Analytical Report n19

Analytical Report 19

UNDERSTANDING SUPPLY AND DEMAND IN DATASET SEARCH ON THE EUROPEAN DATA PORTAL

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, 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

European Data Portal

Gianfranco Cecconi, European Data Portal Lead
Email: gianfranco.cecconi@capgemini.com

Written by:

Luis Daniel Ibáñez
Email: l.d.ibanez@soton.ac.uk

Elena Simperl
Email: e.simperl@soton.ac.uk

Reviewed by:

Eline N. Lincklaen Arriëns
Email: Eline.lincklaen.arriens@capgemini.com

Last update: 18.02.2021
www: https://europeandataportal.eu/
email: 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, 2021

OA-BF-21-001-EN-N ISBN: 978-92-78-42413-8 ISSN:2600-0601 doi: 10.2830/105287

Table of Contents

Executive Summary

Following our initial study of dataset search through the lens of the digital traces collected via web analytics tools, we dug deeper into the data to understand the relationship between supply and demand of datasets on the European Data Portal (EDP). This gives us more detailed answers into common user journeys in dataset discovery and into the information needs that EDP should aim to meet. In the second part of the study, we specifically look at supply and demand for COVID-19 data, and the impact auxiliary content, such as data stories, topical articles, and promotion campaigns have on search behaviour.

Our findings lead to three core recommendations for dataset publishers and open data platforms software:

1. Prepare curated content about datasets of interest and highlight them on a dedicated section of your site. For example, a monthly article that uses several datasets to tell a data story about a current topic. Alternatively, data from external searches could be analysed to understand topics that may be of interest to portal users and drive content production.
2. Design for user journeys and information needs revealed in such analyses. In AR-18 we concluded that user journeys that start on the portal are likely to have very different information needs to the ones that land on the portal from an external site e.g. a search engine. The examples of journeys we provide here are complementary to this basic dichotomy; we believe they could also be considered when studying dataset reuse and impact, as publishers might prioritise some types of users and use cases over others.
3. Perform analyses like these or the ones from AR-18 regularly to confirm and challenge our recommendations and discover categories that are and remain in demand.

.

Introduction

In “Characterising dataset search on the European Data Portal” (Analytical Report 18/AR-18) we took a detailed look at how EDP users search for datasets. We considered four themes, including **success in dataset search**. In the absence of explicit feedback, we used proxies, for example downloads and go-to-source activities logged by the portal to identify search sessions that have likely led to a user finding the data they needed.

In this report, we expand on this theme to understand what datasets people need (and perhaps cannot find), and how this demand evolves over time. To do so, we aim to establish if there are datasets harvested by the portal that:

- **are consistently demanded over time**; or
- **whose demand is periodical** (i.e. demand goes up and down in cycles e.g. elections, seasonal activities such as tourism etc.); or
- **episodical** (i.e. demand is tied to specific events e.g. natural disasters, COVID-19).

To understand differences in **datase**t **supply and demand**, we analyse the interaction logs of the EDP from the beginning of April 2018 to end of October 2020. We answer the following questions:

Q1.1: What categories of datasets are in high demand?

- **Method:** Count the occurrences of each category in sessions that click on the categories facet. From keyword queries (both issued directly toEDP via the search box or from Google), identify keywords that match to categories.

Q1.2: What datasets are most demanded by users?

- **Method:** Count the number of downloads of datasets and identify the most downloaded. From keyword queries (both issued directly to EDP or from Google), identify keywords that match to specific datasets.

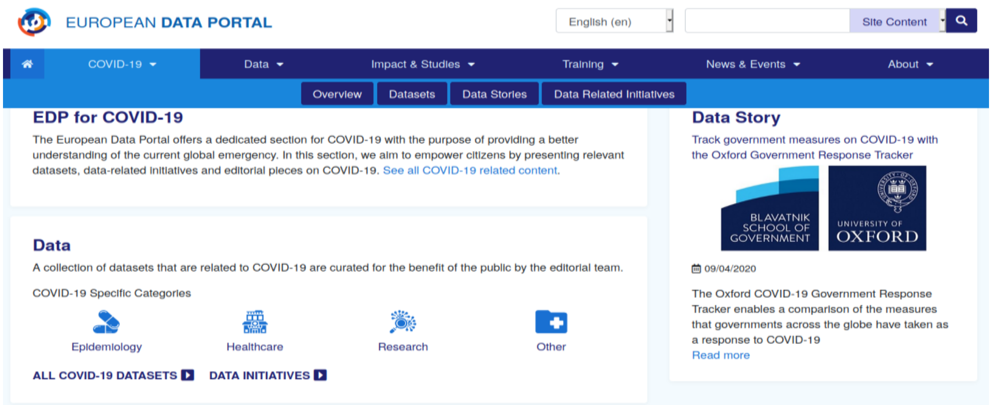
Q1.3: What datasets are used together?

- **Method:** From the sessions with more than one download, compute the top-5 datasets that were most downloaded in combination with other datasets. For these top-5 datasets, we take a closer look at the list of datasets downloaded together with them, and the keywords or facets used in those sessions.

Q1.4 How do these dimensions vary over time? Are there any periodical or episodical effects in demand over time? Can we link the latter to specific events, be that EDP-related (e.g. introduction of a new feature) or external (e.g. a political event that happened at that time)?

- **Method:** Time-series analysis of the data prepared for Q1.1 to 1.3.

On April 6, 2020, **the EDP introduced a dedicated COVID-19 section**, featuring a list of datasets, data initiatives, and data stories curated by the EDP COVID-19 team. This section was featured on the front page and regularly updated until the end of July, when the stories were transferred to a "COVID-19" subsection under "Impact & Studies". In Analytical Report 18, we found that **the COVID-19 section became more popular than the dataset search section**. We discussed that the popularity of this content could have been 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 the event, we stated that **this approach hinted at the role of additional curated material on datasets use**. In this report, we perform a more in-depth analysis of the COVID-19 content performed, in terms of its relative importance compared to the rest of the portal, and what happened after it was archived.



EDP for COVID-19

The European Data Portal offers a dedicated section for COVID-19 with the purpose of providing a better understanding of the current global emergency. In this section, we aim to empower citizens by presenting relevant datasets, data-related initiatives and editorial pieces on COVID-19. See all COVID-19 related content.

Data

A collection of datasets that are related to COVID-19 are curated for the benefit of the public by the editorial team.

COVID-19 Specific Categories

Data Story

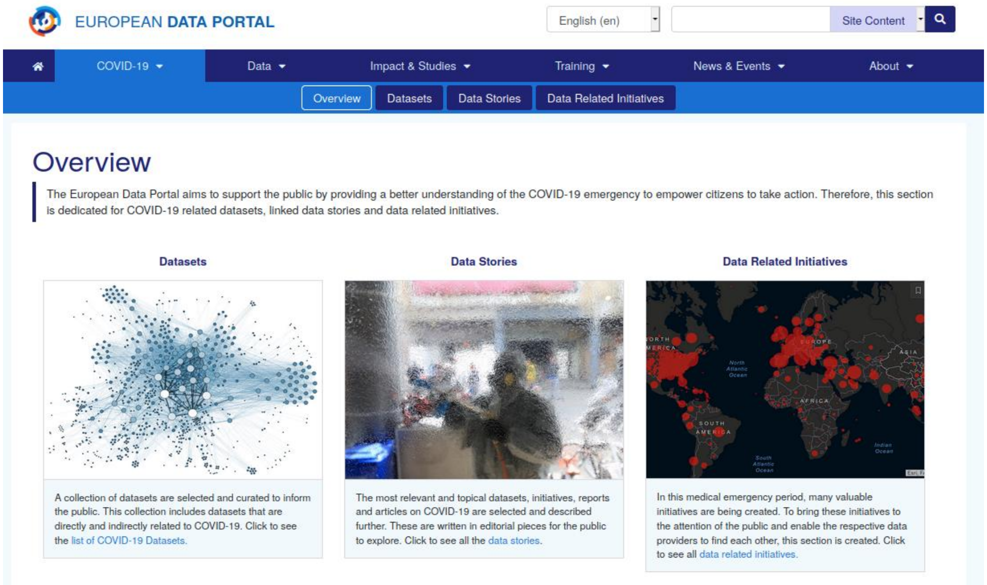
Track government measures on COVID-19 with the Oxford Government Response Tracker

09/04/2020

The Oxford COVID-19 Government Response Tracker enables a comparison of the measures that governments across the globe have taken as a response to COVID-19

Read more

Figure 1: Screenshot of EDP homepage featuring the COVID-19 menu item (April – July 2020)



Overview

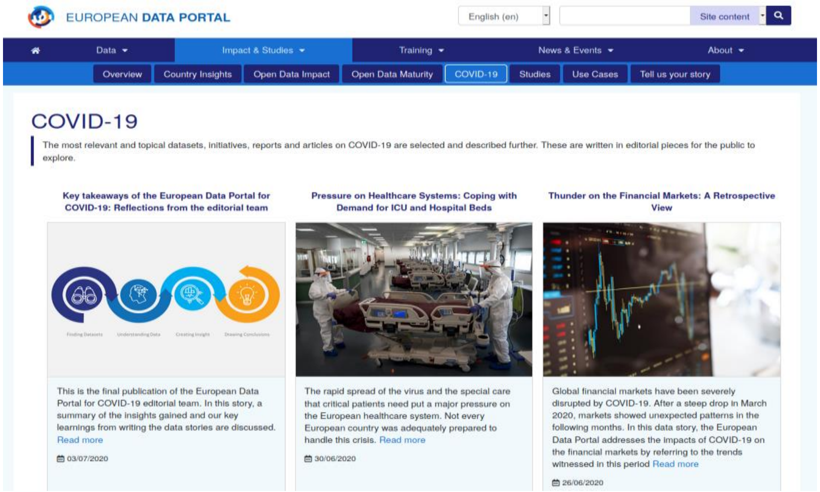
The European Data Portal aims to support the public by providing a better understanding of the COVID-19 emergency to empower citizens to take action. Therefore, this section is dedicated for COVID-19 related datasets, linked data stories and data related initiatives.

A collection of datasets are selected and curated to inform the public. This collection includes datasets that are directly and indirectly related to COVID-19. Click to see the list of COVID-19 Datasets.

The most relevant and topical datasets, initiatives, reports and articles on COVID-19 are selected and described further. These are written in editorial pieces for the public to explore. Click to see all the data stories.

In this medical emergency period, many valuable initiatives are being created. To bring these initiatives to the attention of the public and enable the respective data providers to find each other, this section is created. Click to see all data related initiatives.

Figure 2: "Overview" (initial page) of the COVID-19 section as it was live



COVID-19

The most relevant and topical datasets, initiatives, reports and articles on COVID-19 are selected and described further. These are written in editorial pieces for the public to explore.

Key takeaways of the European Data Portal for COVID-19: Reflections from the editorial team

This is the final publication of the European Data Portal for COVID-19 editorial team. In this story, a summary of the insights gained and our key learnings from writing the data stories are discussed. Read more

03/07/2020

Pressure on Healthcare Systems: Coping with Demand for ICU and Hospital Beds

The rapid spread of the virus and the special care that critical patients need put a major pressure on the European healthcare system. Not every European country was adequately prepared to handle this crisis. Read more

30/06/2020

Thunder on the Financial Markets: A Retrospective View

Global financial markets have been severely disrupted by COVID-19. After a steep drop in March 2020, markets showed unexpected patterns in the following months. In this data story, the European Data Portal addresses the impacts of COVID-19 on the financial markets by referring to the trends witnessed in this period Read more

26/06/2020

Figure 3: COVID-19 section after archival under the Impact & Studies section

We aim to shed light on **the role of curated content (e.g. featured datasets or categories of datasets, data stories) in driving user activities.** To do so, we perform a more in-depth analysis of the period of time when the COVID-19 section was active and the immediately preceding and subsequent periods. Figure 1 shows the homepage of the EDP during the period of activity of the COVID-19 section; Figure 2 shows the overview of the COVID-19 section, and Figure 3 shows the COVID-19 section after its archival under the Impact & Studies section of the portal. This leads to the following questions:

Q2.1 Was the EDP used more during the life of the COVID-19 section than before?

- **Method**: Compare the number of visits to the different sections of the portal for three time windows: April-July 2020 (COVID-19 section life), December 2019-March 2020 (previous comparable period)., August-October 2019 (previous comparable summer period).

Q2.2 The COVID-19 section of the portal was archived by July 2020. Did traffic return to pre-April levels? If traffic remained the same, what parts of the portal accounted for it?

- **Method**: Compare the number of visits to the different sections of the portal in the April-July 2020 period against August-October 2020 (next period of comparable length).

Q2.3 How did users reach the COVID-19 section? Did they look specifically for it, or found it thanks to its prominent position on the homepage during its lifetime?

- **Method**: Compare references to sessions that include a visit to the COVID-19 section from April to October 2020. Direct entries from search engines would suggest users were looking for COVID-related information and landed on the EDP. Reaching the COVID-19 section after visiting EDP's homepage would suggest that users came to the portal organically, and then moved on to the COVID-19 section.

Q2.4 Did the publication of COVID-19 data stories increase traffic on the portal? Did it lead to variations in acquisition channels (e.g. more visits from news sites or social media)?

- **Method**: Analyse sessions dated immediately after the publication of COVID-19 data stories from April to July 2020, and for the August to October 2020 period (when stories were archived in the Impact & Studies section of the EDP). We aim at identifying variations in visits to the different sections of the portal.

Q2.5 How high was the demand for COVID-19 datasets on the portal? How does it compare to other datasets?

- **Method**: Similar analysis to 1.1 and 1.2 for the sessions that referred to datasets featured in the COVID-19 section.

Search and Interaction Logs Corpus

We use the same search and interaction logs used for Analytical Report 18, adding data from July, August, September and October 2020, with a different partitioning: instead of separating data according to the date of new releases of the EDP, we consider two subsets:

- the "old normal" from 1st April 2018 to 5th April 2020, the period before the launch of the COVID-19 section, shortly after the time the pandemics was declared in Europe; and
- the "new normal", from 6th April 2020 to 31st October 2020 for the period after the COVID-19 section was launched.

The "new normal" period is further divided in two:

- the lifetime of the section including curated content about the pandemics, from 6th April 2020 to 2nd August 2020, labelled "COVID-19 section"; and
- the period after the section was retired to "Impact and Studies" and stopped being actively updated, from 3rd August to 31st October 2020.

We also selected two periods from "old normal" of comparable size with "COVID-19 section":

- "previous immediate" from 2nd December 2019 to 5th April 2020; and

- "previous" from 12th August to 15th December 2019. The latter was chosen to include logs that did not include Christmas and New Year, where visits are low.

We summarise below the characteristics of the search and interaction logs, which were discussed in greater detail in AR-18: The European Data Portal uses the Matomo Web Analytics tool to log the actions of users of the portal for each of their visits. The EDP team uses these logs to create aggregated, anonymised analyses of user behaviour, with the objective of improving the site, ensuring its correct functioning, informing further design decisions, customising the information or e-services of interest, and detecting and addressing any abuses or security issues.

A description of all data Matomo can log is available on this link. We detail the subset of fields that are strictly necessary for our analysis below:

- **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 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. For type = "search dataset", blank.
    - **"siteSearchKeyword"**: If type = "Search Dataset" and user consented when starting the visit, contains the keywords typed on the dataset search box. If the action is not "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 the website from which the user clicked a link to the EDP.
- **'referrerUrl'**, URL of referrer website or Social Network
- **'referrerTypeName':** Type of referrer. 'Search Engine', 'Website' or 'Social Network'. When a referrer cannot be identified, a 'Direct Entry' (User typed the landing URL directly on the bar
- **'referrerSearchEngineUrl'**: If referrerTypeName = Search Engine, its URL
- **'referrerKeyword':** If referrerTypeName = 'Search Engine', and 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 them available due to privacy concerns.

We report on three values 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, it will be recorded as a new session. Since its inception, EDP had this value set as 30 minutes. An accurate estimation of the optimal value of this parameter for dataset portals is out of scope of this report.
2. **Exclusion of bots**: Bots are automated agents that crawl websites. EDP's Web server proxy configuration provides a first line of defence against malicious bots. Matomo itself, with its default configuration, is able to filter out most bots that get to the portal. Periodic analysis of this dataset for internal EDP

reporting found no indication of skewing due to bot traffic. We exclude these sessions, as our focus in this report is on understanding people's search behaviour.

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 done following these instructions.

In our analysis we split sessions as shown in Table 1:

TABLE 1: DATA CORPORA USED IN THIS ANALYSIS

Codename	Description	Date range (times as GMT)	Number of sessions
"old normal"	Previous to introduction of COVID-19	02 April 2018 00:00 to 05 April 2020 23:59	742382
"new normal"	After introduction of COVID-19 section, COVID-19 outbreak already declared in most Europe	06 April 2020 00:00 to 31 October 2020 23:59	262035
"COVID-19 section"	EDP COVID-19 section lifetime	06 April 2020 00:00 to 02 August 2020 23:59	140635
"previous immediate COVID-19"	Comparable size period before the launching of COVID-19 section	02 December 2019 00:00 to 05 April 2020 23:59	119353
"previous COVID-19"	Comparable length period to COVID-19 section avoiding Christmas downpeak	12 August 2019 00:00 to 15 December 2019[1] 23:59	103651
"after COVID-19 section"	Comparable size period after the retirement of COVID-19 section	03 August 2020 00:00 to 31 October 2020 23:59	118220

Search keywords data come from two sources:

1. **Matomo**, which registers the queries input on the dataset search box of the EDP, we selected the top 500 queries in the order of appearance. We refer to this dataset as *"internal search queries"* ; and
2. **Google Search Console (GSC)**, which provides a list of keywords input by users of the Google search engine for which an EDP page was shown as part of the list of results. GSC only makes available daily data for the last 3 months, limited to the 1000 queries with the highest 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 April 2018 using the SearchEnginePerformance Matomo plugin. For the "old normal" and "new normal" corpora, we selected the top 500 queries in the number of clicks. We refer to this dataset as **"external search queries"**.

Data Supply and Demand

Results

Q1.1 What categories of datasets are in high demand?

In AR-18, we reported that the most popular facet filter among users is the "category" one. In the EDP, dataset categories are aligned to the "Data-Theme" controlled vocabulary[2], in compliance with the DCAT-AP 1.1 and 2.0

[1] There is an overlap with the previous corpus, as December is often a rather quiet time because of the holidays.

[2] http://publications.europa.eu/resource/authority/data-theme

specifications. Table 2 describes the categories names, example datasets and aliases we use for the rest of this report.

TABLE 2: NAMES, ALIASES AND EXAMPLES OF DATASET THEMES/CATEGORIES DEFINED BY THE EU DEVELOPED DATA-THEME CONTROLLED VOCABULARY

Name	Alias	Dataset examples
Agriculture, fisheries, forestry and food	Agriculture	Agricultural and Vegetable Catalogue; The Community Fishing Fleet Register; Pan-European Map of Forest Biomass Increment; Food composition database for nutrient intake: selected vitamins and minerals in selected European countries.
Economy and Finance	Economy	Tenders Electronic Daily (TED) - public procurement notices from the EU and beyond; General government deficit (-) and surplus (+) - quarterly data.
Education, culture and sport	Education	European Skills, Competences, Qualifications and Occupations (ESCO); EU Member States and international human rights obligations; Participation in any cultural or sport activities in the last 12 months by sex, age and educational attainment level.
Energy	Energy	European gas market reports; Electricity prices by type of user.
Environment	Environment	Attitudes of European citizens towards the environment; Pollutant emissions from transport.
Government and public sector	Government	Candidate countries and potential candidates: Government statistics; Transparency Register.
Health	Health	COVID-19 Coronavirus data; European Cancer Information System
International Issues	International	Consolidated list of persons, groups and entities subject to EU financial sanctions; European Commission — DG DEVCO – development and humanitarian assistance to Afghanistan.
Justice, legal system and public safety	Justice	EU case-law; Information on Member States Law; European Data Protection Supervisor register of processing operations.
Regions and cities	Regions	NUTS - Nomenclature of territorial units for statistics classification; UDP - GDP per capita by metro regions, 2000 - 2060.
Population and Society	Society	Population density by NUTS 2 region; Violence against Women: An EU-wide survey
Science and Technology	Science	CORDIS - EU research projects under Horizon 2020 (2014-2020); Take-up of mobile broadband (subscriptions/100 people).
Transport	Transport	Total length of motorways; Airport traffic data by reporting airport and airlines.

We extracted from "old normal" and "new normal" corpora the sessions that included at least one click on any of the 13 category facets and for each session, extracted the set of categories explored (that is, categories are only counted once per session even if there are multiple clicks), and aggregated the number of sessions per category. Sessions that explored more than one category were counted once per each category.

Figure 4 compares the number of sessions per category for "old normal" (49776 sessions), and Figure 5 the number of sessions per category for "new normal" (2879 sessions).

For "old normal", we can classify categories in four groups based on number of sessions:

1. The first group is composed of the **transport** category, that has 4340 sessions. Over 20% more than the next most demanded category.
2. The second group is composed of the **economy, environment, regions and agriculture** categories, with between 3000 and 3500 visits, approximately 20% more than the next group.
3. The third group is composed of the **health, society, government, energy and education** categories, with between 2200 and 2800 visits, between 30 and 50% more than the members of the last group
4. The fourth and last group is composed of the **science, justice and international categories**, with between 900 and 1800 categories

In the external search keywords corpus, we only find a mention to the **health** category as well as some queries that may be considered as subcategories of health: "Depression", "Mental Health", "Suicide", "Healthcare", "Heart disease", and variations of COVID-19 (but in this dataset less frequent than all the others).

For keywords of internal queries, we found 74 mentions to categories (or possible sub-categories). This is consistent with the findings of Analytical Report 18, where we found that **users often use the search box as a way to filter by country**, in a similar way to the corresponding facet.

TABLE 3: CATEGORIES OR SUB-CATEGORIES IDENTIFIED IN THE INTERNAL SEARCH DURING THE "OLD NORMAL" PERIOD

Category	Number of searches	Category	Number of searches
Tourism	596	Transport	207
Traffic	579	Water	206
Crime	278	Football	203
Weather	256	Energy	183
Population	230	Health	134

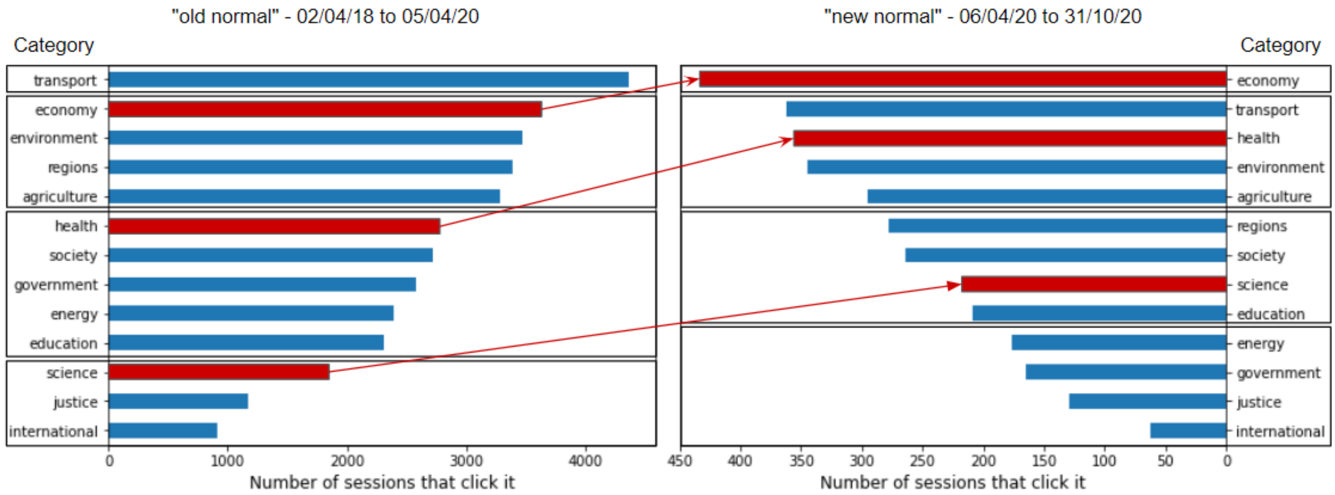


FIGURE 4: COMPARISON BETWEEN NUMBER OF SESSIONS PER CATEGORY IN "OLD NORMAL" (LEFT) AND "NEW NORMAL" (RIGHT), CATEGORIES WITH SIMILAR NUMBERS ARE ENCLOSED IN RECTANGLES. RED BARS AND ARROWS HIGHLIGHT CATEGORIES WITH MORE VISITS DURING NEW NORMAL

For "new normal", we can classify categories in a similar way than we did with "old normal", but the relative order of categories within the demand groups changes:

1. The first group is composed of the **economy** category, with 456 visits, 10-15% more than the members of the next group
2. The second group is composed of the **health, transport, environment and agriculture**, between 300 and 400 visits, 10-15% more than the members of the next group
3. The third group is composed of the **regions, society, science and education** categories**,** with between 200 and 300 visits
4. The fourth group is composed of the **energy, government and justice** categories, with between 100 and 200 visits.
5. Finally, the fifth group, composed of the **international** category, with 69 visits.

We highlight the **increased relative demand of the health, economy and science** categories and the **decrease of transport and energy**. We also note that both **justice and international appear to be less demanded than the others**.

When the search started on Google, we found that **223 out of the 500 search keywords (45%) included the words "covid" or "corona"** (in any case combination). Besides that, we found mentions of **health, transport and**

energy-related topics. For transport and energy, the combination of keywords were "open data transport/energy", "public data health/transport", or simply "transport data", "energy data". This dataset also shows that the EDP is highly ranked for generic "open data" and "european data" queries, when connected to a keyword that can be mapped to a section of the portal (the category), Google shows more impressions of the EDP. There are two hypotheses to explain the difference with the results of the categories section: (1) There are other websites that are higher ranked than the EDP for keywords like "Economy open data" and "economy data" (2) the interests of users that were searching on Google were different to those that landed directly in the portal.

When the queries were issued straight in the EDP search box, **only 52 out of the 500 keywords (10%) included the words "covid" or "corona"**. Unfortunately, the decrease in use of the datasets section (and therefore, the search box) during this period makes the overall number of searches very low (only 4400). The only identified categories with more than 10 searches were **"tourism", "traffic", "population" and "crime"**.

Q1.2 What datasets are most demanded by users?

To answer this question we counted the number of downloads per dataset for "old normal" and "new normal" without taking into account the datasets on the COVID-19 section, but taking into account when those datasets were moved to the regular dataset section after the retirement of the COVID-19 section in July.

For each period, we computed the distribution of dataset downloads and the top-10 most downloaded datasets. Figure 6 (left) shows the distribution of dataset downloads for "old normal" and Figure 6 (right) shows the distribution of dataset downloads for "new normal", both in logarithmic scale. **For "old normal", the majority of datasets have at most 5 downloads, with only a few tens having over 200 downloads.**

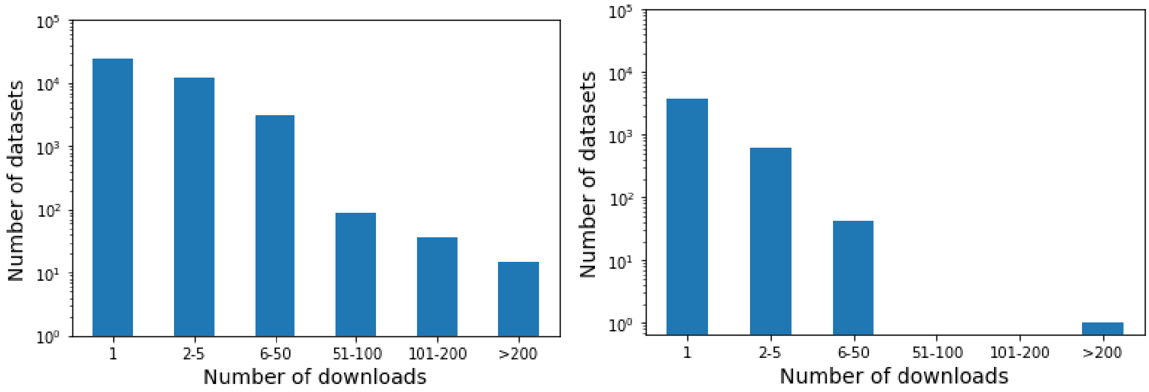
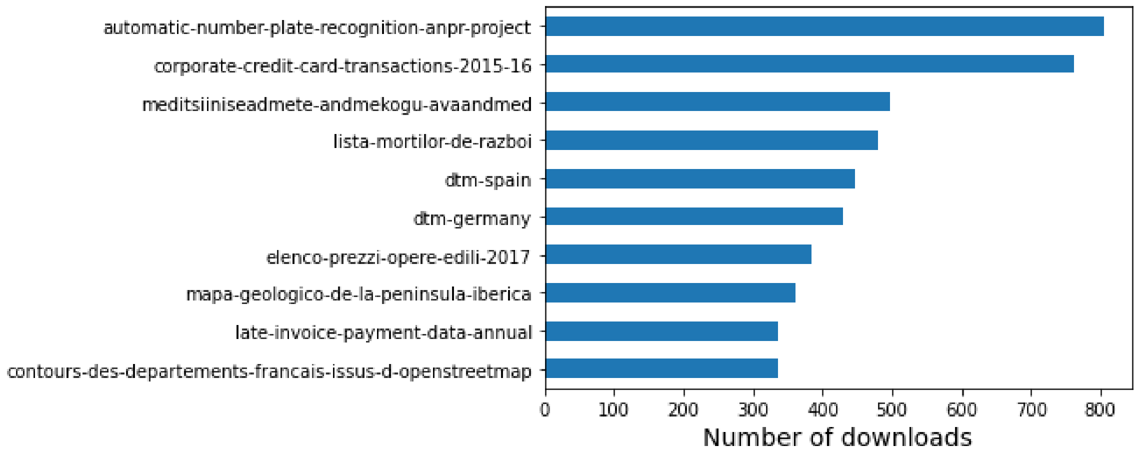


FIGURE 5: DISTRIBUTION OF NUMBER OF DOWNLOADS PER NUMBER OF DATASETS FOR OLD NORMAL (LEFT) AND NEW NORMAL(RIGHT). SCALE IS LOGARITHMIC



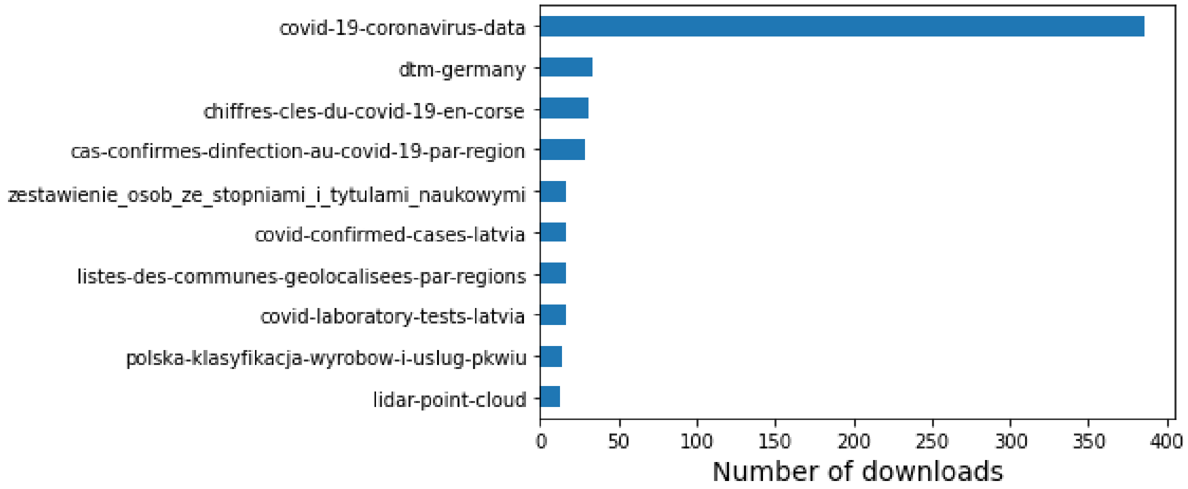


FIGURE 6: NUMBER OF DOWNLOADS OF THE TOP-10 MOST DOWNLOADED DATASETS FOR "OLD NORMAL" (TOP) AND "NEW NORMAL" (BOTTOM). DATASETS ARE DETAILED ON TABLE 4 AND 5 RESPECTIVELY

Tables 4 and 5 show the top-10 most downloaded datasets for "old normal" and "new normal" respectively. **For old normal, we highlight the presence of 4 datasets of geospatial nature (two digital terrain models, one geographical map and one geographical contours map)**. We also highlight that **3 of the datasets (the two digital terrain models and the geographical contours) were contributed by members of the public or non-public-sector organisation**s. For "new normal", **the COVID-19 dataset published by the European Centre for Disease Prevention and Control was the most demanded, 10 times more than the second most downloaded dataset**. This dataset was harvested from the EU open data portal and was also listed in the dataset subsection of the COVID-19 section. From the other 9 datasets, **5 are related to COVID-19, and only one from the "old normal" list appears again (dtm-germany)**. As with "old normal", **three datasets were contributed by members of the public or non-public-sector organisations**.

TABLE 4: TOP-10 MOST DOWNLOADED DATASETS DURING "OLD NORMAL"

# downloads	Dataset	Source and catalog	Description
798	automatic-number-plate-recognition-anpr-project	Source: Leeds City Council Catalog: data.gov.uk	A dataset providing information of the vehicle types and counts in several locations in Leeds. The aim of this work was to examine the profile of vehicle types in Leeds, in order to compare local emissions with national predictions.
761	corporate-credit-card-transactions-2015-16	Source: London Borough of Barnet Catalog: data.gov.uk	All transactions on all corporate credit cards of the London Borough of Barnet in the financial year 2015/16
497	meditsiiniseadmete-andmekogu-avaandmed	Source: Health Board Estonia[3] Catalog: opendata.riik.ee	The open data of the medical device database are available in digital and machine-readable form pursuant to § 28 (1) 30) and § 29 (4) of the Public Information Act. The MSA kit includes the

[3] Not available on metadata harvested by EDP, checked manually at publisher's web page

			following datasets: Clinical Trials, Incidents, Medical Devices (Including Manufacturers, Distributors, and Professional Users)[4]
480	lista-mortilor-de-razboi	Source: Romanian Ministry of National Defence Catalog: data.gov.ro	Statistics released by the National Office for the Cult of Heroes regarding the Romanian and foreign soldiers killed in the War of Independence (1877-1878), the First World War (1914-1918) and the Second World War (1939-1945).
444	dtm-spain	Source: Member of the public Catalog: opendataportal.at	Digital Terrain Models of Spain, Andorra and Gibraltar
396	dtm-germany	Source: Member of the public Catalog: opendataportal.at	Digital Terrain Models of Germany
383	elenco-prezzi-opere-edili-2017	Source: Province of Bolzano - Alto Adige Catalog: dati.gov.it	List of reference prices for public building works commissioned by the province of Bolzano - Alto Adige
362	mapa-geologico-de-la-peninsula-iberica-baleares-y-can-1995	Source: Spanish Institute of Geology and Mines Catalog: datos.gob.es	Geological map of the Iberian peninsula, Balearic islands and Canary islands. 1995 edition (first one to be available in machine readable format)[5]
336	late-invoice-payment-data-annual	Source: North Yorkshire City Council Catalog: data.gov.uk	Number of invoices paid by the council and proportion paid promptly and paid late
334	contours-des-departements-francais-issus-d-openstreetmap	Source: OpenStreetMap France Catalog: data.gouv.fr	Geographic contours of French departments from OpenStreetMap

TABLE 5: TOP-10 MOST DOWNLOADED DATASETS DURING "NEW NORMAL"

# downloads	Dataset	Publisher and catalog	Description
407	covid-19-coronavirus-data	Source: European Centre for Disease Prevention and Control Catalog: EU open data portal	Latest available public data on COVID-19 including a daily situation update, the epidemiological curve and the global geographical distribution (EU/EEA and the UK, worldwide).
34	dtm-germany	Source: Member of the public Catalog: opendataportal.at	Digital Terrain Models of Germany
33	cas-confirmes-dinfection-au-covid-	Source: Member of the public	Aggregation of number of daily confirmed COVID-19 cases by

[4] Automatic translation from original description in Estonian

[5] Translated from original description in Spanish by one of the authors.

	19-par-region	Catalog: data.gouv.fr	region in France as per published daily by Public Health France[6].
31	chiffres-cles-covid-10-corse	Source: Region of Corsica Catalog: data.gouv.fr	Extract of data published by OpenCOVID19-fr corresponding to the region of Corsica[7]
19	listes-des-communes-geolocalisees-par-regions-departements-circonscriptions-nd	Source: NosDonnées.fr Catalog: data.gouv.fr	Compilation of useful administrative codes (INSEE, postal) and geographical information (capital, latitude, longitude) for each of the 36000 french municipalities (communes)[8]
17	zestawienie_osob_ze_stopniami_i_tytulami_naukowymi	Source: Polish Ministry of Science and Higher Education Catalog: dane.gov.pl	List of people who have been awarded the academic title of doctor or habilitated doctor or the academic title of professor
16	covid-confirmed-cases-latvia	Source: Latvian Centre for Disease Prevention and Control Catalog: data.gov.lv	Operational information on confirmed cases of COVID-19 submitted to the Centre for Disease Prevention and Control. Data are collected daily from laboratories and medical professionals.
16	covid-laboratory-tests-latvia	Source: Latvian Centre for Disease Prevention and Control Catalog: data.gov.lv	Number of laboratory tests performed to detect SARS-CoV-2 virus, number of laboratory confirmed case reports received, number of fatal reports received, number of disease outcome reports according to defined criteria.
14	polska-klasyfikacja-wyrobow-i-uslug-pkwiu-201501	Source: Polish institute for statistics Catalog: dane.gov.pl	Polish classification of products and service 2015
12	lidar-point-cloud	Source: UK Environment agency Catalog: data.gov.uk	Light Detection and Ranging (LIDAR) is an airborne mapping technique, which uses a laser to measure the height of the terrain and surface objects on the ground such as trees and buildings. Our LIDAR point cloud product is a collection of hundreds of millions, or sometimes billions of highly accurate 3-dimensional x,y,z points and component attributes

Table 6 shows the top-10 datasets identified in the external search keywords dataset for the "old normal". **We were only able to identify 5 of the datasets from Table 6 in queries in the internal search logs, LIDAR, PKWIU, LPIS, and CORINE Land Cover** (all present in Table 6), alongside EU-wide NUTS dataset (Nomenclature of territorial units for Statistics)

[6] Translated and summarised from original in French by one of the authors

[7] Translated and summarised from original in French by one of the authors

[8] Translated and summarised from original in French by one of the authors

Table 6: Top-10 datasets identified in the external search keywords dataset for the "old normal" period

Keyword	Matched dataset	#clicks	#impressions
aqma	Air Quality Management Areas (UK, but exists for many counties)	1389	34493
pkwiu	Polish Classification of Products and services	434	2705
kampfmittelbelastungskarte	Map of unexploded ordnance (Exists for several countries)	302	520
estacom	Spanish Foreign Trade Statistics	253	2638
lpis	Land Parcel Identification System (Exists for all EU countries)	143	19527
CORINE Land Cover	CORINE Land Cover (EU wide)	126	2119
cod postal romania	List of Romanian postal codes	124	13869
European medical device database	EUDAMED database	97	170
aqma gis data	Air Quality Management Data (UK; same as first of this list)	89	726
katastarske čestice	Cadastral map of the Czech Republic	88	3808

Table 7 shows the top-10 datasets identified in the external search keywords dataset for the "new normal" period. **4 COVID-19 datasets top the list**. **It is also interesting to highlight that the German and Croatian dashboard datasets come from queries written in Dutch language, suggesting the EDP is highly ranked by Google services in the Netherlands**. We were only able to identify in internal searches **the same 5 datasets as for "old normal": LIDAR, NUTS, LPIS, CORINE Land cover and PKWIU**.

Table 7: Top-10 datasets identified in the external search keywords dataset for the "new normal" period

Keyword	Matched dataset	#Clicks	#Impressions
corona dashboard duitsland	COVID-19 dataset for Germany	781	1787
covid map europe	Map of COVID-19 cases in the EU	526	14456
Oxford covid-19 government response tracker	Oxford covid-19 government response tracker	1389	34493
corona dashboard kroatie	COVID-19 dataset for Croatia	198	323
lpis	Land Parcel Identification System (Exists for all EU countries)	434	2705
pkwiu	Polish Classification of Products and services	302	520
postleitzahlen europea	List of European postcodes	298	1985

digital elevation model germany	Digital elevation model of germany	253	2638
Lidar europe	Lidar (exists for many European countries)	143	19527
european cdc situation update worldwide	ECDC global update	38	1161

Q1.3 What datasets are used together?

To answer this question, **we selected for "old normal" and "new normal" the sessions with more than one download and computed the top-5 datasets that were most downloaded in combination with other datasets**. For these top-5 datasets, we took a closer look at the list of datasets downloaded together with them, and the **keywords or facets** used in those sessions in an attempt to uncover patterns.

For "old normal", the top-5 datasets are:

1. **Real-time information for the vehicle's positions of the Brussels Intercommunal Transport Company in Belgium**[9]. The dataset was often downloaded in combination with the real-time API for *"Travelers information"* (planned works, diversions, incidents) from the same publisher[10]. A common query keyword was *"real-time"*, but queries tended to start outside the EDP platform, which meant that we did not have all their details to gain a more thorough understanding of the use case.
2. **The Global City Data index**[11], a range of indicators for a selection of cities produced by the United Nations. Re-published by London's city council because London is one of the cities included in the UN dataset. Often downloaded in combination with other *"Global City"* datasets, also re-published by the London City Council based on the UN data, like population estimates and comparison indicators. Groups of datasets including this dataset were downloaded following query keywords such as *"tourism"*, *"hotels"*, and *"global cities"*. In this case, we also noted the use of the format facet "XLS" that filters data published as Excel spreadsheets.
3. **New apartment prices by agency by year**[12]. Average prices of new apartments for which loans were approved by different types of financial institutions in Ireland. Often downloaded in combination with the related dataset on *"Second hand house prices by agency"* from the same publisher. In terms of query keywords and facets, we detected that multi-download sessions that include this dataset used the country facet *"Ireland"*. Interestingly, **less than 10% of the sessions had an explicit query keyword related to housing** (e.g. *"house-price"* or *"housing"*). Most of the sessions used seemingly unrelated keywords such as *"income"*, *"transport"*, *"school"* in combination with the country facet *"Ireland"*. **This suggests that this dataset was often downloaded as part of an information need about Ireland in general, and not about house pricing specifically.**
4. **Child obesity and excess weight**[13]. From the National Child Measurement Programme (NCMP, published by Public Health England). This is an annual survey of children attending state schools. The data shows children at risk of obesity and excess weight, which includes overweight and obesity. This dataset was mostly downloaded in combination with related health datasets about obesity in adults, or children obesity in other regions. In this particular case, we detected that most multi-download sessions including this dataset include the query keyword *"obesity"* or *"excess weight"*. **This suggests a very precise information need.**

[9] https://www.europeandataportal.eu/data/datasets/1fe9c5a2-7af3-4b15-a356-07801393b34f

[10] https://www.europeandataportal.eu/data/datasets/95527279-2be7-4a6b-9e00-bd829c504777

[11] https://www.europeandataportal.eu/data/datasets/global-city-data?locale=en

[12] https://www.europeandataportal.eu/data/datasets/https-data-usmart-io-org-ae1d5c14-c392-4c3f-9705-537427eeb413-dataset-viewdiscovery-datasetguid-009dcd9f-5410-4dbf-b740-90d642304e49

[13] https://www.europeandataportal.eu/data/datasets/child-obesity-and-excess-weight

5. **Administrative limits of French municipalities from OpenStreetMaps**[14]. Often downloaded in combination with other datasets from the same publisher, notably, the geographic contours of departments identified as one of the top-10 most downloaded datasets in the old normal corpus discussed earlier. Mostly accessed directly from an external search engine without using internal search keywords or facets.

Q1.4 How does the previous dimensions vary over time? Are there any periodical or episodical effects in demand over time? Can we link the latter to specific events?

In this section we study the **change over time of the indicators analysed in questions**. To help frame the analysis, we first provide some context on the variation over time of the total number of visits to the EDP and of the different types of user sessions that may include a dataset download:

1. **Internal dataset search sessions** are sessions that land on an EDP page that is not dataset-search related (e.g. the EDP homepage), then continue to the dataset search interface, then issue a query using keywords or facets.
2. **External dataset search sessions** are sessions that land directly on a dataset search result page, referred to from another website or shown as a result by a web search engine. Web search engines crawl and index result pages (e.g. https://www.europeandataportal.eu/data/datasets?keywords=karte).
3. **Dataset page sessions** are sessions that visit at least one dataset page, which are not internal or external dataset search sessions. 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

Understanding how these three sessions types perform helps us connect trends on demand with trends on the numbers of visits of certain types, facilitating our next task: associating demand up-peaks and down-peaks with internal or external events. Previous analysis of data from national portals suggested that visiting open data portals is work-related[15], with most activity happening during office hours and significantly fewer traffic during weekends. Our analysis of the EDP data from AR-18 confirmed that observation.

As we assumed that activity levels are consistently low over the weekend, we could aggregate data by week when visualising it (Figure 7).

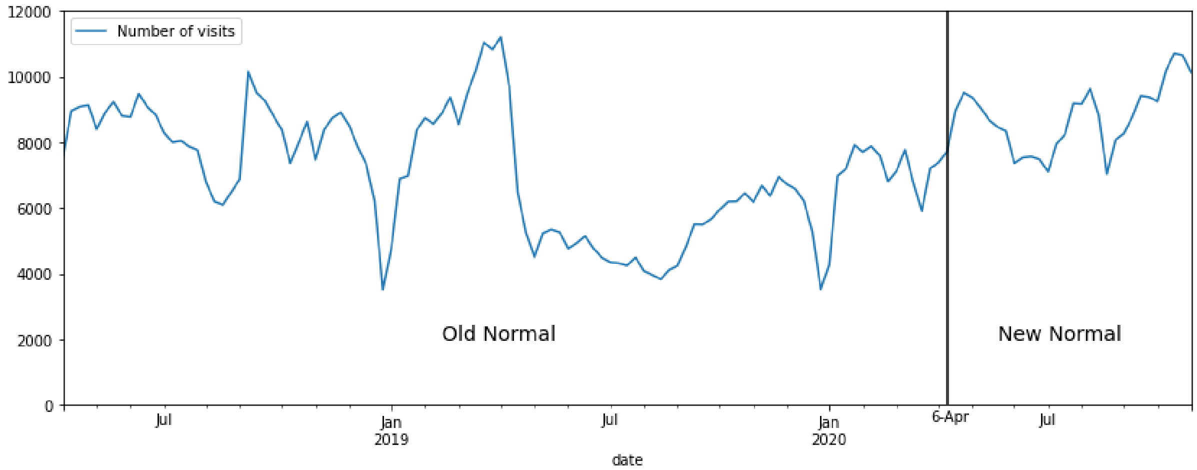


FIGURE 7: WEEKLY VISITS TO THE EDP

Figure 7 shows the number of total weekly visits to the EDP. The black line separates the Old Normal and New Normal corpora, which we analysed separately, as explained earlier.

[14] https://www.europeandataportal.eu/data/datasets/536998c8a3a729239d205051

[15] E. Kacprzak et al. *Characterising dataset search—An analysis of search logs and data requests*, Journal of Web Semantics, Volume 55, 2019.

We note the following trends:

1. There are **significant down-peaks around Christmas and New Year**, consistent with previous observations about open-data-search being a work matter.
2. There was a **sharp increase of visits during the first week of September 2018**, **suggesting an external or internal event. We analyse this week closely later in this section.**
3. There was an **increase in the number of visits in March 2019, followed by a sharp decline from April 2019**. This sharp decline was discussed at length in AR-18; it was due to the change in URL scheme of the dataset section, introduced by a new release of the EDP portal. This led to a period when web search engines had to re-index those pages, and in most cases, with a lower ranking than before the change. **Visits to all other sections maintained or increased their previous trends**, which suggests that overall, the portal remained popular. In particular, the **number of visits in March 2019 is due to the promotion of the new release**.
4. After the sharp decrease, **the number of visits started to recover steadily starting from August 2019** (aside the quiet times around the December holidays). This is the result of the Search Engine Optimisation (SEO) work by the EDP team, following the drop in traffic from April 2019. **From 29 March 2020, we observe one low peak in the month of June and one in the first week in August. After that second low peak, the upward trend continued, and by the end of October, the number of total visits was very close to the historical maximum of 11000 weekly visits**.

Figure 8 shows the total weekly visits to the EDP when searches start on the EDP page rather than landing here from an external site. We make several observations:

1. **The same down-peaks around the end of December apply to internal searches as well.**
2. **The same increase-decrease pattern is associated with the release of the new version of EDP.**
3. There is **no visible increase in the first week of September 2018, suggesting that visit up-peak observed in Figure 3.4 was not related to an increase in the number of internal search sessions**.
4. The number of internal search sessions during the new normal period remained low, indicating the increase in total number of visits observed in Figure 4 was not due to a recovery in the use of the internal dataset search.

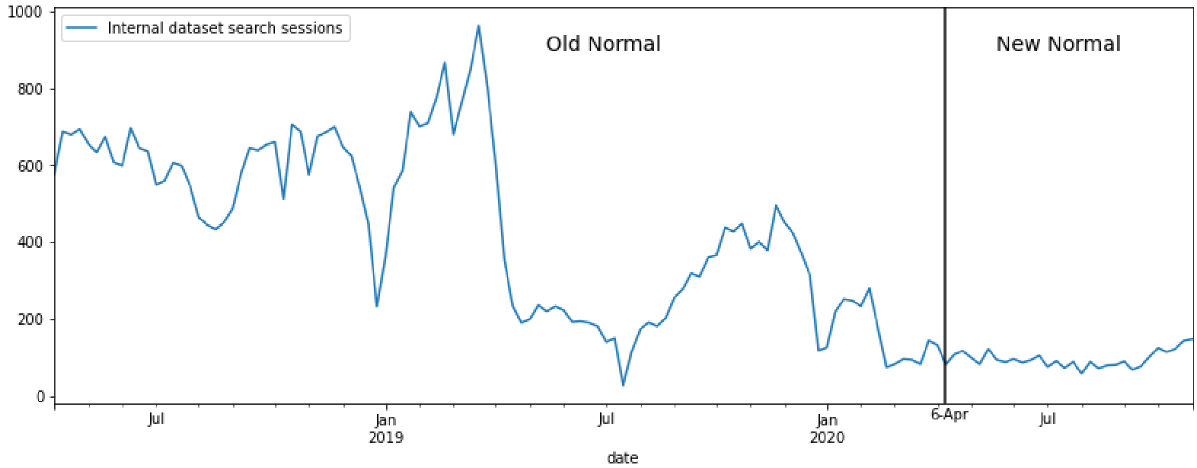


FIGURE 8: WEEKLY INTERNAL DATASET SEARCH SESSIONS. THERE ARE LESS VISITS OF THIS TYPE AFTER NEW NORMAL, MEANING THE OVERALL INCREASE SHOWN ON FIGURE 7 COMES FROM OTHER SECTIONS

Figure 9 shows the same data, but for external search sessions. **Sessions of this type have stayed relatively steady during all the analysed period, with the notable exception of a sharp increase during January 2020 and early February 2020**. Further analysis of the sessions involved revealed that there was not a single external search keyword associated to the peak. The two most significative changes compared to previous and precedent periods are:

1. **The keyword *"Portugal"*, acquired from a link to the EDP with that keyword on a webpage in Portuguese about forestal GIS[16]. This webpage always drives a few monthly visitors to the EDP, but during January 2020 this number significantly increased.** We don't have enough data to pinpoint why this happened.
2. The phrase *"Insurance Application"* that previously had almost zero hits but had over 50 visits in this period. **All visitors came from the USA and almost all of them exited the portal immediately.** The general health insurance application period in the USA expires mid-December, but certain types are still open in January[17]. We speculate that from the large number of people looking for these keywords, a few were shown results from the EDP, clicked on them, and quickly found out that this was not what they were looking for.

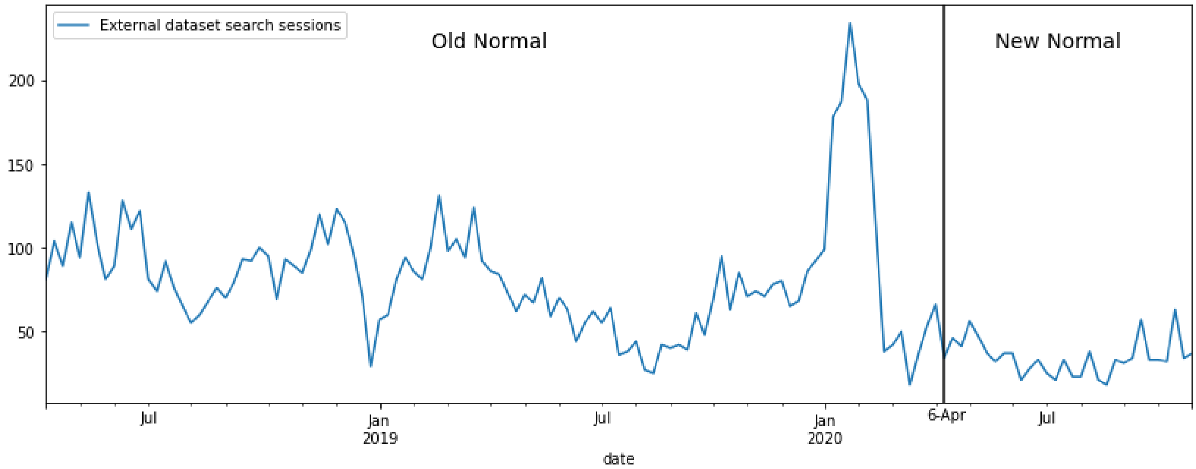


FIGURE 9: WEEKLY EXTERNAL DATASET SEARCH SESSIONS. THERE ARE LESS VISITS OF THIS TYPE AFTER NEW NORMAL, MEANING THE OVERALL INCREASE SHOWN ON FIGURE 7 COMES FROM OTHER SECTIONS

Figure 10 shows the number of weekly dataset page sessions. **We observe a sharp increase in the first week of September 2018 of almost 3000 visits**, matching the peak from Figure 3.4 for the same period. **This means that the up peak was caused almost entirely by this type of session.** An analysis of the referrer type and URL of these visits suggested that the reason was the **launch of the Google Dataset Search engine**. We believe **the novelty of the tool attracted many users to try it and were redirected to the EDP after doing a search there. However, as the novelty wore off, the number of visits reverted to previous numbers.**

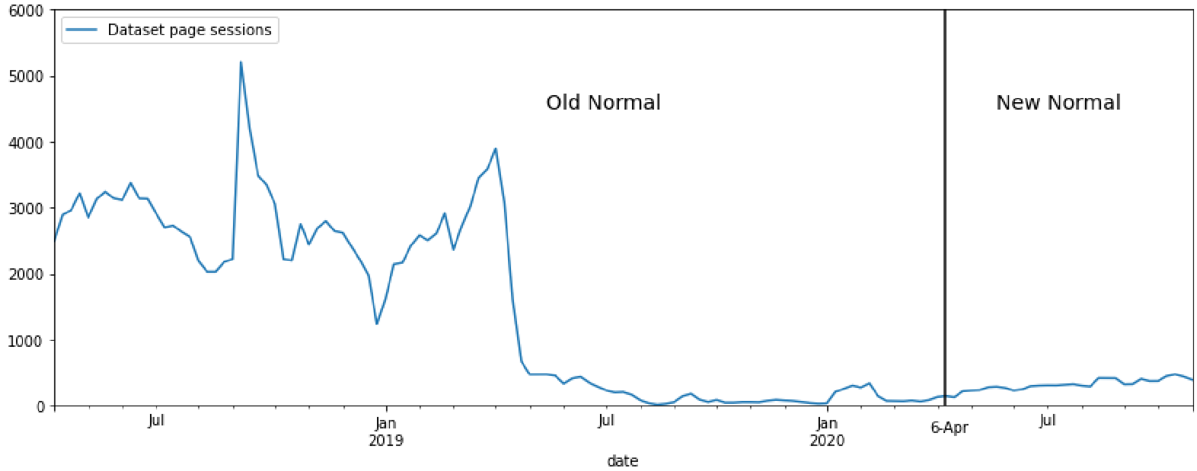


FIGURE 10: WEEKLY DATASET PAGE SESSIONS. THERE ARE LESS VISITS OF THIS TYPE AFTER NEW NORMAL, MEANING THE OVERALL INCREASE SHOWN ON FIGURE 7 COMES FROM OTHER SECTIONS

[16] http://forest-gis.com/2012/01/portugal-shapefiles-gerais-do-pais.html/

[17] https://www.healthcare.gov/coverage-outside-open-enrollment/your-options/

We then studied **the variation in time of the demand for datasets and dataset categories**. For categories, we plotted the number of sessions that include at least one click on a category, aggregated weekly. We split the line graphs into three charts, following the category groupings identified in Section 3.2.1 as follows: groups 1 and 2 (Transport, economy, environment, regions and agriculture) in Figure 11, group 3 (Health, population, government, energy, education) in Figure 12, and groups 4 and 5 (Science, Justice and International) in Figure 13. Note that these Figures are not directly comparable with Figures 8, 9 and, 10, as we are now analysing a very specific subset of sessions to answer the question: what are the most demanded categories?

Across all groups, we note the following:

1. An **increase in the use of categories starting from the first week of April 2019, which is when the new release of the EDP came out.** The effects were long-term, suggesting that as **the categories featured more prominently in the new release, they were also used extensively**.
2. Demand down-peaks are evident at mid-July and at the end of December of 2019, consistent with holiday periods in Europe.
3. A general down-peak around the beginning of March 2020, consistent with the start of the COVID-19 pandemics in Europe.
4. **There were no significant up-peaks of demand in any category**. This suggests that there was no period or event where a particular category was more demanded than others. In Section 3.1 we discussed those categories that are in high demand throughout the entire time interval we analysed.

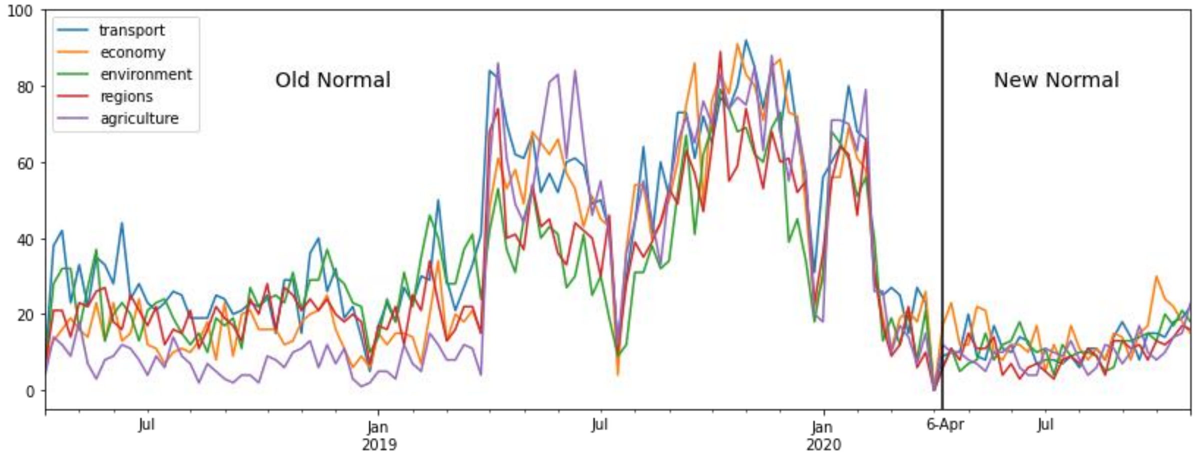


FIGURE 11: WEEKLY DEMAND OF TRANSPORT, ECONOMY, ENVIRONMENT, REGIONS AND AGRICULTURE CATEGORIES

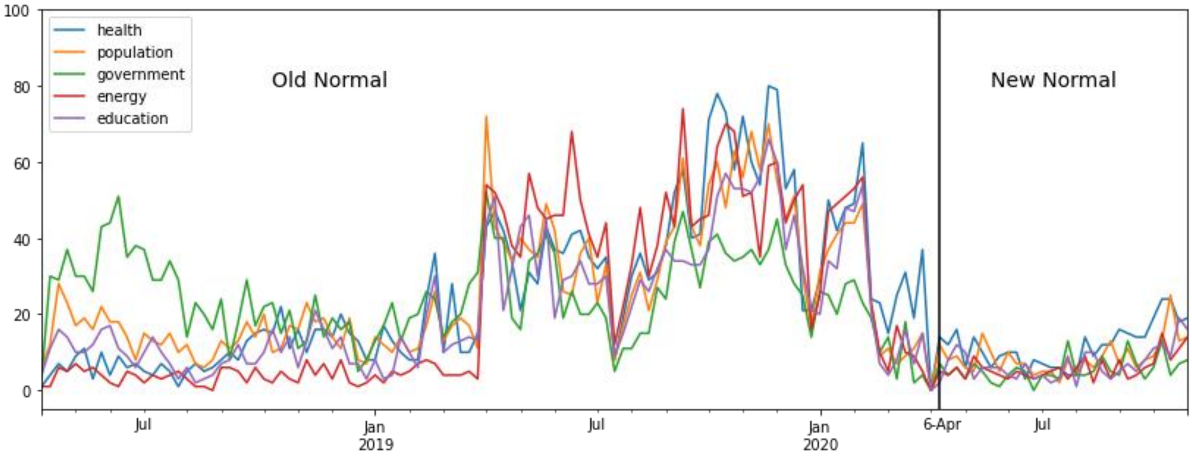


FIGURE 12: WEEKLY DEMAND OF HEALTH, POPULATION, GOVERNMENT, ENERGY AND EDUCATION CATEGORIES

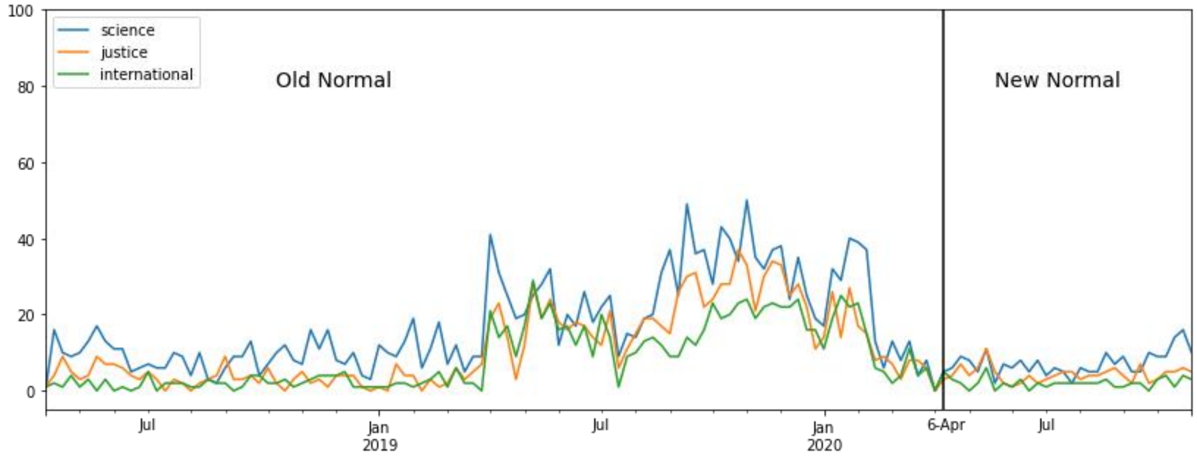


FIGURE 13: WEEKLY DEMAND OF SCIENCE, JUSTICE AND INTERNATIONAL CATEGORIES

Next, we study the **variation in demand of datasets in terms of number of downloads** for two groups of the datasets:

1. The **10 most demanded datasets** from the "old normal" corpus identified in Section 3.2.2, plus the COVID-19 dataset that was the most demanded during "new normal". We decided to not consider the rest of the most demanded datasets in the "new normal" corpus because they have much fewer downloads than the COVID-19 dataset.
2. The **top-10 datasets by maximum download up-peak**. We define **a download up-peak of a dataset as a number of weekly downloads that is at least 3 standard deviations greater than the mean number of downloads of that dataset**[18]**. We computed the maximum download up-peak for each dataset with at least 50 total downloads and reported the 5 datasets with the highest up-peak. We note that only 147 datasets have at least one download up-peak, less than 0.001% of the total number of datasets indexed.**

For each up-peak of the 5 selected datasets, we **took a closer look at the sessions that led to those downloads, including their country of origin and their type.** We used this information to add context to the up peak observed in the data and possibly link it to an internal or external event which would deliver an explanation for the change.

Figure 14 shows the weekly downloads of the top-5 most downloaded datasets in the "old normal" corpus alongside the COVID-19-data dataset from "new normal". **There were four significant up-peaks**:

1. The **automatic number plate recognition dataset** on the last week of May and first of June 2018. Sessions were almost all of the type dataset page. We did detect a few referrals from the mail server of a large consultancy company. In terms of origins, the downloads varied a lot, with a majority from outside Europe, specifically South-East Asia. We were not able to link this to a particular event, but we presume that it might be related to a **machine learning educational challenge in that region**, as the detection of number plates is a common task in computer vision.
2. The **corporate credit card transactions** in the third week of May of 2018. Similar to the previous dataset, a very diverse origin of sessions, mostly from India. In this case, we were fortunate enough that two users allowed the sharing of their Google keywords of "credit card dataset" and one "sample credit card dataset". Again, it was not possible to link to a real-world event, but **we also believe this might be another machine learning challenge, as credit card fraud detection or turnover prediction are also common data science exercises**.
3. The list of **Romanian war casualties** during the last week of November 2018. Sessions connected to this download were almost all from Romania, of type dataset page (coming from a web search engine). After the analysis of Romanian news archives for that week, we found the event that we believe sparked the interest: The inauguration of a war cemetery for Romanian soldiers fallen in WW2 in Krasnodar -

[18] In statistical terms, this is the z-score of the number of weekly downloads

Russia[19]. We presume this might also be related to activities around the centenary of the Great Union Day from 2018 in Romania.

4. The **COVID-19 coronavirus dataset**, with an up-peak end of March, and another one mid-August. The end of March peak had mostly visitors from Europe, and two main referrer sites, the homepage of data.europa.eu, where the EDP was linked as an index of COVID-19 datasets, and the website of the EU Datathon 2020[20], where this dataset was featured. Both referrer sites were approximately equally contributing to the peak. **The mid-August peak does not have any of these referrals, with almost all sessions coming from web search engines, and from which we could not link to a particular event.** Most sessions come from Europe, and we were fortunate that three users shared their Google search keywords: "COVID 19 EU stats", "covid data Europe", and "covid cases europe". Further analysis of search console data suggests that this dataset page ranks highly for combinations of keywords including the terms *"covid"*, *"europe"* and *"data"*. We believe this is due to how the original publisher, the open data portal of the European Union, was pointing to the EDP dataset from one of its main pages, thus increasing its rankings. We also note **a down-peak during May and June**; we assume this is because this dataset was featured in the Covid-19 section of the EDP. We test this hypothesis in Section 4.

The **Estonian medical device database** does not exhibit a clear up peak, but there are 7 different weeks in the June 2020 to November 2020 period with more than 20 downloads. Available internal and external searches are a combination of the keywords *"medical device"*, *"database"*, and *"European"*. The lack of a clearer link to the country of origin in keywords and facet filters suggests that users may have been searching for an European-level database, which was not available to the public - as per November 2020, this dataset is being developed under the codename EUDAMED[21]. The Estonian dataset does mention following *"European legislation"* - we believe this is the reason why it was matched to those queries.

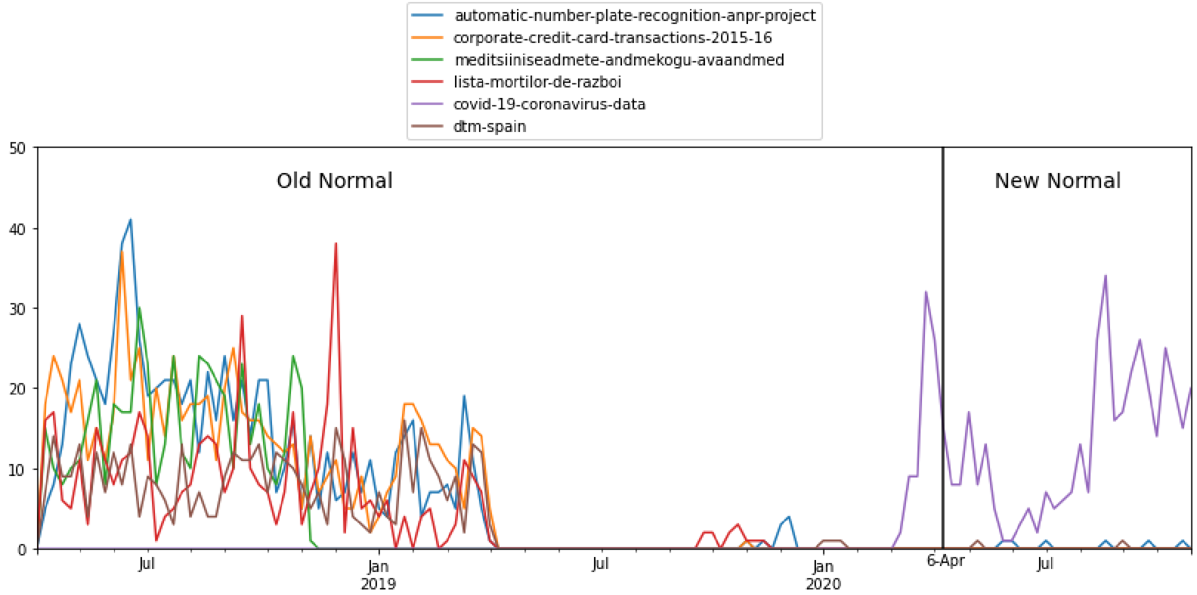


FIGURE 14: WEEKLY DEMAND OF TOP-5 MOST DOWNLOADED DATASETS DURING OLD NORMAL, PLUS THE COVID-19 DATASET, BY FAR THE MOST DOWNLOADED DATASET DURING NEW NORMAL.

Figure 15 plots the weekly downloads of the 5 datasets with the highest download up-peaks. The datasets and their up-peaks are:

[19] https://moscova.mae.ro/en/local-news/1329

[20] https://op.europa.eu/en/web/eudatathon/covid-19

[21] https://ec.europa.eu/tools/eudamed/#/screen/home

1. The **database of road accidents in France**, published by the Ministry of Interior. Further analysis revealed that this up-peak was down to a single user connecting from Tunisia, which in the course of a single day downloaded all relevant files. At that time, this dataset was broken down in 40 distributions, including csv files for different years and different views of the data, e.g., one file for vehicles and another for description of causes. This single user downloaded all distributions.
2. **Digital Elevation Model of Ireland Digital Elevation Model of Ireland**, from NASA's Shuttle Radar Topography Mission (SRTM), published by Dublin City Council, with an up-peak on the second week of December of 2018. Further analysis revealed the exact same situation as with the previous dataset, **a single user downloaded all distributions of a dataset because the distributions were not different formats of the dataset, but different slices of it.**
3. **Invoices paid by the Basildon and Thurrock University Hospitals NHS Foundation Trust** in late November 2018. Inspection of sessions revealed that visits to this dataset were **referred by a question posted on the Quora website, that asked "Where can I get a data set of 1,000 different PDF invoices?"**. Today, the question appears empty, possibly due to the original posters closing their Quora accounts. The session data we analysed confirmed that a direct link to the dataset was available at that time. **Note that similar to the automatic number plate recognition and corporate credit card datasets, users arrived through this dataset not because they were specifically looking for it, but because they were looking for generic examples of datasets, for which this one is an instance.**
4. The **Czech CORINE land cover database** (identified in the figure with the URL 5b7a9ba5…) has an up peak in March 2019 that spans over 3 weeks. Most interested users came from either Czech Republic or Slovakia, and most of the sessions were of dataset page type, coming from web search engines. Unfortunately, there are not search keywords available and we were not able to link this up-peak to an event.
5. **Location of parking meters in Dublin** has an up-peak in February 2019. 7 of the 10 sessions that contributed to this up-peak used the internal search field on the EDP after landing on the EDP homepage. Keywords used are a combination of *"Ireland"* and *"Dublin"*, and all of them have multiple downloads, **suggesting an information need about the country and not about the specific dataset**. The remaining three sessions reached the dataset from a web search engine. It was not possible to accurately link this up-peak to an event or information need as the queries were not available.

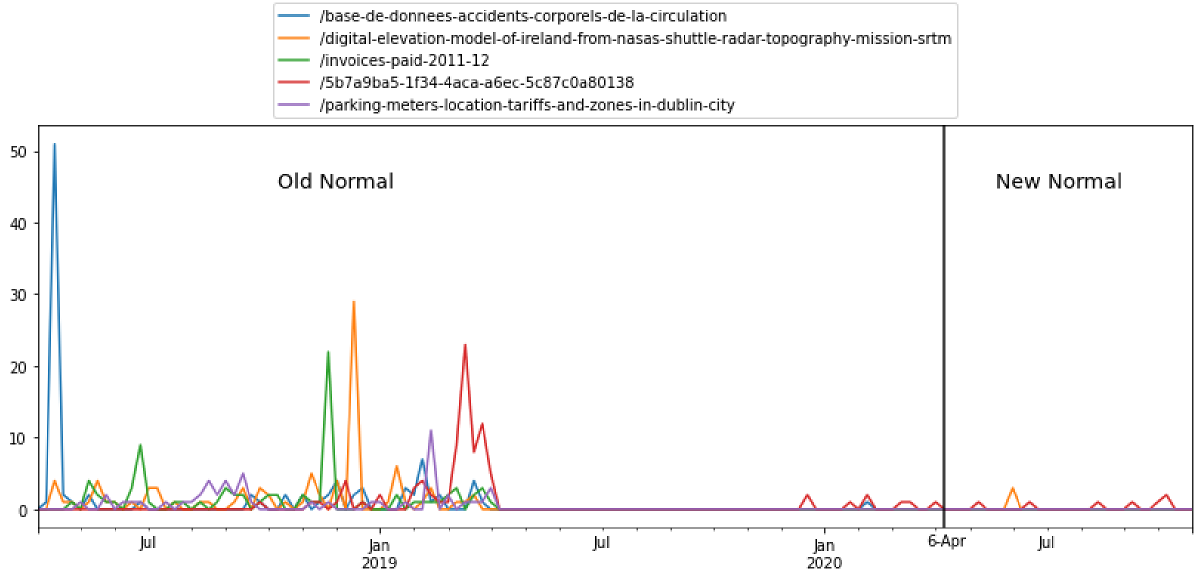


FIGURE 15: WEEKLY DEMAND OF TOP-5 DATASETS WITH LARGEST UP-PEAKS

3.2 Summary of Results

Q1.1 What categories of datasets are in high demand?

- Transport, economy, environment, and health. Health became more popular past March 2020 for understandable reasons. Transport declined.

Q1.2 What datasets are most demanded by users?

- The list of top-10 most demanded datasets is quite varied, during the old normal period there was a prevalence of **geospatial datasets**. Past April 2020, COVID-19 related datasets were the most downloaded, in particular, the **compilation of updates published by the EU Centre for Disease Control**. We noticed that the top 2 most downloaded datasets during old normal are likely to be popular because they were used by people participating in data science training (e.g. challenges, hackathons).

Q1.3 What datasets are used together?

- The list of datasets most used together is as varied as the one of most demanded datasets. From the analysed sample we identify cases where datasets with the same publisher are downloaded together, datasets that are part of a complex information need (datasets about a country or city), and datasets that are updated in real-time.

Q1.4 How do these dimensions vary over time? Are there periodical or episodical events?

- In general, dataset demand is quite **stable, with relatively few peaks** (only 147 datasets showed a download up peak).
- The only periodical event we found affects demand are December holidays, where the overall number of visits to the EDP significantly decreases.
- In terms of external factors that increased demand, we identified that the **launch of Google Dataset Search drove a large number of general dataset page visit**s. However this did not sustain in time. We also observed the **EU Datathon 2020** as an event that increased the demand for health and COVID-related datasets during March and early April 2020.
- We were only able to identify one news event that affected demand: the inauguration of a war cemetery and memorial for Romanian fallen soldiers in Russia increased the demand for an official dataset about the list of soldiers fallen in war.
- Most substantial high up-peaks were identified as being caused by a few users downloading datasets that were distributed in many pieces, something that we may consider as single downloads.

Role of Curated Content

Results

Q2.1 Was the EDP used more during the lifetime of the COVID-19 section than before?

Figure 16 compares the total visits between the COVID-19 section and the two previous periods defined in Section 3.1: immediately previous and a previous period without the Christmas and New Year break. We observe that during the lifetime of the EDP COVID-19 section, **there were ~20% more visits than in the immediately previous period**. We also highlight that despite the immediately previous period running through the end of the year break, it has slightly more visits than a previous period that does not run through it. Combined with the evidence of Figure 4, **this further suggests that the EDP had recovered from being removed from web search engines indexes**.

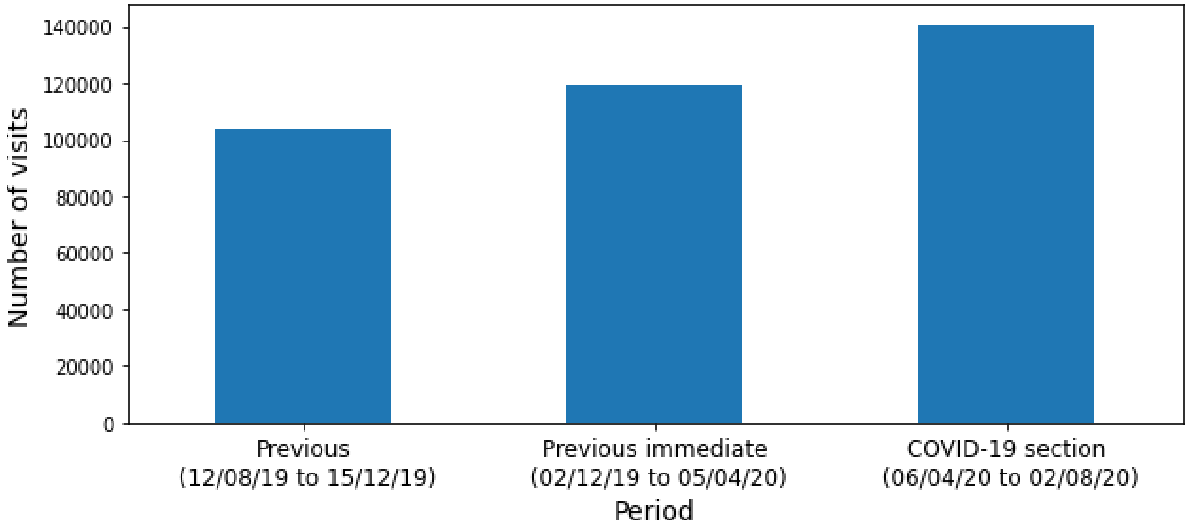


FIGURE 16: COMPARISON OF NUMBER OF VISITS BETWEEN LIFETIME OF COVID-19 SECTION (APRIL-JULY 2020) AND PREVIOUS PERIODS OF THE EDP (AUGUST 2019 - DECEMBER 2019 AND DECEMBER 2019 TO APRIL 2020)

To answer if the COVID-19 section is the main factor for the traffic increase, for each period, we broke down visits according to where on the portal they go. The areas are: COVID-19, Dataset, Training, News & events, Impact & Studies and homepage bounces. Figure 17 shows the results for the three periods (August 2019 - December 2019, December 2019 to April 2020, April 2020 - July 2020).

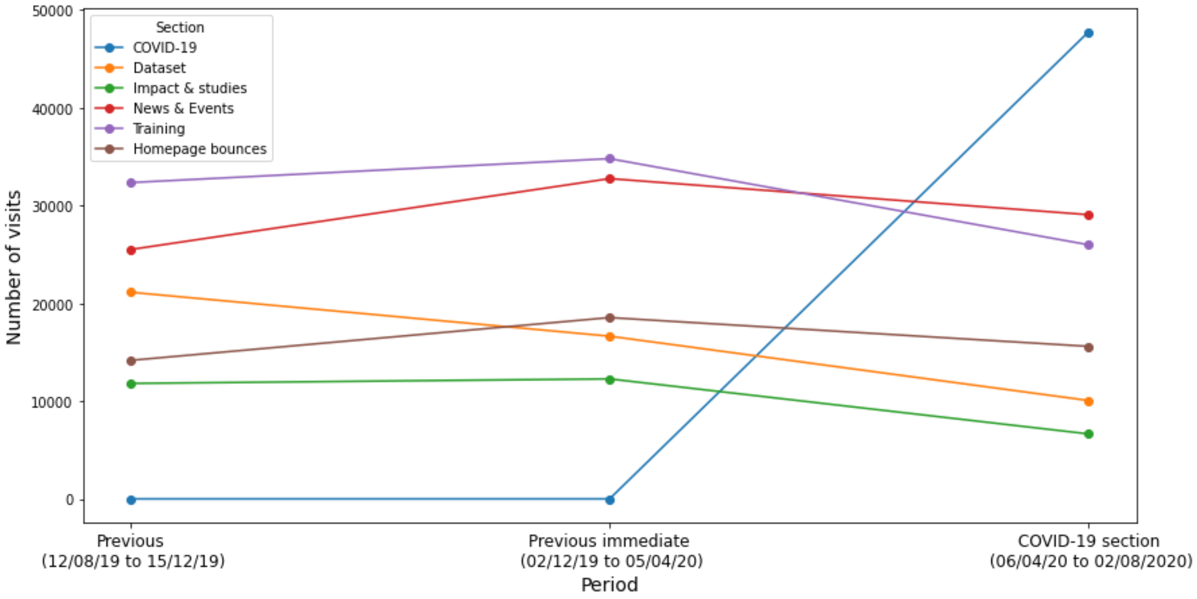


FIGURE 17: NUMBER OF VISITS PER SECTION FOR THE COVID-19 SECTION LIFETIME AND EARLIER PERIODS

The COVID-19 section was divided in three subsections: stories, initiatives and datasets, was any of them more popular than the others? Table 8 summarises the visits per sub-section of the COVID-19 section during the relevant period from 6th April to 2nd August 2020. The initiatives sub-section was the most popular, followed by stories and datasets. Most sessions visited exclusively one section, a behaviour similar to the one observed for other parts of the portal and reported in AR-18. There is also a high rate of sessions that visit at least one item of the subsection. This suggests specific information needs, but also the opportunity to invest more in user retention, for instance with recommendations and cross-links.

TABLE 8: NUMBER OF VISITS PER SUBSECTION OF THE COVID-19 SECTION DURING THE "COVID-19 SECTION" PERIOD (APRIL - JULY 2020)

Sessions \ Subsection	Stories	Initiatives	Datasets
All visits (including bounces and browsing only)	16741 (35.0% of total visits)	19160 (40.1% of total)	12772 (26.7% of total)
Exclusive visits	15017 (89.7% of stories visits)	17383 (90.7% of initiatives visits)	10395 (81.3% of datasets visits)
Visited at least one item	15711 (93.8% of stories visits)	17711 (92.4% of initiatives visits)	7579 (59.3% of datasets visits)

Q2.2 What happened after the COVID-19 section was retired?

To answer this question, we added the period immediately after the retirement of the COVID-19 section to the analysis (August - October 2020). Figure 19 shows the comparison of this period against all the previous ones.

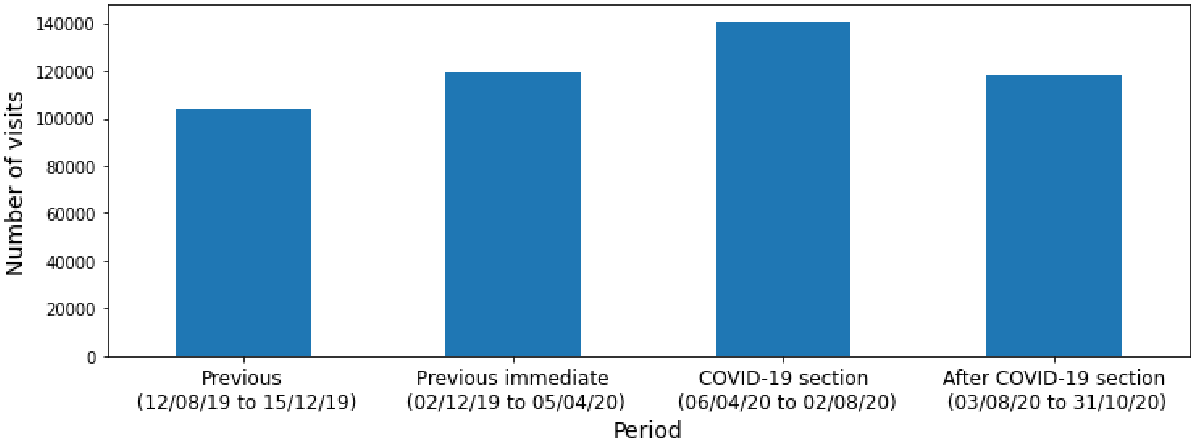


FIGURE 18: COMPARISON OF NUMBER OF VISITS FOR DIFFERENT PERIODS OF THE EDP AROUND THE COVID-19 SECTION LIFETIME

We observe that **after the moving of the COVID-19 section under Impact & Studies, the overall number of visits decreased to numbers close to the immediately previous period**. Was this the result of users not coming anymore to the COVID-19 section in its new location? To find out, we broke down the visits of the "After COVID-19 section" period in a per-section basis in the same way we did with the other periods. We refer to the sessions that visited the new location of the COVID-19 content as *"COVID-19"*, and those that exclusively visited one of the other pages in the Impact & studies section as *"Impact & Studies"*. Figure 20 shows the results.

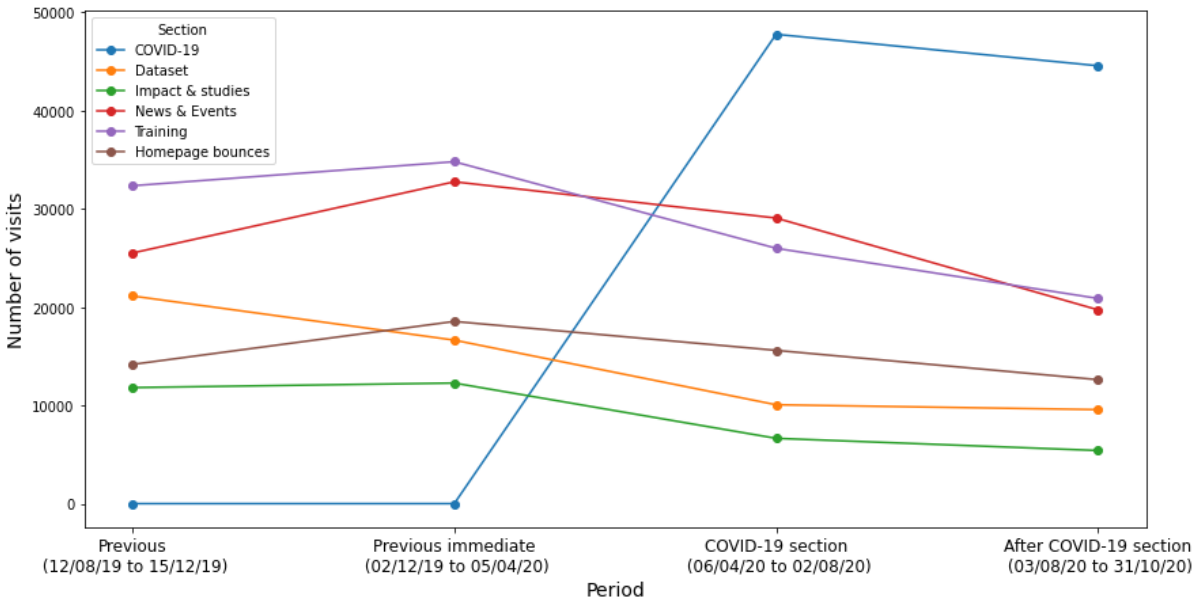


FIGURE 19: NUMBER OF VISITS PER SECTION FOR THE COVID-19 SECTION LIFETIME, PREVIOUS PERIODS AND SUBSEQUENT PERIOD

Interestingly, despite not being in the spotlight anymore, visits to the COVID-19 section in its new place were relatively more than during its lifetime. There is, however, a decrease in the number of visits to all sections of the portal except for datasets. The decrease is sharper for news/events and training.

Q2.3 How did users reach the COVID-19 section?

When the COVID-19 section was featured on the front page, 42029 sessions landed directly on it from an external referrer (81%), 3832 after landing on the homepage (8%) and 5714 visited it after landing somewhere else on the portal (11%),. After being moved under Impact & studies, **43151 sessions landed directly on it from an external referrer (+2.7%) and 1393 sessions visited it after landing somewhere else on the portal (-75.7%), from which 345 landed on the homepage (-91%)**. Most users reached the COVID-19 section through external referrers (mostly search engines), however, the demotion of the content meant a sharp decrease in the number of users that reached the section after landing in the homepage.

Compared to other sections of the portal during its lifetime, the COVID-19 section did not get much more referrers from websites (5.7% vs. 5.3%) but did get a slightly larger ratio of referrers from social networks (2.6% vs 1.5%). We also note an increase in the number of referrals from Facebook with respect to the other sections of the portal during this period.

The websites that drove more traffic to the COVID-19 section during its lifetime were:

1. The Austrian open data portal, that linked to the EDP from its page referring to COVID data https://www.data.gv.at/covid-19/
2. The European Commission page on Coronavirus research and innovation: https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/coronavirus-research-and-innovation_en
3. A story featured on the Austrian newspaper "Der Standard" about open data on COVID-19[22].

After the archival of the COVID-19 section the percentage of referrals from websites and social networks decreased to 2% and 0.5% respectively. The Austrian open data portal remained as the main referrer, followed by a similar link provided by the Irish open data portal.

[22] https://www.derstandard.at/story/2000116864448/oesterreichische-covid-19-statistikdaten-als-open-data-verfuegbar

Q2.3 Did the publication of COVID-19 data stories increase traffic to the portal?

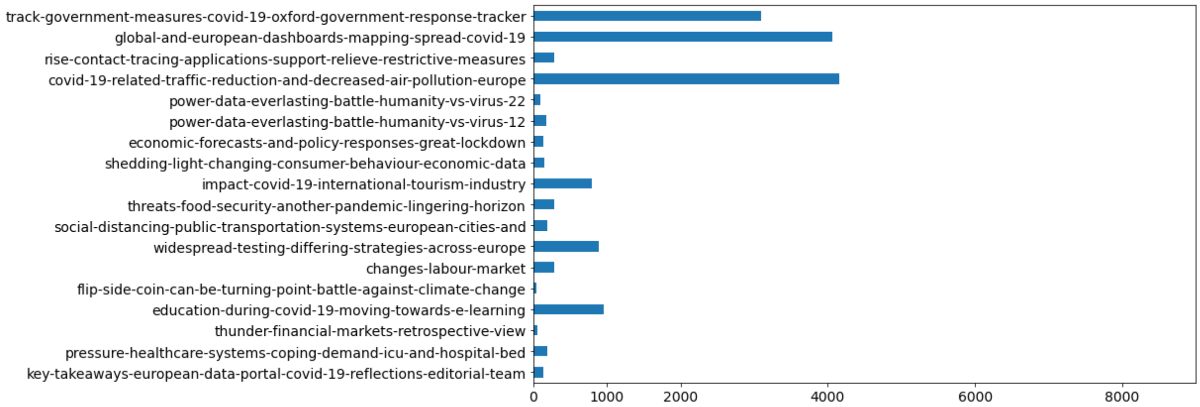


FIGURE 20: NUMBER OF VISITS FOR EACH COVID-19 STORY DURING THE COVID-19 SECTION PERIOD. STORIES ARE ORDERED BY PUBLICATION TIME FROM OLD TO NEW

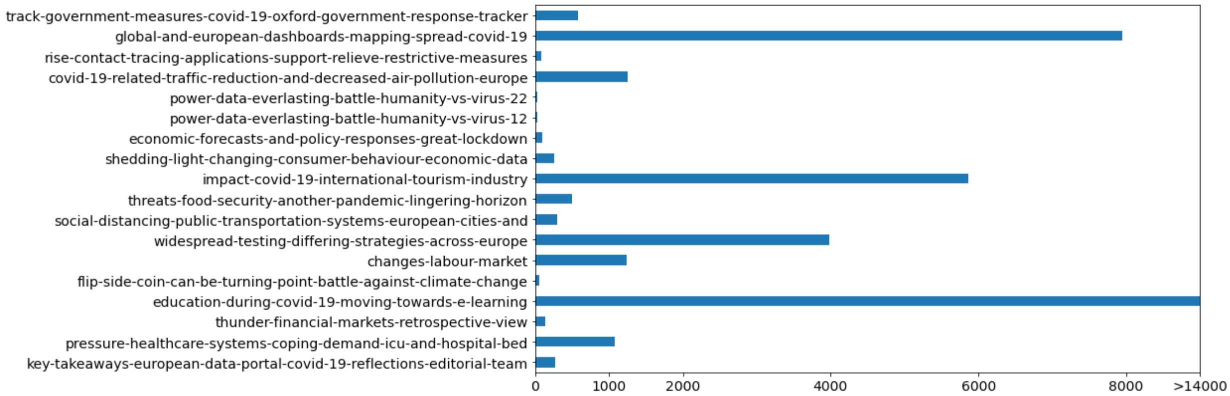


FIGURE 21: NUMBER OF VISITS FOR EACH COVID-19 STORY DURING THE "AFTER COVID-19" SECTION PERIOD. STORIES ARE ORDERED BY PUBLICATION TIME FROM OLD TO NEW

Stories had varied success. Figure 22 shows the number of visits to each story during the COVID-19 section period (April - July 2020). **Three stories got more than one thousand visits: the Oxford response tracker, the one listing the European COVID-19 dashboards and the one on the decrease of air pollution in Europe.** These three were among the first published, therefore, had more time to accumulate more visits than the ones at the bottom of the chart.

Figure 23 shows the number of visits to each story during the "After COVID-19 section" period (August - October 2020). The story on dashboards was even more popular than before. **We also observe a high number of visits to the stories about impact on the tourism industry and about testing strategies across Europe**. An interesting highlight is the story about education moving to eLearning during the pandemic, which received over 14000 visits over the period. By comparison, as discussed in Section 4.2.1 (Table 8), there were 16 thousand visits to all stories from April to July, while the eLearning story achieved almost the same on its own. **The top-3 stories during the "After COVID-19 section" made up for most of the traffic lost when the datasets and initiatives version were deleted.**

To understand if the stories were responsible of the increase, we analysed the sessions that included them. We found that this was indeed the case: users landed in the stories directly from a search engine more than 90% of the times.

We further analysed the time series of the top-5 most visited stories:

1. The **global and european dashboards** story (Figure 24) did not have a lot of traction after its publication. Interest rose by the end of July and remained approximately constant up to the end of October.
2. The **widespread testing** content (Figure 25) had two up-peaks in mid-August and mid-September. For mid-August, almost all traffic came from search engines, the only registered keywords were *"what country makes more pcr tests"* and *"pcr test for EU"*. For mid-September, almost all visitors came from

search engines as well, the only registered keywords were: *"testing rate Europe covid19"*, *"is covid testing in the EU working"* and *"countries managing test and trace"*.

3. The **education moving to eLearning** story (Figure 26) started getting traction in early August, reaching more than 1000 weekly visits by September and keeping the same levels until the end of October. For this story, we looked for the keywords *"learning"* and *"e-learning"* in the external search keywords dataset analysed in Section 3. **We found 18 out of 500 entries including combinations of *"learning/e-learning"* and *"COVID/Corona/Pandemics"*, totalling over 800 clicks with a relatively high clickthrough ratio of 27%.**
4. The **traffic and air pollution reduction** story (Figure 27) had an up peak the third week of May. Almost all traffic came from search engines and no search keywords were registered. Since August, traffic is marginal.
5. The **impact on tourism industry** story (Figure 28) had a similar performance to the ones on dashboards and eLearning: **traffic starts picking up long after the publication of the story and then remains more or less constant**.

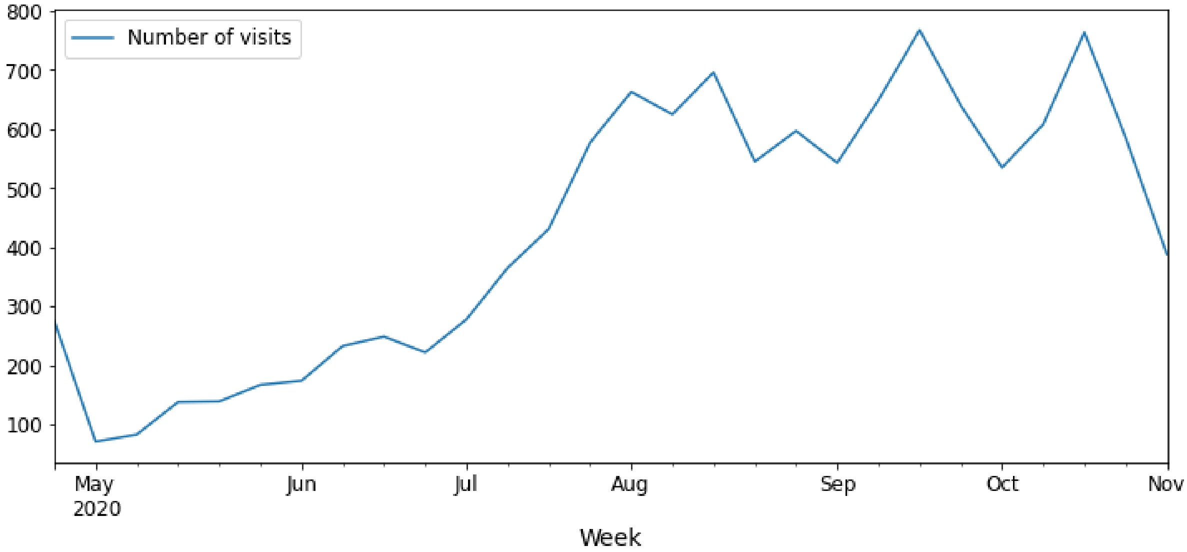


FIGURE 22: WEEKLY NUMBER OF VISITS TO THE GLOBAL AND EUROPEAN DASHBOARDS STORY

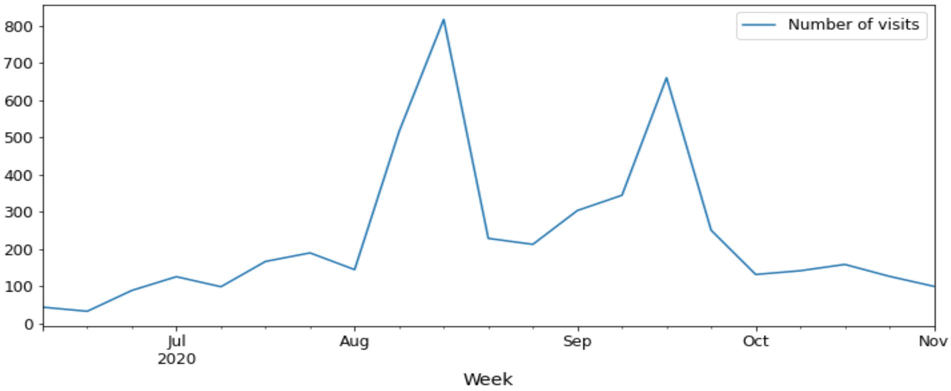


FIGURE 23: WEEKLY NUMBER OF VISITS TO THE WIDESPREAD TESTING STORY

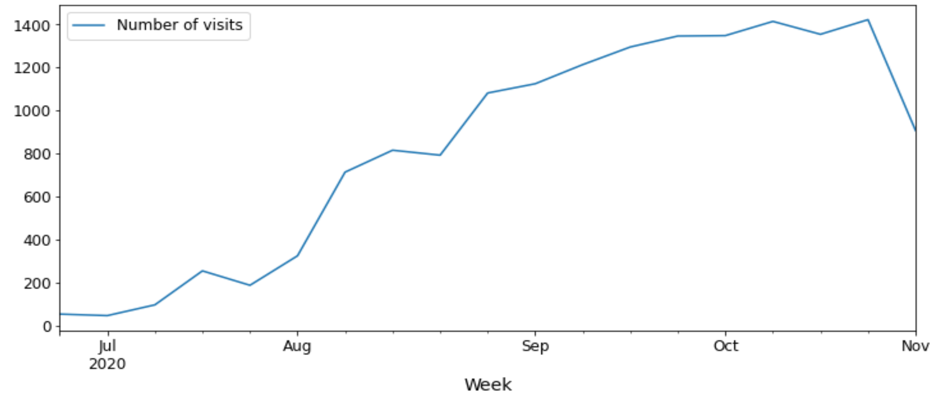


FIGURE 24: WEEKLY NUMBER OF VISITS TO THE EDUCATION MOVING TO ELEARNING STORY

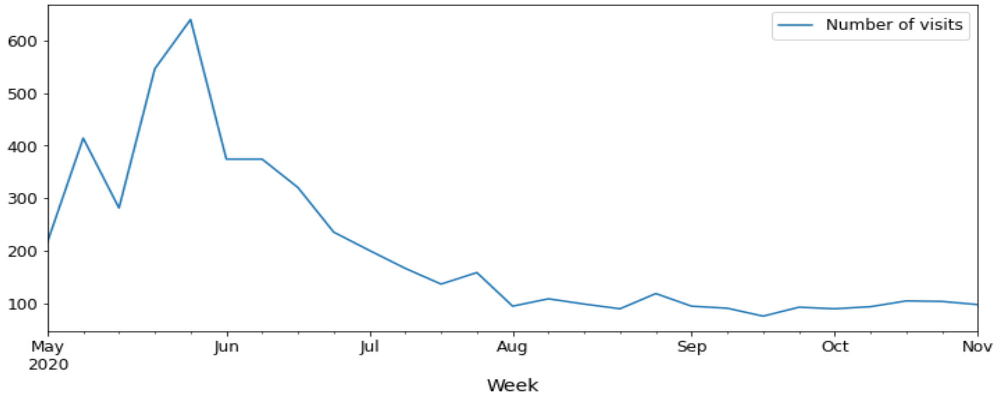


FIGURE 25: WEEKLY NUMBER OF VISITS TO THE TRAFFIC AND AIR POLLUTION REDUCTION STORY

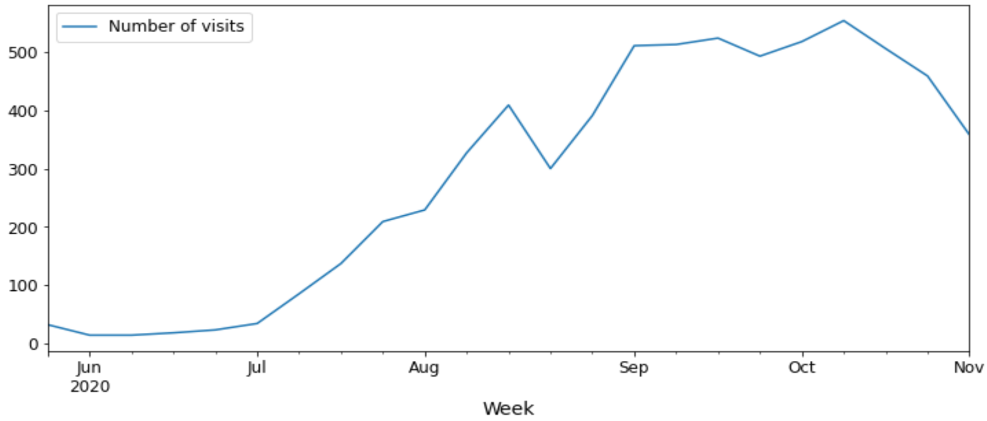


FIGURE 26: WEEKLY NUMBER OF VISITS TO THE IMPACT ON TOURISM INDUSTRY STORY

Q2.5 How high was the demand for COVID-19 datasets on the portal?

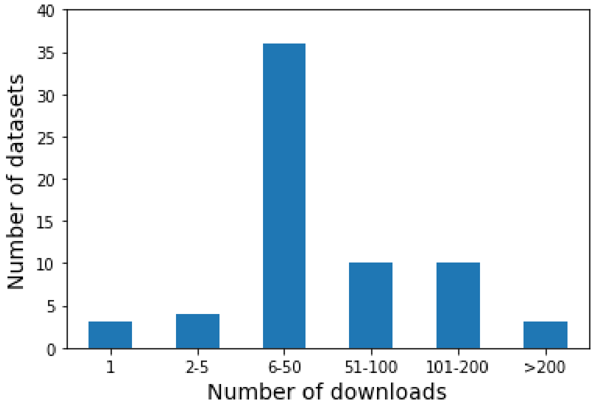


FIGURE 27: DISTRIBUTION OF NUMBER OF DOWNLOADS OF COVID-19 DATASETS

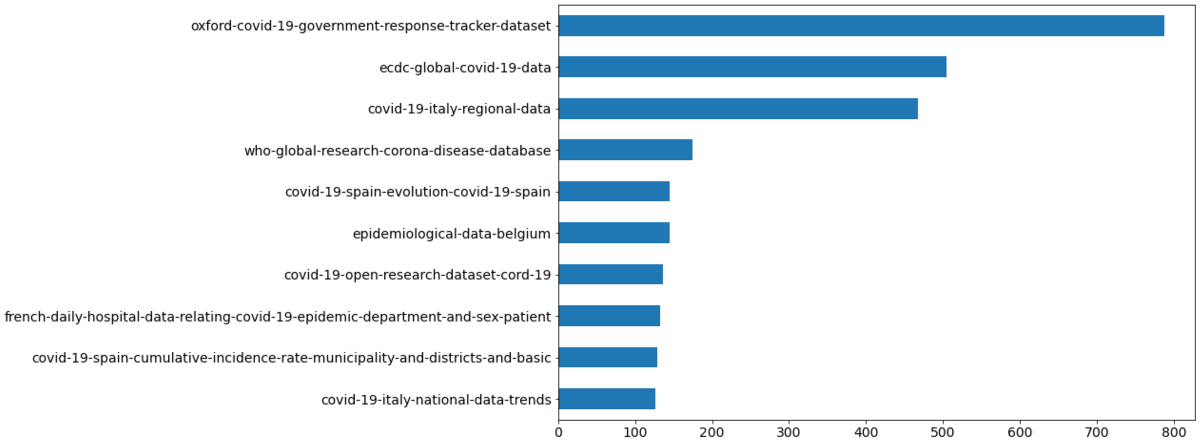


FIGURE 28: TOP-10 MOST DOWNLOADED COVID-19 DATASETS

Figure 29 shows the distribution of the number of downloads of COVID-19 datasets. All COVID-19 datasets were downloaded at least once. Figure 30 shows the top-10 COVID-19 datasets by number of downloads. We took a closer look at the 3 datasets that had more than 200 downloads.

1. The top dataset is the **Oxford COVID-19 government response tracker**, a resource that documents government responses to COVID-19 in different countries across the globe, enabling users to review, analyse, and compare what the governments' responses to the pandemic were and how these have evolved over the full period of the disease's spread. **This tracker was also featured in one story**. This dataset had 788 downloads - when compared with the most downloaded datasets of the whole EDP since March 2018, **this dataset is the second most downloaded in the whole analysis, despite the fact that it was only available since April 2020**. 75% of the visits to this dataset page came from search engines, captured search keywords were combinations of the words *"oxford" "covid-19" "tracker"* and *"government"*. 20% of the visits to this dataset page came from internal pages, of which 68% came from the story about it. A notable feature of this dataset is that it is not published by any portal harvested by the EDP but included manually by the editorial team of the COVID-19 section. Further on, the queries suggest knowledge of the dataset publisher; the users were searching for this particular tracker released by Oxford, searched for it online and landed on the EDP.
2. The second dataset is the **worldwide COVID-19 situation** published by the European Centre for Disease Prevention and Control (ECDC). This dataset was also listed in the regular dataset section, and **we identified it as the most downloaded one during the "new normal" period** (Section 3.1.2). On the

COVID-19 section, it has 505 downloads[23]. Adding those downloads to the ones from the regular dataset section, **this dataset is the all-time most downloaded dataset**. Contrary to the Oxford tracker and the non-COVID-19 datasets, visits to this dataset page mostly came from **the main page of the COVID-19 section and not from external search engines (65%-35%).**This is a search user journey where the user visits the COVID-19 content and then explores the relevant data, hence emphasizing the importance of curated content as a vector of impact for published datasets.

3. The third dataset is the **COVID-19 regional data from Italy**. The featured dataset is an aggregation from all provinces prepared by the department of Protezione Civile. It has 467 downloads. Compared with the non-COVID-19 datasets, **it would have been the 5th most downloaded (6th if we consider the Oxford Tracker datasets)**. **The dataset was not indexed by the Italian open data portal.** Similarly, to the ECDC dataset, **there was a larger number of visits coming from the main page of the COVID-19 section vs external search (55%-45%)**. We also noted that **there was only a marginal number of visits originating from Italy.**

4.2 Summary of Results

Q2.1 Was the EDP more used during the life of the COVID-19 section than before?

- Yes, by approximately 20%, and almost all the increase was due to the COVID-19 section.

Q2.2 What happened after the end of the COVID-19 section?

- Traffic reverted to previous numbers. However, this was due to a decrease across all sections, and not because users stopped visiting the COVID-19 section, which remained the most popular even after being archived. After the COVID-19 section was retired, the top-3 stories almost made up for the number of contributions of the datasets and initiatives subsections of the COVID-19 sections that were deleted. However, the number of users that reached it from the homepage after archival decreased by more than 90%.

Q2.3 How users reached the COVID-19 section?

- Mostly from search engines. The acquisition distribution is almost identical to the rest of the sections of the portal.

Q2.4 Did the publication of COVID-19 data stories increase traffic to the portal?

- Yes, five of the stories proved very popular, attracting more than 5000 visits, with the one on education reaching more than 14000 visits. Popular stories did not have immediate success. However, they addresses relevant topical issues that people search for online, they became popular over time, as web search engines pointed people at them.

Q2.5 How high was the demand for COVID-19 datasets?

- In terms of number of downloads, it was higher than for non-COVID 19 datasets. The most downloaded COVID-19 datasets are in the top-10 overall most downloaded datasets despite the fact that they were only available for a short period of time.

Conclusions and Recommendations

Dataset demand is relatively low. Our results suggest that only a small fraction of the indexed datasets is being downloaded, and therefore, potentially used. Furthermore, COVID-19 cases apart, demand appears to be not spread out in time, with very few episodical events and without periodical events. We were only able to link one news event with a download up-peak of a dataset. **However, it would be beneficial to cross-reference these findings with data from national portals, as we believe that as search engines become more effective in identifying the original sources of datasets, national portals would see more traffic for specific datasets.** We

[23] This dataset had one page the regular dataset section and one page in the covid-19 section, thus, the two different download counts.

would also expect that in case of events of regional or national importance, queries on search engines would be issued in languages other than English. As discussed in AR-18, there is a prevalence of English queries on the EDP.

For users who reach the portal directly, placement is important. The COVID-19 section was much less visited after its archival.

Most demanded datasets are "examples of". 3 out of the 10 most popular datasets appear to have been reached by users who were looking for a dataset of a certain type or with certain property (credit card transactions, license plates, invoices). They were unlikely looking for information about license plates or invoices, they just needed a dataset on that subject to work on.

Curated content was successful, and stories remained successful after archival. The COVID-19 section proved to be a success, becoming the most popular section of the portal in these difficult times. During the lifetime of the section, stories, highlights and datasets were approximately equally popular. After the archival of the section, three stories proved to be immensely popular.

Following these conclusions, we are able to issue three core recommendations for the EDP and other open data portals:

1. Prepare curated content about datasets of interest and highlight them on a dedicated section of your site. For example, a monthly article that uses several datasets to tell a data story about a current topic. Alternatively, data from external searches could be analysed to understand topics that may be of interest to portal users and drive content production.
2. Our qualitative analysis of keywords and facets associated with search sessions (for popular datasets) hints at a set of user journeys with varying information needs and requirements in terms of search experience. To the best of our knowledge this is the first time such journeys are proposed, and we hope portal designers will find them useful in understanding their users and improving their user experience. In AR-18 we concluded that user journeys that start on the portal are likely to have very different information needs to the ones that land on the portal from an external site e.g. a search engine. The examples of journeys we provide here are complementary to this basic dichotomy; we believe they could also be considered when studying dataset reuse and impact, as publishers might prioritise some types of users and use cases over others.
3. National data portals should perform similar analyses to the one carried out by the EDP in AR-18 and -19 to confirm or challenge our findings. This would also help discover popular categories of datasets across several European countries, informing decisions on what content to curate or highlight at the EDP level.