386 queries across the three EDP versions, which corresponds to 1.5% of all 25,566 multi-word queries. 80% of these 386 queries were identified to be in English, with German, Spanish, Polish, French and Italian ranging between 1 and 5%. In 2.5% of the cases we could not determine the language (e.g. the query "corona virus" is valid in many languages).

While the sample we used for the analysis is small, the results clearly show that an English-speaking audience of the EDP, though multilingual and cross-lingual support remains relevant.

### 3.3.2.3.3 Keyword types

Following the methodology proposed by (Kacprzak et al., 2019), we looked for the following types of words in queries:

- 1) **Temporal queries:** years (1000 to 2017), names of months, and the words week(ly), year(ly), month(ly), day(ly).
- 2) **Format queries:** file types: csv, pdf, xls, json, wfs, zip, html, api.
- 3) **Data-related queries:** Type of dataset related keywords: *data, dataset, average, index, graph, table, database, indice, rate, stat, map*.
- 4) **Countries queries:** Name of a country.
- 5) **Location queries:** Geospatial locations that are not countries (cities, regions, etc)

Countries and locations are difficult to detect automatically – our experience from previous studies show the queries are prone to errors, inconsistent spelling etc. In addition, there are also challenges around the languages used to denominate each of them.<sup>5</sup> Therefore, we manually labelled two samples: one-word queries used in 20 or more sessions (467 queries in total), and the same sample of multi-word queries used in 5 or more sessions that we used to estimate language distributions (386 queries in total).

Figure 13 summarises the results. Single-word queries have less than 3% of temporal, format, and data types, but more than 10% of both country and locations. The relatively high usage of countries and locations in single-world queries is noteworthy, as this is conceptually similar to using a country or

<sup>5</sup> [https://en.wikipedia.org/wiki/Names\\_of\\_European\\_cities\\_in\\_different\\_languages:1%E2%80%9393L](https://en.wikipedia.org/wiki/Names_of_European_cities_in_different_languages:1%E2%80%9393L)



location facet. Multi-word queries have a slightly higher usage of temporal types, data-related keywords, and fewer country names. Compared to the results reported for the UK portals analysed in (Kacprzak et al., 2019), we note less temporal, format and data-related keywords are much less frequent, while geospatial (cities, regions etc.) ones are more popular. While the samples are small, further studies should explore whether this is due because of the nature of the EDP, which harvests across multiple portals from different levels of public administration in different countries. By comparison, the 2019 study looked at portals publishing their own datasets (Office of National Statistics in the UK, data.gov.uk).

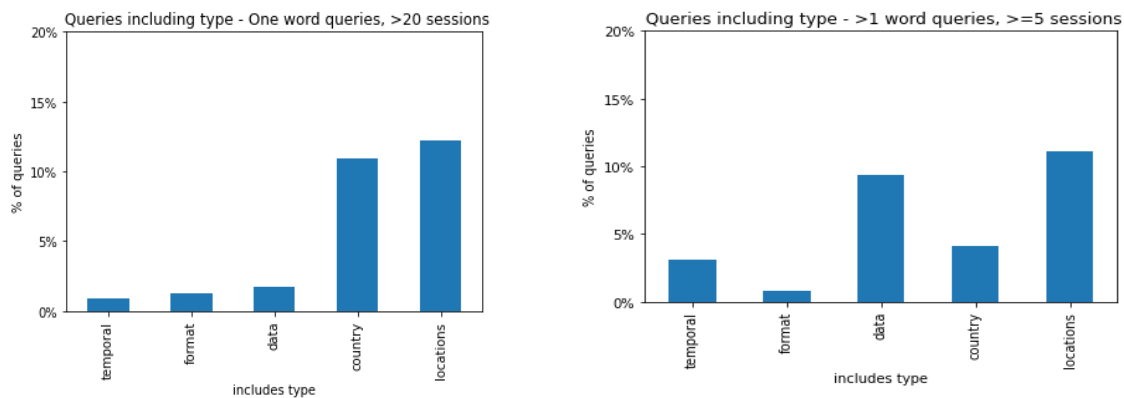


Figure 13: Percentage of queries including keyword types, for single-word queries (left) and multi-word queries (right)

### 3.3.2.4 Summary of findings

*How do people search for datasets on the EDP? Can we identify particular search strategies that are more popular than the others?*

- Yes, the use of only facet filters is more common than the two other strategies combined.

*What are the most popular facet filters? What are the most popular combinations of facets? Is there any difference in the use of facets when the user issues queries via the dataset search box?*

- Country and category are the most popular, and also the most popular combination, followed by combinations of each of these two with the keywords facet. Format and licence are the least popular. Sessions that use the dataset search box use more combinations of facets than facet-only sessions.

*What characteristics do the queries that are issued via the dataset search box have? How do they compare to previous studies on national open data portals?*

- Most queries are single-word. Compared to previous search log analyses, the EDP queries show less use of temporal, format and data-related keywords, but include more often references to various types of locations such as countries, regions etc. Single-word queries have a surprisingly large proportion of country and other location types, suggesting users might be using the search box as a facet to filter results instead of for querying.

### 3.3.3 EDP vs. web search engines in dataset search

Previous research (Koesten et al., 2017) found that dataset searchers tend to use web search engines to look for data. Setting aside the fact that the first impulse for many people for searching anything is to go to a web search engine, these general-purpose tools also have become more and more efficient in crawling, indexing, and identifying datasets, or even developing their specific dataset search engines

(e.g. Google Dataset Search). In this section, we explore the relationship between sites such as the EDP and web search engines, through the lens of the following research questions (Figure 14).

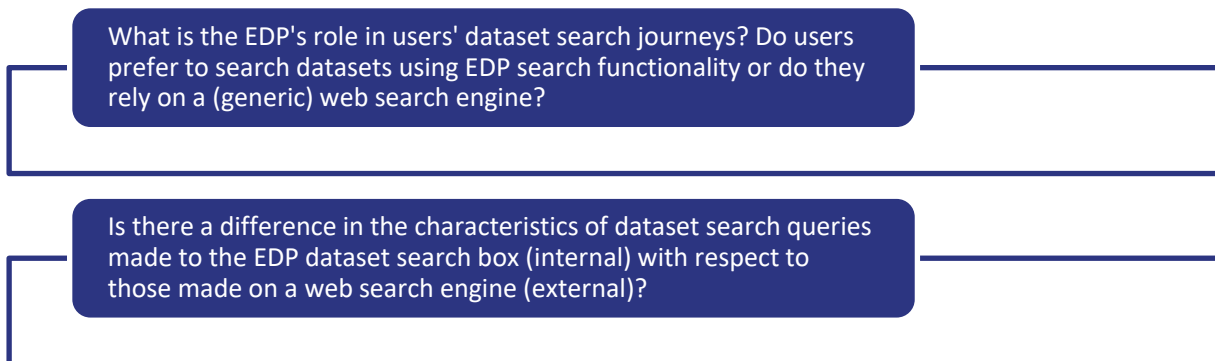


Figure 14: Research questions pursued in the third theme of the study

### 3.3.3.1 Typology of user journeys

To understand the relation between EDP and other external sites which link to its content, we first defined a set of typical user journeys during dataset search:

- 1) **User journey (1):** "I go to the EDP, I use their native dataset search engine e.g. the search box or facets". This is good news for the EDP: it means that users recognise it as a useful tool for searching datasets.
- 2) **User journey (2):** "I go to my preferred web search engine. One of the search results is a dataset page or dataset search result on the EDP. I click on it". Not as good as (1), as users rather use general web search engines, but the EDP dataset section is ranked high enough and is appealing enough to drive traffic.
- 3) **User journey (3):** "I go to my preferred web search engine. I end up on some website, forum, blog or social network page that includes a link to a dataset on the EDP. I click on it." This type of journey is facilitated by the fact that the EDP is linked to other web sites. At the same time, it also means native search affordances are not used and traffic depends on external sites and search engine algorithms.
- 4) **User journey (4):** "I go to my preferred web search engine. One of the search results is a dataset page or dataset search result on the EDP, but I don't click on it, or there is no EDP's dataset page or dataset search result in the search results". Bad news for the EDP dataset search section, users rather use general web search engines, and the section is not well ranked or appealing enough to drive traffic.

### 3.3.3.2 Estimating the number of user journeys of each type

As a next step in our analysis, we estimate user journeys of types (1) and (2) and (3) using the same method described in Section 3.3.2.1. Type (1) journeys correspond to genuine internal search sessions. Types (2) and (3) journeys correspond to sessions that start on a dataset section page and have been referred by a search engine (for type 2) or a website or social network (for type 3), which we included in our analysis of dataset search strategies earlier. We note that our method possibly overestimates types (2) and (3) journeys, as we cannot be 100% sure of the user's intent when being referred to the EDP from outside. For example, a user might be doing an unrelated search on the web and accidentally lands on a dataset page or they read an online article that links to an open government dataset indexed by the EDP. In such cases, we cannot be certain that the user's intent when clicking that link was to look for data.

We estimate user journeys of type (4) and get the external queries from the Google web search engine using data from Google's Search Console (GSC). According to data from [StatCounter](#), Google has had more than 93% per cent of the market share for the last 3 years, suggesting that GSC data is significant. For further confirmation, we counted the number of sessions referred by each search engine for dataset section sessions for the three versions of the portal. We found that 94% were referred by Google Web Search, with a further 4% coming from the Google Dataset Search Engine.

The GSC registers:

- a list of queries entered into Google's web search engine where a page from the EDP appeared as a hit page viewed by the user (called an **impression**);
- the **average position** of the topmost EDP page on the search results;
- the number of **clicks** on an EDP page shown on the search results, and
- the ratio of clicks over impressions (called **clickthrough rate**).

The inverse of the click-through rate provides a lower bound estimate for our type (4) journeys. It represents the percentage of times that users were shown an EDP page on a list of search result but chose not to click on it. We do not have data on queries that intended to find a dataset and were not shown an EDP page in the search results.

GSC only makes available daily data for the last 3 months, limited to the 1000 queries with the number of clicks, as well as aggregates for the last 6, 12, and 16 months. To circumvent this limitation, the EDP has been collecting daily data since the beginning of EDPv1 using the *SearchEnginePerformance* Matomo plugin. For each version, we selected the top 1000 queries in the number of clicks. However, GSC does not make available the link between queries and the actual page, meaning that we cannot know if the page shown to a user is from the dataset section or not. As a consequence, our method would overestimate journeys of type (4), as it may include impressions of pages of other sections of the portal. To reduce this error, we manually label the queries as "dataset" or "other" using the following heuristic:

1. Select the subset of queries that contain data-related and format keywords, using the same method we used for internal queries (Section 3.3.2.3.3). From this subset we discard:
  - a. Queries that contain the keyword "data portal", like "European data portal", "European Open Data Portal" or "German Data Portal". We also removed queries including URLs or URL fragments (e.g., 'mapy.geoportal.gov.pl'). These are called "proxy queries" and in most cases, they have the intention of reaching a specific website for which the exact URL is not remembered.
  - b. Queries referring to data-related events, e.g. 'International Open Data Conference' or 'Data Day 2019'.
  - c. Queries referring to known open data reports, e.g. "Open Data Maturity".
  - d. Queries with keywords that suggest the user was looking for something different to a dataset, e.g. "Open Data definition" "How to clean data" "data scraping software", "data exchange platform", "SPARQL tutorial pdf".
  - e. We chose to keep queries like "open data", "EU open data" and "[countryname] open data".
2. From the subset of queries that do not contain data-related or format keywords, we applied the following criteria:

- a. For queries where an acronym could be identified, we searched for the acronym on Google and verified if in the first page of results we could find a result suggesting that the acronym is from a dataset. If so, we include the query, otherwise, we discard it. Examples of dataset acronyms we found are “LIDAR”, “nuts3” and “srtmlg1”, examples of acronyms that could not be verified as corresponding to a dataset or were found to be from something else are “IODC”, “iQuanta”, “CKAN”.
- b. We removed the remaining queries, for example, “wind power” (we did include “wind power data”), or “postal codes” (we did include “list of postal codes France”). We made one exception for queries referring to the Oxford Covid-19 government tracker. “Tracker” is not in our list of data related keywords, but we know this particular tracker is a dataset.
- c. Note it is possible to judge certain queries as asking implicitly for a dataset. For example, the query “limits of Italian territorial waters” may be interpreted as implicitly asking for a map or shapefile of the limits, and the query “EU agricultural subsidies by country” may be understood as expecting a table as output. We chose not to label these queries as dataset ones.
- d. We used the queries of type “dataset” for the characteristic analysis. As we used data and format keywords to construct the dataset, we focused our comparison on the temporal, country, location, and language dimensions.

### 3.3.3.3 Results

Figure 15 shows the user journey type for each of the three versions of the portal. We show the percentage of dataset section sessions (left), and a number of sessions (right). The red bar (N/A) represents sessions for which Matomo could not determine the referrer (mostly due to user privacy settings). During v1, we observe that more than 60% of user journeys are type (2), which is consistent with the use of the EDPv1 as a tool to find data. For v2 and v3 this number falls to between 30-40%, similar to type (1) journeys. We believe this change was not due to users switching from web search engines to the EDP but is merely a consequence of the re-design of the site which disrupted how external search engines indexed EDP datasets.

To dive deeper into these findings, we considered three possible scenarios:

- A. Users started ignoring results from the EDP, that is, journeys of type (2) changed to journeys of type (4);
- B. Web search engines lowered the ranking of EDP pages, or took them from their index altogether; or
- C. a combination of (A) and (B).

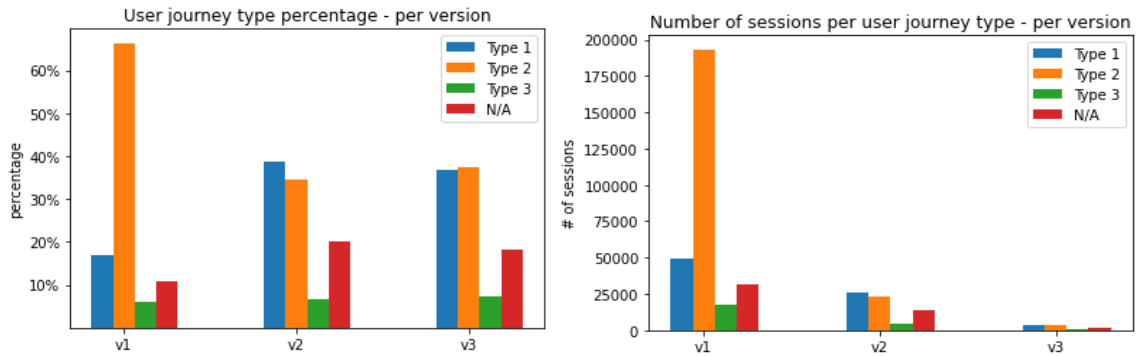


Figure 15: Estimation of user journey types per EDP version

To estimate (A) and (B), we make use of the GSC dataset. A decrease only in clickthrough rate between versions would suggest (A), while a decrease in the number of dataset queries, impressions and average position with a stable click-through rate would suggest (B).

Figure 16 shows the number of dataset queries (left) and the sum of impressions (right) for EDPv1, v2 and v3. For reference, we include the sum of impressions for all queries in the version sample. Figure 17 shows the average position per query type (left), the overall click-through rate CTR (right). Again, for reference, we compare between dataset and non-dataset queries. We observe a **decrease in the number of dataset queries and impressions from v1 to v2**, while the average position and CTR remain approximately the same. This suggests that **pages in the dataset section were removed from the index or their ranking was decreased**. The overall CTR between 5 and 7% is consistent with industry reports from [BackLinko](#) and [AWR](#) for the average position of EDP pages on Google's result pages (between 6th and 7th). The latter also implies that there is a **large number of type (4) journeys, as more than 90% of users that are shown an EDP page, do not click on it**. We believe this is a consequence of the average position of EDP pages on results. We discuss in section 4.4 possible improvements from a Search Engine Optimisation perspective.

We also point to an **increase in dataset queries and their impressions from v2 and v3**. We believe this is due to the general interest in COVID-19 related datasets during this period. To validate our hypothesis, we counted the number of dataset queries in v3 that include the words 'COVID' or 'Corona' with any combination of lowercase and uppercase. 46% queries (234 out of 506) contained those keywords.

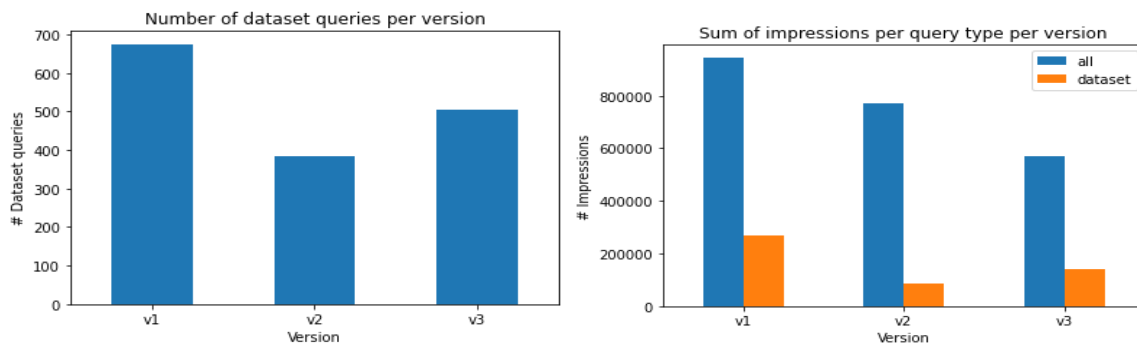


Figure 16: Number of dataset queries per version (left) and the sum of impressions (right)

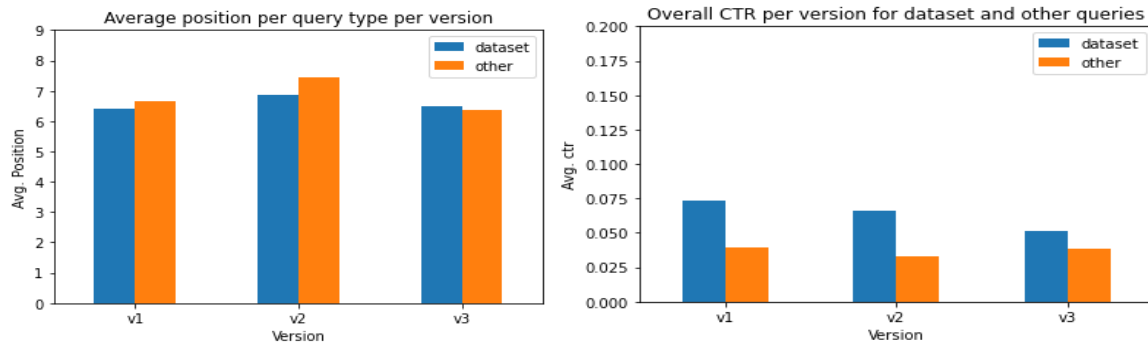


Figure 17: Average position per query type (left), overall CTR (right)

We then analysed the queries in the same way we did with internal queries (that is, queries entered in the EDP search box) in Section 3.3.2.3.

Figure 18 shows the mean, median, and mode of the **length of the relevant queries**, for the three releases of the portal. **The statistics are very similar, with mean, mode and median around 3.** This is consistent with results observed for the UK open data national portals cited earlier, which reported a mean of 2.76 for external queries.

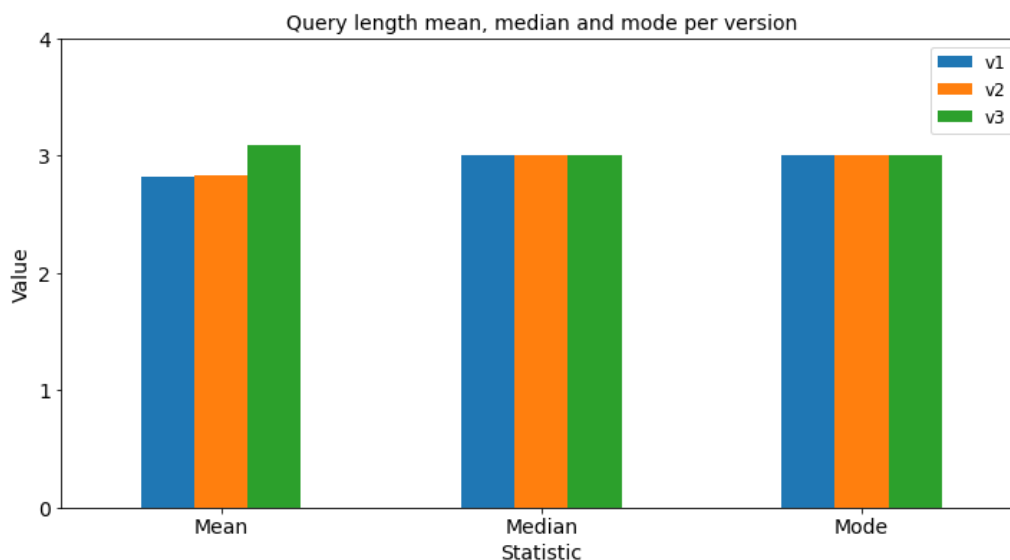


Figure 18: Mean, median and mode length for queries in the GSC dataset, per EDP version

Figure 19 compares the **query length distribution of internal and external queries**. There is a very low number of queries with one word, most of them correspond to acronyms of known datasets such as PKWIU or LIDAR. This shows that people tend to use external search engines and native search tools offered by the EDP differently. It would be interesting to run a follow-up study to understand why, one possible reason being different intents, different expectations in terms of search performance, the lack of faceted search in web search, or different groups of users altogether.

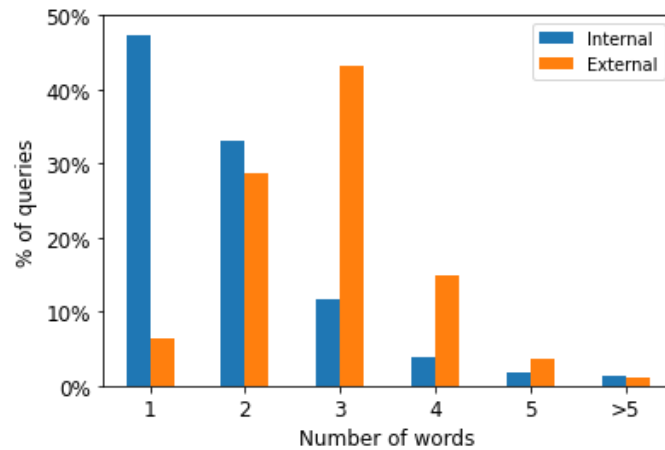


Figure 19: Query length distribution comparison between internal and external queries

Similar to the method followed in Section 3.3.2.3, we then looked at **query topics**. Figure 20 shows the percentage of queries containing a country, geolocation or temporal keyword. As the number of single-word queries is much smaller than for internal queries, we **analysed queries of more than one word only**. Compared with multi-word internal queries, for which we reported small percentages of queries (between 3 and 12%) using countries, other locations or temporal words, for external queries, queries are different. A much larger share includes geospatial keywords (between 20 and 25%), while temporal keywords are rare. As hinted at earlier, we believe this can be due to a range of reasons, including, among other things, the additional capabilities to use filters on the EDP.

In terms of **language distribution**, we identified **65% of queries as English**, significantly less than for our sample of internal queries. German had 10%, with Polish, French, Italian, and Romanian ranging between 1 and 5%. 9% of queries were judged as "can't determine precise language", three times more than for our sample of internal queries. This suggests that users feel compelled to ask questions to the EDP in English, we hypothesise that this is because users perceived the EDP as an international portal for which most content is already in English.

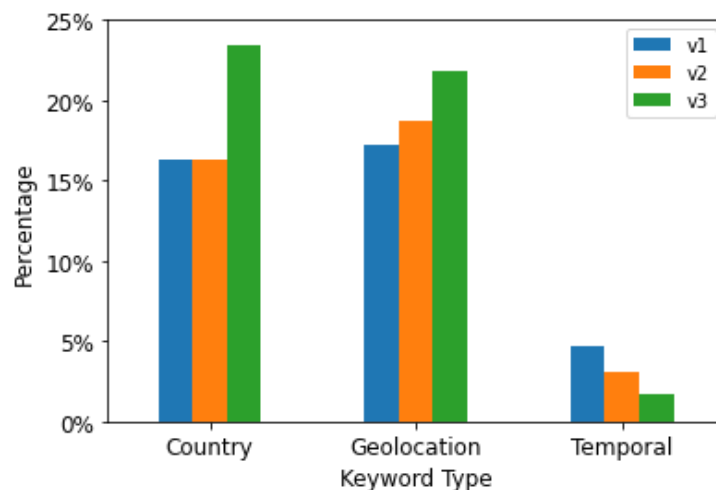


Figure 20: Percentage of queries containing query type, per EDP version

### 3.3.3.4 Summary of findings

*What is the EDP's role in users' dataset search journeys? Do users prefer to search datasets using EDP search functionality or do they rely on a (generic) web search engine?*

- Our data suggests that users preferred web search engines. For EDPv1, more than 60% of dataset searches came from web search engines. From EDPv2 onwards, the number of dataset searches dropped, an issue we were able to empirically track as a consequence of EDP dataset pages and SERPs being removed from search engine indexes. This highlights the importance of web search engines for dataset portals and the need to develop a strategy for this alongside improving the performance and user experience on native search capabilities.

*Is there a difference in the characteristics of dataset search queries made to the EDP dataset search box (internal) with respect to those made on a web search engine (external)?*

- Yes, queries to web search engines use more keywords. We believe this is a consequence of not having facet filters available, prompting users to be more descriptive about their needs. We also found more internal queries in English than external. However, our sample is biased towards the most popular queries, further analysis on the tail of the distribution should be conducted for more in-depth insights.

### 3.3.4 Success in dataset search

In this final theme of the study, we explore the following research questions (Figure 21).

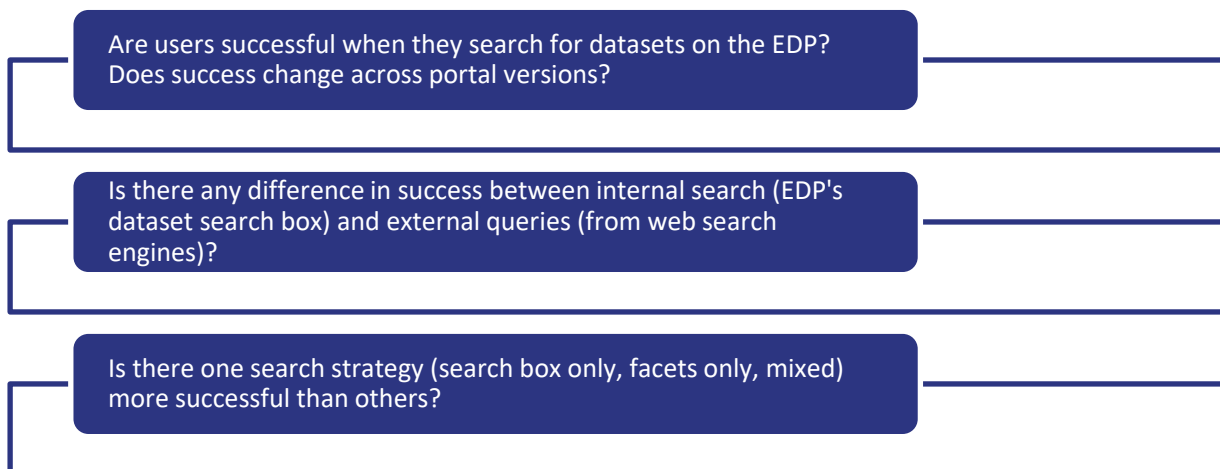


Figure 21: Questions addressed in the success in dataset search theme

We assume that users who visit the dataset section of the portal have information needs of the form "Find a dataset(s) that is(are) relevant to my criteria". It is best practice to ask users who visited an online site to provide feedback on the purpose of their visit and whether they found what they were looking for. **In the absence of such explicit user feedback, an alternative is to look for explicit actions (or lack thereof) that signal the success (or failure) in satisfying the information need of interest.** As EDP did not collect explicit user feedback, we assume that **sessions that include download or go-to-source activities were successful.** In Service 3 WP4 we have proposed proxies for open data re-use which could be considered for follow-up studies of alternative portal architectures that put a stronger emphasis on community building around datasets.



We acknowledge that **our assumption may overestimate the number of successes**. In the same time, without more detailed information or means to track re-use, finding alternative metrics will remain challenging. For example, a user may download an EDP dataset but realise, after manually inspecting it on their local computer, that it is not what they are looking for. Moreover, we consider here only atomic information needs expressed through a sequence of queries or facet selections. Information retrieval approaches still have challenges supporting users with complex information needs – for instance, a user may be looking to multiple datasets to use in combination. Finding one, but not the other, or finding both, but not being able to integrate them, ultimately impacts on user's perception on what constitutes a successful dataset search.

Concerning unsuccessful sessions (failures), we consider two scenarios:

1. "only SERP": The user only looks at Search Result Pages and does not click on any of the result dataset pages.
2. "Dataset Page View" (DPV): The user clicks on at least one dataset page shown to them SERPs but does not download anything.

Figure 22 shows for internal search sessions the percentage of successful, "Only SERP" and "DPV" sessions per version. According to our classification, 37% of relevant sessions were successful on EDPv1. That share dropped to 22% on EDPv2 and rebounded on EDPv3 to 40%. For EDPv1 and EDPv3, the proportion of "only SERP" and "DPV" failures is approximately similar, with a spike on "Only SERP" failures for v2. **To verify if the spike is due to a temporal effect, we computed the distribution of "Only SERP" failures per month for versions v1 and v2, but we did not find any significant change.** This suggests that the **dataset search engine introduced in EDPv2 is less effective than in EDPv1, but the changes introduced in EDPv3 improved the previous situation.**

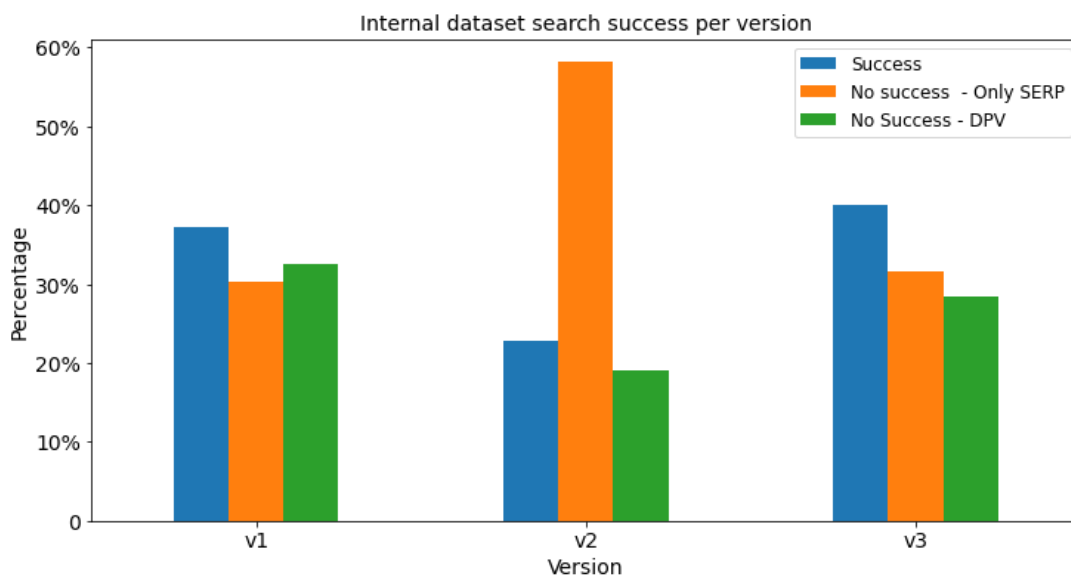


Figure 22: Internal dataset search success per EDP version

Figure 23 shows the same data for external search sessions. We note a similar trend as observed for internal search, but with lower rates: **EDPv1 and EDPv3 have around 25% of successful searches, but there is a drop in EDPv2 to less than 10%, combined with a spike on the number of "Only SERP" failures.**

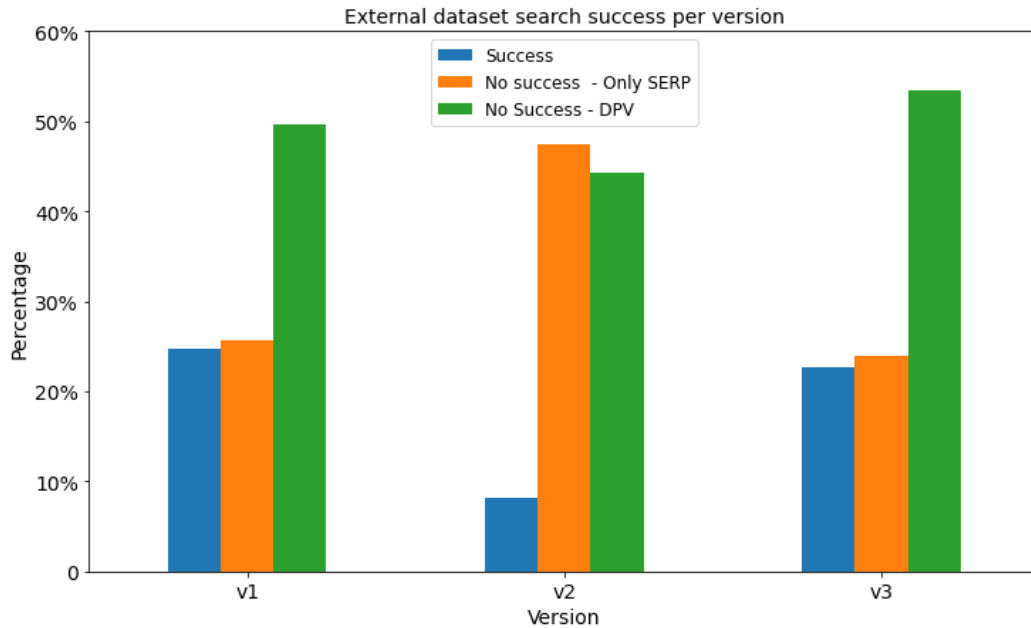


Figure 23: External dataset search success per EDP version

As we did with internal queries, we went on to **analyse the distribution of "Only SERP" failures per month for all versions to discover any temporary effects**. This time **we did find a larger number of failures on the first three months of v2**. We analysed the titles and URLs of the landing pages of these sessions and found a high rate of **"404 missing page"** titles, suggesting that hits were linked to the wrong resources. Very likely this is a direct consequence of the change in the dataset section URL scheme introduced in EDPv2: many pages referred by web search engines that led to a failed search were facets. In particular, **78% were 'tags' facets** (e.g. <http://www.europeandataportal.eu/data/en/dataset?tags=lidar>). The tags facet was replaced by the keywords facet on EDPv2, but no redirection rules were made. By contrast, redirections to dataset pages and main categories were set up; this meant that the number of DPV failed search remained constant from EDPv1 to EDPv2. **After July 2019, the number of "Only SERP" failures decreased to negligible levels**; we believe this was the time it took web search engines to completely remove the old pages from their indexes. This explains why the success rate in EDPv3 is similar to EDPv1.

Figure 24 compares the success rate between internal and external searches. We observe that for EDPv1 and v3, internal searches are around 50% more successful than external ones. In EDPv2 the difference is more than 100%, however, this is due to the 404 pages issue already described.

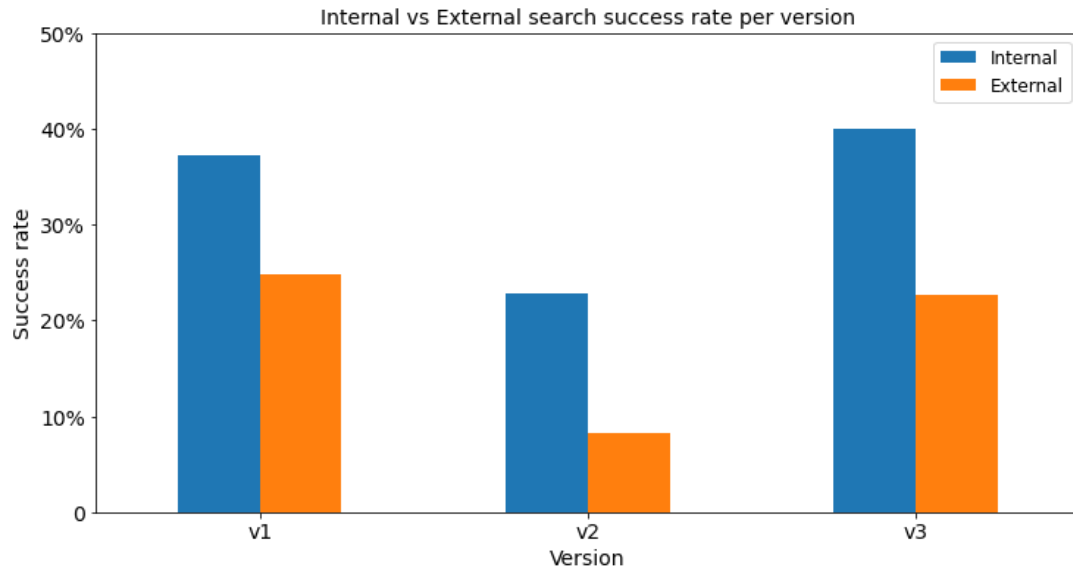
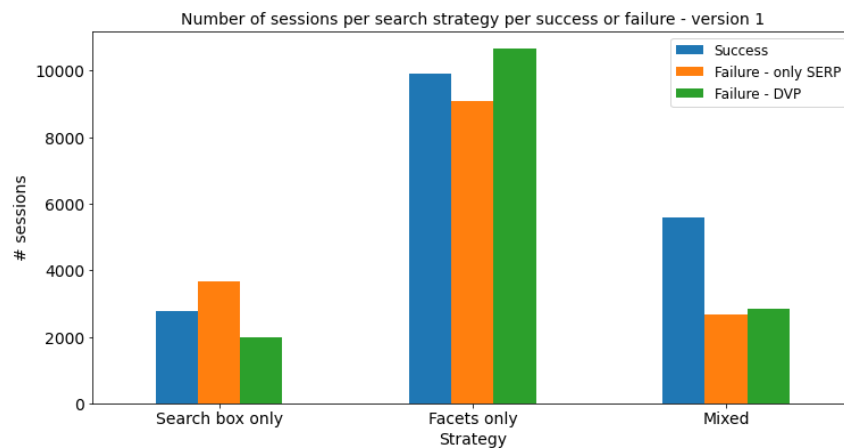


Figure 24: Success rate comparison between internal and external search, per EDP version

Figure 25 shows a breakdown of successful and failed searches for each of the search strategies introduced in Section 3.3.2; as a reminder, these strategies were concerned with the use of queries issued via search boxes and/or facets. Each bar chart corresponds to one version of the EDP: v1 top, v2 middle, v3 bottom. We notice that for EDPv1 and EDPv3 a mixed strategy led to more successful searches than using just queries or facets alone. In EDPv2, the mixed strategy works better as well, but this is because of the very large number of only SERP failures for the other two strategies.



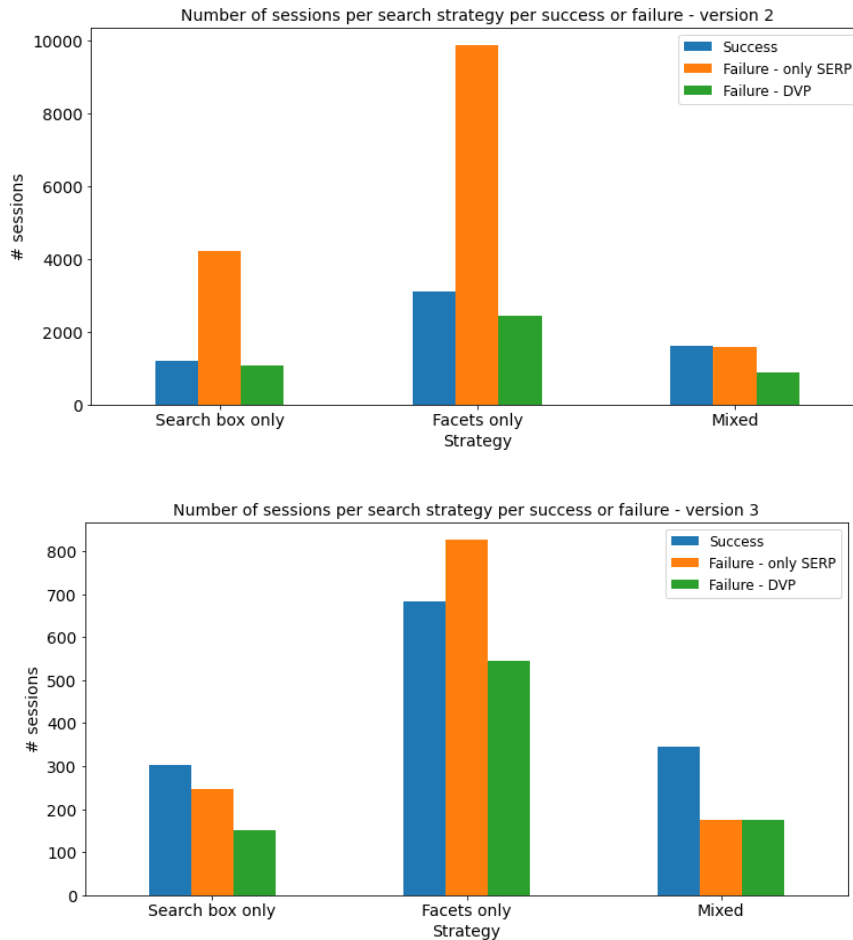


Figure 25: Success rate comparison between internal and external search, per EDP version

### 3.3.4.1 Summary of findings

Are users successful when they search for datasets on the EDP? Does success change across portal versions? Is there any difference in success between internal search (EDP's dataset search box) and external queries (from web search engines)?

- The percentage of successful sessions varies between 20 and 40% for internal search, and between 8 and 25% for external search. Success was indeed different across portal versions, with performance dropping in EDPv2. Success rate between v1 and v3 is approximately similar.

Is there one search strategy (search box only, facets only, mixed) more successful than others?

- A mixed search (search box + facets) appears to be more effective than searching using only the search box or only the facets.

## 4 Discussion

### 4.1 Search experience on EDP

Our analysis shows the importance of filters when using the EDP as a tool to find data. 60% of dataset search users rely exclusively on facets, without using keywords and between 15 and 20% mix keywords and facets while searching. This is in line with recommendations in prior work of ours in which the following facets are suggested: **location, provenance, format, licence, time frame and date, publishing date, location of publication, and data schema.**

Location-based queries and facets were particularly prominent when inspecting query types, even more so than reported in previous work. One reason may be that the portals indexed by the EDP have a broad geographic spread and include data resources from different administrative levels. Our current analysis supports specifically **the importance of location-based filtering and the results suggest the recommendation of even more fine-grained filters of location.** That would allow users to search for local datasets at different levels of granularity, for instance not just for countries but also for counties, cities, boroughs, etc. The importance of geospatial search in dataset search has been pointed out in prior work as discussed in Section 2. We know that the wrong granularity in terms of both location and time can easily result in the data not being usable for a task (e.g., Koesten et al., 2017). The user would perhaps download the datasets, but end up not using it as the data is not aggregated at the right level and changing that, if possible, is costly, especially without appropriate technical skills and tools.

Less prominent in contrast to prior work was the prevalence of temporal information in the queries. This is partially because no filtering based on timeframes is possible on the EDP and the majority of portal-based search is done via facets. **However, it would be interesting to explore the usefulness of time-based facets in respect to user needs in future qualitative work, as related literature suggests it is a core dimension people considered when looking for or selecting data to use.**

Our findings also show that **a large portion of filtering was done using categories of datasets.** As this applies mostly to EDPv2, we attribute the popularity of this facet to changes in User Interface (UI) design. By showing a category panel on the landing page of the EDP dataset section, users are primed to explore the collection via this facet. It is possible to switch to a “search datasets by term” setting; however, it is not prioritised in the UI (Figure 26).

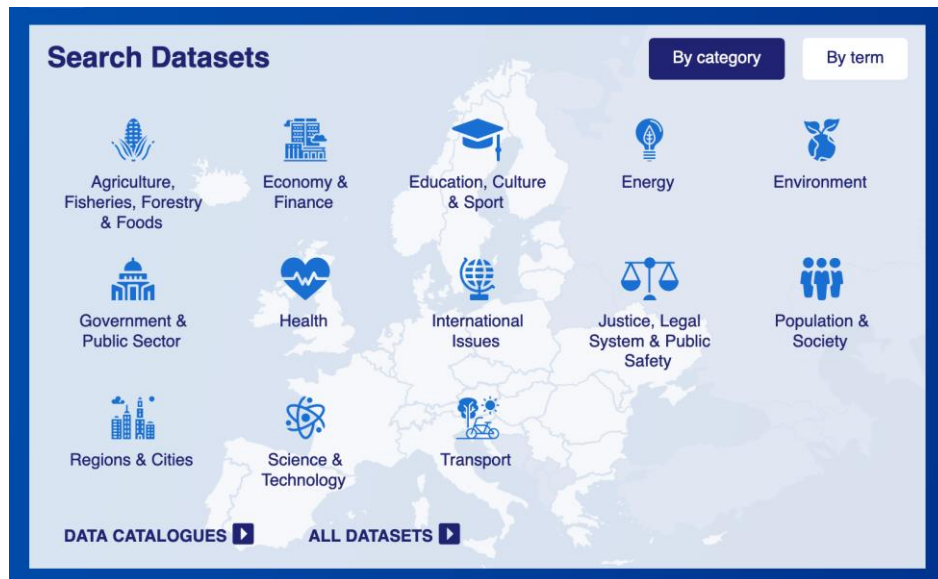


Figure 26: Search interface in EDP v3, category facet highlighted

This is a design choice that directly affects user interaction in dataset search on the portal and would therefore be **an interesting starting point for user research, both in terms of validating the general search strategy through categories but also regarding potential subcategories to make filtering actions more precise**. The results suggest that **the success rate of a search seems to be higher if both keyword search and filtering strategies are applied, which has implications for how to support this interaction through the UI**.

## 4.2 Search through web search engines

**Our analysis confirmed previous findings on web search engines being the main tool for dataset search, next to human recommendations**. The results show that the majority of users arrive at the dataset section through web search engines (more than 60% in EDPv1). They also document how important the link between the EDP and the search engine is, which led to a drastic decrease in traffic after EDPv2. A change in the URL scheme needs to be considered and timed carefully to mitigate effects.

Google's general Web search engine refers to 94% of the external visits to the dataset section, while Google dataset search only accounts for 4% per cent of them. Improving the SEO of EDP results in general web search would improve the performance of external searches.

The other factor to consider when users land on the dataset pages directly is the **importance of the dataset preview page and the contextual information shown there**. In Section 2 we reviewed existing literature on selection factors and criteria which make it more likely for users to re-use a dataset; **EDP implements only a subset of them when displaying datasets**.

**We found that users accessing the portal via web search engines tended to be less successful in their search activity**. This may be because those who actively use the internal search functionality are a more informed user group who are more familiar with the EDP context and have matching expectations rather than landing on the site by chance. At the very least, this hints at different user groups, potentially with different intents and information needs. **Given the importance of referred searches for the EDP, we believe future versions of the EDP UX must consider this demographic directly**.

**Queries issued through web search engines tend to be longer and contain more temporal and geo-spatial keywords.** This is a trend observed more broadly, as search engines improve their capabilities to serve very specific queries and answer questions. **In the same time, it also offers valuable information on the needs of that user group, which should be supported more explicitly by the EDP.**

### 4.3 Curated datasets with context

An interesting change of pattern could be noted in EDPv3, which contains COVID-19 datasets. We observed a **decreased use of the dataset section with increased use of the COVID section.** 46% of external queries include terms such as COVID or Corona during that time frame. The EDP provided specific COVID-19 related pages, including a collection of key datasets and editorial pieces on the pandemic. The popularity of this content can be explained by the urgency of the crisis, alongside design decisions by the EDP team: the content was prominently advertised during the first months of the pandemic, included high-quality material, and was linked to other sites. Putting aside the unique character of this topic, **we believe this approach shows the role of additional curated material related to datasets.** This could mean that one of the future roles of the EDP as a meta-portal is not merely providing good search results to any data search query, but providing access to resources that add value to datasets e.g. training, tools to use the datasets, stories etc. The success of this section of the EDP sites means that the COVID-19 datasets are likely to become more popular and enable easier re-use by people, which in turn allows improving SEO for specific topics rather than for the whole dataset corpus.

### 4.4 The role of SEO

**Our analysis suggests that data portals should take care of Search Engine Optimisation (SEO), like everyone else on the web.** What is less clear is how this should be done effectively for a data portal, which is different than other online sites. For commercial websites, the goal is to be above your competitors in the search engine results list for most meaningful queries. For data portals, the notion of competitors is much fuzzier. One way to look at it is to aim to position a dataset indexed by the EDP higher than the same resource published on the provider's site. This is problematic for various reasons. Consider the query "trees geolocation dataset Cáceres, Spain" entered onto a web search engine. The dataset that answers this query is published by the [Municipality of Cáceres](#), indexed by [the Spanish Open Data Portal](#), and [harvested by the EDP](#). A web search engine needs to decide the ranking of the results; it tends to prioritise original sources (or what they can identify as the original source). As the portals downstream increase their web presence and compliance with standards like DCAT-AP, and search engines become more effective in identifying datasets and original sources, it is very likely that for queries such as the one above, the provider's portal will receive more traffic than intermediaries such as the EDP. In more general terms, any dataset query that includes some sort of geospatial information (a city, a region, a country etc) is more likely to be redirected to the publishing portal rather than the EDP.

This raises questions about the relationship between EDP, the source open government data portals and the degree to which they need to compete for more traffic from web search engines despite following very similar aims – to make that data easier to find and use by everyone. While SEO is important, it is perhaps too narrow of a view to define success. In the same time, EDP adds value to downstream data portals and could explore ways to quantify (or even monetise) that value in the long term.

## 5 Summary and conclusions

Our analysis was long overdue. It points to findings which more often than not can be directly implemented on the EDP to improve search performance and experience, and prompts to further user research to answer follow-up questions around user groups, the use of specific facets etc.

The EDP is a complex resource, with resources for different purposes and with different types of content. This is confirmed by the log analysis, which confirmed that the various sections of the portals are mostly independent of each other in terms of visitors. Combined with the effects observed for COVID datasets, **this hints at an opportunity to improve the experience by cross-linking between data and non-data content systematically**. For instance, training resources could use EDP datasets. At the same time, the logs show steady demand for open data training, reports, news and information –the number of visits for these types of content increased over time, **confirming the role of the EDP as an open government data hub, which we consider at least as important as its data harvesting one**.

Moreover, we found evidence that a **number of users recognise the EDP as a tool for searching datasets**. Better understanding who these users are, and their specific goals, would be tremendously useful. This could include **qualitative approaches using in-depth interviews or a targeted survey with the segment of active dataset search users on the EDP** to get a deeper understanding of the types of tasks they are involved in that make them engage in data discovery.

In terms of further improving the search experience in the portal, we found that **a majority of users search for datasets using only the facet filters, in particular, country and category**. This leaves room for improvement of the **geospatial facet** by expanding it to enable more granular location queries. We also flagged the need for further studies to understand the use of the category facet. Such studies could include **applying qualitative methods from user experience research to this target group**. This could be done via user interviews to obtain a different perspective and understand the users' mental models during data discovery, but also more targeted approaches to identify meaningful subcategories to improve the EDP facets through card sorting or contextual inquiry.

Dataset search literature suggests that better dataset descriptions would aid data discovery and re-use. This can be done by tailoring them to user needs. This is discussed in prior work of ours in Koesten et al., (2020) with suggestions to **support publishers in summary creation via guidance and templates and potentially invest in semi-automatic approaches for the creation of dataset summaries that are meaningful to the user**.

Given the multilingual nature of the EDP translation services (as suggested in a [prior report](#)) are another direction to improve the accessibility of dataset summaries for a wider audience. While our analysis could be expanded to a larger query sample, there is evidence that suggests an English-speaking audience alongside some other languages such as German or Polish.

One way to add value during dataset discovery would be to aid the user through **recommendations**. Links between datasets, based on the similarity of their content (e.g. rows or columns) or to the related context of editorial nature, additional documentation or reference points (e.g. standard vocabularies) are likely to enable re-use.



**Monitoring search performance in the dataset search is still early days.** Besides adding feedback features widely used in other sectors, we need research to understand how we could define success beyond the relatively crude measures we could apply on search and interaction logs. These nevertheless show that the EDP needs to do more to understand and support external searches.

**The importance of web search engines to the EDP cannot be underestimated.** Our analysis shows one way to monitor and reflect upon it with data, though we believe we need a much broader conversation around the value chains between web search engines, content providers such as the national portals, and the EDP which adds value while at face value having to compete for traffic with them as well to succeed.

## 6 References

- Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E. and Groth, P., 2020. Dataset search: a survey. *The VLDB Journal*, 29(1)
- Carevic, Z., Roy, D. and Mayr, P., 2020. Characteristics of Dataset Retrieval Sessions: Experiences from a Real-life Digital Library. *arXiv preprint arXiv:2006.02770*.
- 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*, 55, pp.37
- Kacprzak, E., Koesten, L., Tennison, J. and Simperl, E., 2018, April. Characterising dataset search queries. In *Companion Proceedings of the Web Conference 2018* (pp. 1485-1488).
- Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S., 2020. Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*.
- Grubenmann, T., Bernstein, A., Moor, D., Seuken, S.: Financing the web of data with delayed-answer auctions. In: *Proceedings of the 2018 World Wide Web Conference, WWW '18*, pp. 1033–1042. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2018).
- Jiang, L., Rahman, P., Nandi, A.: Evaluating interactive data systems: workloads, metrics, and guidelines. In: *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, pp. 1637–1644. ACM, New York, NY, USA (2018).
- Koesten, L.M., Kacprzak, E., Tennison, J.F. and Simperl, E., 2017, May. 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*
- Koesten, L., Kacprzak, E., Tennison, J. and Simperl, E., 2019, May. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).
- Koesten, L., Simperl, E., Blount, T., Kacprzak, E. and Tennison, J., 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies*, 135, p.102367
- Neumaier, S. and Polleres, A., 2019. Enabling spatio-temporal search in open data. *Journal of Web Semantics*, 55, pp.21-36.
- Nunes, S., Ribeiro, C., David, G., 2008. Use of temporal expressions in web search. In: *European Conference on Information Retrieval*. Springer, pp. 580–584.
- Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment and evolution of open data portals. In: *2015 3rd International Conference on Future Internet of Things and Cloud*, pp. 404–411 (2015).
- Noy, N., Burgess, M., Brickley, D.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: *28th Web Conference (WebConf 2019)* (2019)
- Reynolds, P.: DHS Data Framework DHS/ALL/PIA-046(a). Technical Report, US Department of Homeland Security (2014)
- Sansone, S.A., González-Beltrán, A., Rocca-Serra, P., Alter, G., Grethe, J., Xu, H., Fore, I., Lyle, J., E. Gururaj, A., Chen, X., Kim, H., Zong, N., Li, Y., Liu, R., Burak Ozyurt, I., Ohno-Machado, L.: Dats, the data tag suite to enable discoverability of datasets. *Sci. Data* 4 (2017).