A comparison of dataset search behaviour of internal versus search engine referred sessions

Luis-Daniel Ibáñez
University of Southampton
Southampton, UK
l.d.ibanez@southampton.ac.uk

Elena Simperl
King's College London
London, UK
elena.simperl@kcl.ac.uk

ABSTRACT

Dataset discovery is a first step for data-centric tasks, from data storytelling to labelling for supervised machine learning. Previous qualitative research suggests that people use two types of search affordances to find the data they need: they either go to a data portal that probably contains the data and search there; or they start on a regular web search engine, which sometimes returns results that are datasets. For the first type of search, prior works have analysed logs from different data portals to understand basic tenets of search behaviour such as query length or topics. In this paper, we advance the state of the art in dataset search behaviour with a comprehensive transaction log analysis study (n = 236441 sessions) of an international open data portal, in which we compare sessions straight on a data portal (internal searches) against sessions that land on a dataset or SERP (search engine result page) through a referral from a web search engine (external). Using dataset downloads as a proxy for successful searches, we find a statistically significant, though weak relationship between the use of keyword search and session type and between the use of search facets and session type (moderate). We also discover and discuss behavioural patterns and user profiles across session types.

CCS CONCEPTS

• **Information systems** → **Environment-specific retrieval**; *Search interfaces.*

KEYWORDS

dataset search, information seeking, log analysis, search behaviour

ACM Reference Format:
Luis-Daniel Ibáñez and Elena Simperl. 2022. A comparison of dataset search behaviour of internal versus search engine referred sessions. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3498366.3505821

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHIIR '22, March 14–18, 2022, Regensburg, Germany
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9186-3/22/03…$15.00
https://doi.org/10.1145/3498366.3505821

1 INTRODUCTION

Data has become a fundamental resource to improve organisational processes, inform decisions, and train AI algorithms. For many data-centric tasks, the first step is *data discovery*, a term which refers to all activities around finding, making sense, and evaluating data for reuse [4].

When looking for datasets, potential users tend to follow several strategies [16]: (i) they type keywords in a web search engine, which may lead them to an online data repository; (ii) they have an idea where the data might be located and use the search affordances of that site to find datasets matching their needs; (iii) they ask other people for suggestions; or (iv) where the publisher is a public authority, they issue a freedom of information or a data request to that authority. Data discovery technology is heavily reliant on the availability of metadata, which records key information about datasets, for instance the authors, their affiliation, the domain, important keywords etc [15]. Tools such as Google Dataset Search (GDS) promote the use of standardised metadata vocabularies, such as DCAT and schema.org, to improve how they match people's information needs to datasets.

Previous work in understanding user search behaviour in data discovery has compared strategies of type (ii) to (iv), using a mix of descriptive and qualitative methods on corpora with aggregated search sessions and data requests [12]; as well as strategies of type (ii) across different verticals, for instance, document against dataset search in the context of a digital library with both types of artifacts [3].

In this paper, we go a step further and present a transaction log analysis to compare sessions representative of strategies of types (i) and (ii). We tackle the following research questions:

(1) What are the behavioural patterns of dataset search users? Do these patterns change depending on the type of session?
(2) Are users more successful using one strategy or another? Is there a statistically significant relationship between strategy type and success?
(3) Do users use more keywords and facets when they start on the portal or when they land on the portal from a web search engine? Is there a statistically significant relationship between strategy type and the use of facets and keywords?
(4) What are the user profiles of dataset search users?

We use a transaction log of 236441 sessions over a one-year period from the European Data Portal (EDP, meanwhile re-named to data.europa.eu).[1] This is a portal for openly available government data. It aggregates and curates metadata of 1.4 million such datasets, held in over 80 national and regional repositories from 36 countries.

[1] https://data.europa.eu/en

On top of the metadata records, the EDP implements keyword-based and faceted search. The log in our analysis includes the *referrer* field for all sessions, allowing us to separate sessions that land on an EDP dataset or dataset search-engine-result-page (SERP) from a web search engine (that we assume to be type (i)); from those that use the native EDP search (that we assume to be type (ii)).

Working with data has become central to many professions. Data portals, whether openly available on the web or internal to an organisation, are the de facto means to provide access to datasets. They use metadata about these datasets to facilitate search and browsing. Our study helps managers of data portals, but also providers of portal technology understand data user behaviour, enhance search experience, and improve the performance of discovery tools and algorithms.

2 RELATED WORK

Search log analysis is an established method to understand information needs and user behaviour in information retrieval. There are many use cases, including: monitoring and predicting user satisfaction, enhancing user experience through contextualisation and personalisation; and improving system performance [11, 23]. Recent studies used logs to e.g. extract user interests in digital libraries [1, 10]; detect email search patterns [20]; or understand how users utilise visual and image queries on e-commerce sites [5].

In the context of dataset search, [12] analysed search behaviour on four open data portals (three national government data portals in the UK, Canada and Australia, and the UK's Office for National Statistics portal), building on their initial study from two years' earlier [13]. Their work drew from multiple sources: interviews with data practitioners, data requests, and search logs. For the latter, they reported descriptive statistics pertaining to query characteristics such as length, structure, topics, and types. The authors did not have access to click-through data, which meant that their session-level analysis was limited to attributes such as search duration, time of day, number of pages viewed, device and browser used etc. They suggested that dataset search is a work-related activity, following from the time distribution of the queries, with shorter queries that are topically and structurally different to those issued in general web search. Compared to this work, our research answers a broader set of questions using methods of analysis that go beyond descriptive statistics; in addition, the EDP search logs cover a much higher number of data portals in different countries.

[3] analysed search sessions from a digital library for social sciences that stores documents and datasets. They extracted and compared document search sessions (those that look only for documents), dataset search sessions (those that look only for datasets), as well as mixed sessions. In line with [12] and [13], they found that dataset search queries are shorter than document queries, with a higher prevalence for numeric values, and similar interaction paths.

[27] analysed a six-month log (105k sessions) from three US city-level open government data portals. The authors looked at the number of sessions that reached each portal directly or through a referral; this matches our search types (i) and (ii) in Section 1, though they distinguish between sessions that include the use of a web search engine and other web sites that link to data on one of the portals. They studied the link between type of session and success, using downloads as proxies for success. Further on, they reported the share of high-level behaviours across the two types of sessions: only browsing, only keyword search, or both. Compared to [12], they found their portals to have a higher proportion of direct accesses than referred by search engine; for one of the portals, these internal search sessions seemed also to be more successful than the ones starting on a general-purpose search engine. While their research questions are not that different than ours, the scale of their study and methods of analysis are: (i) in addition to descriptive statistics, they resort to web content mining to topically categorise datasets; we have access to such information in our corpus, as the EDP already categorises datasets across all 80 portals uniformly into a set of XXX categories used for faceted search. (ii) by contrast, we build maximal sequential patterns (see Section 3) to detect more granular search patterns; establish statistically significant links between session types and downloads; and use clustering to identify seven types of user profiles. We discuss how the findings of their and our work compare in Section 5.

3 METHODOLOGY

3.1 The European Data Portal

The European Data Portal (EDP) offers access to more than 1.4 million open government datasets in many domains. These datasets have been originally published on various data portals managed by public sector organisations at different administrative levels. Each of them has a metadata record; the metadata is harvested by the EDP and create an index for data discovery.

The portal has four main parts: (1) News, events, and highlights from the open data community; (2) eLearning courses about open data and related technologies; (3) Analytical reports and thought leadership pieces on topics such as open data impact, business models for open data publishing, open data and digital inclusion etc. (4) A list of organizations publishing open data harvested by the EDP; and (5) Dataset search engine.

From February 2017 to April 2019, the EDP search engine was based on CKAN,[2] a popular open-source software for data portals; from May 2019, they switched to a custom-made system based on linked data and semantic technologies [14]. As of February 2021 the portal has been renamed to https://data.europa.eu portal, but the underlying data discovery capabilities remained the same as the version from May 2019.

A dataset has an EDP page that displays its metadata (description, date of harvesting, last date modified) and the list of distributions. These refer to how the dataset can be accessed, for instance as files for download or via an API. A dataset may have multiple distributions, for instance the publisher may decide to support files in different formats or split the data into files according to geospatial or time attributes.

The dataset search engine interface is comprised of a keyword search box and the following facet filters: (1) **Catalog**: Data portal from which the dataset was harvested, *e.g.*, Czech Geoportal, Trentino region open data portal. The notion of a catalog is similar to the organisations mentioned earlier when we talked about the main structure of the EDP site. (2) **Countries**: Provenance of the

[2] https://ckan.org

dataset, including, but not restricted to EU Member States, as well as the EU as an data publisher in its own right. (3) **Tags**: provided by the data publisher or automatically extracted from the dataset description provided by the publisher. (4) **Format**: Format of the dataset, *e.g*, xls, csv, pdf etc. (5) **Location**: Approximate geographic area covered or referred to by the dataset. Users control location by drawing a rectangular area on top of a minimap to approximate the location of interest, or provide the name of the location (e.g. city, region etc).

When a user inputs a query in the keyword search box or clicks on a facet, the system returns a paginated list of datasets (n=10). Users can sort Search Result Pages (SERPs) by name and last modified date.

When inspecting a dataset page, users can directly download a distribution, or inspect the distribution page, which displays more detailed metadata. Alternatively they can show similar datasets, show the categories a dataset belongs to; or show a list of dataset activities e.g. downloads, updates etc.

3.2 The EDP search logs

For this study we started from a corpus consisting of all searches from April 2018 to April 2019, collected using the web analytics tool Matomo[3]. We chose this slice because as noted earlier, at that point the EDP team swapped the original CKAN-based backend against a new, bespoke implementation. Among others, this included a new URL scheme for the pages, and a slightly different set of search affordances. The corpus did not include any IP addresses or other digital identifiers. Matomo implements session identification based on a configurable timeout parameter T: if an user returns to the portal within T, their actions are recorded as part of the same session; otherwise, the activity will considered a new session. For our corpus, T was set to 30 minutes. We did not have access to raw HTTP logs to implement a different session identification approach.

For the 12 months worth of logs, each session is identified by an unique id, and includes its duration in seconds and a list of actions of the following types: (1) **Page visit**: a page of the portal was loaded in the user browser. This includes clicking on facet filters. (2) **Outlink click**: the user clicked on a link to a non-portal page (3) **Download file**: The user downloaded a file hosted on the portal (4) **Dataset search**: The user asked a query using the search box.

Matomo was configured with a default parameter that does not allow the measurement of time spent on the last page of a session. This means that the recorded duration of a session is a lower bound of the real-time spent by the user. For this reason, we decided not to use duration in the rest of the study. We acknowledge this as a limitation, as we were not able to use dwell time either as a signal of success or as a feature for clustering.

3.3 Processing the logs

We selected sessions that either visited at least one dataset page or used at least one of the dataset search capabilities (keyword search or facet filtering). Furthermore, we partitioned these sessions in three *session type* subsets:

[3]https://matomo.org/

Table 1: Descriptive statistics of the corpus

Characteristic	Internal	External SERP	External Dataset	Combined
Number of sessions	33323	77303	125815	236441
Average length	11.84	4.89	2.95	4.78
Median length	7	3	1	2
Bounces	0	30817	67245	134996
Average duration (s)	506.9	267.18	204.4	298.3
Median duration (s)	165	91	61	88

- **Internal dataset search sessions** are sessions that start on an EDP page that is not dataset related (e.g., the homepage), then use the dataset search interface. We assume these sessions correspond to users who start their search on the EDP.
- **External SERP sessions** are sessions that start with a dataset search result page, referred from a web search engine. We assume these sessions correspond to users who use web search first, then click on a **result page** from the EDP.
- **External dataset page sessions** start with a dataset or distribution page, referred from a web search engine. We assume these sessions correspond to users who use a web search engine and then click on a **dataset or distribution page** from the EDP.

We undertook an exploratory data analysis and we discovered that 12801 external dataset page sessions were referred by Google Dataset Search (GDS) [2]. We decided to discard them fro two reasons: first, because more than 60% of them arrived within the first week of the launch of GDS. We believe most of these sessions were from people trying out the service due to its novelty; second, we are interested in the interaction between data portals and web search engines. We defer the study of the interaction between dataset search portals at different levels, e.g, GDS and EDP, or EDP and national portals as future work.

3.4 Analysing the logs

In our analyses, we considered downloads and clicks on dataset source links as proxies for successful searches.

To detect patterns in search sessions we used **alphabet encoding** [7]. According to their approach, actions in a session are encoded as symbols of an alphabet and the resulting sequences are fed into a pattern mining algorithm. In our case, we first built an alphabet with 32 possible actions supported by the EDP, which are detailed in Table 2. Note that for the sake of readability, we grouped each of the five facets described in Section 3.1 in rows $AF - X$ and $RF - X$, respectively.

As discussed in [7], the size of the alphabet directly impacts the quality of the patterns mined from the logs. Longer alphabets could lead to a more diverse range of patterns, which occur less frequently. To prevent this, related actions should be aggregated meaningfully and encoded with the same symbol. For example, all actions adding a facet, no matter what the facet is, are encoded in our case as AF rather than having one symbol per facet addition. Conversely, if the symbols are too coarse, the mining algorithm

Table 2: Full alphabet encoding of user actions (n=32)

Symbol	Description
HH	EDP homepage
DH	Dataset section homepage
DP	Dataset page
RP	Dataset distribution page
DO	Click on dataset download or source link
SQ	Issue search keyword query
AF-X	Add facet X X ∈ [Categories,countries,tags,catalog, location,format]
RF-X	Remove facet X X ∈ [Categories,countries,tags,catalog, location,format]
AS-X	Add sort by condition X X ∈ [metadata_modified,name]
RS-X	Remove sort by condition X X ∈ [metadata_modified,name]
NRP	Next SERP page
BRP	Previous SERP page
NDP	Non-dataset section page
OH	Organisation homepage
OP	Organisation page
OF-X	Add organisation section facet X X ∈ [Countries, location]
OR-X	Remove organisation section facet X X ∈ [Countries, location]
OO	Non-dataset outlink or download
SS	Show similar datasets
SC	Show categories a dataset belongs to
SA	Show dataset activity e.g. updates

might return only a few patterns that occur frequently and miss more interesting, nuanced patterns.

To make a principled decision about how to cluster actions to manage the size of the alphabet, we applied **maximal sequential patterns (MSP)**, a technique used for exploratory analysis of log sessions [18].

Definition 3.1. A **sequence** $S = [A_1, ..., A_i, , A_n]$ is an ordered list of actions. We denote i as the index of the sequence. We call a set of sequences D a *sequence dataset*

Definition 3.2. A sequence $P = [A_1, ...A_n]$ is a **subsequence** of sequence $S = [B1, ..., B_n]$ iff there exists integers $1 \leq i < ...j < n$ such that $A_1 = B_i, ...A_n = B_j$. Conversely, S is a **supersequence** of P

Definition 3.3. A sequence $P = [A_1, ...A_n]$ is a **contiguous subsequence** of $S = [B1, ..., B_n]$ iff P is a subsequence of S and the integers $1 \leq i < ...j < n$ are contiguous. Contiguous subsequences are also known as *n-grams* in natural language processing.

For example, consider the sequence $S = [X, SQ, AF, NP, DP, RP, DO]$. $P = [SQ, AF, DO]$ is a subsequence of S, and $Q = [SQ, AF, NP]$ a contiguous subsequence of S.

Definition 3.4. A **sequential pattern** P is a subsequence of one or more sequences in a sequence dataset. The set of sequences d for which P is a sequential pattern is called the *support set*. The *support* of P is the ratio $\frac{|d|}{|D|}$, in other words, the percentage of sequences in D where pattern P appears.

Definition 3.5. Given a minimum support m, a sequential pattern P is **maximal** if there is no sequential pattern P' that is supersequence of P and has support greater than or equal to m.

Similar to [18] we used the Vertical Mining of Maximal Sequential Patterns (VMSP) algorithm implemented in the Sequential Pattern Mining Framework 2 (SPMF2) [6] to compute for each of the three session type subsets the MSPs with a support greater than 10%. We considered symbols that were not found in the MSPs as candidates for being grouped together in a coarse partition. For example, if the **Add facet 'location'** and the **Add facet 'format'** actions do not appear in the MSPs, that means that they appear in less than 10% of the sessions in our corpus and should be grouped together in to the same **Add facet** symbol in the alphabet. Additionally, MSPs gave us insight into the most popular relative ordering of actions. We applied VMSP on the subset of sessions sets excluding bounces.

After reducing the size of the alphabet, we re-encoded all sessions according to it and computed the frequency of the first symbol of the session. We then computed their contiguous subsequences using the n-grams functionality of the NLTK package in Python. NLTK is designed for text analysis, so we appended to each sequence a separator symbol $ and joined them all in a single string, to which we applied the *everygrams* function to generate subsequences from length 3 to 7. We chose the lower bound 3 following from [7]. They did not specify the upper bound they used, so we chose the maximum median length among all session types, 7. From the output, we filtered out the subsequences containing the $ symbol.

In our analysis, we determined for each session type the most common patterns and for each pattern the percentage of sessions that include it, the median length of the sessions that include it, and the percentage of successful sessions that include it.

To answer RQ2 and RQ3 we conduct three Chi-Square tests of independence with level of significance of 0.05. To control Type-I errors, we applied Bonferroni correction, and rejected the null hypothesis with a p-value $< 0.05/3 = 0.01$. Each test has 3 rows, one for each type of session, and 2 columns, hence, each test has 2 degrees of freedom and a critical value of 9.21. We describe for each test the hypotheses and columns below.

(1) $H1_{null}$: Search success is independent of session type. $H1_{alt}$: Search success is dependent of session type. Two columns: session is successful/unsuccessful.
(2) $H2_{null}$: The use of facets is independent of session type. $H2_{alt}$: The use of facets is dependent of session type. Two columns: session contains at least one facet action, session does not contain a facet action
(3) $H3_{null}$: The use of keyword queries is independent of session type. $H3_{alt}$: The use of keyword queries is dependent of session type. Two columns: session contains at least one keyword search action, session does not contain a keyword search action

Table 3: Session features for clustering

1	ratio keyword searches
2	ratio usage of country facet
3	ratio usage of tags facet
4	ratio usage of category facet
5	ratio usage other facets
6	ratio visited dataset pages
7	ratio visited non-dataset pages
8	ratio visited distribution pages
9	ratio visited EDP home or dataset home pages
10	ratio next SERPs clicked
11	ratio back SERPs clicked
12	ratio facets removed

We use Cramer's v as measure of association strength. We included bounce sessions in the population. We use the SciPy implementation of Chi Square to compute the tests [26].

To mine user profiles we cluster sessions according to the 12 session features from Table 3. We chose to use features that characterise the affordances of the portal: the search actions (keywords and facets) and actions for inspecting results (SERPs, dataset pages, distribution pages and home pages visited), leaving deliberately out the number of downloads, which are our success proxy. We defer to future work the inclusion of downloads as a feature, which would allow us to characterise successful sessions rather than distinguish between different search strategies. Following from the results of our MSP analysis, we decided to group the a actions Add format facet, Add geo-location facet and Add catalog facet into a single 'Add other facet' feature; and all facet removals into a single 'Remove Facet' feature. We consider bounce sessions a separate cluster, therefore, we do not input them to the clustering algorithm.

We use the k-medoids algorithm with Euclidean distance implemented in the PyClustering library[21]. To choose the number of clusters k, we applied the silhouette method in a similar way as [1], which measures the separation between the clusters with values ranging from −1 to 1, the higher values indicating a better clustering. For each session type, we computed clusters with values of k ranging from 3 to 10 and selected the k with highest average silhouette width up to two digits of precision. If two values of k were within 0.01 silhouette width we chose the smaller.

4 RESULTS

4.1 Maximal Sequential Patterns

The internal session type has 18 MSPs with at least 10% of support. The 14 actions involved are DH, HH, DP, RP, DO, SQ, NRP, BP, OO, $AF-Category$, $AF-Country$, NDP(see Table 2 for the description of the codes). Some of the MSPs we found were singletons. Others were probably a function of the EDP search experience rather than indicative of a distinct user behaviour. For instance, the pattern [DP,RP,DO] can be mapped to a visit of the dataset page, followed by a visit of the distribution page, followed by a download of the distribution. We highlight the following MSPs :

(1) $[HH, AF-Category, AF-Country]$ (12.8% support) indicates that the country facet is often applied after the category facet.
(2) $[NDP, NDP]$ (14.6% support) is for sessions that visit more than one non-dataset pages on the EDP.
(3) $[AF-Category, HH]$ (10% support) is for sessions where a category facet is used, followed by a return to the main page.
(4) $[NDP, DH]$ (10% support) is for sessions that visited other sections of the portal, and then reached the dataset search home.

The external SERP session type has 10 MSPs with at least 10% of support. The 10 actions involved are SQ, DP, RP, DO, NRP, BRP, BP, $AF-Tags$, $AF-Country$, $RF-Country$. We highlight the following interesting MSPs:

(1) $[AF-Tags, DP]$ (39.8% support) for sessions using the tags facet prior to landing on a dataset page.
(2) $[AF-Tags, AF-Country]$ (19.3% support) for sessions where the country facet is applied after tags.
(3) $[RF-Country]$ (13.7% support) for sessions, in which users remove country facets to have a broader selection of similar datasets from different countries. As noted earlier, the EDP harvests data from 36 countries.

The external dataset session type has 6 MSPs. The 5 actions involved are DP, RP, DO, BP, $AF-Tags$. The only interesting pattern to highlight is the use of tags as a facet. On the other hand, we also note the absence of keyword search queries SQ compared to the other two session types.

Based on the observed MSPs we generated a reduced alphabet of 21 symbols (Table 4). We grouped together all actions pertaining to the organisation section of the EDP site. We did the same for all sort and show actions, and for the additions and removals of the catalog, location, and format facets.

4.2 Behavioural patterns

For each session type, we found that starting actions are strongly biased. 74% of internal sessions started in the EDP homepage or dataset home page, and a further 17.5% started on a Non Dataset Page. We believe this is due to users directly entering to the portal, or inputting navigational queries to web search engines.

83% of the external SERP sessions started with a tags facet, with no other facet having a frequency larger than 5%. This suggest that tags facet SERPs were EDP's best ranked pages for web search engines. For external dataset sessions, 97% started with a dataset page and 3% with a distribution page.

Table 5 compares the most frequent behavioural patterns (contiguous subsequences) extracted for each of the session types from Section 3.3. For completeness, we add the singleton patterns of Search Query (SQ), all facet additions, Non-Dataset-Page (NDP), and Download (DO).

We were only able to detect patterns in sessions with length larger than the median of the session type. This is due to the large number of short sessions in our dataset. For all session types, patterns appear at most in 12% of the relevant sessions. At the same time, patterns that are common in one session type are much less common in the other two session types. This suggests that there is a difference in behaviour among the session types. We highlight the following: The sequence $NDP; NDP; NDP$, that is, three EDP pages that are not directly related to datasets, appears 7 times more

Table 4: Reduced alphabet encoding of user actions after application of MSPs (N=21)

Symbol	Description
HH	EDP homepage
DH	Dataset section homepage
DP	Dataset page
RP	Dataset distribution page
DO	Click on dataset download or source link
OO	Non-dataset outlink or download
SQ	Issue search keyword query
AF-Tags	Add facet tags
RF-Tags	Remove facet tags
AF-Country	Add facet country
RF-Country	Remove facet country
AF-Category	Add facet category
RF-Category	Remove facet category
AF-X	Remove facet X X ∈ [catalog, location, format]
RF-X	Remove facet X X ∈ [catalog, location, format]
NRP	Next SERP page
BRP	Previous SERP page
NDP	Non-dataset section page
ORG	Organisation section actions (OH,OP,OF,OR)
SO	Sort actions (AS,RS)
SH	Show actions (SS,SC,SA)

frequently in internal sessions than in external sessions. This suggests users are less likely to engage with other sections of the portal when landing on a dataset page referred from a web search engine.

The sequence [Homepage;AF-Category;AF-Country] appears in 12% of the internal sessions, and in less than 0.5% of the internal sessions. The high frequency could be explained by the fact that the homepage is the most common starting point, and that it featured links to Category facets very prominently. By definition, external sessions do not start in the homepage, therefore, the low frequency of this pattern suggests that users from external sessions don't start a new search using facets, or even at all, as suggested by the absence of a pattern including HH or DH among the most common.

The [AF-Tags;DP;RP] sequence appears in almost 10% of external SERP sessions, and in 0.34% of internal sessions. The high frequency in external SERP sessions is explained by the fact that AF-Tags is the most common landing point for this session type, while for internal sessions Tags is the least used facet, and furthermore, users are very unlikely to inspect any of the presented results.

4.3 Relationship between session type and success, facet use and keyword use.

We explored the relationship between session types and other variables such as search success, the use of facets, and the use of keywords. We report the results of each of the statistical tests described in section 3.4. To help with interpretation, we describe in Table 6 the counts of each session belonging to each pair of categories of each test, together with the percentual diference wrt. expected counts according to the χ^2 test. We highlight that we observed a count of Internal sessions that use facets 182.9% higher than expected, and a count of sessions that use keywords 257% higher than expected. Conversely, observed use of facets for External Dataset sessions was 61.8% less than expected and use of keywords was 77% less than expected. In terms of success, differences are less dramatic, with the most important highlight being the less than expected success of External SERP sessions. Recall the purpose of the test is to evaluate is these differences between observed and expected are significant.

For the association between session type and success, we got a statistic of $838.34 > 9.21$, therefore, we rejected the hypothesis $H1_{null}$ that success is independent of session type. The Cramer's $v = 0.05$, suggesting the association is statistically significant, but weak.

For the association between session type and use of facets, we got a statistic of $48075.1 > 9.21$, therefore, we rejected the hypothesis $H2_{null}$ that use of facets is independent of session type. The Cramer's $v = 0.45$, suggesting the association is statistically significant and moderate.

For the association between session type and use of keyword queries, we got a statistic of $43385.5 > 9.21$, therefore, we rejected the hypothesis $H3_{null}$ that use of keywords is independent of session type. The Cramer's $v = 0.42$, suggesting the association is statistically significant and moderate.

4.4 User profiles

Tables 7, 8 and 9 show the statistical description of the clusters found for internal, external SERP and external dataset session types respectively. The first four rows of each table show the size of the cluster, the percentage of the sample that the cluster represents, the median length of the sessions in the cluster and the number of sections that include at least one download (i.e., successful). The remaining rows provide statistics about the percentage of sessions that include at least one action symbol, e.g., $\%SQ$ shows the percentage of sessions within the cluster that include at least one keyword query search action. We highlighted high indicator values in bold font over light cell shade, and low values in white font over dark cell shade.

For internal sessions our method yielded $k = 4$ (average silhouette width = 0.40). Cluster $C1$ presents a larger median length and success rate. There is also larger rate of Dataset Page (*DP*), Distribution Page (*RP*) and Next REsult Page (*NRP*) actions and we could not note a strong bias towards a specific search affordance (e.g. more keyword queries than facets). Note that the use of the tags facets across all the sessions is this type of search sessions is lower than the use of the 'other' facets combined. Due to this higher rate of visiting, we label this cluster *diggers*.

Cluster $C2$ main characteristic is that all sessions visited at least one non-dataset section page. Sessions have shorter length and four times less success rate than the *diggers*. They also have significantly lower rates of result and item inspection actions (*NRP*, *DP*, *RP*) On the other hand, they have a similar search affordance distribution. We believe users in this cluster come to the portal to explore all

Table 5: Comparison of behavioural patterns extracted from different session types, plus singletons, per session type in terms of (1) percentage of sessions per type that include the pattern, sessions' median length and percentage of successful sessions that include the pattern

Session type	Internal			External SERP			External dataset		
Pattern	%Sess.	Median len.	%Success	%Sess.	Median len.	%success	%sess.	Median len.	%success
From internal									
NDP NDP NDP	8.17	18	28.45	1.19	14	37.86	1.58	11	41.25
ORG ORG ORG	6.06	19	47.89	1.26	18	49.57	1.17	16.0	54.09
HH AF-Category AF-Country	12.87	9	34.67	0.43	15	47.26	0.29	14	52.07
SQ SQ SQ	5.55	14	35.66	3.01	12	36.21	0.87	15	51.67
NRP NRP NRP	6.57	21	48.9	3.84	17	43.75	0.85	21	60.4
From External SERP									
AF-Tags,DP,RP	0.34	18	78.57	9.83	7	79.37	1.07	10	66.83
AF-Tags DP BRP	0.34	23	56.25	8.77	5	21.89	1.65	7	25.65
AF-Tags AF-Country RF-Country	0.09	11	9.68	5.11	3	11.87	0.3	7.0	28.41
AF-Tags AF-Country RF-Tags	0.1	14	56.25	3.03	7	26.35	0.11	13	46.03
From External DP									
DP RP DO	10.29	19	100	13.68	9	100	16.98	5.0	100
DP AF-Tags DP	0.86	19.0	60.49	2	11	48.33	8.03	5.0	26.73
DP DP DP	2.84	24	70.48	2.57	13	68.79	3.1	5	32.95
Singletons									
NDP	37.85	10	29.61	8.75	10	44.63	9.75	6	41.21
SQ	45.45	9	31.08	24.59	8	33.97	6.18	12	51.4
AF-Tags	5.15	13	43.21	85.22	5	35.26	15.3	5	30.43
AF-Category	48.87	7.0	25.56	6.24	11	36.64	1.93	12	46.19
AF-Country	29.51	10.0	34.87	32.27	6	27.53	3.72	11.0	46.74
AF-X	10.77	14.0	43.09	20.66	7	33.49	1.99	12.0	49.06
DO	26.58	15.0	100	34.95	8	100	56.86	4.0	100

Table 6: Session counts of each type and category and percentage of difference wrt. expected counts according to χ^2 test

	Successful	Unsuccessful	Facet	No Facet	Keyword	No Keyword
Internal	8858 (+7.6%)	24465 (-2.5%)	21439 (+182.9%)	11884 (-53.8%)	15145 (+257%)	18178 (-37%)
External SERP	16249 (-14.9%)	61054 (+4.8%)	21412 (+21.8%)	55801 (-6%)	11321 (+15.1%)	65982 (-2.2)
External Dataset	33304 (+7.15%)	92511 (-2.3%)	10911 (-61.8%)	114904 (+18.2%)	3617 (-77%)	122198 (+11.28%)

features, either as primary goal, or because after fulfilling their initial information need in one section, they explore others. We label this cluster the *section switchers*.

Clusters $C3$ and $C4$ tend to have short sessions with low success rate. The difference is that $C3$ has a significantly larger use of keyword queries than facets, while $C4$ is the opposite. Apart from that, we observe that sessions in $C3$ visit at least one extra result page six times more often than sessions in $C4$. Based on this, we call $C3$ *keyword-and-done* and $C4$ *facet-and-done*.

For external SERP sessions, our method yielded $k = 6$ with silhouette width = 0.35.

The $C1$ cluster has longer sessions with a high success rate. This cluster has the second highest rate of dataset pages, next result pages and home pages. It has the highest rate of distribution page visit and non-dataset page. We don't observe a noticeable bias in terms of search affordances. Due to the similarities with the first cluster from the internal sessions, we use the same label *diggers*.

The $C2$ and $C3$ clusters are quite similar, though $C3$ shows a slightly higher search affordance usage rate. The main difference is going back to a search result action (*BRP*) rate, very low for $C2$ and very high for $C3$, which is probably what led our method to suggest two clusters instead of one. Upon inspection of the sessions, we believe the difference is not relevant in practice; given their low rate of search affordances compared to the other clusters, we merge them into a cluster with the label *land-inspect-done*.

Cluster $C4$ has the highest usage of search queries (*SQ*) (70%), click on additional search result lists (*NRP*) (almost 80%) and home visits (*HH* or *DH*) (26%). Compared to $C1$ in the same session type, it has a higher rate of facets' usage but a lower rate of dataset and distribution page visits. It also has a significantly lower rate of success (18% vs 71%). We believe users in these sessions are not satisfied with the results they are getting. We refer to this cluster as the *unsuccessful diggers*.

Clusters $C5$ and $C6$ have very short sessions, with a very low success rate. There are two noticeable differences among them: $C5$

Table 7: Clusters found in internal search sessions. Bold font on light shade indicates a high value. White font on dark shade indicates a low value.

cluster label	C1	C2	C3	C4
size	13096	6105	5466	8656
perc	39.3	18.32	16.4	25.97
median length	**12**	8	4	4
%DO	**55.93**	13.46	6.38	4.18
%SQ	51.71	33.97	**100**	9.62
%AF-country	37.42	27.98	7.01	**32.82**
%AF-category	44.22	35.45	19.61	**83.84**
%AF-tags	9.04	4.16	1.21	2.44
%AF-X	17.04	10.04	2.69	6.90
%RF-X	5.04	1.20	0.24	1.26
%NRP	**44.65**	14.43	**32.56**	5.24
%BRP	**47.22**	19.33	11.21	14.24
%DP	**85.92**	42.39	21.41	19.71
%RP	**41.17**	9.03	2.25	2.09
%Homes	79.93	79.97	96.76	99.97
%NDP	34.02	**100**	13.19	15.38

Table 8: Clusters found in external SERP sessions

cluster label	C1	C2	C3	C4	C5	C6
size	15929	6002	6737	7127	7099	3592
perc	34.26	12.91	14.49	15.27	15.33	7.72
median len.	**8**	3	4	**7**	3	2
%DO	**71.76**	30.84	22	18.66	2.03	0.31
%SQ	21.86	2.97	7.44	**70**	6.42	**50.84**
%AF-country	26.43	6.11	10.14	40.61	**93.38**	6.01
%AF-category	8.11	0.28	1.87	11.74	7.47	2.76
%AF-tags	78.18	**99.92**	90	75.57	86.34	99.83
%AF-X	27.26	2.18	9.04	29.31	25.96	16.43
%RF-X	5.45	0.13	1.31	8.36	13.44	0.67
%NRP	15.10	2.85	9.28	**79.54**	2.79	1.20
%BRP	46.57	5.30	**100**	47.48	21.09	17.85
%DP	93.31	**100**	79.68	37.95	9.56	2
%RP	**59.48**	7.53	5.49	9.44	0.77	0.56
%Homes	16.52	1.65	1.84	**26.70**	6.94	13.56
%NDP	**17.90**	2.43	3.95	7.13	1.90	4.45

has the highest usage of country facets among all clusters while $C6$ has the lowest; $C6$ has the second highest usage rate of keyword queries, while $C5$ has the lowest. Upon inspection of the data, we noticed that for both clusters the use of search query and country facets happens after landing on a SERP result. This means that most users in these sessions follow the pattern: land on a tag facet, click on a country facet ($C5$) or make a keyword query without further SERP inspection ($C6$), leave. Because of the short session lengths and low success rates, we label this cluster as the *two-step bouncers*.

For external dataset sessions, out method yielded $k = 7$ with silhouette width = 0.56.

Clusters $C1$ and $C2$ are almost identical except that $C2$ has a very high rate of sessions including at least one visit to a distribution page (*RP*) and a slightly lower success rate (79% vs 67%). Manual inspection revealed that both sessions land on a dataset page, and the difference is simply that some users clicked on one or more distribution pages. As such, we believe these clusters should be merged into one. Due to the high rate of success, we assign them the label *land-and-grab*.

Cluster $C3$ shows a relatively large use of further search affordances after landing. Similar to what we did with the previous two session types, we label this group as the *diggers*.

Cluster $C4$ has a high rate of sessions including at least one distribution page (*RP*) and a relatively low rate of sessions including at least one dataset page (*DP*), but otherwise is quite similar to the *land-and-grab* cluster from before. Further inspection revealed that in this cluster, 40% of landings are on a distribution page, from where users move to the parent dataset page. As such, we believe this cluster also needs to be merged with $C1$ and $C2$. Future studies should compare behaviours depending on when a user lands on a dataset page or a distribution page.

Cluster $C5$ uses only tags facets. Note that across all other clusters except $C3$, further usage of search affordances is very low. However, the short median length suggest that users (after 11% of them downloaded the dataset they landed on) click once on a facet tag and then leave. We believe this is related to the fact that at the time period we analysed, links to the tags of the dataset leading to the corresponding tag SERP were available on dataset pages. We assign to this cluster the label *land-tag-leave*.

Cluster $C6$ is for all sessions visiting at least one non-dataset section page - that is, users who after landing on a dataset, proceeded to another section. However, the low median length (only 3 actions) suggests that there is only a single Non-Dataset section page (*NDP*). We believe this might be users that simply click on another part of the portal before leaving. Note that barring *NDP*, this cluster is very similar to $C1$. Further studies should further distinguish between sessions that simply 'click before leaving' (that could be added to $C1$) and those that pay a meaningful visit to another section (that could be considered *section switchers* as cluster $C2$ of the internal session type).

Finally, cluster $C7$ has all values very low except visits to dataset pages (*DP*). Further analysis revealed that more than 90% of these sessions are comprised of two visits to the same dataset page. We believe this is simply a user reloading or opening the same page in a second tab. As such we consider this cluster as a *bounce*.

5 SUMMARY AND DISCUSSION

In this section we summarise our main findings and discuss their implications for dataset portal management and design and outline hypotheses for further studies.

Behavioural patterns are short, and differ according to session type. We found differences in the behavioural patterns most prevalent for each session type, owing to the starting point of each session. Internal sessions often started from the EDP homepage and followed with a category facet from a selection that was prominently featured on the page at the time. By comparison, external sessions rarely used any categories whatsoever. However, external SERP sessions often started with a tags SERP, while this facet was the least used for internal sessions. Some of these findings are due to how the EDP

Table 9: Clusters found in external dataset sessions

	C1	C2	C3	C4	C5	C6	C7
size	16026	15868	9511	5204	5082	2297	4582
perc	27.09	27.09	16.23	8.88	8.67	3.92	7.82
median len.	2	4	**6**	5	3	3	2
%DO	**79**	66.98	56.15	69.41	11.33	18.46	1.35
%SQ	1.26	1.35	**31.11**	0.75	2.18	4.01	0
%AF-ctry	0.81	0.81	**17.39**	0.46	3.74	2.35	0
%AF-categ	0.23	0.37	**10.10**	0.37	0.24	1.83	0
%AF-tags	5.88	5.44	19.87	1.35	**100**	4.09	0.41
%AF-X	0.62	0.65	**8.73**	0.23	2.05	0.78	0
%RF-X	0.16	0.16	2.82	0	0.87	0.17	0
%NRP	0.51	0.60	**17.93**	0.37	1.10	1.18	0.02
%BRP	3.78	2	20.51	0.29	15.19	2.22	0.02
%DP	100	99.91	95.27	64.87	100	95.08	**100**
%RP	4.98	**100**	20.50	**100**	1.99	10.40	0.07
%Homes	4.16	2.50	**32.94**	2.98	1.87	11.49	0.22
%NDP	3.39	5.41	17.12	6.11	1.16	**100**	*0.17*

implements search, and what web search engines decide to index from the EDP. The link between the two is yet to be explored in depth - dataset portals are different than regular web sites because their mandate is to facilitate data discovery; at the same time, we observe portals draw a substantial share of their traffic from search engine referrals, but there is very limited evidence of SEO (Search Engine Optimisation) specific to dataset search. We did not find evidence in our corpus of dominant search patterns. One reason for this is that sessions are in general very short. For internal sessions, the pattern HH;AF-Category;AF-Country is probably common due to the design of the interface that prominently presented category filtering in the home page. The other patterns suggest that issuing multiple queries or browsing many result pages in succession is more common than using some facet in between.

Session type is associated with success (weak), use of facets and use of keywords (moderate). We observed that external SERP sessions are less successful than expected if session type and success were independent. Considering the previous observation that most sessions of this type landed on a tags dataset, we believe that users clicked on an EDP result after a broad search on the web search engine (e.g "Data about wind"), some of them were not happy about a new list of results, and abandoned more often than people starting the search from the EDP.

In terms of use of affordances, results are mainly due to the large number of bounces in the external searches. Our interpretation is that users that started their search on a web search portal are less likely to continue their search on the EDP.

User profiles of Internal and External SERP are similar, External Dataset is very different. In both internal and external SERP sessions we managed to identify *diggers* , users with longer sessions that use several affordances and/or are willing to browse several result pages. This suggests these users are either conducting longer tasks (exploration, satisfying multiple information needs), or are motivated enough to go through many result pages. As facet usage is relatively high in this profile, portal designers may introduce faceted query suggestions [22] to help these users to become more effective.

Following the assumption that External SERP users are continuing a search in the EDP, a possible improvement could be to ask them to share their immediate search history in the Web search engine with the dataset search engine, in the style of systems like [17, 19]. We believe this might also reduce the number of bouncers and leavers, at least for those users that can afford the extra time and cognitive effort required to consent and select the history to share. Previous works for long search tasks [8] and for multi tasking search [25] should be revisited in the context of ongoing dataset search paradigms and their ongoing realisation in common portal software to better support these users. Works in related areas, such as databases [24] and entity-relationship graphs [28] could be explored to understand how they transfer to dataset contexts.

We discovered another profile in Internal sessions: *section switchers*, users that visit other sections of the portal. These might be users that are just exploring what the portal has to offer, but more studies are required to understand if there are broader information needs, including but not limited to datasets that could be better supported. For example, the EDP offers educational courses on Open Data that are quite popular (more than the dataset search), creating and exploiting links between those courses and the dataset search engine for educational purposes might be an interesting direction for portal managers.

On the other hand, for External Datasets most users leave without using the search affordances, their decision can be simplified as to download or not to download. This suggests they either consider the portal as a static content provider, or that their query is navigational, looking for a dataset. In any case, it is worthy to further study if users don't realise they can search in the EDP (that is, an interface issue) or if they choose not to, and why.

Comparison with analysis of local portals. It is worth comparing the results from the three US-based city-level data portals from [27] to discuss potential differences between a high level metaportal like the EDP and portals that are at the bottom of the pyramid. The authors confirm a higher prevalence of browsing, that is, sessions that only use facets. Whole, the most popular facets depend on the nature of the portals and their search interface, there are parallels between the two studies. For the EDP case, countries (corresponding loosely to the public authority releasing the data) and categories (corresponding to high-level domains such as energy, agriculture, education etc) are the most important facets. For the three portals the same facets correspond to organization (the data publisher) and groups/topics (which are equivalent to EDP categories).

The most noticeable difference is that the local portals appear to have much less access through organic search (matching our external sessions) than the EDP, and much more through direct access (matching our internal sessions) or referrals from other websites (that we did not consider in our study). We believe this is due to the large difference in number of datasets hosted/indexed, less than 500 for each of the local portals versus more than 700 thousand for EDP (at the timeframe considered the analysis), the EDP has much more chances to appear in a result list of a query to a web search engine.

5.1 Limitations

Our study has several limitations. Using dataset download as signal of success is an overestimation, as users may find after inspection that it was not what they needed. We did not have information on returning users, therefore we assumed all visits were independent. For search sessions that start in a web search engine, we did not have data on the behaviour of the user before and after landing on the EDP, therefore we could only assess the success of the part of the session that visited the EDP. We decided not to use dwell time as a signal or feature due to the dataset having the measurement of the time spent in the last page.

6 CONCLUSION

In this paper, we analysed logs of an open government data portal (the EDP) to discover and compare behavioural patterns and user profiles from internal and external search sessions: (i) *internal* searches start on the EDP and use the search affordances it implements; while (ii) *external* searches start with a general-purpose search engine and reach different types of EDP pages. In the second category, we distinguished between so-called *external SERP* sessions, where the user clicks on an EDP result page, which may include multiple datasets; and *external dataset page* sessions, where the user clicks on a different type of page that pertains to a specific dataset or dataset distribution (i.e. the same dataset in different formats or split across different files).

As a dataset discovery tool, the EDP is unique in the sense that it acts as a primary source of public sector data - it hosts all datasets published by the European Union - and at the same time provides unified access to data from many other national and regional public authorities, harvesting, curating, and building an index over their metadata records. Like prior works [12, 27], we looked at different types of search sessions, in particular at native searches starting on the EDP vs external searches. We focused on four themes, comparing and contrasting between the session types: behavioural patterns based on sequential patterns in search sessions; successful searches; the use of keywords and facets; and user profiles clustered from the most common patterns.

As future work, we would like to conduct a deeper analysis of the clusters we found, for example, trying to identify more granular search strategies as in [9]. We would also like to apply our method to transaction logs from other vertical search engines that are indexed by web search engines in order to compare behavioural differences (e.g. open digital libraries or marketplaces). The methodology we followed in this paper could be complemented by various types of user studies. While log analysis is unobtrusive and can be carried out at scale, it does not capture the context of dataset search in great detail. For instance, we would like to run a user study where we directly observe participants undertaking different dataset searches using different tools (such as web search engines, Google Dataset Search, and the EDP) to understand more about search intent and satisfaction as a session progresses and explore other factors that impact on the latter, for instance the results' presentation via the dataset SERP.

Acknowledgements. Research was supported by the *data.europa.eu* portal, an initiative funded by the European Union. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant Data Stories (EP/P025676/2).

REFERENCES

[1] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen. 2019. Searching for Old News: User Interests and Behavior within a National Collection. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 113–121. https://doi.org/10.1145/3295750.3298925

[2] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1365–1375. https://doi.org/10.1145/3308558.3313685

[3] Zeljko Carevic, Dwaipayan Roy, and Philipp Mayr. 2020. Characteristics of Dataset Retrieval Sessions: Experiences from a Real-Life Digital Library. In *Digital Libraries for Open Knowledge (Lecture Notes in Computer Science)*, Mark Hall, Tanja Merčun, Thomas Risse, and Fabien Duchateau (Eds.). Springer International Publishing, Cham, 185–193. https://doi.org/10.1007/978-3-030-54956-5_14

[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2019. Dataset search: a survey. *The VLDB Journal* (Aug. 2019). https://doi.org/10.1007/s00778-019-00564-x

[5] Arnon Dagan, Ido Guy, and Slava Novgorodov. 2021. An Image is Worth a Thousand Terms? Analysis of Visual E-Commerce Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 102–112. https://doi.org/10.1145/3404835.3462950

[6] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. 2016. The SPMF Open-Source Data Mining Library Version 2. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*, Bettina Berendt, Björn Bringmann, Élisa Fromont, Gemma Garriga, Pauli Miettinen, Nikolaj Tatti, and Volker Tresp (Eds.). Springer International Publishing, Cham, 36–40. https://doi.org/10.1007/978-3-319-46131-1_8

[7] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems* 23, 2 (April 2005), 147–168. https://doi.org/10.1145/1059981.1059982

[8] Ahmed Hassan, Ryen W. White, Susan T. Dumais, and Yi-Min Wang. 2014. Struggling or Exploring? Disambiguating Long Search Sessions. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 53–62. https://doi.org/10.1145/2556195.2556221

[9] Jiyin He, Pernilla Qvarfordt, Martin Halvey, and Gene Golovchinsky. 2016. Beyond actions: Exploring the discovery of tactics from user logs. *Information Processing & Management* 52, 6 (2016), 1200–1226. https://doi.org/10.1016/j.ipm.2016.05.007

[10] Daniel Hienert. 2017. User Interests in German Social Science Literature Search: A Large Scale Log Analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 7–16. https://doi.org/10.1145/3020165.3020168

[11] Daxin Jiang, Jian Pei, and Hang Li. 2013. Mining search and browse logs for web search: A Survey. *ACM Transactions on Intelligent Systems and Technology* 4, 4 (Oct. 2013), 57:1–57:37. https://doi.org/10.1145/2508037.2508038

[12] Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, and Elena Simperl. 2019. Characterising dataset search—An analysis of search logs and data requests. *Journal of Web Semantics* 55 (March 2019), 37–55. https://doi.org/10.1016/j.websem.2018.11.003

[13] Emilia Kacprzak, Laura M Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *International Conference on Web Engineering*. Springer, 429–436.

[14] Fabian Kirstein, Simon Dutkowski, Benjamin Dittwald, and Manfred Hauswirth. 2019. The European Data Portal: Scalable Harvesting and Management of Linked Open Data. In *Proceedings of the International Semantic Web Conference 2019 Satellite Tracks*. CEUR-WS.org, 2.

[15] Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *Int. J. Hum. Comput. Stud.* 135 (2020). https://doi.org/10.1016/j.ijhcs.2019.10.004

[16] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1277–1289. https://doi.org/10.1145/3025453.3025838

[17] Ian Li, Jeffrey Nichols, Tessa Lau, Clemens Drews, and Allen Cypher. 2010. Here's what i did: sharing and reusing web activity with ActionShot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*.

Association for Computing Machinery, New York, NY, USA, 723–732. https://doi.org/10.1145/1753326.1753432

[18] Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, and Alan Wilson. 2017. Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 321–330. https://doi.org/10.1109/TVCG.2016.2598797

[19] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. SearchBar: a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1207–1216. https://doi.org/10.1145/1357054.1357242

[20] Kanika Narang, Susan T. Dumais, Nick Craswell, Dan Liebling, and Qingyao Ai. 2017. Large-Scale Analysis of Email Search and Organizational Strategies. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 215–223. https://doi.org/10.1145/3020165.3020175

[21] Andrei Novikov. 2019. PyClustering: Data Mining Library. *Journal of Open Source Software* 4, 36 (apr 2019), 1230. https://doi.org/10.21105/joss.01230

[22] Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2020. Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement. *Journal of the Association for Information Science and Technology* 71, 7 (2020), 742–756. https://doi.org/10.1002/asi.24304 _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24304.

[23] F. Silvestri. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends® in Information Retrieval* 4, 1–2 (2010), 1–174. https://doi.org/10.1561/1500000013

[24] Manish Singh, Michael J Cafarella, and HV Jagadish. 2016. DBExplorer: Exploratory Search in Databases.. In *EDBT*. 89–100.

[25] Amanda Spink, Minsoo Park, Bernard J. Jansen, and Jan Pedersen. 2006. Multitasking during Web search sessions. *Information Processing & Management* 42, 1 (2006), 264–275. https://doi.org/10.1016/j.ipm.2004.10.004 Formal Methods for Information Retrieval.

[26] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[27] Fanghui Xiao, Zhendong Wang, and Daqing He. 2020. Understanding users' accessing behaviors to local Open Government Data via transaction log analysis. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020). https://doi.org/10.1002/pra2.278 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.278.

[28] Sivan Yogev, Haggai Roitman, David Carmel, and Naama Zwerdling. 2012. Towards expressive exploratory search over entity-relationship data. In *Proceedings of the 21st International Conference on World Wide Web*. 83–92.