UNIVERSITY OF SOUTHAMPTON

Improving Searchability of Datasets

by

Emilia Kacprzak

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
Electronics and Computer Science

March 2022

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Emilia Kacprzak

Data is one of the most important digital assets in the world thanks to its business and social value. As is becoming increasingly available online, in order to use it effectively, we need tools that allow us to retrieve the most relevant datasets that match our information needs. Web search engines are not well suited for this task as they are designed for documents, not data. In recent years several bespoke search engines have been proposed to help with finding datasets, such as Google Dataset Search crawling the whole web or DataMed focused on creating an index of biomedical datasets. In this work we look closer into the problem of searching for data on the example of Open Data platforms. We first applied a mixed-methods approach aimed at deepening our understanding of users of Open Data portals and types of queries they issue while searching for datasets accompanied by analysis of search sessions over one of these data portals. Based on our findings we look into a particular problem of dataset interpretation - meaning of numerical columns. We propose a novel approach for assigning semantic labels to numerical columns. We conclude our work with the analysis of the future work needed in the field in order to potentially improve the searchability of datasets on the web.

Contents

List of Figures

List of Tables

Research Thesis: Declaration of Authorship

Print name: Emilia Kacprzak

Title of thesis: Improving Searchability of Datasets

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

    - A query log analysis of dataset search; E. Kacprzak, LM. Koesten, LD. Ibáñez, E. Simperl, J. Tennison Proceedings of the International Conference on Web Engineering (2017)
    - Characterising dataset search queries; E. Kacprzak, L. Koesten, J. Tennison, E. Simperl Companion Proceedings of the The Web Conference (2018)
    - Making Sense of Numerical Data-Semantic Labelling of Web Tables E. Kacprzak, JM. Giménez-García, A. Piscopo, L. Koesten, LD Ibáñez, J. Tennison, E. Simperl Proceedings of the 21st International Conference EKAW (2018);
    - Characterising dataset search—An analysis of search logs and data requests E. Kacprzak, L. Koesten, LD. Ibáñez, T. Blount, J. Tennison, E. Simperl Journal of Web Semantics (2019);
    - Characterising Dataset Search on the European Data Portal: an Analysis of Search Logs; LD. Ibáñez, L. Koesten, E. Kacprzak, E. Simperl Analytical Report 18 for the European Data Portal (2020)

Signature:

Date:

Acknowledgements

I would like to express my gratitude to my supervisors, Dr Jeni Tennison, Prof. Elena Simperl and Dr Luis-Daniel Ibáñez without whom none of this would be possible.

I would also like to thank the people who have supported and helped me through this journey – especially my parents and my friends without whom it would be impossible to complete this work. For your continuous support I will be eternally grateful.

Finally, I would like to thank the Marie Skłodowska-Curie Innovative Training Networks (the Horizon 2020 programme) for enabling this work under the Answering Questions using Web Data (WDAqua) project.

Chapter 1

Introduction

Data has become one of the most important digital asset in the world. The ability to produce business value from data analytics has become a differentiating factor between top and lower performers among companies from more than 30 industries in 100 countries (Lavalle et al., 2011). In the public sector, through initiatives such as Open Government Data, it can also generate great social impact, optimise public services, and increase transparency (Verhulst and Young, 2016).

As we advance in the digital age, more and more of the data we generate can be accessed (or purchased) online. A study by Cafarella estimates in excess of one billion sources of data on the web as of February 2011, counting structured data extracted from Web pages (Cafarella et al., 2011). The Web Data Commons project recently extracted 233 million data tables from the Common Crawl; they are freely available for download (Lehmberg et al., 2016). A growing number of organisations have set up their own data portals to publish datasets related to their activities. A large share of these organisations are public administrations at local, regional, and national levels. In this context, the European Data Portal[1] indexes to date no less than $1,159,953$ datasets from 28 EU countries, while the official US data portal[2], lists $217,693$. Similar trends can be observed in the commercial sector, with specialised vendors (for example in finance and marketing) co-existing alongside data marketplaces (e.g., Quandl[3], Data.world[4], Enigma[5], BigDataExchange [6], Dawex[7]) that connect supply and demand.

With this trend, of accelerated growth of data available on the web, searching for it becomes a pressing issue if one wants to increase the value extracted from it. Data search and discovery is researched in a range of complementary disciplines but despite

[1] https://www.europeandataportal.eu/data/en/dataset
[2] http://www.data.gov
[3] http://www.quandl.com
[4] https://data.world/
[5] http://www.enigma.io
[6] http://www.bigdataexchange.com
[7] https://www.dawex.cometc

that, it is by far not as advanced as related areas such as searching for documents, both technologically (Cafarella et al., 2011) or from a user experience point of view (Koesten et al., 2017). Prior studies investigating search strategies for datasets amongst users have shown that personal recommendation (Koesten et al., 2017), as well as links from literature to datasets still play a big role in a discovery process (Gregory et al., 2019).

The dataset search problem can be addressed at various levels. Services such as Google Dataset Search (Brickley et al., 2019) and DataMed (Chen et al., 2018; Sansone et al., 2017) crawl across the web, build an index of resources and facilitate a global search across them. Existing data search approaches use tags found in metadata mark-up, expressed in vocabulary terms from schema.org[8] or DCAT[9], to structure and identify the metadata considered important for datasets. The same approach is used in data portals including open government data portals on national (such as data.gov.uk) or local (such as opendata.bristol.gov.uk) level, scientific repositories such as Elsevier's and data markets (e.g. (Grubenmann et al., 2018)).

Currently, there is a gap between what dataset resources are being made available, and what a user can actually find, trust and use according to their needs. To fill this gap datasets need to become more searchable, and by the searchability of datasets we mean the quality of the searching process – that being easiness in navigating, performing and evaluating the search for datasets on the web – for the end user when searching for datasets. One way of improving the searchability is to provide the tools (such as data portals or crawling and ranking algorithms) with the informative metadata. We consider that to know the more useful metadata, we need to understand the characteristics of the dataset search process, the key struggles and identify how users explain their information needs for datasets.

In this work we consider datasets as first class citizens and study dataset search as a search vertical (search for specific subset of the web based on the structure or the topic) on its own, aiming to characterise the differences with other search verticals. In order to improve the information retrieval process (and more specifically it's searchability), the first step is to understand the characteristics of the process, the key struggles and identify how users think of it and explain their needs. With this aim in mind, in the first part of this thesis, we investigate how users search for datasets, what are the most important information towards an effective dataset search process and how it could be improved, using existing search infrastructure. We identified the lack of sufficient metadata (specifically on temporal and geospatial coverage, their granularity and the semantics of the data - textual and numerical - allowing to interpret it) to be one of the crucial problems in order to advance towards improving search algorithms, introducing recommendation systems or improving general web search indexation process of data portals. As Mitlöhner et al. (2016) in their analysis of 200K tabular resources from

[8] https://schema.org/

[9] https://www.w3.org/TR/vocab-dcat-2/

Open Data portals reported numeric columns to be the most popular column type, we devote the second part of this thesis to the automatic generation of semantic labels for numerical data. We analyse the problem of disambiguating the meaning of numerical columns based on data from a Knowledge Base, and propose a novel approach aimed at this issue.

1.1 Problem Space & Motivation

While searching for data, the main goal of a user is to satisfy their information needs. Whether it is a data journalist writing an article that compares government transparency in different countries, an app developer trying to expand into new markets, a business analyst searching for evidence to substantiate their report, or a scientist replicating an experiment, the first (and foremost) step all these professionals will have to take is to find, or *retrieve* the most relevant datasets for their needs.

Users can encounter a variety of issues when attempting to find datasets. Those can be reusability factors (e.g. format, licence or cost associated with use of the dataset), reliability (e.g. how the data was collected, were any outliers removed) or they may not be able to find a dataset with the specific set of keywords they use in the search process (Koesten et al., 2017). Previous work has looked at special-purpose search engines, tailored for datasets in a given domain, including hydrology (Ames et al., 2012); earth sciences (Devarakonda et al., 2010); and geography (Walker et al., 2004). In the context of e- and open science, researchers have proposed technologies to find datasets from scientific experiments (Lu et al., 2012; Singhal et al., 2013). Kunze and Auer (2013) have introduced the concept of *dataset retrieval*, as a branch of information retrieval applied to data instead of documents focused on determining the most relevant datasets according to a user query. However, though not limited to a particular domain or application context, they focus on graph data stored as RDF.

There are several unresolved issues that could be applied both to open data and to discoverability of the datasets on the web in general. The European Commission defined six barriers for open public data in 2011[10]. Our work focuses on the first two barriers 1) A lack of information that certain data actually exists and is available; 2) A lack of clarity of which public authority or organisation holds the data; which are directly connected with the the dataset search process on both general search engines and through internal search capabilities of data portals.

In work by Koesten et al. (2017), authors asked people to share their experiences when carrying out search to satisfy their information needs. They interviewed 20 data practitioners with different professional backgrounds and skills, to better understand the

[10] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm

challenges they face in retrieving the data they need. They found out that in the absence of data search engines, they rely on general purpose search engines or data portals, or ask other people for recommendations. Existing search technologies do little to help. Using conventional search engines is not ideal, as these have been designed primarily for documents, not data (Cafarella et al., 2011). In face of those findings Koesten et al. (2017) argued that expanding metadata, describing datasets could aid search process on both, specialised search platforms (such as open data portal) and on general web search engines.

Currently, open datasets are typically catalogued on official portals under defined lists of categories, and accompanied by short metadata descriptions according to standards such as DCAT, which define attributes such as title, description, language or licence[11]. Despite the search functions provided by such catalogues, it is often not possible for an ordinary user to find relevant pieces of information quickly (that being single or multiple datasets). This can be caused by non-intuitive or limited data descriptions, misleading naming conventions, incorrect assignment of categories to datasets, the user's lack of in-depth knowledge of the subject, or simply because the search is only conducted over the metadata records rather than the data itself (which might be equally not sufficient in comparison to more descriptive metadata as it is not the focus of structured data document to provide context). The metadata describing datasets can be incomplete or outdated, as maintaining it is frequently manual and expensive. In many cases the metadata does not describe the full potential of the data, so some relevant datasets may not be presented as a result of a query simply because appropriate keywords were not used in the description or within the data, this further motivated our research on automating metadata generation which will allow generation of search indexes representing the content of the dataset in greater detail and with a greater chance of matching the description provided by a potential dataset user.

1.2 Aim & Research Contribution

The aim of this work is to extend our understanding of the dataset search process and look into potential improvements to this process in order to aid users on their data journeys. The overall **research question** of this work is "*how can we improve searchability of datasets that are published on the web?*" with the following specific research questions:

- [RQ1] Understanding Dataset Search; How do users search for data? What search strategies do they use?

- [RQ2] What information should be included in the metadata to improve the dataset search process?

[11] https://www.w3.org/TR/vocab-dcat/

- [RQ3] How to automate the generation of metadata? In this work we focus on one of the most under-investigated areas: semantic labelling of numerical columns.

The objectives and contributions of this work are connected with different steps of the data discovery process. Our initial study, based on search logs from four data portals, provides a better understanding of users of the data and the ways they structure their queries. We have extended this work with a follow up study focusing on more elaborate user interactions with data portals - namely dataset requests. Finally we conduct a session analysis of users of one of the portals. We conducted the above studies in order to define potential improvements to the dataset search process.

Based on our finding on the types of information which could aid the searching process we looked into one of the research domain which showed itself as one of the most under-explored. We analysed existing approaches from the research field of Semantic Web aiming at improving understanding of datasets columns with use of semantic labels assigned to each column. We propose a novel approach targeting this issue, focusing on the problem of labelling of numerical columns, which we believe is the most understudied. Effective assignment of semantic labels to columns could aid dataset search process on many levels: from query based information retrieval through vertical search engines as well as general search engines, faceted search functionalities or dataset recommendation systems.

The presented research has lead to a number of per reviewed publications in addition to publications published as part of collaborations:

- *A query log analysis of dataset search*
  E. Kacprzak, LM. Koesten, LD. Ibáñez, E. Simperl, J. Tennison
  Proceedings of the International Conference on Web Engineering (2017); (Kacprzak et al., 2017)

- *Characterising dataset search queries*
  E. Kacprzak, L. Koesten, J. Tennison, E. Simperl
  Companion Proceedings of the The Web Conference (2018); (Kacprzak et al., 2018b)

- *Making Sense of Numerical Data-Semantic Labelling of Web Tables*
  E. Kacprzak, JM. Giménez-García, A. Piscopo, L. Koesten, LD Ibáñez, J. Tennison, E. Simperl
  Proceedings of the 21st International Conference EKAW (2018); (Kacprzak et al., 2018a)

- *Characterising dataset search—An analysis of search logs and data requests*
  E. Kacprzak, L. Koesten, LD. Ibáñez, T. Blount, J. Tennison, E. Simperl
  Journal of Web Semantics (2019); (Kacprzak et al., 2019)

- *Typology-based semantic labeling of numeric tabular data*
  A. Alobaid, E. Kacprzak, O. Corcho
  Semantic Web Journal (2019)
  Personal contribution: Help with the design of typology of numerical values, implementation of detection process and editing support; (Alobaid et al., 2019)

- *Collaborative Practices with Structured Data: Do Tools Support What Users Need?*
  L. Koesten, E. Kacprzak, J. Tennison, E. Simperl
  Proceedings of the CHI Conference on Human Factors in Computing Systems (2019)
  Personal contribution: contribution to study design, review of tools supporting collaborations and editing support; (Koesten et al., 2019)

- *Characterising Dataset Search on the European Data Portal:*
  an Analysis of Search Logs
  LD. Ibáñez, L. Koesten, E. Kacprzak, E. Simperl
  Analytical Report 18 for the European Data Portal (2020)
  Personal contribution: quantitative analysis of anonymous user sessions data collected from Web Analytics package; (Ibáñez et al., 2020)

1.3 Thesis Structure

The remaining chapters of this work are structured as follows:

Chapter 2 provides background on Information Retrieval in general, other search verticals, current state of dataset search and relevant related work regarding studies conducted in this work. Chapter 3 presents methodology: definitions, problem statements and research questions. Chapters 4, 5, 6 and 7 shows our work on understanding dataset search using quantitative and qualitative methods on Web analytics data collected from open data portals. Chapter 8 outlines the state-of-the-art in semantic labelling of numerical columns. It further presents a novel approach for assigning semantic labels. At last, Chapter 9 summarises findings of previous chapters and concludes the work.

Chapter 2

Background & Related Work

Information retrieval is the process of finding relevant resources that satisfy an information need from a collection of information resources (Belkin and Croft, 1992). To enable this process, resources gathered in a corpus are represented through structured metadata, describing its content. In this chapter we review the current research in the area of information retrieval in the context of its specialisation – dataset retrieval. We further discuss the literature in relation to each of the research questions. Looking at [RQ1] "Understanding Dataset Search; How do users search for data? What search strategies do they use?" we discus search log analysis and general session log analysis as they are a common approach for evaluation of information retrieval systems. We than discus the metadata to introduce further the topic which we want to touch on in [RQ2] "What information should be included in the metadata to improve the dataset search process?". Finally, we summarise the literature in relation to specific area of metadata generation – assigning semantic labels to the data in relation to [RQ3] "How to automate the generation of metadata?".

2.1 The Dataset Search

The basic dataset search process consists of four steps as defined by Chapman et al. (2020): 1) querying, 2) query handling, 3) data publishing and metadata generation and 4) results presentation, each of which an area that specific research efforts can focus on: Figure 2.1 outlines the dataset retrieval process.

Querying. Querying for resources can be performed with use of different methods allowed by a search functionality. It can be performed with use of keywords along more advanced functionalities such as filters (a.k.a facets), structured query languages (such as SQL or SPARQL); or question answering type functionality. In terms of open data platforms querying occurs commonly in the form of keywords, to which filters can be

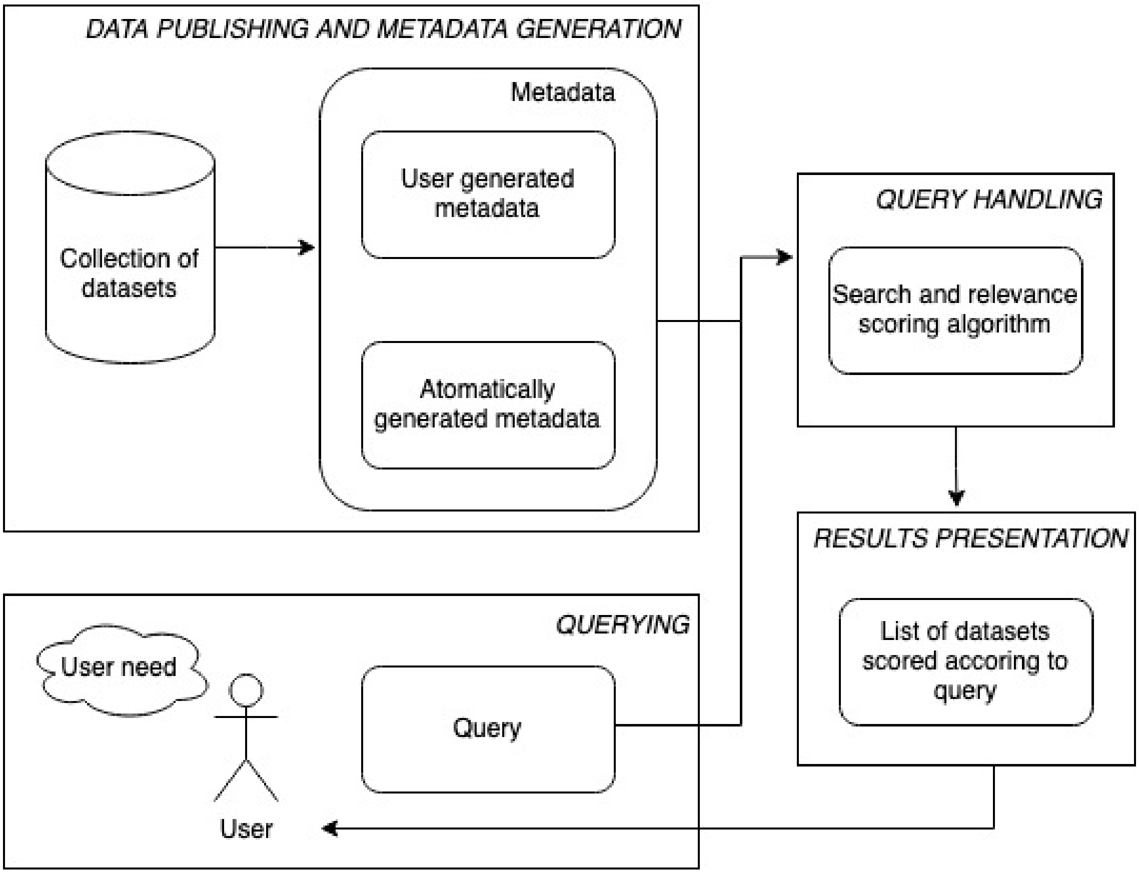

DATA PUBLISHING AND METADATA GENERATION
Collection of datasets
Metadata
User generated metadata
Atomatically generated metadata
QUERY HANDLING
Search and relevance scoring algorithm
RESULTS PRESENTATION
List of datasets scored accoring to query
QUERYING
User need
User
Query

FIGURE 2.1: Schema of dataset retrieval flow

applied. Such facets can differ from portal to portal (examples of facets are: countries, catalogues, formats, licenses etc.). In this work in Chapter 6 we analyse queries issued through various data portals as well as queries issued to search engines leading to the portals' dataset section.

Query handling. Most dataset search algorithms use as an input the datasets' metadata which in most cases is produced manually. Search results are then produced based on the similarity of the metadata describing the data, to the search terms issued within a query. Unfortunately, low metadata quality (or missing metadata) affects searchability of the data - both the discovery and the consumption of the datasets within open data portals (Umbrich et al., 2015). The success of the search functionality depends on the publisher's knowledge of the dataset and the quality of the descriptions they provide. *Ranking datasets* is a separate research problem. Traditional general web search approaches such as PageRank are not best suited at this problem due to limited links between datasets (Brickley et al., 2019). The applicability of IR models built mainly for document retrieval is questionable. Datasets might require different approaches to ranking due to their unique characteristics both in terms of their structure as well as concerning the types of search tasks users engage in (Chapman et al., 2020).

Data publishing and Metadata generation. In a data publishing process, publishers are asked to provide relevant metadata describing the data that they are releasing. Presence of unified vocabularies such as DCAT, 'CSV on the Web', schema.org used by a particular portal allow for a consistency in-between metadata which allows for a search (query or facets based) functionality to be developed. The process of metadata generation is in general resource-intensive and problematic for a publisher as it is mostly manual and there in insufficient guidance. As a result often dataset descriptions could be considered incomplete with not enough detail, as has been shown in study of dataset summaries (Koesten et al., 2020). Not sufficient metadata cripples the potential of query handling approaches aimed at matching user queries with the relevant datasets. This also points at the direction of approaches targeted at achieving better metadata generation in automatic and semi-automatic way.

Results presentation. In manner similar to the way the dataset relevance and ranking functionality judges the relevance of a particular dataset to a specific query, a user chooses datasets and decided on their usefulness to their task base on the information provided alongside the dataset – dataset's metadata. The Search Engine Result Page (SERP) provides the user with a number of links pointing to datasets in an order of their calculated relevance, this is shown in a manner analogous to the general web search result SERP (Chapman et al., 2020). When selecting a particular search result, the user is taken to a dataset page showcasing the metadata and containing links allowing to download a particular dataset. Google Dataset Search[1], which aims to show results from many different repositories, in comparison to a traditional SERP allowing to navigate to a specific dataset page proposed an interface which allows to see the SERP while showcasing the dataset page on the right part of the screen.

2.2 General Web Search

Dataset are published on the web, however, algorithms proposed to search for web pages might not be best suited at the dataset search as the general web search engines targeted at the web pages take advantage of the specific structure of the web besides known in text mining techniques (e.g. tf-idf (Zhang et al., 2008)). Web pages can be modelled as the nodes of a graph, with hyperlinks as edges. Information retrieval in web search engines is performed in three phases: crawling, indexing and ranking (Croft et al., 2009). *Crawling* is the process of data discovery. That means looking for new or updated content so that the graph of all the information available on each page is always up to date. *Indexing* is the process of creating a list of the words and phrases that describe the resource in order to be able to access it faster. Last, the *ranking* phase is responsible for search accuracy. Every web page is assigned a rank based on which the results are ordered and

[1] https://datasetsearch.research.google.com/

presented to the user (Brin and Page, 2012). The basis of this process which since were improved on greatly is described shortly in the paragraphs below.

Page et al. (1999) introduced the PageRank algorithm which main innovation is taking into account the relative importance of individual web pages. Based on the assumption that every page has forwardlinks and backlinks the importance of every web page can be determined based on importance of web pages pointing to it. This importance is then split between the web pages that this web site itself, points to. Such graph based importance measure allows to determine more important resources with the idea that more pages (which themselves are considered valuable) will point to them than to the less important resources in a particular domain. As a result, a query submitted by a user to a search engine will serve to determine the subset of relevant pages, however, the PageRank of those pages will be used to rank them on the search results list page.

In parallel to PageRank, Kleinberg et al. (1999) presented the HITS algorithm which describes the task of ranking of web pages based on link analysis, scoring web pages according their authoritativeness (estimation of the value of the pages content) and their hub weights (estimation of the value of their links pointing to the other web pages). Given a particular user query, the most relevant web pages are retrieved (with use of standard web mining methods (Grover and Wason, 2012) such as text-based search), this set is than expanded with the web pages which are linked within this set and their hyperlinks. This approach allows to form a query dependant set, which is scored based on their authoritativeness and hub weights and returned ordered to the user who issued the query. Authority weight is the sum of the hub weights that are pointing to the given node whereas hub weight is the sum of authority weight that a given node points to. The user who submitted a search query obtains a list of web pages with the highest authoritativeness and hub weights.

General web search has evolved beyond both of these basic algorithms using other techniques such as machine learning (Agichtein et al., 2006; Balaji et al., 2021), question answering systems (Kwok et al., 2001; Wang et al., 2018) or personalisation of results for each user (Sieg et al., 2007; Sharma and Rana, 2020). However, it still is not a good fit for the activity of dataset search. Cafarella et al. (2008) pointed out that for structured data content on the web, relations contain a mixture of structural and related content elements which cannot easily be mapped to unstructured text scenarios. Relations also lack the incoming hyperlink anchor text that helps general web search. For this reasons PageRank-based algorithms are not to the same extent applicable to table search (tables within web pages), particularly as tables of widely-varying quality can be found on a single web page. Table search is further discussed in Section 2.4.

2.3 Vertical Search

Vertical search is search in which the subject of the search is a specific subset of content, as opposed to general web search where the aim is to include all types of resources or over a repository of specific type of content. The subject of vertical search can be a collection which is distinct based on its topic, data type or context. Typically, web vertical search engines use similar crawlers to general web search engines to index the content of the web. Crawlers for vertical search are more focused and generally work based on predefined topics. Because of the limited scope of resources that are taken into account in vertical search engines, they often offer greater precision, more complex schemas or ontologies to match specific searching scenarios, and tend to support more complex user tasks (Li et al., 2010).

Some examples of vertical search engines are: for instance email search is an example in which Ai et al. (2017) noticed that when searching for emails, users know the precise attributes of a resource they are looking for. The authors point out that one of the key differences to general web search is that the set of emails is a personal set unique for each user. On top of that email search offers additional metadata (e.g. sender address, subject or timestamp) which can help both organizing and searching through the results.

People search (Weerkamp et al., 2011; Guy et al., 2012) is gaining more importance with portals like LinkedIn or Facebook. In this search vertical, the most relevant factors are the first and last name of the person but search could also be dependent on the relations of two people that could be expressed through the same educational background, home town or common friends. People search is also of importance in enterprises, which is slightly different as e.g. phone number, email or the organisation employing the person can also be relevant.

In search for research publications (Li et al., 2010; Yu et al., 2005) argue that general web search does not fully take advantage of the potential of this specific subset of resources - temporal information attached to each of the publication. Algorithms like PageRank and HITS calculate the relevance of each resource and include this information while ranking the relevance of resources to a user query. In publication search, the reputation of the resource, in addition to its content relevance, citation count and reputation of its authors and journals are the most influential parts. However, in this scenario it is important to take the publication time bias into account when generating search results for user query, as the most recent resources, for example, if the task of the user is to compile the state of the art of a particular topic, recent relevant resources are more useful to the task, however, more recent resources may not have been cited enough to have a high enough PageRank score to be easily found.

Each of the verticals is heavily dependent on the specific set of metadata properties which take advantage of the structure of their resources and the nature of the search that is performed. In our work we aim to identify the specifics of dataset search with aim of furthering the improvements in systems build to handle datasets.

2.4 Dataset (Table) Retrieval

With the growing presence of tables and datasets within web pages, company repositories or dedicated platforms there is a growing body or research targeting search and exploration of this kind of resources. However, dataset search and retrieval is still a relatively unexplored area compared to document search and retrieval - below we present the overview of work in this field.

In the works on web table search, tables present within web pages' DOM, contain additional information withing HTML tags on the page. Such unstructured information such as the table title, page title, paragraphs proceeding and following the table could provide a context for the meaning of the table, guide additional metadata creation (e.g. assignment of semantic labels to columns), and therefore provide more information to a table search functionality than it could be provided to dataset in a CSV or TSV format published on the web with publisher provided metadata to accompany it. Recently, Zhang and Balog (2020) conducted an extensive survey on table retrieval providing an overview of the topic on various stages: table extraction, table interpretation, table search, question answering, and knowledge base and table augmentation.

When searching for tables published as part of the web pages it is natural to treat it at first as an extension of the general Web search, for example, Cafarella et al. (2008) proposed an approach for searching through web pages that contain a table with structured data. They explore different approaches including *NaiveRanking* - which matches the general web search, showing the top k results with some structured content to the user; *FilterRanking* going down further into search results list to show more web resources that include structured data; *FeatureRank* approach in which the ranking is performed not based on general web search engine but based on a set of features including: number of rows, columns, nulls in table, hits on headers and the most left column and hits on the table body in addition to the document search rank of the page containing the table. Finally, they include scores measuring the coherency of a table; this ranking was called *SchemaRank*. This is then used as a starting point for web table search that take advantage of the web structure. The above methods, especially the *SchemaRank* propose the utilisation of the dataset (here a table) item itself for the search purposes. However, the terminology used within the data or their granularity (e.g. cities vs countries) might not be sufficient for finding the relevant data by the user.

In 2010 Google developed Fusion Tables which is a data management system (Gonzalez et al., 2010). Users of the platform can make their table public which will include them in general Google web search. Google creates for it a corresponding HTML page that is then crawlable by general search engines in the same way that other web pages are crawled and indexed. Fusion Tables also recognised the need for an internal search functionality that will allow the user to search only among tables that are managed by the system. They also found that existing techniques for dataset retrieval are not applicable for datasets in a web search scenario.

In terms of more specialised approaches efforts have been focused on developing systems to support search for datasets as part of the data sharing process in scientific communities like Hydrometeorology (Ames et al., 2012), Earth Science (Devarakonda et al., 2010) and datasets coming from Geographical Information Systems (Walker et al., 2004). More general approaches to encompass any type of research datasets were developed by Lu et al. (2012) and Singhal et al. (2013). In both cases, information extracted from the papers in which the datasets are mentioned is used to construct a knowledge base to power the search. Kunze and Auer (2013) who defined the problem of *dataset retrieval* as determining the most relevant datasets according to a user query, restricted their scope to RDF datasets and propose a retrieval mechanism inspired by faceted search, where dataset relevance is checked against a set of semantic filters. In all these works, the process is data-focused, the starting point are datasets from specific domains in formats, from which the search strategies are developed.

Recently, Brickley et al. (2019) introduced Google Dataset Search, a search system which crawls data portals, indexes available metadata and is allowing users to search through it with keyword queries. Building upon their experience as a commercial search engine, they identified multiple challenges for further advances of dataset search as a separate search vertical: varied metadata quality, metadata duplication on search result list, multiple copies of datasets metadata instances available through various portals, relatively low visits rates of dataset pages impacting their crawl time, difficulty in judging relevance and ranking datasets, and at last various metadata standards used.

As could be seen in the above literature, the dataset search (and table search in some sense similar manner) is not a task different from general web search for web pages. Dedicated methods are needed for the retrieval processes along the metadata describing the data which could support the process. Expanding the information in the data with more information on scope and meaning of the dataset (table) in addition to description provided by a data published conveying the context of the data would be beneficial to the user and the search functionalities (Koesten et al., 2020).

2.4.1 Open Data Portals

In this work to better understand the dataset search vertical we analyse the data on users from five open data portals. Open data portals are a point of free access to the governmental and institutional data for both commercial and non-commercial purposes through cataloguing with common metadata, which allow search and explorative activities.

The most popular platform in the governmental open data domain is CKAN[2]. CKAN is an open source data management system, which provides tools for publishing, sharing, finding and using data. It is one of the most popular data platforms used by many national data publishing entities (used by for example the UK, USA, Canada, Australia and until recently European Data Portal).

CKAN is built using Apache Solr[3], which uses Lucene to index the documents. In this scenario the documents are the datasets' metadata provided by the publishers. CKAN integrates the DCAT metadata schema which is an RDF vocabulary facilitating interoperability between data catalogs published on the web[4].

The search functionality in Solr is composed of two main operations: finding the documents that match the user query and ranking those documents. By default, the documents are ranked based on relevance, which means that after the final set of matching documents has been found, an additional operation is necessary to calculate a relevance score for each of the matching documents. As Solr is a search engine designed to search across large amounts of unstructured text and assign relevance to each of the results, its search algorithm focuses mostly on natural language provided by the publisher of the data. In this scenario, the success of the search functionality depends on the publishers knowledge of the dataset and the quality of the descriptions they provide.

Lucene[5] is a text search engine library. In open data portals it maps DCAT metadata fields into an inverted index consisting of a list of terms and ids of documents in which given term appears. In order to do that they use a list of predefined metadata fields that describe the document (the most relevant of them being title and description). An inverted index is built based on terms from completed metadata fields which serve as a base of the search process.

Calculation of the relevance of a document to a user query is performed using the *term frequency–inverse document frequency (TF-IDF)* algorithm. It calculates the weighting of a term through a composition of two statistical approaches. The TF is the frequency with which a word appears in a single document, whereas IDF indicates the inverse proportion of the word's frequency in the whole document corpus. The basic idea of this solution is that the more often the word appears in a document, the more accurately it describes its content (Sebastiani, 2002). On the other hand, if the word appears in more documents, it became less representative for a single document and should be given less weight (Zhang et al., 2008). Each document and query are represented as a vector in

[2] https://ckan.org/
[3] http://lucene.apache.org/solr/
[4] https://www.w3.org/TR/vocab-dcat/
[5] https://lucene.apache.org/core/

a vector space model; the similarity score between them is the result of calculating a cosine between the query vector and document vector.

As for the datasets published on open data portals, Mitlöhner et al. (2016) analysed the metadata and CSV files from such portals. They found that the average open data CSV file contains 365 rows and 14 columns and that the most values are numerical. They also analysed what notation is used in headers; checking for the following features: underscores, single words, multiple words and use of camel case. In 40% of cases, headers contained underscores; 33% were single words; 17% consisted of multiple words and 9% were expressed in camel case. The form in which the header of the column is written can influence machine readability, potential further work with a dataset (e.g. making it harder to transform such a file into RDF format) but also the way in which metadata should be extended with the information from the dataset itself.

2.5 Search Log Analysis

Analysing query logs serves as a proxy to analyse the search behaviour of users (Kaur and Aggarwal, 2018; Kelly, 2009; White et al., 2016) and can serve as a way to understand the users intent when interacting with a search functionality (Clark et al., 2012). The first query log analysis on the web was published in 1999, for the Altavista search engine (Silverstein et al., 1999), and the approach has since been used to study many aspects of web search (see Jiang et al. (2013) for a survey). Vertical search engines have also used query log analysis, for example, in people search engines (Weerkamp et al., 2011), electronic health records (Yang et al., 2011) and digital libraries (Jones et al., 2000). Search patterns have unique characteristics in different search environments, for instance Ortiz-Cordova et al. (2015) analyse patterns in search behaviours within two sets of logs: internal and external search logs. Those sets were collected for the ecommerce site *www.BuenaMusica.com*, listing traffic coming from general search engines (external logs) and search activity within internal search function (internal logs). It is important to remember that results of different search log analysis are not directly comparable as pointed out by Jansen and Spink (2006) in their transaction logs analysis of nine search engines. However, they are significant measure of user behaviours and search functionalities performance. In our studies we conduct analysis of search logs from four national open data portals and one supranational portal, and more freely phrased data requests - a different form of users communicating their data needs (Chapters refchapter:users-7) with the aim of understanding current state of dataset search and exploring it as a separate search vertical. Based on our findings we explore potential improvements to those solutions. Below we present a summary of metrics used in studies presented in Chapters 4-7.

Query Length and Distribution. Average length, distribution, percentage of 1, 2 and 3 words queries. These are the most commonly (e.g. Silverstein et al. (1999); Jansen et al. (2000); Bendersky and Croft (2009); Zhang et al. (2009); Taghavi et al. (2012)) presented statistics and are part of the analysis in our study. Taghavi et al. (2012) have shown two trends in web search query length and its distribution: increase in query length through the years. This might be connected with a fact that with the development of search functionalities users ask more detailed queries making it a relevant metric for a much younger domain of dataset search on the web. Ortiz-Cordova et al. (2015) show the difference in average length and length distribution between internal and external queries. The results show that internal queries are shorter than external queries (on average 2.76 words for external and 2.25 words for internal). They used this information further to analyse the differences in consecutive search activities and to define search patterns.

Topics. The topic categorisation used in other studies gives an overview of topics asked in web search queries. Below we present two of such classifications. Spink et al. (2002) classification: *Commerce, travel, employment or economy*, *People, places or things*, *Unknown*, *Computers or Internet*, *Sex or pornography*, *Health or sciences*, *Entertainment or recreation*, *Education or humanities*, *Society, culture, ethnicity or religion*, *Government*, *Performing or fine arts*. Beitzel et al. (2004) classification: *Personal Finance*, *Computing*, *Research & Learn*, *Entertainment*, *Games*, *Health*, *Holidays*, *Home*, *US Sites*, *Porn*, *Shopping*, *Sports*, *Travel*, *Other*, *Government*, *Movies* and *Music*. Classification of such type cannot be directly applied in our case, as data portals have a more narrow scope containing specific set of information. In our case we analysed topics based on the data categorisation proposed on the data.gov.uk portal[6].

Query Types Classification. Broder (2002) created a taxonomy of web search queries based on user needs. They propose three classes of queries: *informational* - get information about something, *navigational* - to reach a particular site and *transactional* - to perform a transaction. Informational queries intend to access particular information. *Navigational* queries represent the intention of reaching particular web pages. *Transactional* queries represent activities which a user wants to perform on the web, e.g, shopping or downloading content. Classifying queries according to their intent is often challenging, as queries are short, ambiguous and their meaning may change over time or with a change of location (Jiang et al., 2013). We believe this taxonomy is not directly applicable to dataset search overall as the information need is *finding data* and could so be seen as predominantly informational. We chose to classify queries containing specific types of information that have been studied in other search contexts: acronyms, geographic location, temporal indication, numeric values or question queries (starting with words that indicate questions, e.g. what or how). Understanding the amount of queries related to these dimensions can help shape indexing strategies in dataset search engines.

[6] https://www.data.gov.uk

We, however, split queries collected from external search engines based on the fact if they are believed to be navigational or not. Further details are presented in Chapter 6.

User and Session Statistics. Information retrieval studies for web search also use behavioural characteristics, which are metrics not directly concerned with the search query itself but with search behaviour. These can give additional insights about the user population which perform dataset search that cannot be obtained by just analysing the query itself, in the void from other actions. Under this category fall both metrics such as session length (Jansen and Spink, 2005); browser statistics or distribution of queries for a time frame (e.g. per day, week or month) (Taghavi et al., 2012); but also analysis of query reformulation (Odijk et al., 2015) or user satisfaction – which we discuss below.

User Satisfaction. Search log data have been proposed to analyse the effectiveness of search functionalities using implicit information. Odijk et al. (2015) in their analysis of struggling in web search conducted an analysis of session tasks, which they split into successful and unsucessful based on users clicks on result in a result list and time spend on it. Fox et al. (2005) defined a list of alphabet characters, mapping user actions. They further used those mapping to mine the most frequently occurring sequences and collide it with information collected from users on satisfaction. Hassan et al. (2010) followed similar approach with satisfaction for each task judged by humans. Due to unavailability of information on user satisfaction with performed search in our analysis we followed approach similar to (Odijk et al., 2015) defining a successful session as the one containing external resource click or resource download. Further details are provided in Chapter 7.

2.6 Metadata

Defining a set of important and unique metadata features for dataset search is important in order to treat dataset search as a vertical (which e.g. can be later included as one of the verticals included in the results of aggregated search which is more and more popular in general web search) (Zhou et al., 2013).

There are several metadata standards applicable to data on the web. One of them is DCAT[7] mentioned earlier as the standard used by the CKAN platform. DCAT is an RDF vocabulary used to describe datasets in data catalogues and allows interoperability between different data catalogues. It can be used to describe data on the web presented in the form of structured data to Linked Data in RDF format. This includes descriptions of keywords, theme, frequency, spatial and temporal coverage. Numerous extensions to DCAT have been developed to include additional properties that are considered relevant by their designers, e.g. DCAT-AP (for public sector data), GEO-DCAT-AP (geospatial properties) or Data-ID (versioning, technical descriptions of datasets).

[7] https://www.w3.org/TR/vocab-dcat/

To accommodate one of the most popular data format on the web, the 'CSV on the Web' Working Group has developed a standard[8] for expressing useful metadata about tabular resources and CSV files specifically. Their goal is to provide a standardised way of ensuring consistency of data types and formats (e.g. uniqueness of values within a single column) for every file, which can provide basis for validation and prevent potential errors.

Finally, schema.org (Guha et al., 2016) is a schema for describing structured data on the web. It is applicable to a wide variety of data formats. It can be used as markup to describe structured content (e.g. tables within web pages) or as a metadata schema describing specific data with a defined list of metadata attributes - for example a dataset[9].

Umbrich et al. (2015) in their work provide an assessment of the quality of open data portals focusing mostly on the metadata aspect. They point out that low metadata quality (or missing metadata) affects both the discovery and the consumption of the datasets.

They propose a list of metrics to evaluate the quality of metadata on a CKAN platform.

- *Retrievability*: "The extent to which metadata and resources can be retrieved."
- *Usage*: "The extent to which available metadata keys are used to describe a dataset."
- *Completeness*: "The extent to which the used metadata keys are non empty."
- *Accuracy*: "The extent to which certain metadata values accurately describe the resources."
- *Openness*: "The extent to which licences and file formats conform to the open definition."
- *Contactability*: "The extent to which the data publishers provide contact information."

These metrics were applied to 82 different CKAN platforms following which the authors concluded that the heterogeneity of metadata across different portals results in difficulties for integration of those portals. They also noticed a steady growth in the number of datasets on a majority of the portals since 2014 and that the predominant format for dataset publication is CSV. In addition, metadata is often missing information about the licence, contact information and format of the dataset (Umbrich et al., 2015).

Existing descriptions of data are often not extensive enough to provide sufficient background for search and discovery processes. As the manual generation of metadata that is sufficient for helping users find the data is time-consuming and often requires domain-knowledge, we explore ways to automate it as much as possible, providing human annotators with hints about the properties of a knowledge base that better match the content

[8] https://www.w3.org/TR/tabular-data-primer/

[9] http://schema.org/Dataset

of certain columns of a dataset. With developments in the Semantic Web community one of the viable direction for automating the process of metadata generation is assigning the dataset with semantic labels. We look into this process in the next section.

2.7 Semantic Labelling

As pointed out in previous section of this chapter, additional metadata could provide itself beneficial for dataset retrieval process. With other works focusing on generation of metadata by the data publisher and supporting them in this process (Koesten et al., 2020); there is also a need for automatic approaches providing context without the publisher, which could place a dataset in a broader sense of other datasets and resources. In this work we are interested in the later. There are various manual and automatic approaches targeted at expanding the metadata describing the dataset and interconnecting it with other resources further deepening out understanding of it. Approaches such as Naumann's propose joining two datasets based on their features, similar to a reverse engineering process which reveals the possible relations between datasets based on constraints identified in the data (Naumann, 2013). Cell-level and column-level analysis enables effective schema matching between datasets. Initially unrelated schemas can thereby be mapped and semantically correct correspondences can be revealed (Rahm and Bernstein, 2001). Integration of seemingly unconnected datasets can provide additional insights to the original data.

Structural analysis (e.g. the number of columns, rows and their data types, along with information about the uniqueness and completeness of the data (Abedjan et al., 2015; Naumann, 2013)) of a dataset for which no external resource analysis is needed can be of use for scoping the content of the dataset. To generate additional insights, topical analysis provides connections to different resources, for example DBpedia, by evaluating the dataset's topic coverage (Fetahu et al., 2014).

Several approaches have been proposed to assign semantic labels to structured data. Some of them focus on tables embedded in web pages (in HTML `<table>` elements (Wang et al., 2012; Ritze et al., 2015); others analyse any type of structured data with a specific focus on tabular or comma-separated data (Mulwad et al., 2010; Syed et al., 2010; Taheriyan et al., 2014; Knoblock et al., 2012). Humans can be involved in the process to achieve better results (Ermilov et al., 2013). Many approaches make use of content descriptions associated with the table (*e.g.*, information in an HTML page (Venetis et al., 2011; Adelfio and Samet, 2013; Wienand and Paulheim, 2014), headers within tables (Wang et al., 2012; Adelfio and Samet, 2013)) or rely on data in textual columns within the table to assign semantic labels (Ermilov and Ngomo, 2016). Others match table rows to KB entities, leaving out of the scope matching the table columns to KB properties (Efthymiou et al., 2017; Bhagavatula et al., 2015). It is important to point out that only a few of the existing approaches propose solutions specifically targeted towards numerical values in structured data.

Numerical Values. Numerical values present different challenges than textual information when assigning semantic labels to columns in tables. An approach targeting specifically the problem of labelling numerical values in structured data was first shown by Ramnandan et al. (2015). They propose an algorithm that learns a semantic labelling function. The authors introduce a list of features, differentiated between those targeted at numerical and at textual values in structured data. They propose testing the distributions of numerical values corresponding to semantic labels based on the idea that the distributions of values for each semantic label are expected to be different (*e.g.*, the distribution of population of cities will be different from the distribution of population density). They used three different tests: the Welch's t-test, the Mann-Whitney U test, and the Kolmogorov-Smirnov test (KS test). Their results show that the latter achieved the best results. Neumaier et al. (2016) and Pham et al. (2016) used similar metrics. Below we present a more detailed dive into those two approaches of which we perform a replicability with the aim of understanding the state of the art within this specific research problem.

Semantic labeling: A domain-independent approach. Pham et al.'s (2016) approach, named DSL (Domain-independent Semantic Labeler), targets semantic labelling of both textual and numerical values by introducing a list of similarity metrics between bags of values. DSL uses a Logistic Regression classifier to infer the semantic type of a new bag of values based on the similarity scores among previously labelled bags. A total of 6 similarity metrics are used as features, each one covering a different aspect of the values. For instance, some metrics apply for only textual data, and others only for numerical data or for both. In Table 2.1 we present a short description of features used to train DSL.

Feature name	Data type	Description
ATT NAME	any	Jaccard similarity on attribute name
TEXT JACCARD	textual	Jaccard similarity on textual values
TF-IDF COSINE	textual	TF-IDF cosine similarity on textual values
NUM JACCARD	numerical	Modified Jaccard similarity for numerical values
NUM KS-TEST	numerical	Kolmogorov-Smirnov test on numerical values
MW HISTOGRAM	any	Man-Whitney test (MW test)

TABLE 2.1: List of DSL features with the data type they are applicable to and description

Bags of values can be mono-type (either numbers only or text only), or multi-type, (mix of text and numbers). In order to assign higher importance to specific features for a given attribute, authors introduce adjustments to their importance. Depending on the distribution of textual and numerical values in bags of values the authors adjust values in a feature vector. The adjusted value is the product of the harmonic mean over the distribution of data in the pair of attributes, and the original value. This step gives a higher importance to numerical similarity metrics over textual similarity metrics for bag of values with numerical data and vice versa. In our running example, values in the `weight` column have more numbers than text, the adjusted value will give more importance to numerical similarity than to textual similarity metrics.

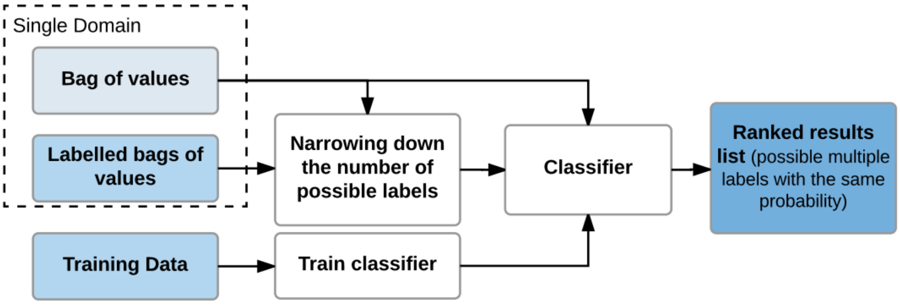

Single Domain
Bag of values
Labelled bags of values
Training Data
Narrowing down the number of possible labels
Train classifier
Classifier
Ranked results list (possible multiple labels with the same probability)

FIGURE 2.2: Workflow in the DSL approach

The workflow of assigning semantic labels is depicted in Figure 2.2. First, the classifier is trained, and bags of values in the training data are stored as a domain data. In the next step the approach narrows the number of possible semantic labels for a bag of values with the use of labelled bags of values that originate from the same domain. In our running example, for the bag {`London, Paris, Warsaw, Madrid`}, candidate properties are {`dbo:city, dbo:capital, dbo:birthPlace, foaf:name,...`}

The trained classifier model takes as an input a bag of values and a set of labelled attributes, and outputs a set of top-k labels for the bag. Labels are ranked according to the probability assigned by the classifier to each label to be appropriate for the bag. Labels with the same probability are considered tied.

Multi-level semantic labelling of numerical values. Neumaier et al.'s (2016) approach, referred to as MSL, targets the problem of assigning semantic labels to the bags of numerical values. Figure 2.3 shows the workflow of MSL approach. The background knowledge graph is a hierarchical clustering over an RDF knowledge base containing information about typical numerical representatives of contexts, i.e., grouped by properties and their shared domain. Each cluster becomes a root node in the graph and is then split into subgroups first based on type hierarchy, and then on shared attribute-value (p-o) pairs. For example, the *height* property could be split into *height of a person* when considering type hierarchy and into *people born in Southampton* when considering p-o hierarchy. The type hierarchy is based on the class hierarchy of DBpedia whereas the p-o hierarchy is built with property-object pairs (excluding rdf:type).

Subgroups in the background knowledge graph are selected based on their distance to the parent node. The subgroup with the furthest Kolmogorov-Smirnov distance is selected in order to assure high diversity of nodes.

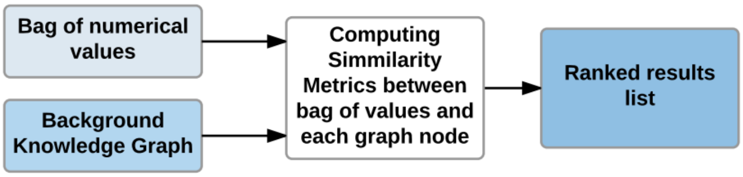

Bag of numerical values
Background Knowledge Graph
Computing Simmilarity Metrics between bag of values and each graph node
Ranked results list

Figure 2.3: Workflow in the MSL approach

The approach computes the Kolmogorov-Smirnov distance between the input bag of numerical values and each node in the background knowledge and aggregates them with 5 different methods for determining top-k closest neighbours (nodes): *property level*, *exact type level*, *root type level*, *all types level* and *p-o level*. The final ranked result list is generated through *majority vote* or *aggregated distance*.

In Chapter 8 we showcase an analysis of the current issues in the area of assigning semantic labels to numerical columns and we propose our approach tackling this problem. Our approach builds upon these by analysing numerical columns in two ways: first, in terms of the similarity (measured in terms of the KS test) of distribution of values with respect to those of properties in a target KB; second, by calculating the relative difference between numerical values in a column and numerical values of properties associated to entities of the type identified from the column that holds the main entities of the table. The main observation is that tables often include a column that identifies the entities described by the table (called the *subject column*), while the rest of the columns hold values of properties linked to those entities (or *objects*). Venetis et al. (2011) reported 75% of the tables in their corpus of web tables had a single main subject column, and that the accuracy of semantic labelling increases when first determining a subject column.

Subject Column Identification. To determine which of the columns in a table is a subject column, Venetis et al. (2011) suggest two methods: taking the left-most column that is not a number or date column, or treating it as a binary classification problem. They propose learning a classifier for subject columns with features that are dependent on the name and type of the column and the values in different cells of the column. Wang et al. (2012) and Ermilov and Ngomo (2016) proposed similar characteristics of the column: (1) the *connectivity* of a column (*i.e.*, how it is connected with other columns of the table by means of properties mapped to the KB) and (2) *support* of the column (*i.e.*, ratio of cells disambiguated to KB entities in the column). A combination of both connectivity and support is then used to determine which of textual column is the most likely the subject column. In our work we focus on labelling columns with numerical data, assuming a subject column has been previously identified. Further details are provided in Chapter 8.

Semantic Labelling Benchmarks. Few benchmarks were proposed to test the efficiency of the semantic labelling process. The benchmark dataset created by Limaye et al. (2010) comprises 400 tables mapped to DBpedia and YAGO at instance– and schema-level. Efthymiou et al. (2017) introduce a benchmark including 485K tables from Wikipedia, which were mapped to DBpedia by leveraging the links in their label column. Neither of these two benchmarks was suitable for our experiment, which aims to map columns to properties. Instead, the dataset in (Limaye et al., 2010) contains cell-to-entity mappings, while that in (Efthymiou et al., 2017) row-to-entity. T2D (Ritze et al., 2015) is a set of 1,748 tables[10] with schema and instance-level mappings to DBpedia. However, it does not contain a sufficient number of numeric columns to be suitable for our case (the large majority of disambiguated column were textual columns). Therefore, we decided to create a new benchmark, that we detail in Section 8.8.1.

[10] http://webdatacommons.org/webtables/goldstandard.html#toc0

Chapter 3

Methodology

3.1 Definitions

In Information Retrieval (IR), documents are transformed into their representation (e.g *Vector space model* or *Boolean model* (Manning et al., 2008)) or index themselves depending on the applied retrieval strategy. In the context of dataset search documents (here datasets) are retrieved based on the representation in the metadata accompanying the data which is often more descriptive of the data than the data itself. We define a dataset as a collection of tabular data. Our definition of dataset results from the fact that most data that can be found on the web is tabular, rather than published as RDF or Linked Data (Umbrich et al., 2015). Dataset in our work is defined as structured (tabular) data represented by single data matrix (e.g. a Web table or a CSV file). A dataset consists of columns and rows (see the relational tables in (Lehmberg and Bizer, 2016)). The first row is a header row which describes the values of a given column (except a special case in which there is no header row). Each column consists of one type of information, whereas each row represents a set of related data. Every row in the table has the same structure. Each dataset should be accompanied by a metadata holding all the information necessary to it's interpretation, describing it's context, rights of the potential data reuse, etc.

We formally define a *dataset*, and *metadata*, and further define *search log*, *query* and *data requests* analysed in this study as:

Definition 3.1. Dataset
A dataset $DS = (H, D)$ is a tuple consisting of a header H and data D, where:

- the header $H = (h_1, h_2, ..., h_n)$ is a vector of size n which contains header elements h_i. In special case of dataset without a header some of the values in this vector could be null.

– the data D is a (m, n)-matrix consisting of n columns and m rows.

$$D = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \multicolumn{5}{c}{\dotfill} \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Definition 3.2. Metadata
A dataset metadata M is a set of data that describes and gives information about a dataset. It can contain various types of metadata as shown by Beyene (2017), such as information on the dataset content, it's structure, technical details (e.g. format), provenance, etc. Schema.org or DCAT are sample vocabularies facilitating the metadata about datasets. Sample DCA-AP (DCAT extension for public administration) can be seen on an simplified example below:

```
@prefix dcat:   http://www.w3.org/ns/dcat# .
@prefix dct:    http://purl.org/dc/terms/ .
@prefix foaf:   http://xmlns.com/foaf/0.1/ .
@prefix owl:    http://www.w3.org/2002/07/owl# .
@prefix rdfs:   http://www.w3.org/2000/01/rdf-schema# .
@prefix schema: http://schema.org/ .
@prefix time:   http://www.w3.org/2006/time# .
@prefix xsd:    http://www.w3.org/2001/XMLSchema# .

http://data.europa.eu/88u/dataset/9fpg-67a4
        a                  dcat:Dataset ;
        dct:description    "This file contains \"categorisation data\" from the
                           ERDF/ESF/Cohesion Fund programmes as adopted.  [...];
        dct:issued         "2016-03-14T09:45:37.000Z" ;
        dct:modified       "2021-11-15T06:56:26.000Z" ;
        dct:publisher      [ a           foaf:Organization ;
                             foaf:name   "European Commission - Directorate-General for
                                         Regional and Urban Policy"
                           ] ;
        dct:title          "ESIF 2014-2020 categorisation ERDF-ESF-CF - planned"@en ;

        dcat:distribution  http://data.europa.eu/88u/distribution/
                            dc9b22a1-2aaa-43c4-945e-6d3705c11833 , [...];
        dcat:keyword       "social inclusion"@en , "education infrastructure"@en ,
                            "tourism"@en , "fire"@en , "labour market"@en .

http://data.europa.eu/88u/distribution/3f0e632c-796b-4f77-923c-34c70a45d728
        a               dcat:Distribution ;
        dct:format      http://publications.europa.eu/resource/authority/file-type/CSV ;
        dct:identifier  "Data as csv" ;
        dct:title       "Data as csv" ;
        dcat:accessURL  https://cohesiondata.ec.europa.eu/resource/9fpg-67a4.csv .

http://data.europa.eu/88u/record/9fpg-67a4
        a                  dcat:CatalogRecord .
```

Definition 3.3. Search (a.k.a. Transaction) Log

As defined by Jansen (2006), "A transaction log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine.". Below we present a sample search log from Matomo Web Analytics tool used by the European Data Portal.

```
{"lastActionTimestamp": 1522619552,
"location": "Paris, France",
"referrerName": "Google",
"browserName": "Chrome",
"actionDetails": [{
    "subtitle": "london population",
    "bandwidth_pretty": "0 M",
    "pageId": "2232972",
    "generationTime": "1.19s",
    "serverTimePretty": "Apr 1, 2018 23:47:32",
    "generationTimeMilliseconds": "1193",
    "timestamp": 1522626452,
    "type": "search",
    "title": "Site Search",
    "siteSearchKeyword": "London population"}],
"referrerSearchEngineUrl": "http://google.com",
"firstActionTimestamp": 1522619552,
"referrerKeyword": "Keyword not defined",
"daysSinceLastVisit": "0",
"deviceType": "Desktop",
"visitorType": "new",
"searches": "1",
"actions": "1",
"idVisit": "556",
"interactions": "1",
"referrerTypeName": "Search Engines",
"operatingSystemName": "Windows",
"portal_version": 1}
```

Definition 3.4. Query

A query q is a string represented by a list of keywords that a user issues to a search engine in order to satisfy their information needs. A result of issuing a query is a list of datasets ordered by their relevance to the query with the aim of satisfying the information need of the user. Sample query – 'London population' – could be seen in a search log example presented above.

Definition 3.5. Data Request

A data request is an unstructured description of a dataset or datasets made by a citizen directly to open data portals, it can be collected through web forms (*e.g.*

data.gov.uk or datos.madrid.es), email (*e.g.* danepubliczne.gov.pl), or regular/quarterly platform meeting inviting the community (*e.g.* open.wien.gv.at). A sample data request could look as follows: *I wish to find out up to date figures of numbers of adults with moderate, severe and profound disabilities (in particular learning disabilities) who are currently working in the UK, either part time or full time; where in the UK they work and in what numbers; and at what occupations.*

Metadata can be used to improve user understanding of a dataset during the selection process on a search engine result page (SERP) or after entering a detailed dataset page before downloading it for further analysis. The metadata also includes additional information regarding the content of the dataset (e.g. the fact that a given column consists of European countries) to search functionalities to improve their capability to match queries to specific datasets and rank them (e.g. in case of a query regarding European countries).

In our work we identified that specific types of information about a dataset are of higher importance than others. We recognised that data is often described (and searched for) using specific search criteria: location and time frame with specific granularities are primary criteria along with other data-specific dimensions dependent on the domain, which could be represented by the datasets' columns descriptions.

In the process of the creation of metadata we can benefit from addition of statistical information introduced previously in the databases community or in different search verticals but also from full or partial semantification of the data of which the majority is not yet available in Linked Data format. As discussed in in Chapter 2 various approaches have been proposed to link the contents of structured data files to the Web of Data. They mostly focus on textual information, however, in many datasets values are often *numerical* (Mitlöhner et al., 2016). Numerical values raise different challenges than entity matching from strings. In our work we analyse existing techniques in the field of assigning semantic labels to numerical values in CSV files and propose a novel approach for matching them to the knowledge base.

3.2 Problem Statements & Research Questions

The aim of this work is to understand the dataset search and its characteristics to improve the searchability of datasets published on the web. Firstly, we look into the search log data and data requests to see what kind of information is the most relevant while searching for datasets. Secondly, we focus on the problem of generation of missing descriptive metadata about numerical columns in datasets. The aim of this work is to provide information of the state of the dataset search as a vertical, potential metadata expansion, and directions for further improvements to information retrieval process and user experience.

The main **research question** of this work is *how can we improve searchability of datasets that are published on the web?* We plan to answer this question by focusing on the following areas:

- [RQ1] Understanding Dataset Search; How do users search for data? What search strategies do they use?
- [RQ2] What information should be included in the metadata to improve the dataset search process?
- [RQ3] How to automate the generation of metadata? In this work we focus on one of the most under-investigated areas: semantic labelling of numerical columns.

In order to answer our two first research questions ($RQ1$ and $RQ2$) we conduct a four part analysis: in what context users do the search (Chapter 4), qualitative study on free form data requests to understand what are the controls they need (Chapter 5), quantitative query log analysis to five data portals to understand how the search for data is conducted (Chapter 6), and quantitative session log analysis of how they use the controls currently provided (Chapter 7). As for the last research question ($RQ3$) we look into the issue of expanding the existing metadata with automatic methods. We investigate existing approaches to assign semantic labels to numerical columns which we follow by our own approach tackling this issue.

3.3 Research Methodology & Experiment Data

In order to answer the research questions posed in this work we conduct a number of experiments. In terms of first two research questions ($RQ1$ and $RQ2$) we perform a **qualitative** and **quantitative** data analysis with use of logs from five open data portals with different levels of depth. We start from the quantitative analysis of users (Chapter 4) of three of the portals focusing on their prerequisites when coming to the data platform (such as equipment and tools they use to access the portal or if they are new or a returning user of a portal), next to dig further into the users we conduct an in-depth thematic analysis of data requests submitted to one of the portals with an aim of finding data which will fulfil their needs (Chapter 5). Diving more into behaviour of users on open data portals we conduct a quantitative analysis of query log – queries submitted to the open data portals – with aim of understanding current state of search on such portals and reviewing findings from previous chapter where users were asking for data in a more free form (Chapter 6). Concluding this part of the presented work we conduct a more high-level analysis of users behaviour on on of the data portals with use of session logs, showing not only how users query for data but how they use the overall search environment (Chapter 7). Looking at the last research question ($RQ3$)

we analyse a problem of creation of additional metadata which could be utilised by a search functionalities in the future. After analysis of different approaches we recognise that semantic labelling if numerical values in one of the most under-investigated areas. We conduct an in-depth analysis of this problem and approaches which were proposed up-to-date to solve this issue. This allowed us to understand the scope and the existing problems. Next, we proposed our approach for assigning semantic labels to numerical columns and we compared the achieved results against state of the art approach in this domain. Methodology details for each part are detailed in the corresponding chapters.

Below, we present a overall description of the data used in our experiments, from the open data platforms used throughout the experiments regarding analysis of the users of such platforms to the knowledge bases used in our experiments in the problem of assigning semantic labels to numerical columns. The following datasets were used due to the availability of the data, here query and session logs. The knowledge bases were selected as they are state of the art benchmark in the research domain for the type of a problem we proposed our approach for.

UK Governmental Data Portal[1] (DGU) is a repository of data published by central government, local authorities and public bodies available to everyone. One can search the portal to find data on topic such as student loans, food standards or road traffic statistics etc., in case of data being not available, they can submit a *data request* for this data to be published.

The Office For National Statistics[2] (ONS) is a UK's governmental body which collect, analyse and disseminates statistics about the UK's economy society and population. The main difference to other data portals is the analysis of the data published along the data.

Canadian government Open Data portal[3] (CAN) and **Australian government Open Data portal**[4] (AUS) are repositories of data published by central governments, local authorities and public bodies available to everyone in the similar manner as in case of DGU.

The European Data Portal[5] (EDP) crawls national data portals (such as DGU) and harvests the metadata in order to show one central point of access to open data from 36 European countries. Users can search this portal in different languages, however, as the data is often not translated to other languages it might not present accurate results. Session data for this portal is collected with use of Web Analytics package. Further details on the log structure will be provided in Chapter 5.

[1] https://data.gov.uk/

[2] https://www.ons.gov.uk/

[3] https://www.open.canada.ca

[4] https://www.data.gov.au

[5] https://www.europeandataportal.eu/en

Wikidata[6] is a community-created knowledge base of Wikipedia. It is composed of structured data presenting items (e.g. concepts or objects) in a form of triples. Each triple consists of subject, property and object. It stores data in multiple languages. Each item is assigned an unique identifier. It can be queried using SPARQL as the query language. Since it launch in 2012 to the date it gathered 90 million data items.

DBpedia[7] is an automatically extracted structured data knowledge base originating from Wikipedia. Analogically to Wikidata it consists of data in form of triples and can be queries with use of SPARQL query language. It contains links to other datasets in the Linked Open Data cloud. DBpedia contains descriptions of 38.3 million things (across 125 languages).

[6] https://www.wikidata.org/wiki/Wikidata:Main$_Page$

[7] https://wiki.dbpedia.org/about

Chapter 4

Users of Data Portals

In this chapter and the subsequent Chapters 5, 6 and 7 we present our work on understanding search for datasets, here, focusing on the high-level information we have on those users. We focus on our first research question: *How users search for data? What search strategies they use?* In order to better understand the dataset search we first analyse who are the users of the portals and how do they access it. This chapter presents the analysis of such context in which users access the data portals. We look into the type of device and browser users are using to access the portal, the time when they accessed it, how they got to it (e.g. via search engine, social media etc.) and weather they are new or returning users to the portal. Some of the results presented in this chapter appeared in (Kacprzak et al., 2019).

4.1 Methodology

In this chapter we analyse the general information on users of data portals. User statistics were available for the users of the following portals: DGU (data.gov.uk), ONS (the Office for National Statistics) and EDP (the European Data Portal). The portals show the information about a specific regions of the world, in our case it is UK and Europe. The users of those portals can be a vast range of users who are interested in the data on those areas, that being citizens, or companies. More information on the who are the users of the DGU portal was available in user questionnaire submitted by user while submitting a data requests - details are provided in Chapter 5. The presented results on user statistics were collected via Google Analytics (in case of DGU and ONS) and accessed through their online interface, and via Matomo Web Analytics (in case of EDP) through log analysis. In case of Google Analytics only aggregated version of the data is being made accessible. The data shows the users information for the time period of 30/01/2013 - 31/08/2016 in case of DGU; 28/02/2016 - 31/08/2016 in case of ONS; and 01/04/2018 - 30/06/2020 in case of EDP, which is the same time-period as for the following chapters of this work. Below we present the statistics on the devices and browsers that were used by the users of the data portals, the time they access the portal, the channel that lead them to accessing the portal. Finally, we look into the statistics regarding new and returning users of the portals. The presented analysis is the high-level overview of the users of the open data portals, providing a background information to the further analysis presented in the following chapters.

4.2 Devices

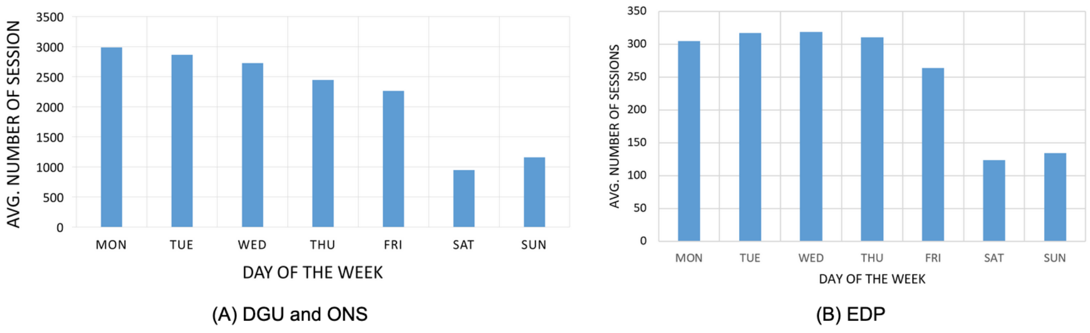
Table 4.1 shows how people access the portals. We distinguish between *desktop computers*, *mobile devices*, and *tablets* and list the relevant share of sessions for each of them, as well as the average number of pages viewed per session and the average session duration. For the EDP portal we have additional category ‘Others’ which was present in the data. This category includes devices such as gaming consoles, portable media players, tv etc. An overwhelming majority are desktop computers ($\sim$ 86% on average for all portals portals), compared to mobile devices ($\sim$ 9%) and tablets ($\sim$ 4.5%). Both the number of pages viewed and session duration are highest for desktops. We believe the high percentage of desktop users can be explained by the fact that data search is mostly a working-hours related activity; this is confirmed by the time of the day people access the portal (see below). Looking for datasets is a first step in a much more complex workflow, in which a relevant dataset is subsequently downloaded and then inspected and visualised using exploratory data analysis tools. These activities are typically performed on desktop computers due to their larger screens and additional processing power.

Device	% Sessions			Pages viewed			Session duration		
	DGU	ONS	EDP	DGU	ONS	EDP	DGU	ONS	EDP
Desktop	79.81	90.95	88.34	3.42	3.04	7.77	02:35	03:41	04:59
Mobile	12.93	4.94	9.03	1.78	2.65	4.50	00:57	02:14	02:28
Tablet	7.27	4.11	2.07	2.24	2.29	5.83	01:22	01:53	03:18
Other	N/A	N/A	0.56	N/A	N/A	5.34	N/A	N/A	03:07

TABLE 4.1: Devices used to access data portals. The parentage of sessions, pages viewed per session and session duration.

4.3 Time of access

Users are mostly active during weekdays, as can be seen in Figure 4.3. Monday has the highest level of activity for DGU and ONS, which falls slightly every day until Friday, to reach the lowest point on Saturday and grow slightly again on Sunday. Similar behaviour



(A) DGU and ONS

(B) EDP

Figure 4.1: Distribution of sessions with search per weekday

can be observed for EDP with highest number of sessions in the middle of the week, followed by a drop over the weekend. Activity during weekends is approximately half or a third of that during week days. Users access the portals during working hours (9am to 6pm) and issue most queries between 9am to 11am for DGU and ONS and between 10am and 4pm for EDP. This pattern, in combination with the prevalence of desktop computers, further suggests that dataset search is a work time activity.

4.4 Channels

As can be seen in Table 4.2, the majority of users (62.32% for DGU; 74.33% for ONS; 64.99% for EDP) reach portals through the result page of a web search engine (a scenario which we refer to as *external*); by accessing the portal directly through its URL (*direct* - 14.3% for DGU; 16.72% for ONS; 23.47% for EDP); or by following a link from a different website that is not a social network or a search engine (*referral* - 9.62% for DGU; 8.52% for ONS; 10.23% for EDP). Less than 1% of visits are generated through social networks. The *Other* row in the table groups together traffic coming from email links, advertising, and paid search. The high share of externally driven traffic suggests that most users either resort to common web search engines to search for data and are then directed to the portals, or they use web search engines as proxies - this means, instead of going directly to for example data.gov.uk and issuing a search there, they start with a regular web search engine with additional keywords to their queries, for example "data UK" which lead them to a portal. We discuss this type of query in the next chapters of this work.

4.5 Browsers

The majority of data search sessions used Chrome (41.35%), followed by Internet Explorer (IE) (30.50%), Safari (13.97%), and Firefox (9%). Interestingly, Internet Explorer

Channel	% Sessions		
	DGU	ONS	EDP
External	62.32	74.33	64.99
Direct	14.30	16.72	23.47
Referral	9.62	8.52	10.23
Social	0.86	0.43	1.30
Other	4.30	<0.01	<0.01

Table 4.2: Channels through which users access portals

(under this name we mean combined data for Internet Explorer and Microsoft Edge browsers) was much less popular among EDP portal users. In case of the UK based portals, compared to general web browser usage,[1] both worldwide and from the UK, we note a higher share of IE users by almost 10%. As discussed earlier, people seem to be accessing data portals during weekdays and during office hours. In corporate and government environments, the use of IE is still widespread, which might help explain its relatively high popularity in the search logs. As for the data from EDP portal we hypothesise that lower popularity of Internet Explorer might be a result of the EDP portal being more often accesses by a wider audience (such as researchers, scientists, social workers etc.) than other portals and the data being relatively newer (see Table 6.2 for comparison of timeframes) – popularity of IE browser was decelerating with time[2]. In addition we noticed that out of almost 13% of Other devices for EDP portal, around 9.35% are visits from mobile devices. This can be connected with the fact that Internet Explorer is less popular and therefore strengthening out hypothesise of wider audience of this portal or that this portal is more often that DGU and ONS explored in contexts different than to find, download and analyse the data.

Browser	% Sessions		
	DGU	ONS	EDP
Chrome	37.95	44.76	55.98
Internet Explorer	29.87	31.13	9.28
Safari	16.09	11.86	5.19
Firefox	10.93	7.18	16.80
Other	5.16	5.07	12.75

Table 4.3: Browsers used to access portals

[1] Using statistics from 2013 to 2015, from http://gs.statcounter.com/

[2] http://gs.statcounter.com/

4.6 New and returning users

Table 4.4 shows the the percentage of new and returning users and compares the two cases in terms of the average number of pages viewed and average session duration. Our main observation is that returning users view on average more pages and engage in longer sessions. Query log analysis from other verticals do not consider this metric except for Weerkamp et al. (2011), which reports 7% returning users out of its 7 million sessions. In terms of new users, they overall spend less time and view less pages which might suggest new users might need more guidance through their data discovery and interpretations journey. This is especially visible in case of DGU and somewhat EDP. We believe these differences suggest that users return with the intent to work with data and spend more time in assessing the relevance of their search results. The higher proportion of returning users to the ONS portal is probably a function of the reputation of the ONS as an established, authoritative source of data, compared to the much newer initiative around data.gov.uk. This was also confirmed by interviewees in (Koesten et al., 2017), who said that they trust the ONS to deliver high-quality data that is useful in various scenarios. This trust is probably the case also for new users of the ONS portal, who on average spend more time on this portal than new DGU users.

User	% Sessions			Pages viewed			Session duration		
	DGU	ONS	EDP	DGU	ONS	EDP	DGU	ONS	EDP
New	76.47	58.38	77.47	2.63	2.92	6.84	01:38	02:50	04:00
Returning	23.53	41.62	22.53	4.54	3.1	9.43	04:08	04:32	07:10

Table 4.4: Percentage of new versus returning users per portal

4.7 Conclusion

Our findings based on the analysis of user search behaviour when accessing data portals suggest that dataset search is a working-hours related activity (working professionals, students, academics etc.). We note that the percentage of returning users for ONS portal with 1.5 times more compared to other portals. This could mean that sources of reliable information draw users towards themselves, encouraging returning to the portal and driving more interaction from the get-go (which is visible in similar number of pages viewed and closer session duration for new and returning users of ONS portal). ONS portal showcases analysis and visualisation of the datasets, which could have been another factor in time spent. However, combined with the previously reported lack of context when searching for datasets this could suggest potential direction for improvement for other data portals. We noticed differences between UK based portals and European Data Portal in terms of browser and device used which we hypothesise is due to more diverse audience of this portal. This chapter showcases a preliminary analysis to the work presented in next chapters, which will show more in-depth analysis of a dataset search based on data requests, queries and sessions.

Chapter 5

Data Requests

In this chapter we present an in-depth thematic analysis data requests submitted to the DGU portal. Data requests are unstructured descriptions of datasets made by citizens directly to open data portals, they can be collected through web forms (*e.g.* data.gov.uk or datos.madrid.es), email (*e.g.* danepubliczne.gov.pl), or regular/quarterly platform meeting inviting the community (*e.g.* open.wien.gv.at). In this work we focus our analysis on the title and description of the data request with the main aim to answer following questions: *How do people request data when they are allowed to formulate their information needs with no restrictions? What properties do they consider the most important and how are these used?* This analysis was performed on different levels, defined as *data attributes* and *request context* level, which are discussed in detail below. In addition, we present statistics over the different data request's multiple choice fields. These give us an overview of the people who issued the requests, of the most popular themes, and of the intention of use of the requested data to give context to the findings presented in Sections 5.6 and 5.7. The presented work also provides more insides into the users of open data portals presented in previous chapter of this work. Providing more detailed information on their background and aims. It is however, important to notice that the the population of users accessing data portals as presented in the previous chapter is different than the ones submitting the data requests, shown in this chapter. Some of the results presented in this chapter appeared in (Kacprzak et al., 2019).

5.1 Methodology

Data requests are a representation of information needs submitted by users of a data portal in order to get a specific dataset that they usually could not find. We used a set of 1600 data requests from the United Kingdom governmental open data portal (DGU) which are partially available as a dataset on the portal[1]. Requests represent information

[1] https://data.gov.uk/dataset/data-requests-at-data-gov-uk

needs for data. As they are submitted in natural language and are much longer than search queries they contain additional information of the task that people aim to do with the requested data. After contacting the portal owners we were able to acquire 800 additional requests that were tagged to remain confidential, however only publicly available elements of the data requests are included in the examples cited in this work.

5.1.1 Structure

Data requests are submitted to the portal via a semi-structured contact form available on the portal website by users who aim to satisfy their information needs. The form is targeted at citizens who can ask there for particular government datasets to be made public, and if possible, to have an open licence attached to it. Users are required to provide a description of the data along with a reason and context for their requests. This results in long descriptions of the data that is needed - providing us with additional context of the connected task. A sample description of the data request is for instance: *I wish to find out up to date figures of numbers of adults with moderate, severe and profound disabilities (in particular learning disabilities) who are currently working in the UK, either part time or full time; where in the UK they work and in what numbers; and at what occupations.* We present an overview of the structure of all fields in a request form in Figure 5.1.

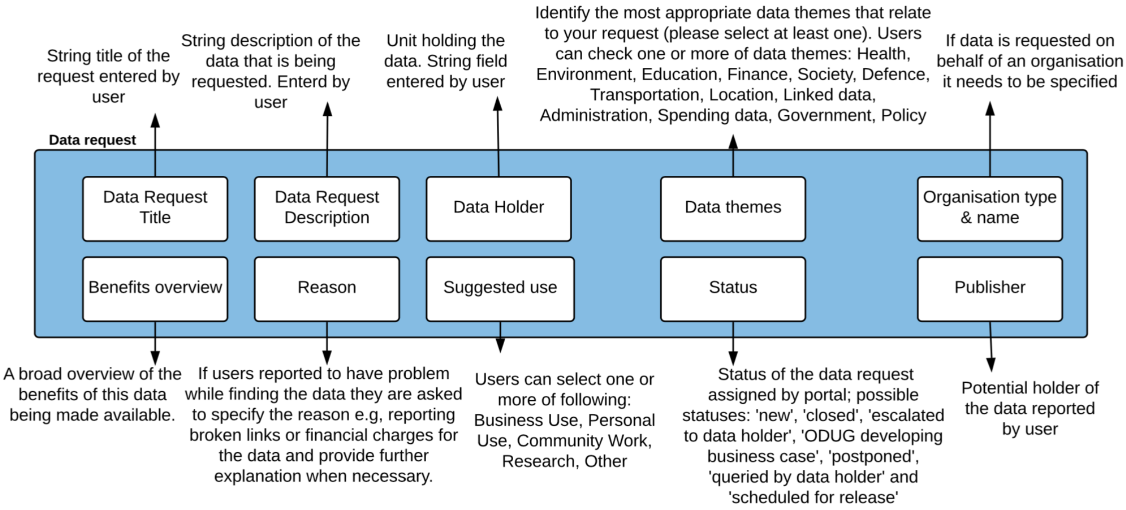

String title of the request entered by user
String description of the data that is being requested. Enterd by user
Unit holding the data. String field entered by user
Identify the most appropriate data themes that relate to your request (please select at least one). Users can check one or more of data themes: Health, Environment, Education, Finance, Society, Defence, Transportation, Location, Linked data, Administration, Spending data, Government, Policy
If data is requested on behalf of an organisation it needs to be specified
Data request
Data Request Title
Data Request Description
Data Holder
Data themes
Organisation type & name
Benefits overview
Reason
Suggested use
Status
Publisher
A broad overview of the benefits of this data being made available.
If users reported to have problem while finding the data they are asked to specify the reason e.g, reporting broken links or financial charges for the data and provide further explanation when necessary.
Users can select one or more of following: Business Use, Personal Use, Community Work, Research, Other
Status of the data request assigned by portal; possible statuses: 'new', 'closed', 'escalated to data holder', 'ODUG developing business case', 'postponed', 'queried by data holder' and 'scheduled for release'
Potential holder of the data reported by user

FIGURE 5.1: Data request fields and their descriptions; Description obtained from the data request form on DGU. This form is no longer used. The difference between 'data holder' and 'publisher' is unclear, both fields were used in the same way by the users of the system.

5.1.2 Pre-processing

Data requests in our sample were selected manually, we excluded requests that do not define a clear data need or that require complex data analysis. In addition, we filtered out the same request that were accidentally submitted twice as a separate requests to the portal. Our analysis was conducted over a set of 200 data requests which were randomly selected and met our inclusion criteria.

5.1.3 Qualitative Analysis - Thematic analysis

We analysed 200 data requests qualitatively, using thematic analysis - a method to identify patterns or themes within qualitative data (Robson and McCartan, 2016). Coding was done using NVivo (version 11), a qualitative data analysis package. Two of the authors of the study individually coded a sample of the data requests inductively (Thomas, 2006), compared code lists with each other and discussed conflicting results with two senior researchers. This process was repeated twice until there were no conflicting codes between the two researchers and there were no new emerging codes. These were then used to code the remaining data requests. We grouped emerging codes related to the *attributes* of the data and those related to the *structure* of the request into these two high level categories. For each of the categories we applied two layers of coding (Robson and McCartan, 2016). The *data attributes* layer allows an understanding of how users are talking about data when describing information needs to another person (the receiver of the data request is an employee of data portal). These are grouped into subsets of: geospatial content, temporal content, restrictions on the requested data (for instance specific formats or licences), mentions of the required granularity. The *request context* layer includes the prevalence of common features to get an overview of the composition of the data requests. This includes mentions of expected representation and structure of the data, the unit of interest (whether a data point, a dataset or the result of an analysis is requested), rationale for the data request or mentions of quality issues with existing datasets with a request for the same data, but in better quality.

5.2 Organisation type

When making a request, users could self-report the organisation they belonged to. Table 5.1 shows that the largest group issuing data requests were individuals, making up over 45% of all requests. The next largest groups were users requesting data for academic and research purposes (15%) or users representing small to medium businesses (13%).

Organisation type	% of requests
Individual	45.43
Academic or Research	15.27
Small to Medium Business	12.81
Start up	7.79
Large Company (Over 250 employees)	7.30
Public Sector Organisation	6.56
Voluntary sector or not-for-profit organisation	4.84

Table 5.1: Possible options to select in order to define the type of organisation that is requesting the data

5.3 Suggested use

Users were asked to specify how they would use the data that they requested. Table 5.2 shows the list of options users could choose from. Research was the most popular declared use of the data (52%) which, combined with the fact that only 15% of requests were declared to be made on behalf of research and academic institutions, suggests that data is used for non-academic research purposes. Taking into account the high proportion of requests made by individuals (45%), much fewer - 24% of the requests - were declared to be for personal use. This indicates that individuals may look for data for business use, which was the second most popular use option.

Suggested Use	% of requests
Research	52.12
Business Use	37.46
Personal Use	24.52
Community Work	13.43
Other	8.09

Table 5.2: Options for suggested use of the data

5.4 Request motivation

In Table 5.3 we present the list of motivations that users could choose from when requesting the data. The inability to find the required data is the most popular reason: this justification is given for more than 40% of the requests. It is not possible to determine if the data was available and users could not find it, or if it was indeed missing from the portal. However, this could be an indication that portals which offer search need to improve their search functionalities; this is supported by (Koesten et al., 2017). We analyse the ways users talk about data in their data request (in Section 5.6) to find commonalities pointing to relevant areas of improvement in such search functionalities. The second most popular reason given by users for issuing a data request was to request existing, published data in another format more suitable for their purposes. Other reasons for issuing data requests (e.g. financial charges for the data, broken links, or restrictive licences) were less frequent.

Data Theme	% of requests
Not able to find the data	41.75
The data is published but not in a format I can download and use (e.g. only displayed on-screen or only downloadable as a PDF rather than CSV)	8.15
The data is supposed to be published but the download links don't work	5.03
The data is not up-to-date	4.78
A version of the data is published but I need a different version	3.99
There are financial charges for the data	3.74
The data is available but the licensing terms are too restrictive	2.82
The data is subject to restrictions because of personal confidentiality	2.21
The data is subject to restrictions because of commercial confidentiality	1.59
Other	25.94

Table 5.3: List of reasons for making a data request with percentages of their popularity

5.5 Time of issuing the request

Each data request contained a time stamp of the date and the exact time the data request was issued. We saw that majority of requests are done on weekdays (more than three times in comparison to weekend days). We also noticed that the majority of requests were issued between 9*am* and 6*pm*. This supports the results from user activity in Chapter 4, however, both groups represent different samples of users.

The thematic analysis resulted in two main categories describing different approaches of understanding the data requests. In Section 5.6 we describe the different attributes that were prevalent in the requests in order to present a comprehensive picture of the request content. For example geospatial or temporal information, format, license, etc. In Section 5.7 we present an analysis of the request context, in which people express content beyond the actual data they are looking for. This included for instance representation of motivation, comparisons, references to other datasets, analysis or specific questions they want an answer to.

5.6 Data Attributes

In this section we examine four data attributes that emerged as prominent themes from the data requests: geospatial information, temporal information, restrictions and information about granularity.

5.6.1 Geospatial information

(n = 77.5%) of requests included some reference to geospatial information at varied levels of detail and scope. They were asking either for information about several nations or larger areas such as the whole of the UK. In contrast to that many requests included specific points of location, such as a borough, a street or even a specific address, as can be seen below.

> *Q1: Would you provide me with any groundwater level data you have for the Preston area centred on grid reference SD 546 291?*

> *Q2: Number of yearly conviction for all computer misuse offences in England and Wales for each year since 2006, as defined under the Computer Misuse Act 1990.*

Some users request location in specific granularity, as can be seen in Q3:

> *Q3: Could I request a full dataset of all the UK speed limits per road.*

Location is expressed differently by different users and by their respective information need. We found geospatial information referred to as country or city names, ISO codes, abbreviations, latitude/longitude, grid references or other specialised identifiers. On the other hand people also used very vague terms, such as “overseas”, “near to” or “surrounding area of”. Some users do not seem to know what data is available and therefore try to narrow the scope of the location rather than specifying an exact location. This range of behaviours can be seen in quotes Q4 and Q5:

> *Q4: I am looking for shapefiles on general environmental data near Ferndale (Wales)*

> *Q5: I would like to have access to all the data available up to nowadays regarding fish (where as it is monitoring, surveys, or any other type of data collected regarding fish) for Beane river downstream (lat 51.806014; long -0.066997) upstream (lat 51.981001; long -0.094448)*

People often expressed geospatial needs or requirements by defining the boundaries of the area they are interested in (e.g. "London and surrounding areas"; "from Richmond to west Thurrock". They do this either by defining an area if a "name" is known, or by expressing borders between which their area of interest lies in.

Expressions of geospatial information in the requests are complex and show a large variety. This is partially due to the fact that there is no standard way of expressing geospatial information in natural language. There are also many domain specific geospatial boundaries in use which in addition have changed over time (such as currently unused historical boundaries). The way that we record administrative boundaries has changed and to make historic comparisons people search for a comparable area (e.g. Q6).

Some mentions of locations were focused on locations that are not directly understood as geographical areas. This can be seen in Q7, where a user is requesting information about specific zones.

> *Q6: I am looking at British voting patterns across the past three General Elections of 2005, 2010 and 2015 and comparing the vote shares of the key political parties [..] the shape of parliamentary constituencies changed from '05 to '10. I need both maps to sort data from one to the other. At the moment, I am unable to find the data for the previous, 2005, voting parliamentary constituencies anywhere online.*

> *Q7: I want to do some simple mapping of flood risk and I the require latest flood zone data (zone 2 and 3, and flood defences) to import into a GIS. If possible historic flood zone data from a few years ago (1-10 years ago) would be great to offer a comparison for analysis.*

There is a range of geospatial information needs represented in the requests. From what we could infer from the requests users often aim to obtain data about specific locations to inform their decisions, to integrate the information in an analysis that is focused on one ore several areas, or integrate the data into an existing service or application:

> *Q8: I would like to investigate the relationship between weather conditions and occurrence of potholes. For this study, data on historic pothole occurrence is needed. The last 10 years of data for the occurrence of potholes Birmingham area is requested.*

> *Q9: I have been interested in the split of EU voters by region, however I feel that a more useful statistic would be "by place of birth" as some people vote away from their home in the UK or overseas. is this split possible?*

> *Q10: I am doing a ground investigation report on the jubilee line extension investigation carried out during 1990 and would find the lidar data helpful for my report.*

Note the contrast between the high number of requests including geospatial information and the comparatively low number of keyword queries containing them. This suggests that text search boxes are not appropriate for searches that include geospatial parameters.

5.6.2 Temporal information

Looking at temporal information (which we defined as every mention that includes a reference to a unit of time) – which was represented in (n = 44%) of the requests – among which majority (87.5%) of statements refer to the time period covered by the data, meaning the temporal boundaries of interest. These were expressed either from a point in time to the current date, or between two boundaries, or requesting the most up-to-date data on a topic. These were represented in different levels of granularity. Other temporal information referred for instance to temporal information in the dataset, such as the age of a person or the time of an event.

These requests illustrate statements from a point in time to the present:

> *Q11: The period 20/5/2016 until most recent.*

> *Q12: Number of deaths in the last 20 years where cause of death has been certified as cancer. Counts are required by age of death.*

These requests are an example of two exact boundaries specified in the requests:

> *Q13: I would like to see all comprehensive school terms and holiday dates from every council in England for 2015/2016 and 2016/2017*

> *Q14: I am requesting the microdata of this survey from year 2012 to 2014 (smoking and drinking habits). I am working in a research looking at smoking prevalence by birth cohort, age and year. I am using the General Lifestyle Survey from 2000-2011 and would like to continue the sequence until 2014*

The following examples show temporal information, requesting the most up-to-date time period in which the data was recorded:

> *Q15: Most up to date Stop and Search data in the UK. Including Ethnicity.*

> *Q16: I am interested in obtaining a complete up to date list of every licensed taxi and private hire operator in England, Scotland and Wales*

Other statements containing temporal information specified the required granularity of the data, such as *daily/monthly/yearly* as can be seen in Q17:

> *Q17: Daily Average temperature UK 2014 to 2016*

> *Q18: Inflation, from january/2006 till march/2013 (as monthly)*

> *Q19: The time of the crime (to the second ideally, but just as accurate as we can get)*

Others expressed the required time in vague terms:

> *Q20: Most recent and historical commercial property rents, by postcode, census output area or ward.*

Some requests also mention temporal information in order to answer a question about a specific point in time. In that case temporal information from within the data is required to answer the query:

> *Q21: For which tax year was there the greatest inheritance tax revenue per head of population, in real terms? In this tax year, what were (a) the inheritance tax rates, and (b) the other major differences from this year's rules?*

> *Q22: Amount of deaths in the last three years of those with learning disabilities.*

The following example shows required temporal information from within the dataset when it refers to “children under 18 years”:

> *Q23: Up to date statistics concerning children (under 18 years) smoking, drinking and taking drugs, attendance/ exclusion from school, and anti social behaviour statistics in the Hertfordshire area*

Or the following example in which people asked for the dates of all bank holidays:

> *Q24: Data on all UK bank holidays, past and future. [..] Data should go as far back as reasonably practical - 1970 as a minimum, post war is desirable. Into the future, the data is obviously a prediction, but should cover 40 years in advance generated using current known rules for bank holidays.*

In summary - temporal statements were used in several ways to define boundaries for the requested data: either to ask for the most current data; or from a point in the past to the present; or between two specific dates or for a certain number of years. These were presented at different levels of granularity as some statements specified a certain date and others were more vague mentioning e.g. “historic data”. Another type of temporal information was represented when users were trying to answer specific temporal questions.

5.6.3 Restriction

To get data that will be useful for a specific task, users specify various constraints or restrictions on the requested data ($n = 26.5\%$). Requests include statements specifying restrictions on the format of the data; price; specific data types; licence or a subset of data when a file was too big to use.

Below is an example of users specifying expectations about the price and license of the requested data:

> *Q25: There are licensees for the complete dataset, and queries from the dataset are of a suitable form, but quite expensive (5p per query). The cost would need to be in the fractions of pence or free to make this a viable usage.*

> *Q26: Unique property reference number, and post code(s) are very important data assets (persistent identifiers) when it comes to empowering individuals to take more interest in their personal data, and data about them, and ultimately benefit from doing so. Both are locked up behind barriers created by history, and/ or failure to account for the impact of opening them up.*

> *Q27: The licencing is very restrictive and does not allow for commercial use.*

As Q28 and Q29 show it is crucial to publish data in a way that assures the possibility of it being useful for various tasks.

> *Q28: I guess Bank and IT Companies would be having. It would be helpful if not all the data at least the subset of original data is available.*

> *Q29: Each NHS foundation trusts sends their annual financial data to Monitor in excel format. That data should be made available in excel format. Monitor currently publish a tiny subset in PDF format which is useless.*

When file formats are mentioned they either relate to data being published in a non-machine readable format:

> *Q30: The data is published but not in a format I can download and use (e.g. only displayed on-screen or only downloadable as a PDF rather than CVS)*

Or they specifies a format for the dataset that is needed for the respective task:

> *Q31: The data to be provided in a shape file format with the appropriate address(es) attached to the unique ID number*

5.6.4 Granularity

We define granularity as the level of detail to which the data is broken down (e.g. data could be presented per kilometre, meter or centimetre). Requests often specified the desired level of granularity of the data ($n = 24.5\%$). This was mostly found for temporal or geospatial granularity, but also subject granularity. For instance *"hourly weather and solar data set"*; *"25cm grid data"*; or *"prescription data per hospital"*. Below we present different ways granularity was expressed in the requests.

Users request data with specific granularity as this can be crucial to make the data actually useful for a specific task. For example in the case of Q32, if the data was presented by hospitals in specific boroughs or for all London hospitals it would fully miss the reason of the data requests – in which the crucial part was to have an overview of the data in a per hospital manner.

> *Q32: I would also like to know the number of accident and emergency admissions and births per hospital over a year.*

Q33 is another example where the granularity of the data is highlighted as important, together with additional specifications of the lowest granularity required for the data to be suitable for the task.

> *Q33: I require a data set that shows the average daily temperature for the UK from 1 March 2014 to 31 July 2016. For example with the following columns: Date, Average Temperature (e.g. 01032014, 12). The data doesn't need to broken down any further than that.*

Q34 and Q35 are examples of requests for data with the most detailed possible granularity - which could be particularly challenging for search functionalities to understand as they most often do not search within the dataset itself and granularity is challenging to express in metadata.

> *Q34: The research will identify whether there is a correlation (or not) between Road Traffic Accidents and Accidental Dwelling Fires and Youth Unemployment. The data that we are looking for should break down as much as possible e.g. into post codes.*

> *Q35: A data source with sufficient accuracy to enable marker post references around the M25 to be located on the network. The data should be in the form of OSGRs or XY co-ordinates of sufficient accuracy*

Lastly, another example of granularity information in data requests can be seen in Q36 where users requested data per day and want to be updated with new datasets on a monthly basis.

> *Q36: I want to find out how much it costs to run the London underground network each year. For example, how much does it cost to repair tracks each year? How much does it cost to run each tube station, per day/week/-month (preferably broken down into specific areas for example lighting, heating, maintenance, staff, etc.)*

> *Q37: For our project we need to know the type of individual crimes as far back into the past as possible, and when they happened. Monthly crime records are not granular enough.*

The level of granularity can be crucial for some the data to be useful. For example data that is aggregated to a country level would not be very useful for an analysis in a per city manner, although both datasets cover the same region. In current dataset search solutions defining the desired granularity of the dataset is not possible (unless it explicitly was stated in the description of the dataset).

5.7 Request Context

The analysis of the request context focuses on expressions framing the requirements associated with the information need, rather than the data required to fulfil it. This means statements expressing requirements or justifications, beyond the actual data that is requested. For example, this includes details such as users' motivation, comparisons or references to other datasets, and examples of specific questions users aim to answer with the data.

We grouped codes which were related to the expected structure of the data, and the type of expected outcome of the request (such as looking for a full dataset, looking for a particular data point, or looking for the results of an existing analysis). We also included mentions of specific headers, as well as pointers to other datasets that are similar to the requested one. In addition, we included requests that describe their rationale for seeking a particular type of data in more detail, as well as mentions of data quality. Below, we provide examples of each of these structures.

5.7.1 Representation and structure

Some requests contained detailed description of the dataset they are looking for ($n =$ 32%) This was presented as a list of information –for instance a list of headers in the dataset. Others pointed to another dataset that presents information in a similar way as they are looking for; or pointed to a dataset that already exists but that still does not fulfil their information need (because of insufficient information, insufficient granularity or of different geospatial or temporal boundaries for their task).

In Q38 a user wanted to obtain the same dataset as one that was already published for a different time frame:

> *Q38: The river quality data that is available is limited to 2006. I currently need the same river quality data for the East Anglia region that is more recent, ideally as recent as possible (e.g. post-2010)*

Q39 is an example of a specific list of information that is expected in the dataset:

> *Q39: For each of the schools under the Academy trust - the Head teacher, address, number of students, age range, telephone number.*

Q40 shows a similar scenario, in which a user highlights a specific type of information that is missing in an existing dataset, but that is necessary for their task:

> *Q40: Accident Cause column in the data is missing, for example: Accident cause =“over speeding”, “jumping a red light”, “wrong overtaking”, “lack of safe distance between vehicles”*

In Q41 below we can see an example of a request for dataset that is similar to an existing dataset.

> *Q41: Details of all expenditure over £500 (or some other limit) on a monthly basis similar to the current publication of spend data by central government departments and local authorities.*

When requesting a specific data structure, Q42 specifies data needed in a simple format that will not be challenging to analyse and understand.

> *Q42: I would like the overall cost of the UK Government published in simple format that the non accountancy literate among the electorate can understand. Ideally it would be set out as costs per themes as listed below BUT also show the complete cost to ensure that nothing is omitted.*

The need for data in a specific structure or representation could be an indication of a need for implementing functionalities to search engines that support search for similar datasets to ones that get proposed by the user; or for supporting search over specific headers of the dataset. Further it could indicate the need for dataset recommendation systems that suggest datasets similar to ones already selected by the user, but fit their information need better. Requests in which the required headers or categories expected in a dataset are listed indicates the usefulness of presenting the headers of a dataset to users in a search scenario.

5.7.2 Expected outcome

We further found that requests differ also in terms of their expected outcome. Some users expect specific data points or answers ($n = 5.5\%$), others expect whole datasets ($n = 78\%$) or the results of an analysis ($n = 11.5\%$). This indicates the need for systems that provide a better support of datasets search functionalities. However, we those percentages can be biased by the nature of the data portals which provide uses with datasets only and does not support other functionalities (e.g. onsite data analysing tool or question answering functionalities) which could support a wide range of information tasks and a range of skill-sets amongst users.

Below we see an example of requests in which user express their information need as searching for already performed analysis instead of searching for a whole dataset:

> *Q44: I am currently investigating the number of hospitals, clinics, geriatric residencies pharmacies and laboratories across the UK and was wondering if a study could be done showing them per region and maybe a map of the UK, visually showing where they are gathered. Big circles on those regions with the most of them and smaller*

Q45 illustrates that users can expect an answer to a question which could be a single data point from an existing dataset (assuming that such a dataset exists):

> *Q45: Is there any statistics pertaining to the number or percentage of schools in the UK that are adhering to Prevent Duty in terms of IT/network security and firewall settings?*

Q46 and Q47 present requests for whole datasets:

> *Q46: All parking fines recorded by fine amount and location address of car parking*

> *Q47: listing of all the Academy Trusts with member schools of each trust (Primary and Secondary). For each trust the CEO (trust leader)/ address. For each of the schools under the Academy trust - the Head teacher, address, number of students, age range, telephone number.*

The way people express the desired outcome of their requests might be influenced by the semi-structured request form and the majority of entries are expressed in free-text. Search for data points is currently not supported on governmental open data portals.

5.7.3 Rationale

In 31% of the requests the underlying motivation was specified (e.g. *"I am a PhD student working on aquatic plants"*) or details of the analysis that is planned with the data (e.g. *"in order to show where (in Birmingham) there exists unemployment"*). In some requests, users specified that they want to compare the dataset they are requesting with one that they already have (e.g. to compare the income and expenses; compare spending to other London boroughs) and want to be supported in this process.

In Q48 we see personal reasons mentioned for the data request:

> *Q48: I am looking for a dataset available on all economic sanctions imposed by countries on each other for my master's dissertation. So I can run a regression analysis on the imposition of economic sanctions against rise or fall in GDP per capita of a country.*

Q49 and Q50 illustrate a description of planned analysis with the requested data:

> *Q49: I am trying to find map data for all local authorities in the UK (England, Wales, Scotland, NI) so I can render it on Google maps or OpenStreetMap.*

> *Q50: The research will identify whether there is a correlation (or not) between Road Traffic Accidents and Accidental Dwelling Fires and Youth Unemployment*

We hypothesise that for the majority of requests describing the rationale behind them, reasons are given due to the assumption that those requests will be read and assessed by people working at the data portal, as opposed to being screened automatically. This encourages users to describe their data needs in detail and in natural language. However, when indicated in a request that the data is needed for comparison (or to be combined) with two datasets, this may indicate value in implementing features to automatically assess the potential of combining two datasets (for instance based on the presence of the same header in each of the datasets or through the semantic labels).

5.7.4 Quality

Some requests indicated that a particular dataset has quality issues and they request the data in better quality. This included the one caused by a service providing the data (such as a broken link to a dataset) and by the data itself. Quality can be understood on many different levels and is very dependent on the users' task (Koesten et al., 2017). Different users describe quality in different ways - fitting their information seeking scenario. Some requests mention quality of data indirectly by stating that a dataset has insufficient granularity or by requesting additional columns for a published dataset.

Below we present an overview of different mentions of quality in the data requests. However, the line between quality metrics and restrictions for or granularity of the data is not clear cut. Quality is mentioned mostly in relation to already existing data to criticise or explain why it is not useful for a particular information need, whereas restrictions or granularity are often expressed as requirements for requested data.

Data not being detailed enough (e.g. as in Q51) can be seen as one of the quality issues. Aggregation of data can result in it being not suitable for a task.

> *Q51: Met office publishes current data, but only historic averages, not historic data values. Without this history, I will have to wait years to amass sufficient current data for analysis.*

A similar issue for the usefulness of data is it's format (e.g. quote Q52). Many tasks require data to be of a specific format; for example geospatial data or a dataset saved as a PDF file.

> *Q52: The data set of the PCT boundaries they supplied, in KML format, has data quality issues and they no longer have access to their source of that data.*

Another group of quality issues are missing parts of datasets (e.g. Q53); or specific values of the dataset missing in an existing dataset (e.g. Q54 where the dates are missing); or in some cases errors within the existing data are mentioned (e.g. Q55 and Q56).

> *Q53: I require the IMD data which covers the North East. In this region there are several statistics missing from the 2011 IMD publications which related to the LSOA I was wondering why this data is missing and if you could provide a complete dataset.*

> *Q54: Accident Cause column is missing, for example: Accident cause = "over speeding", "jumping a red light", "wrong overtaking", "lack of safe distance between vehicles"*

> *Q55: There seems to be a gap in detailed LIDAR data available for the area where this golf club is based between hemel hempstead and St Albans. could this be updated?*

> *Q56: OSGR Eastings & Northings require to be 7/8 digit not 6 digit*

Quality awareness can be seen as understanding the state of the dataset; meaning if it is out-of-date or when the next update should happen, or if there are missing values and whether it is still usable for a certain type of analysis. This awareness allows users to judge the relevance of a dataset for an information need. Search functionalities could therefore allow users to judge certain aspects of data quality in the context of their task - potentially before downloading the dataset.

5.8 Limitations

The objective of this analysis was to gather additional insights into how people ask for data when they are not constrained by the limitations of a current search environment.

The set of data used in this study, the data requests, are real natural language articulations of people asking for data which is why we chose them for this study. However, they come with natural limitations. As these requests span over a period of time we had no opportunity to follow up with people to understand what they meant. Some of the requests were relatively short and the topics were relatively domain and UK specific. The generalisability of the results is unclear, however we believe that these requests enable us to get unique insights into how people might articulate their information needs about data. Naturalistic information seeking tasks requiring data are not commonly reported in literature, which is why we believe the data requests are a valuable means to better understand how people search for data. It would be interesting in future work to compare these written requests, both in their structure as well as in their content, to requests for data in different digital environments. Finally, not much of consistent demographic information is collected alongside the requests. It is possible that the users making these requests represent a specific sample of the population.

5.9 Discussion and Conclusion

Data requests are issued by users with specific characteristics and are not directly comparable to the sample of data searchers represented in the log analysis which we present in the next chapter or users statistics shown in the previous chapter of this work. However, they add to our results by adding an in-depth perspective on information needs for data that are explained in more detail. We found that a large proportion of requests were issued by individuals, with most of them classified to be done for research purposes. However, only a small number of requests were issued by declared academics or researchers, which suggests that data is often used for non-academic research purposes potentially including private decision making contexts or business use. This somehow contradicts our hypothesis that majority of portal users as described in Chapter 4 are using the portal in the work related environment. We hypothesise that people who issued data requests chose the *organisation type* on the requests form to be *individual* and not *business or academic/research institutions* as a way of not answering additional questions. This is because the form somehow suggested that declaring to be part of an organisation was not compulsory and might result in having to answer additional questions and so in more laborious. Another possibility is that people consider certain activities as research, even if they are done privately, *e.g.*, research about potholes to file a data-founded complaint to the council. The most common reason for issuing a request was specified as the inability to find suitable data, whereas less common reasons included, for instance, data that was available but in the wrong format or for the wrong time frame.

Over three quarters of all requests included Geospatial requirements. Geospatial requirements are specified through boundaries, which are expressed in varying levels of precision. As the requests were issued in natural language, this also included vague terms such as *overseas* or through use of more informal geographical definitions such as *tube zones*.

Almost half of all requests contained temporal information, the majority of which were requirements for a specific year, a particular time frame or simply for the most up-to-date data. These define the temporal boundaries for the information need. Temporal information is often discussed using non specific expressions in natural language and this is also reflected in the requests, such as *historic, or in the past*. Temporal information can also refer to specific attributes within the dataset, such as *diabetes people over 60 years*.

The high prevalence of temporal and geospatial information indicates the importance of these features for the fitness of use of data for a specific information need, supporting results of prior research (Kern and Mathiak, 2015) that identified importance of the time frame to which the data refers to. However, a new aspect that we identified is the relevance of the granularity of the geospatial or temporal information. Even if data that is topically relevant to an information need is available - if it has the wrong time frame or location - it becomes useless. In current dataset search solutions defining the desired granularity for the dataset is not possible (unless it was explicitly stated in the description of the dataset). Even if granularity is not provided as a facet in the search functionality, an overview of the available granularity of the data in the dataset could be presented in the metadata. Our findings show both the popularity of, and the complexity of expressing, geospatial or temporal boundaries for datasets, which suggests the need for designing more advanced search functionalities to cater for these attributes - this had already become a subject of research in works such as by Neumaier and Polleres (2019).

Other features prevalent in the requests concerned restrictions on the data. These refer for instance to the format or the size of the data, price restrictions or licence. Those can to some extent be resolved in providing functionalities on the data portal to change the data format, or by allowing users to select appropriate subsets of data. Those kind of issues could partially be resolved by providing both publishers and data users with additional information on the data publishing process (e.g. assigning appropriate licences to the data).

Data requests provided further insights into how people expect data that they are requesting to look like. This included specifying the headers that are expected in the dataset, or defining a certain format. This might be due to limited technical skills or data literacy. Their task could also involve comparing the data to other datasets that they already have which is easier in a certain format or with comparable attributes. This suggests a number of potentially interesting directions for further research, such as recommendation systems for datasets based on similarity between datasets or that take a dataset as an input. This can also include the indexing of headers to make them discoverable for search functionalities, as well as the presentation of headers to users in a search scenario. The data publishers based on our finding could be encouraged to include this type of information in their metadata and portals along metadata standards should be build in the way to ease the publisher in a process of relevant metadata generation.

Our findings further point to several quality dimensions that are considered important in this context. This includes access to data, completeness and the amount of data available, characteristics that were covered in literature (Pipino et al., 2002). We believe that to judge the relevance of a dataset for a tasks users need to be aware of these characteristics. This suggests that the inclusion of basic quality dimensions in metadata or search result presentation could support the discovery process.

Chapter 6

Search Log Analysis

In this chapter we present a study of the patterns and specific attributes that data users use to search for data and how it compares with the general web search. We performed a query log analysis based on logs from four national open data portals, one European data portal and queries generated by crowdworkers based on data requests issued on one of the portals (detailed description of a data request was provided in Chapter 5). Search queries issued on data portals differ from those issued to web search engines in their length, topic, and structure. Our findings suggest that portals search functionalities are currently used in an exploratory manner, rather than to retrieve a specific resource. In our study of data requests presented in Chapter 5 we found that geospatial and temporal attributes, as well as information on the required granularity and the content of the data are the most common features. The findings of both analyses suggest that these features are of higher importance in dataset retrieval in contrast to general web search, suggesting that efforts of dataset publishers should focus on generating dataset descriptions including those features. Some of the results presented in this chapter appeared in (Kacprzak et al., 2019, 2018b, 2017).

6.1 Analysis of Search Patterns

In order to understand better behaviours of users who already search for datasets we conducted a study looking at existing data on dataset search. In order to gain deeper understanding on 'How users search for data?', 'What search strategies they use?' and 'What information should be included in the metadata to improve the dataset search process?' as stated in our first and second research question.

In this chapter we present a study based on query logs from four open data portals, of which three belong to the national governments of the UK, Australia, and Canada, one is from the UK's Office for National Statistics, next we compare our findings with the European Data Portal. Together, the logs include more than 2.3 million queries (of which 768 thousands are unique queries), issued between 2013 and 2020[1].

In an open data scenario, data formats and models are heterogeneous – the cost of mapping and transformation of such data into RDF is often too high for publishers. An alternative is to compute or manually fill metadata descriptions of non-RDF datasets using an agreed vocabulary, for example, the Data Catalogue Vocabulary (DCAT) was designed to facilitate interoperability between data catalogues published on the web. This includes descriptions of keywords, theme, frequency, spatial and temporal coverage. Numerous extensions to DCAT have been developed to include additional properties that are considered relevant by their designers, e.g. DCAT-AP (for public sector data), GEO-DCAT-AP (geospatial properties) or Data-ID (versioning, technical descriptions of datasets). Another schema for describing datasets is *Schema.org*, which recently started to gain traction with introduction of Google dataset search platform. However, to the best of our knowledge, there are no systematic studies from the point of view of data users about what properties are more important than others for effective search and discovery of datasets. This is important, as the generation of metadata needs to be done on a property to property basis, which also represents a cost for data publishers. Knowing what are the properties that they need to focus on to satisfy user information needs reduces the time and effort required for the publishing process. Furthermore, current open data portal solutions base their metadata search on indexing free text descriptions of datasets and applying natural text based document modelling and search techniques. We believe that based on our finding we can gain valuable insights into what kind of advanced search functionalities should be explored to support user information needs. Relevant metadata should be encouraged more of publishers and, in fact, in case of DCAT-AP temporal property *dct:temporal* was moved from optional to recommended property as per findings of presented work[2].

As search functionalities tend to fall short for information seeking tasks for datasets (Koesten et al., 2017) and queries issued on portals shown to be too short to provide the basis for an extensive log analysis, we therefore also analyse a set of queries for data (*crowd queries*) which were generated using human computation and data requests issued to one of the data portals. Data requests are unstructured descriptions of datasets made by citizens directly to open data portals, as explained in Chapter 5. We compare crowd queries to those issued to search on data portals in order to understand whether the change of search environment results in different characteristics of queries.

In a nutshell, our goal is to advance towards the understanding of the most important properties of a dataset description from the point of view of data users, by analysing how people search for data on current portals.

[1] https://github.com/chabrowa/data-requests-query-dataset - Account of the thesis author
[2] https://github.com/SEMICeu/DCAT-AP/issues/64

In this chapter we aim to answer the following detailed **research questions** in relation to our first two research questions regarding the understanding of dataset search at it current state. On one side in [RQ1] (*Understanding Dataset Search; How do users search for data? What search strategies do they use?*) we analysed available information on users of the portals, data request submitted by users and in this chapter we look at the available queries. On the other side in [RQ2] (*What information should be included in the metadata to improve the dataset search process?*) along analysis from Chapter 5 we look at the specific characteristics of queries to understand what are the aspect of search when searching for datasets that are used by searchers in their search queries.

- (*a*) What are the characteristics of queries for datasets in terms of their length, distribution, and structure? How this informs the decision of which properties should be prioritised in a description?
- (*b*) How do search queries for data within a data portal differ from those in a less constrained environment?
- (*c*) How does the search queries in dataset search compare to general web search?

When characterising users (shown in Chapter 4) we found that data search is performed via desktop computers on weekdays during working hours. Returning users had longer session durations, which suggests they might benefit from additional, more advanced, search functionalities. The majority of users ended up on data portals via external search engines. This suggests that search functionalities on portals might not be considered sufficient, which was suggested by qualitative analysis of interviews by Koesten et al. (2017).

Queries on data portals are generally short, which might be the result of a lack of trust that longer queries will return useful results. This assumption was supported in study by Koesten et al. (2017). The results suggested that issuing a query is often conceptualised as an activity aiming to narrow down relevant subsets of data that is available on a topic, rather then expecting a matching dataset directly in the result list as we are used to from web search. Both queries issued on a portal and crowd queries - generated in a less constrained environment contained geospatial and temporal information - in both cases relevant keywords are represented in different levels of granularity (month/years, cities/regions). Queries issued on portals included indications of time were five time more frequent than in web search. Furthermore data format and file type were popular amongst all the queries - in case of external queries a fifth of all queries contained such attributes.

The analysis of data requests shown in previous chapter of this work revealed similar features. The most common features mentioned in the requests were temporal and geospatial information together with topic and definitions of their expected granularity.

These features tend to be complex and need to be taken into account when generating metadata for datasets that is utilised by search functionalities. The wrong granularity in terms of both location and time can easily result in the data not being usable for a task. Furthermore, more than one dataset can be equally relevant to a single information need. Requiring information for longer time spans can result in many equally relevant datasets, as each contains a portion of the desired time period. Our findings suggest that publishers should focus their efforts on generating spatio-temporal properties of their descriptions and also improving the descriptions of the columns (expressing the topic of the data), motivating the development of search interfaces for appropriately filtering and joining by them.

6.2 Methodology

In this chapter we conduct a query log analysis over search log data collected via Google Analytics and Matomo Web Analytics, and crowdsourcing which expands our findings from qualitative thematic analysis of data requests shown in previous chapter. In our experiments we used two types of data acquired from different governmental open data portals: internal and external search logs and a set of search queries generated by crowdworkers (crowd queries), which collection we describe later in this section.

Four well-known data portals from three English-speaking countries: United Kingdom, Canada and Australia; along one supranational European portal – provided their query logs to us: the official UK government Open Data portal (DGU), the Office for National Statistics of the UK (ONS)[3], the Australian government Open Data portal (AUS)[4], the Canadian government Open Data portal (CAN)[5] and European Data Portal (EDP)[6].

The logs primarily represent search for structured data. All portals collect log data using Google Analytics except EDP which uses Matomo (in case of internal queries) and Google Analytics (in case of external queries), but might be using different settings. As a consequence, and as they started recording their logs at different points in time the available information and time frames per portal vary.

Three of the portals at the time of query log collection store datasets or links to datasets with use of the CKAN portal software[7] (i.e. DGU, AUS and CAN), which bases its search functionality on the Solr search platform[8]. Search functionality is provided through a search box in the portal. Queries are evaluated against textual descriptions and metadata text fields associated with the datasets. The ONS stores their data on a

[3] https://www.ons.gov.uk

[4] https://www.data.gov.au

[5] https://www.open.canada.ca

[6] https://www.europeandataportal.eu/en

[7] http://ckan.org

[8] http://lucene.apache.org/solr/

custom portal which is more targeted at presenting an analysis of the collected data, partially through visualisations. The EDP is also a custom-build portal aggregating information from the number of data portals around Europe (this includes for example DGU). Both DGU and ONS present 10 results for a query per page. AUS and CAN show 20 datasets as results per page. EDP show 15 datasets as results per page. Through timeframe for which data was collected, the EDP introduced changes to its website and search functionality which we will discuss at more length in next chapter where we analyse full dataset search sessions.

All portals provide facets by which users can filter and browse the results. The summary of collected logs with dataset logs size and time frames for collection can be seen in Table 6.1 for internal queries and Table 6.2 for external queries. We distinguish between three types of **queries**: internal, external and crowd queries. *Internal* - queries issued directly in the search box of a portal; *External* - web search queries that led the user to open a page of the data portal. External queries were only provided for DGU, ONS and EDP. A query object comprises the following fields: *search terms* and *total unique searches*. The search terms of a query are made of the string, i.e., the sequence of search keywords, typed into the search box (of the portal, for internal queries; and of the web search engine, for external ones). Majority of the portals described in this analysis did not have any additional event tracking, e.g., click-through data, configured. The exception to this is EDP for which full session information was available. In order to present comprehensive analysis of queries, session data from EDP was aggregated to match the shape of the data of the other portals. The analysis of the information discarded for the purpose of the study presented in this chapter is shown in Chapter 7.

6.2.1 Pre-processing

Search log data from both internal and external queries was pre-processed as follows:

Step 1 The N-Gram Fingerprint method was used to clean the data as it can detect basic spelling mistakes which could be a swap of two or more letters within a word. For example a 2-gram string for the word *london* would be *do,lo,nd,on* and for 1-gram *d,l,n*. For this purpose functionalities of Open Refine tool were used.

Step 2 Discard outliers in terms of length. 99.9% of all queries had less than 19 words. Based on manual inspection, we considered longer queries to be likely the result of accidental pasting of text into the search box and discarded them from our analysis.

Step 3 Finally, we removed those external queries which were registered, but not specified. They were of two types: *(not provided)* and *(not set)*, according to the eponymous Google Analytics flag. The first are not specified due to the privacy policy of Google

[8] https://openrefine.org/

Portal	Internal		Time Ranges
	All	Unique	
DGU	1,058,197	332,823	30/01/2013 - 31/08/2016
ONS	950,593	342,054	28/02/2016 - 31/08/2016
CAN	231,473	46,661	23/08/2015 - 23/08/2016
AUS	5,311	2,557	01/08/2016 - 31/08/2016
EDP	106,334	44,506	01/04/2018 - 30/06/2020

TABLE 6.1: Summary of internal search log data. Column *all* refers to the total number of internal queries per portal excluding queries eliminated in Step 2, while column *unique* refers to the number of unique queries determined via clustering in Step 1.

Analytics, while the second refers to traffic that did not occur as a result of a search, but via referral sites, direct links, or other search channels such as Google Maps and Google Images.

Internal Search Logs. Queries issued directly to the internal search capacity of a data portal into the search box. We have a total number of $2,245,574$ ($2,351,908$ with EDP data) internal queries excluding queries removed in Step 2, $724,095$ ($768,601$ with EDP data) unique queries determined via clustering in Step 1 (data cleaning). The breakdown of internal queries per portal after pre-processing steps can be seen in Table 6.1. Figure 6.1 outlines structure of internal queries along the definition of particular metrics.

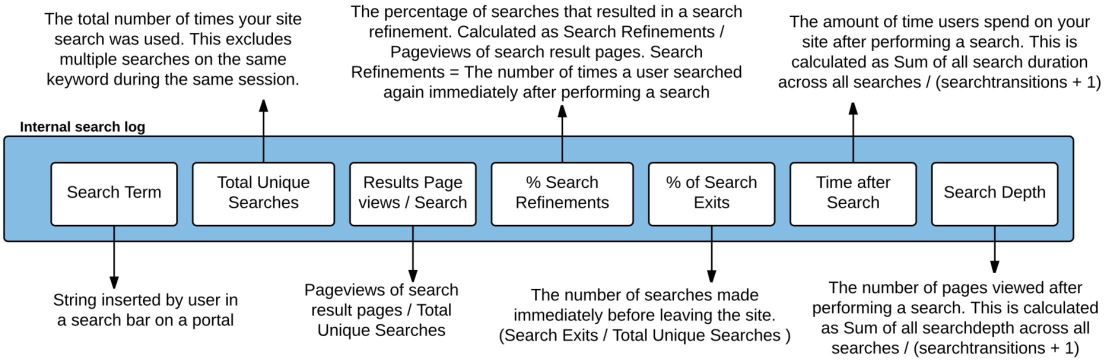

The total number of times your site search was used. This excludes multiple searches on the same keyword during the same session.
The percentage of searches that resulted in a search refinement. Calculated as Search Refinements / Pageviews of search result pages. Search Refinements = The number of times a user searched again immediately after performing a search
The amount of time users spend on your site after performing a search. This is calculated as Sum of all search duration across all searches / (searchtransitions + 1)
Internal search log
Search Term
Total Unique Searches
Results Page views / Search
% Search Refinements
% of Search Exits
Time after Search
Search Depth
String inserted by user in a search bar on a portal
Pageviews of search result pages / Total Unique Searches
The number of searches made immediately before leaving the site. (Search Exits / Total Unique Searches)
The number of pages viewed after performing a search. This is calculated as Sum of all searchdepth across all searches / (searchtransitions + 1)

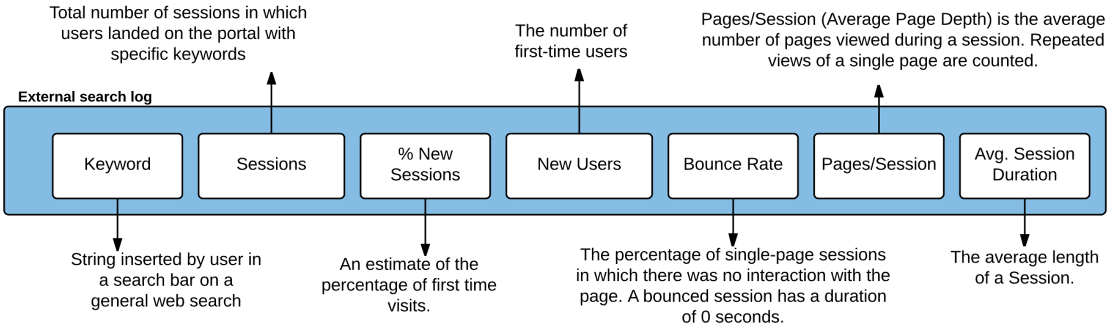
FIGURE 6.1: Structure of internal query logs and details of their meaning

External Search Logs. Queries issued through a general web search engines search as that lead to a page of the data portal. This set consisted of $1,101,201$ ($1,166,864$ with EDP data) external queries, after removing lengthy queries (Step 2) as well as missing values from the data (Step 3). $3,983$ were *not set*, $4,228,602$ were *not provided*, and $18,985$ were *Other* in case of EDP (shown as *not provided* in Table 6.1) in our sample. There were $433,637$ ($437,754$ with EDP data) unique external queries (determined via data cleaning in Step 1). The breakdown of external queries per portal after pre-processing steps can be seen in Table 6.1. We assume that when a user issues a

Portal	External				Time Ranges
	All	Unique	Not Set	Not Provided	
DGU	1,062,937	419,750	3,159	3,902,006	30/01/2013 - 31/08/2016
ONS	38,264	13,887	824	326,596	28/02/2016 - 31/08/2016
EDP	65,663	4,117	-	18,985	01/05/2018 - 31/03/2020

Table 6.2: Summary of external search log data. Column *all* shows the number of queries obtained after removing overlengthy queries (Step 2) and not provided and not set ones (Step 3). Columns *not set* and *not provided* show the number of not set and not provided queries. In case of EDP 'other' queries are shown as 'not provided'. The column *unique* was calculated like for internal queries (Step 1)

query to the search on an external engine, and clicked in the result that direct them to the data portal, their intention was to find a dataset. We acknowledge the limitation that we cannot know if the user just clicked on the dataset as part of an informational query looking for something else from the external portal. A possible more fine-grained heuristic is to collect and analyse the search exits from external queries, and assume that those that immediately exit the portal were not looking for a dataset.



Total number of sessions in which users landed on the portal with specific keywords

The number of first-time users

Pages/Session (Average Page Depth) is the average number of pages viewed during a session. Repeated views of a single page are counted.

String inserted by user in a search bar on a general web search

An estimate of the percentage of first time visits.

The percentage of single-page sessions in which there was no interaction with the page. A bounced session has a duration of 0 seconds.

The average length of a Session.

Figure 6.2: Structure of external query logs and details of their meaning

Crowd-generated Queries. Crowd queries were generated in a crowdsourcing experiment based on data requests (described in Chapter 5 of this work). An example of an excerpt of a data request is *"Request annual return data on total numbers of Sheep & Lambs and Cattle & Calves in the following two North Yorkshire parishes from* 1986 *to the latest available date: for Malham Moor Parish and for Buckden Parish"*. We randomly selected 10% (50 requests) of all the openly published data requests, and manually checked for their understandability concerning language and domain specific terminology - we excluded requests which were potentially difficult to understand and replaced them with other randomly selected requests.

In our experiments we used the title and the description of the request. For each data request we generated 10 queries through human computation. After excluding spam answers (51 of all queries, which were manually detected) the set contained 449 queries in total. An example of a resulting query is *"Businesses in Yorkshire that employ over* 1000 *workers"*.

Participants of the crowdsourcing experiment were users of the crowdsourcing platform CrowdFlower. As the data requests are unstructured English text that could potentially be difficult to understand for people with low English language skills, we limited the experiment to workers in native-English speaking countries; and we restricted the worker pool to a smaller group of more experienced, higher accuracy contributors on the platform. We included 5 short qualification questions, assessing basic reading, reasoning and data literacy skills. Workers were paid $0.15 to generate each search query that they considered suitable for a single data request.

Our open-ended text creation task was formulated as: *We ask you to write a search query which you think would return the requested dataset from a data search engine.* The workers were shown an overview of the task, step-by-step instructions, and a sample data request with examples of corresponding queries. The output was a search query constrained to be between one and twenty words in length. To minimise "spam" answers we prevented pasting of content and validated each word from the query against an English language dictionary, requiring an 80% matching threshold for a query to be accepted. We also rejected answers containing the same word three or more times. Participants were not instructed to generate their queries in a particular structure; however, they were shown five examples, with various compositions of keywords, and a question. The minimum time permitted to generate a single query was 1 minute to allow time for detailed reading of the data requests. No personal data was collected. Despite the workers' lack of in-depth understanding of the information need that is represented in the data request we believe that the resulting queries give us valuable insights into the necessary complexity and characteristics of queries for data. The data collection described in this section was approved by the Ethics Board of the University of Southampton (#30548).

6.2.2 Quantitative Analysis

This section describes the metrics presented in Table 6.3 chosen to analyse the queries. Their selection was based on background literature in web search and other search verticals shown in Section 2.5 and the relevance of commonly used metrics when applied to the analysis of dataset search logs. To analyse the data we first inserted all search logs into a MongoDB[9] database as separate entities. We created separate collections for internal search logs and external query logs (for four national data portals queries, other were analysed directly from the file). Results for the metrics listed in Table 6.3

[9] http://www.mongodb.org

Metric - Method

Average length; number of words in a query Computed for all internal and external queries. Both of these metrics were calculated for *all* queries in the log as well as for the subset of *unique* queries.

Query characteristics Matching queries to keywords describing: location; time frame; file and dataset type; numbers; abbreviations. Keywords used for each of those metrics are specified in Table 6.6. Computed for internal and external queries. The keywords for each category were selected by taking a sample of top 50% of queries and listing the words indicating particular information type (as listed in Table 6.6); we, in addition, used the most popular words that were not found in those top queries (e.g. yearly or quarterly). We compared the list of keywords against all queries to detect how many of them contained particular keyword.

Question queries To recognise question queries we counted queries containing the words: *what*, *who*, *where*, *when*, *why*, *how*, *which*, *whom*, *whose*, *whether*, *did*, *do*, *does*, *am*, *are*, *is*, *will*, *have*, *has* as done in (Bendersky and Croft, 2009). Computed for all internal and external queries.

Query topical distribution Manual categorisation of topics was done by two of the authors for a sample set, representing the 665 most popular unique queries. This sample size was determined using a 99% confidence level, a 5% confidence interval (or margin of error - e = 0.05), z-score equal 2.58 (used for a 99% confidence level), distribution 50% (p = 0.5), which gives the largest sample size, and population size of 2.2 million queries using the following formula[11]:

$$samplesize = \frac{\frac{z^2 * p(1-p)}{e^2}}{1 + \frac{z^2 * p(1-p))}{e^2 N}}$$

We derived 12 topics (plus *other*) from themes used by DGU to tag datasets. We exclusively categorised each query to one of these topics: *Business and Economy*, *Environment*, *Mapping*, *Crime and Justice*, *Government*, *Society*, *Defence*, *Spending*, *Towns and cities*, *Education*, *Health*, *Transport* and *Other*.

Table 6.3: List of metrics performed in the qualitative analysis

were generated using Python code[10] connected to the aforementioned collections unless specified differently.

6.3 Internal & External Queries

In this section we present the results of query analysis with use of the metrics introduced in Table 6.3. We present the comparison of *internal* queries for five different open data portals and crowd generated queries and of *external* queries for the two UK portals (DGU and ONS) and one european portal (EDP) for which they were available. Internal queries were the queries issued directly to the search functionality of the data portal whereas

[10] https://github.com/chabrowa/search-log-analysis - Account of the thesis author

[11] https://www.surveymonkey.com/mp/sample-size-calculator/

external queries are those issued to web search engines that lead users to open a page on the data portal. The assumption is that if a user opened a page in a data portal following a web search hit, the intention of the query was to retrieve a dataset.

Internal queries were analysed as one set and further details using specific measures are presented accordingly in this section. We categorised external queries in two categories. We refer to *proxy queries* when they contain the name of a data portal. All remaining external queries are referred to as *direct queries*. 6.71% of external queries for DGU and 54.82% for ONS are proxy queries. In case of EDP 15.19% were proxy queries. A proxy query indicates that the user wanted to reach a result from the portal in question, but did so through a web search engine instead of going first to the portal and use its search capability. Our initial analysis of proxy queries revealed a high variance of spelling and use of URIs. To avoid skewing results due to noise, we chose to focus on direct queries, excluding queries identified as proxy queries. We split the queries into direct and proxy queries by analysing keywords indicating portal names (i.e. queries containing word groups as: *gov* and *uk*; *office*, *national* and *stat* or *o n s*) or queries in a form of an URL link (i.e. queries containing *www* or *http*). Code used in order to split queries is available on Github[12].

6.3.1 Open Data Portals

6.3.1.1 Query Length

Table 6.4 shows the average query length for both internal and external queries. Internal queries are between one and three words long, with an average of 2.03 words for all queries and 2.67 for the unique ones. The average external query length is 3.98 words for all queries and 4.74 for unique queries. External queries are on average more than one word longer than internal queries. This could be the result of web search queries being generally longer (Taghavi et al., 2012), or a different perception of the internal search functionality by users. External queries were found to be longer than the reported average of 3.08 words in a web search query by Taghavi et al. (2012) in 2012. However, this might not fully apply currently as general web search underwent rapid developments, for instance in answering conversational search queries. These advances might have resulted in much longer queries[13].

[12] https://github.com/chabrowa/search-log-analysis/blob/master/database/externalProxyQueries.py

[13] https://searchengineland.com/google-hummingbird-172816

Portal	Internal		External	
	All	Unique	All	Unique
DGU	2.04	2.78	4.12	4.82
ONS	2.52	3.42	3.83	4.66
AUS	1.63	2.31	-	-
CAN	1.93	2.17	-	-
EDP	1.55	1.97	2.82	3.35

Table 6.4: Average number of words per query

Figure 6.4 shows the distribution of internal queries according to their length, for all national portals, for both all and unique internal queries. When considering all queries, single word queries represent almost half of the entire corpus. When focusing just on unique queries, this number falls to 25%. The distribution for unique queries is very similar to the results reported for web search engines in 2001 by Spink et al. (2001). Figure 6.3 shows the distribution of external queries according to the number of words in a query, for ONS and DGU. The distribution of number of words in external queries is more similar to the one shown for general web search than the distribution of number of words in internal queries. It could be that advances in dataset search will lead to similar behaviour patterns as observed in web search today, which would mean longer queries, that are closer to natural language.

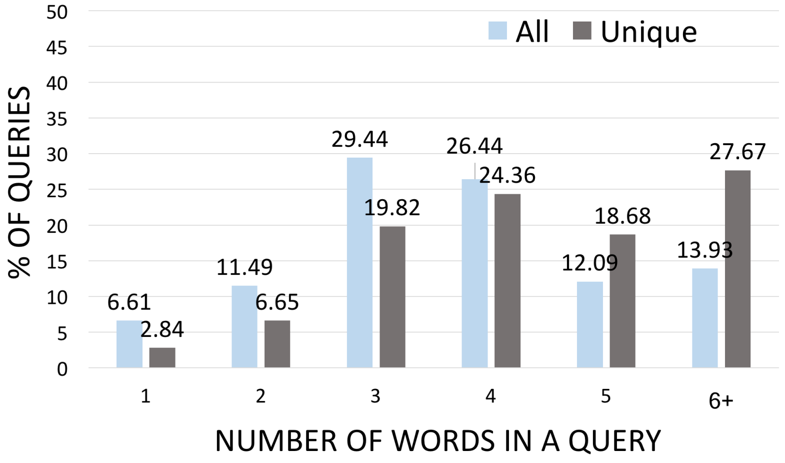


Figure 6.3: Percentage of internal queries by average number of words (all and unique queries, all national open data portals)

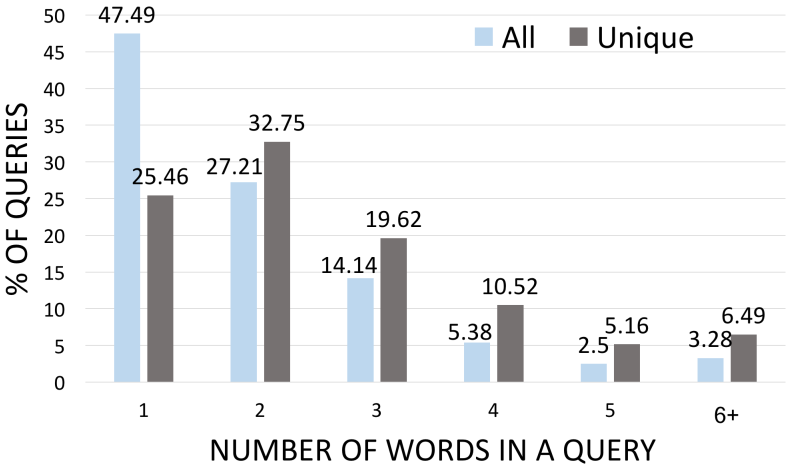


Figure 6.4: Percentage of external queries by average number of words (all and unique queries, ONS and DGU portals)

6.3.1.2 Query Types

People have different strategies while searching for datasets (Koesten et al., 2017). In this work we focus our analysis on the keywords that are used in search queries for data. For most metrics as defined in Section 6.2, we computed the corresponding values automatically, for all internal queries of the four national portals and all external queries for two UK based portals. To classify queries into specific topic we used a sample of the logs, as explained in Section 6.2.2 Table 6.3 under *query topical distribution.* Table 6.6 summarises the percentage of queries representing each metric.

Geospatial. In the set of internal queries we found 5.44% containing location specific keywords, whereas in the set of external queries this appeared to be slightly more popular with 7.93%. A previous study on general web search (Gan et al., 2008) reports this metric at 12.01%. This difference could be caused by the fact that the data portals we studied are already bound to a country, and users do not need to specify the location as frequently as in general web search. In addition, it might be caused by the fact that queries in dataset search are significantly shorter in comparison to web search. With almost 50% of queries being one word queries, users might specify only the general topic of their information need in a query.

Temporal. Keywords indicating temporal information appeared in 7.29% of the internal queries and in almost twice as many external queries (12.26%). This number is much higher than the 1.5% reported for general web search (Nunes et al., 2008). This may mean that users have a higher interest in the temporal details related to the data they are searching as opposed to searching for web pages. This can include the time frame the data represents (data about a particular year) or the creation time of a dataset (the time the data was collected and published, including the frequency of updates).

File or data type. In dataset search, queries included restrictions for the shape the data should have, in order to fulfil the information need. 6.25% of internal and 20% of external queries included an indication of file type or data type (query analysis with use of keywords listed in Table 6.6). We note that the governmental portals represented in this study offer filtering options for file types that are not reflected in our data - the overall number of queries specifying a file format could be higher. However, it could also indicate that filters in current interface design might not be prominent enough. From a data point of view, this figure could be an indication that users search for alternative file types and formats and that publishers need to be able to support different and popular formats for their data. We believe the higher percentage in external queries is due to the fact that users intend to find data and not textual documents. In general web search, users need to indicate this in addition to their query - this step is unnecessary on data portals, which are designed to support search specifically for data.

Numerical. Data is often numerical - it was shown by Mitlöhner et al. (2016) that numerical values are the most popular data type in open governmental datasets. In comparison to documents, which are mostly text, we also computed the number of queries containing numbers (excluding those indicating temporal information). 5.23% of internal queries contain any number and 0.38% contain only numbers. External queries present almost the same statistics with 5.23% for queries including numbers and next to zero for queries containing only numbers (only 0.008% of queries which we believe is so small due to the fact that external queries were much longer on average in comparison to internal queries). Those results show a disproportion in the amount of queries with numbers in comparison to the number of numerical values in the data. This also indicates the need to understand the underlying meaning of numerical columns in datasets e.g. by lifting them to a linked data format which could then provide additional context to the data.

Abbreviations. In our analysis, we also identified that users frequently use abbreviations in their queries, as many datasets use acronyms like *rpi* for *Retail Price Index*. 5.11% of internal and 7.05% of external queries contained at least one acronym. However, we noticed that the full expansion of those acronyms is also used in queries. For the majority of governmental open data portals the main content that is indexed by the platform and searched over is a description of the dataset provided by the data publisher. Therefore some datasets are described in the index only with the full expansion of the acronym related to them. However, some users might only search using acronyms which results in false negative results.

Question queries. Formulating queries as questions is increasingly common in web search (White et al., 2015). In 2009, 7.49% queries were questions in a study on general web search (Bendersky and Croft, 2009). All figures are significantly below the 7.49% reported by Bendersky and Croft (2009) for web search. However, external question queries totalled 5.09% for DGU and 1.52% for ONS, significantly more than internal

Metric	internal	external
Geospatial - the name of a city or geographical area (either town, city, county, region or countries)	5.44%	7.93%
Temporal - years (1000 to 2017), names of months, days of a week and the words *week(ly)*, *year(ly)*, *month(ly)*, *day(ly)*, *date*, *time* and *decade*	7.29%	12.26%
File or data type - file types: *csv, pdf, xls, json, wfs, zip, html, api* and keywords denoting a type of dataset: *data, dataset, average, index, graph, table, database, indice, rate, stat*	6.25%	20.01%
Numbers - the number of queries including numbers excluding those indicating time frames	5.23%	4.46%
Only numbers - queries that contain only numbers	0.38%	0.008%
Abbreviations - 72 most popular, manually identified acronyms	5.11%	7.05%

Table 6.5: List of metrics with definitions and their prevalence in internal and external queries. The keywords lists are available on the Github repository.

question queries and much closer to the results reported to general web search. We believe this is mostly due to users' understanding of the search functionality of dataset search engines (as a source of data to be downloaded for further use, and not as a question-answering engine) and due to the type of service that is currently provided, which might supply relevant search results for keyword queries but does not support question queries. (For instance the CKAN platform, as one of the most common data management systems, up-to-date does not provide such a functionality.)

Query topics. This metric aims to capture the domain of data people are searching for. We present our own classification, as the ones used in web search are not directly relevant (Jansen et al., 2000) - for example, in web search sexual topics/pornography is the most prevalent topic category (25%) which does not fit the content of the platforms like governmental open data portals. In this work we used alternative topic categories as described in Section 6.2. Figure 6.5 shows the distribution of queries according to the data domain. The most popular category is *Business and Economy* (20.03%), followed by *Society* (14.74%). This is in line with our observations earlier about the use of data portals in professional contexts and is influenced by the nature of the portals themselves, which publish official statistics or data produced by different governmental departments. The distribution of topics for external queries differs from the one for internal queries. As can be seen in Figure 6.5 *Business and Economy*, *Environment* and *Other* queries are less frequent, while *Towns and Cities*, *Health*, *Education and Society* are more frequent. These are naturally influenced by the domain specificity of the portals.

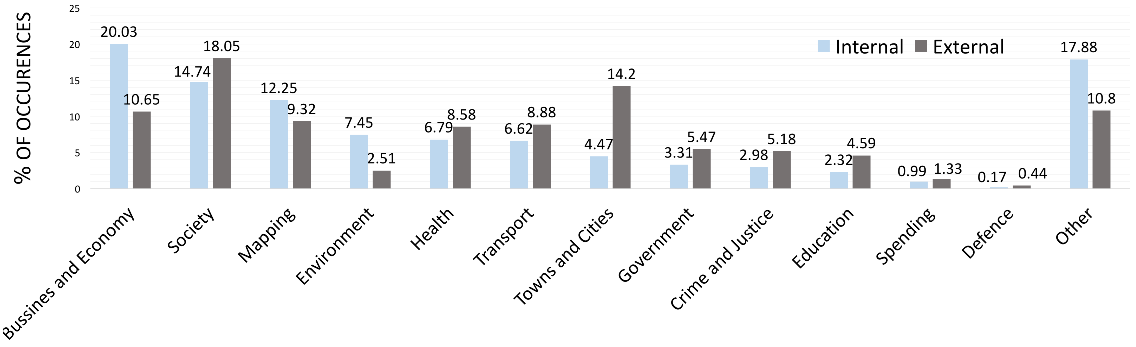


Figure 6.5: Distribution of topics within internal and external queries

6.3.2 EDP Queries

In this section we present the results of our analysis of the EDP queries and compare them to the internal and external queries presented earlier in this work. As the EDP is supranational (it aggregates information from many portals on national level) and its data log is much more recent we decided to analyse it separately to the national portals. In the same manner to external queries of national data portals we excluded proxy queries from further analysis. In terms of query types we followed a different approach to national data platforms due to the multilinguality of the platform and decided to label a sample of data manually.

6.3.2.1 Query Length

The query length of EDP queries was 1.55 words for all and 1.97 words for unique queries in terms of internal search. In terms of external queries they had on average 2.82 words for all and 3.35 words for unique queries. Figures 6.6 and 6.7 shows the query length distribution for all queries and unique queries for internal and external queries. The large proportion of single-word queries is consistent with results for national data portals.

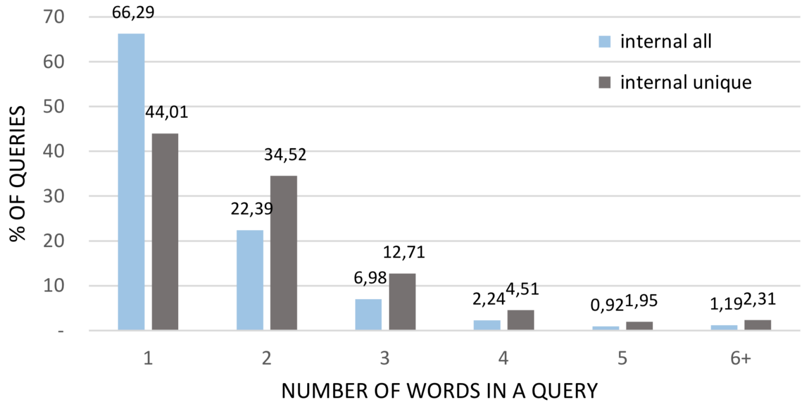


Figure 6.6: Percentage of internal queries by average number of words (all and unique queries, EDP portal)

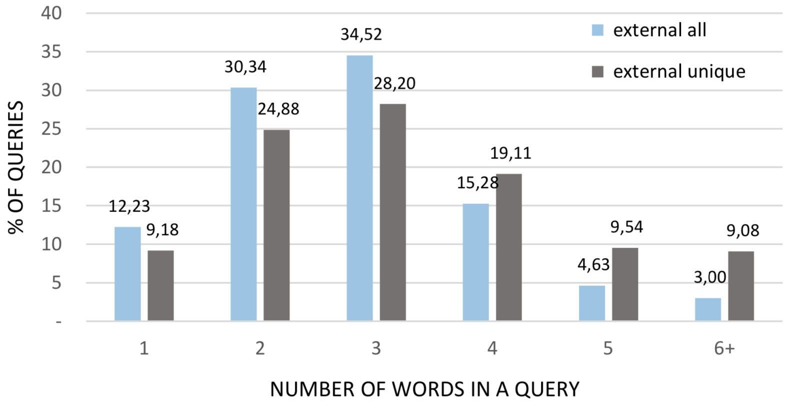


Figure 6.7: Percentage of external queries by average number of words (all and unique queries, EDP portal)

Our results suggest that users may be using the search box in a similar way to a facet, that is, to sift through datasets rather than writing longer, more complex queries or ask questions, as it is often the case for informational queries on the web. It may also indicate users' perception that the search capabilities are limited – hence, issued queries are more general in their substance and results are filtered manually by the user or through filters.

6.3.2.2 Query Types

Countries and locations are difficult to detect automatically – in this set there are also challenges around the languages used to denominate each of them. Therefore, we manually labelled two samples: one-word queries used in 20 or more sessions (467 queries in total), and the same sample of multi-word queries used in 5 or more sessions that we used to estimate language distributions (386 queries in total).

We used a following set of annotation guides while annotating the data:

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

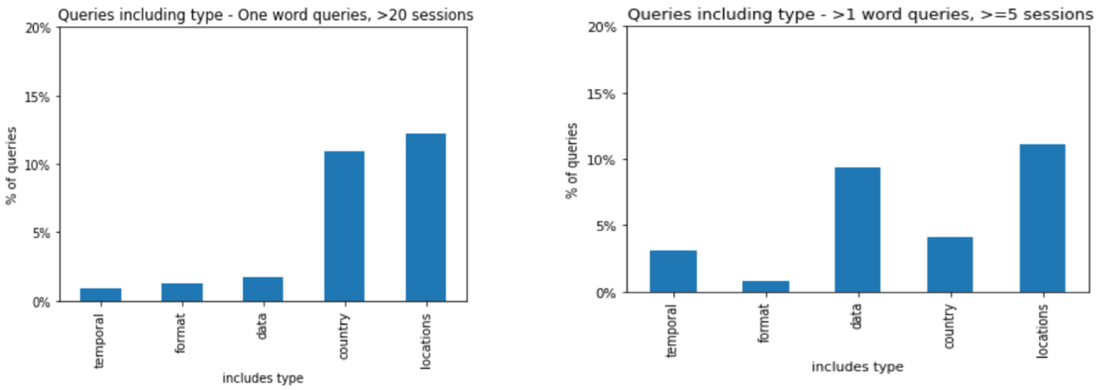


Figure 6.8: Percentage of queries including keyword types, for single-word queries (left) and multi-word queries (right)

Figure 6.8 summarises the results. Single-word queries have less than 3% of temporal, format, and data types, but more than 10% of both country and locations. The relatively high usage of countries and locations in single-world queries is noteworthy, as this is conceptually similar to using a country or location facet - which we analyse in the next chapter. Multi-word queries have a slightly higher usage of temporal types, data-related keywords, and fewer country names. Compared to the results reported for the national portals analysed, we note less temporal, format and data-related keywords are much less frequent, while geospatial (cities, regions etc.) ones are more popular. While the samples are small, further studies should explore whether this is due to of the nature of the EDP, which harvests across multiple portals from different levels of public administration in different countries.

Query language. In addition to above metric, we considered the language used to write the queries for EDP, accounting for the European character of the EDP. We wanted to know whether English is the main language of choice of EDP users to express their data needs or whether other languages are used as well.

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

6.3.3 Crowd Queries

In this section we present the results of our analysis of the *crowd queries* (queries generated by crowdworkers in a CrowdFlower tool based on data requests) and compare them to the internal queries presented earlier in this work.

6.3.3.1 Query Length

The majority of internal queries is between one and three words of length, with an average of 2.03 words per query. We found *crowd queries* to be significantly longer than internal queries with an average of 9.16 words per query.

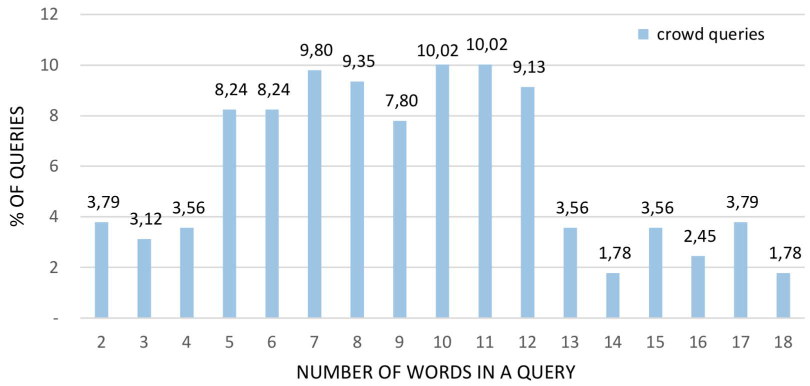


Figure 6.9: Percentage of queries according to number of words in them

Figure 6.9 shows an overview of the percentage of queries by number of words per query, for both the crowd and the internal queries. As discussed earlier we can see single word queries represent almost half of the entire corpus of the internal queries whereas the *crowd queries* had a minimum of 2 words, with the most queries between 7 to 11 words.

We believe this difference in query length indicates that the internal queries might not represent realistic search strategies, but rather expose limitations of current dataset search. Users do not expect search functionalities to fulfil their information need when searching for data, which can lead to underspecified queries (Koesten et al., 2017).

6.3.3.2 Query Types

As for internal and external queries we analyse crowd queries using metrics such as: geospatial, temporal, numerical information or appearances of file type and acronyms in the query. Table 6.6 summarises the percentage of queries for each of the metrics. Geospatial information was much more prevalent in the *crowd queries*: 36.1% of those contained a location in comparison to only 5.4% of internal queries and 12.01% in general web search (Gan et al., 2008). In contrast to searching on a data portal, which is often tied to a specific location or can have national boundaries attached to it, our experiment did not specify a particular location and was so less constrained from a geospatial point of view. Participants may have compensated for this by specifying location keywords. However, the high number of location bound keywords (36.1%) may simply emphasise the high importance of location in data search. Temporal information was seven times more popular in the *crowd queries* 49.2% in contrast to the results achieved by internal queries (7.29%) and 32 times more prevalent than for general web search (1.5% (Nunes et al., 2008)). Users indicate interest in different aspects of temporal information: date of data creation, the frequency of data releases, updates and time frames described in the data. File and dataset types (such as queries containing *csv* or *json*, etc as can be seen in Table 6.6 were much more popular within *crowd queries* (49%) in comparison to the internal queries for which file types were reported for 6.25% of the queries. This could be due to filtering options over file types on the data portals in which the internal queries were recorded. Crowd workers could further be biased in their creation of queries in thinking they need to add the word *data* to a query for data (as would be the case on general web search). Excluding the word data from this analysis we found 26.95% queries including common file types, as can be seen in Table 6.6. The percentage of queries containing numbers, that were not temporal information, were 5.57% and there were no queries containing only numbers. These results are similar to those reported for internal queries (5.23%). Numbers in queries represent mainly sample sizes or desired constraints to the data, such as: *Police spending over £500 local data.* We further report the percentage of queries including abbreviations. We found 2.23% in the *crowd queries* included abbreviations; in comparison to 5.11% reported for the internal queries. Abbreviations were mostly used when they appeared in the data request that the query was based on.

Question queries: In terms of question queries, less than 1% of the internal queries are structured as questions. The low number of question queries in dataset search might be

Metric - Definition	% portal	% crowd
Geospatial - the name of a city or geographical area (either town, city, county, region or countries)	5.4%	36.1%
Temporal - years (1000 to 2017), names of months, days of a week and the words *week(ly)*, *year(ly)*, *month(ly)*, *day(ly)*, *date*, *time* and *decade*	7.3%	49.2%
File and dataset type - file types: *csv, pdf, xls, json, wfs, zip, html, api* and keywords denoting a type of dataset: *data, dataset, average, index, graph, table, database, indice, rate, stat*	6.3%	49%
Numbers - the number of queries including numbers excluding those indicating time frames	5.2%	5.6%
Only numbers - queries that contain only numbers	0.4%	0%

TABLE 6.6: Definition of query characterisation metrics. Percentage of queries for both, internal queries and crowd queries of this study

caused by the lack of question-answering capabilities of the dataset search functionalities. This might be also connected with users issuing only short keyword queries as they might not trust the system to handle longer (or more advanced as questions are) queries. We found 9.35% of *crowd queries* were questions. This follows the web trend of issuing more question queries, but also may be connected with the fact that larger search box used in our experiment could have encouraged longer queries. It is important to remember that crowd-workers were asked to create a hypothetical query for a given scenario meaning the experiment did not mimic fully the natural search scenario which could cause vast over representation of question queries.

6.4 Discussion and Implications

We found that most queries issued directly on the portals (i.e., the *internal* queries) were related to datasets in the area of *business* and *economy*. By contrast, external queries were topically more diverse, with topics such as *society* and *towns and cities* appearing regularly. We also noticed differences in the ratio of question queries - a larger percentage of external queries included question queries. This may indicate that different ways of accessing the portal could be related to different types of information needs (e.g., specific answers versus full datasets). Further analysis is needed to determine whether internal and external queries are indeed authored by distinct user groups and where these differences comes from. In our previous interview study (Koesten et al., 2017) we found evidence that there is overlap between the two groups. From the point of view of description metadata, these results suggest that open data portals should focus on providing business related themes and concepts.

Our findings show that dataset search queries are generally short, on average one word shorter than web search queries, as per the 2011 report by Taghavi et al. (2012). We believe short queries potentially indicate that, currently, users do not expect that the search functionality will be able to provide relevant data for longer and more specific queries. It appears that users currently tend to treat the search box of a data portal as a starting point for further exploration (one of the most popular queries on data.gov.uk were *crime*, *lidar*, *defra*, *flood*, *population*). The categories and metadata attributes used in data portals as well as enabling linking between datasets or metadata properties could be key in improving dataset search functionalities.

We believe both temporal search, which is more prevalent than reported for general web search (Nunes et al., 2008), and geospatial search (Gan et al., 2008) require better support. In both cases, relevant keywords can have different levels of granularity (e.g., months versus years, cities versus regions or countries), which is not always matched by the publishing practices of the data owners. While four of the data portals we analysed are location-bound to a country and most datasets hold national data, supporting question-answering and dataset search scenarios will require more advanced geospatial indexing and reasoning features, and even more so on a supranational portal such as EDP. For national portals queries including some indication of time were almost five times more frequent than in web search (Nunes et al., 2008), suggesting that datasets have a stronger relationship to time than documents. DCAT already includes properties for temporal and geospatial description of datasets, and our findings suggests that providing fine-grained descriptions of these properties could improve search experience.

We found that queries for data differ between those issued on a data portal and those created in an environment with fewer constraints. The queries generated in this study were longer and included approximately seven times more temporal and geospatial information. The higher importance of these information types has been recognised in literature (Kern and Mathiak, 2015; Kunze and Auer, 2013). Structurally we found the *crowd queries* to include a higher percentage of questions and 4 times as many queries included a specific file type or format. The length of these queries suggest that the information need expressed in the data requests are complex; based on literature we believe this is typical for data centric information needs (Koesten et al., 2017). In comparison, the internal data portal queries were short and underspecified. Although in both of the query sets people were looking for data, we believe that neither set necessarily represents how people would like to search for data. These findings emphasise a large design space for data search environments; one possible direction being encouraging users to issue longer queries, for instance by providing tailored search functionalities, larger search boxes or suggesting additional keywords. Search log analyses can illustrate the specific characteristics of a given search vertical. We know that queries on portals are underspecified, but crowd-generated set of queries shows that when asked to search outside of a search environment people issue much longer queries which correspond to complex information needs for data. The high prevalence of geospatial or temporal information in the *crowd queries* should inform the design of dataset search systems, for instance by allowing users to search by specific locations or time frames. It could further indicate a need to extend existing metadata standards to include these two types of information, which could then be exploited by search functionalities. We believe new retrieval models for dataset search, that take the unique characteristics of this information source into account, are needed to make data on the web more discoverable.

6.5 Limitations of the study

6.5.1 Search Logs

Comparisons of different search log analyses present difficulties as concluded by Jansen and Spink (2006) in their study comparing nine search engines by their transaction logs. Even within web search, it is stated that findings resulting from the analysis of one search engine cannot be applied to all web search engines. Following this, any comparison of our results with web search needs to be seen with caution, due to the different nature of the collected data. However, we believe that including data from several countries and different audiences increases the generalisability of our analysis.

Our study is based on dataset search engines that are part of governmental open data portals. Further studies with other kinds of dataset search engines are required before drawing general conclusions. Query topics were annotated with the use of tags from one portal (data.gov.uk). We decided to use those categories as they present an overview of the content of governmental open data portals. Furthermore, as all portals but EDP used the Google Analytics suite, we were subject to its session definition and identification algorithm. We complement our analysis with session log analysis in Chapter 7 of this work.

As we did not have control on the analytics being collected by each data portal, we had different time frames and data for each one. In cases where all queries were considered, there is a bias towards DGU, as the portal with most available data. We had no means of detecting potential automated user agents which might have influenced some of our statistics. Finally, due to privacy considerations, we did not have access to a large number of external queries.

In addition, further limitations resulting from sessions data collection for the EDP will be discusses in Chapter 7 where the full analysis of this log will be presented.

6.5.2 Crowd-generated queries

As with any experiment using human computation, instructions and the experiment design influence the outcome. We tried to take into account that workers might not know what *data* is and used a spreadsheet as well as a product search analogy in the instructions. We had no control of the workers prior experience and their conceptual models of data. However, this is a natural limitation of such experiments. While we acknowledge that the *crowd queries* are created in an artificial setting, without the workers own naturalistic information need, we believe that they give us relevant insights into how queries for data could potentially look in the future. We acknowledge that neither query set can be a representative reflection of how people would search for data in an "ideal" system. However the results of this work can be seen as an approximation that can inform further research.

6.6 Conclusion

In this chapter we presented an analysis of search log data for dataset retrieval, based on internal search logs of four national data portals and one supranational European data portal, external search logs of two of the portals that were based in the United Kingdom along European data portal and queries generated through crowdsourcing using requests for data, in order to understand how data portal users search for data and provide insight about what are the most important features of descriptive metadata from the point of view of data users.

Presented analysis allowed to uncover initial characteristics of dataset search and which features were the most prevalent. We compared our findings with search patterns occurring in the less constrained environment and with the general web search. Our findings can be summarised as: (i) Dataset queries are generally short. (ii) There is a difference in topics, length and structure between dataset queries issued directly to data portals and dataset queries issued to web search engines. (iii) Dataset search logs do not fully represent the behaviour of users searching for data, but rather uncover the limitations of current search functionalities. (iv) Our analysis suggests that the prioritary properties to describe datasets are topical, temporal and geospatial coverage, with varying levels of granularity. All of them already exist in current vocabularies. Our results suggest that efforts on automatic generation of dataset descriptions should be focused on these properties.

In analysis of search queries the context of the whole session is missing. In order to evaluate the system and how it serves it's purpose to the users one needs to analyse whole user journey. For this reason in next chapter of this work we conduct a session log analysis of EDP portal, building on findings of this and previous chapters.

Chapter 7

Sessions Analysis

In this chapter we expand the search queries analysis presented in Chapter 6 where we analysed the dataset search from sole perspective of queries issued on various portals. In order to gain a more complete picture we conducted a session log analysis with the available data for one of the data portals - EDP.

Our analysis suggests that many EDP users land on the dataset section from searching with general web search engines. This means portal designers could focus less on improving the accuracy of search results, and instead understand the implications of people's using external search tools on user journeys (e.g. adding relevant metadata improving the crawling process). In a dataset search context, approaches need to consider aspects such as data provenance, annotations, quality, granularity of content, and schema to effectively allow users to evaluate a dataset's fitness for a particular use (Chapman et al., 2020). The user does not have the ability to introspect over large amounts of data, and they need guidance and tools support. Users are attempting to discover and assess datasets for a particular purpose. Supporting them requires frameworks, methods and tools that specifically target data as its input form and consider the specific information needs of data professionals.

In this chapter we aim to answer the following research questions which related to first main research question [RQ1] *Understanding Dataset Search; How do users search for data? What search strategies do they use?* with the aim of analysis of the users of the portal through the searching process and not only throughout queries they issued on the portal. We take advantage of the availability of a dataset from the EDP containing whole search sessions, to go beyond an analysis of only queries and to an analyse the behaviour of users, how they use the search functionalities available in the portal and how successful they were. The session data allows us to answer the question regarding the search strategies used by users of the portals as the steps taken by users when searching for data are visible throughout their sessions behaviour.

1. Dataset search in the context of the EDP
    (a) How is the dataset search section of the EDP used? Are there variations across portal versions?
    (b) How does dataset search compare to other sections of the portal? Do users visit several sections of the portal in the same session? Are there variations of the above across portal versions?
2. Dataset search strategies
    (a) How do people search for datasets on the EDP? Can we identify particular search strategies that are more popular than the others?
    (b) What are the most popular facet filters?
    (c) What are the most popular combinations of facets?
    (d) Is there any difference in the use of facets when the user issues queries via the dataset search box?
3. Success in dataset search
    (a) Are users successful when they search for datasets on the EDP? Does success change across portal versions?
    (b) Is there any difference in success between internal search (EDP's dataset search box) and external queries (from web search engines)?
    (c) Is there one search strategy (search box only, facets only, mixed) more successful than others?

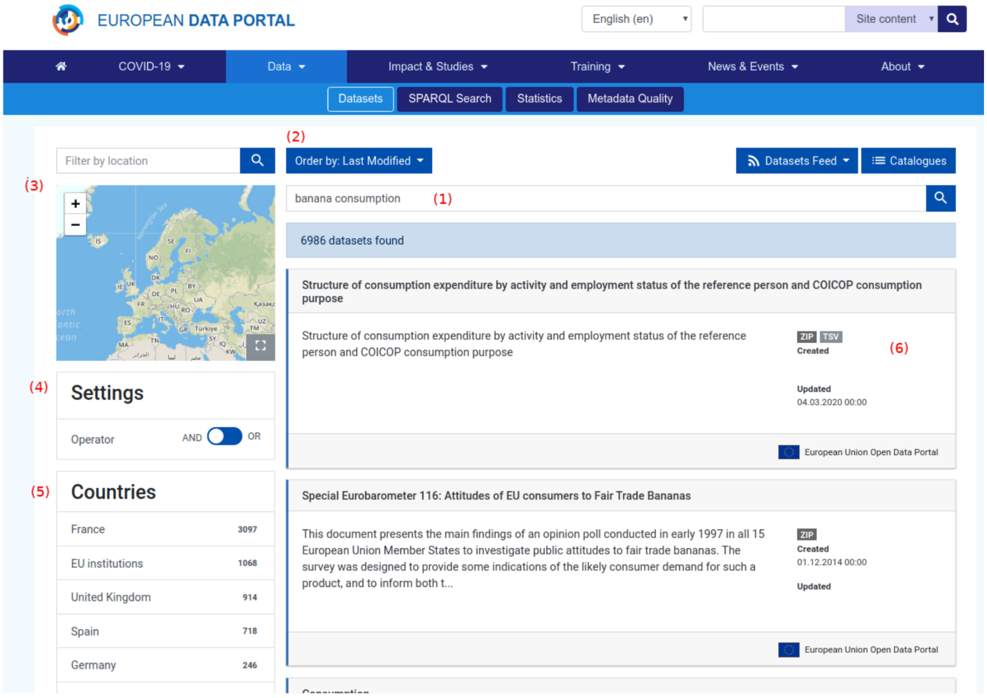
To address these questions, we conducted a quantitative analysis of $844,343$ EDP user session logs from April 2018 to June 2020. A user session log includes, among other things, which website referred the visit to the portal, the pages visited during a session, the queries issued to the dataset search section of the portal, and which datasets were accessed from the portal. As for success in dataset search, no specific studies for satisfaction in dataset search exist yet, we rely on the state-of-the-art techniques for measuring and inferring user satisfaction on document search engines and discuss how they transfer to dataset search.

Following our finding we conclude this chapter with a discussion of the main findings of the search and interaction logs analysis, along findings of previous chapters with aim of defining dataset search as it's own search vertical and gain insights into potential next steps in dataset search development.

7.1 The EDP Dataset Search Interface

EDP's dataset search interface follows a traditional structure in the style popularised by digital marketplaces, and in use by most national data portals across the world. Figure 7.1 shows relevant components: (1) Dataset search box: Where users type their queries; (2) Order by selector: Allows re-ordering results according to one of the criteria, such as: relevance to the query keywords; descending date of modification; descending date of creation; ascending alphabetically by dataset name; descending alphabetically by dataset name.

The leftmost column lists the available "Facets". Facets are filters that can be combined by users to narrow down the number of obtained results. Figure 7.1 shows three of the facets available in the EDP (3) location, (4) operator, and (5) country.



banana consumption

6986 datasets found

Structure of consumption expenditure by activity and employment status of the reference person and COICOP consumption purpose

Structure of consumption expenditure by activity and employment status of the reference person and COICOP consumption purpose

Created

Updated
04.03.2020 00:00

European Union Open Data Portal

Special Eurobarometer 116: Attitudes of EU consumers to Fair Trade Bananas

This document presents the main findings of an opinion poll conducted in early 1997 in all 15 European Union Member States to investigate public attitudes to fair trade bananas. The survey was designed to provide some indications of the likely consumer demand for such a product, and to inform both t...

Created
01.12.2014 00:00

Updated

European Union Open Data Portal

Settings

Operator: AND ● OR

Countries

France	3097
EU institutions	1068
United Kingdom	914
Spain	718
Germany	246

FIGURE 7.1: EDP's dataset search interface

The facets available on the EDP are:

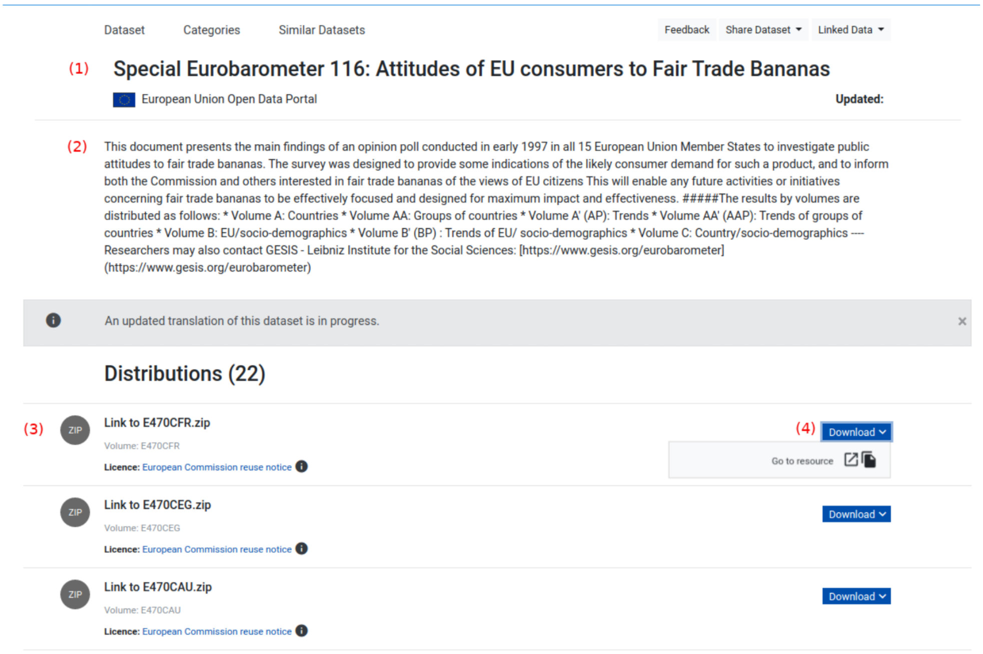
1. Location: Approximate geographic area covered or referred by the dataset: Control: Mini map where user can draw a rectangular area to approximate the location of interest, or name of the location (city, region, country). The metadata used is DCAT-AP geospatial information and if not available, a region or city of a portal on which the data was published originally is used.

2. Operator: Sets how multiple facets should be combined. Options: logical AND (show results that match all facets), logical OR (show results that match any of the facets)

3. Country: Country of the dataset. Options: one can select European Union countries
4. Catalogues: Data catalogue from which the dataset was harvested. Options: All catalogues linked to at least one dataset in the result set.
5. Keywords: Dataset keywords according to the DCAT-AP description of the dataset. Options: All keywords detected in the datasets in the result set.
6. Licences: Licence(s) of distribution of the dataset. Options: All licences detected in the current result set.
7. Formats: Format(s) on which dataset distributions are available. Options: All format types detected in the current result set.

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

When a user clicks on a summary, they are redirected to the page of the corresponding dataset (Figure 7.2), with the full details about the dataset: (1) title, (2) description (3) distributions, and (4) for each distribution, a link to download or go to the page of the resource in the catalogue of origin.



Special Eurobarometer 116: Attitudes of EU consumers to Fair Trade Bananas

European Union Open Data Portal

Updated:

This document presents the main findings of an opinion poll conducted in early 1997 in all 15 European Union Member States to investigate public attitudes to fair trade bananas. The survey was designed to provide some indications of the likely consumer demand for such a product, and to inform both the Commission and others interested in fair trade bananas of the views of EU citizens This will enable any future activities or initiatives concerning fair trade bananas to be effectively focused and designed for maximum impact and effectiveness. #####The results by volumes are distributed as follows: * Volume A: Countries * Volume AA: Groups of countries * Volume A' (AP): Trends * Volume AA' (AAP): Trends of groups of countries * Volume B: EU/socio-demographics * Volume B' (BP) : Trends of EU/ socio-demographics * Volume C: Country/socio-demographics ---- Researchers may also contact GESIS - Leibniz Institute for the Social Sciences: [https://www.gesis.org/eurobarometer] (https://www.gesis.org/eurobarometer)

An updated translation of this dataset is in progress.

Distributions (22)

Link to E470CFR.zip
Volume: E470CFR
Licence: European Commission reuse notice

Link to E470CEG.zip
Volume: E470CEG
Licence: European Commission reuse notice

Link to E470CAU.zip
Volume: E470CAU
Licence: European Commission reuse notice

FIGURE 7.2: Dataset page

7.2 Methodology - Quantitative analysis

In this section we describe the analysed data in the session log and its collection, the limitations and introduce portal versions - changes introduced to the portal which each new version roll-out to the user. Data was collected by the EDP consortium and provided stripped of online identifiers (ip and location guessed from IP). The secondary analysis described in this section was commissioned by the EDP consortium as part of an analytical report (Ibáñez et al., 2020) and approved by the Ethics Board of the University of Southampton (#57118).

7.2.1 The search and interaction logs corpus

In this chapter we use a following nomenclature: **Actions and interactions on the portal**: For the web analytics package we use, actions are a superset of interactions. An interaction is a page view or a site search, while an action also includes downloads, outlinks and other events. **Search session**: A logged session that includes one or more queries or facets use. **Search sessions vs. visits**: We differentiate between search sessions on the portal and visits, the latter being triggered by external queries leading to dataset pages.

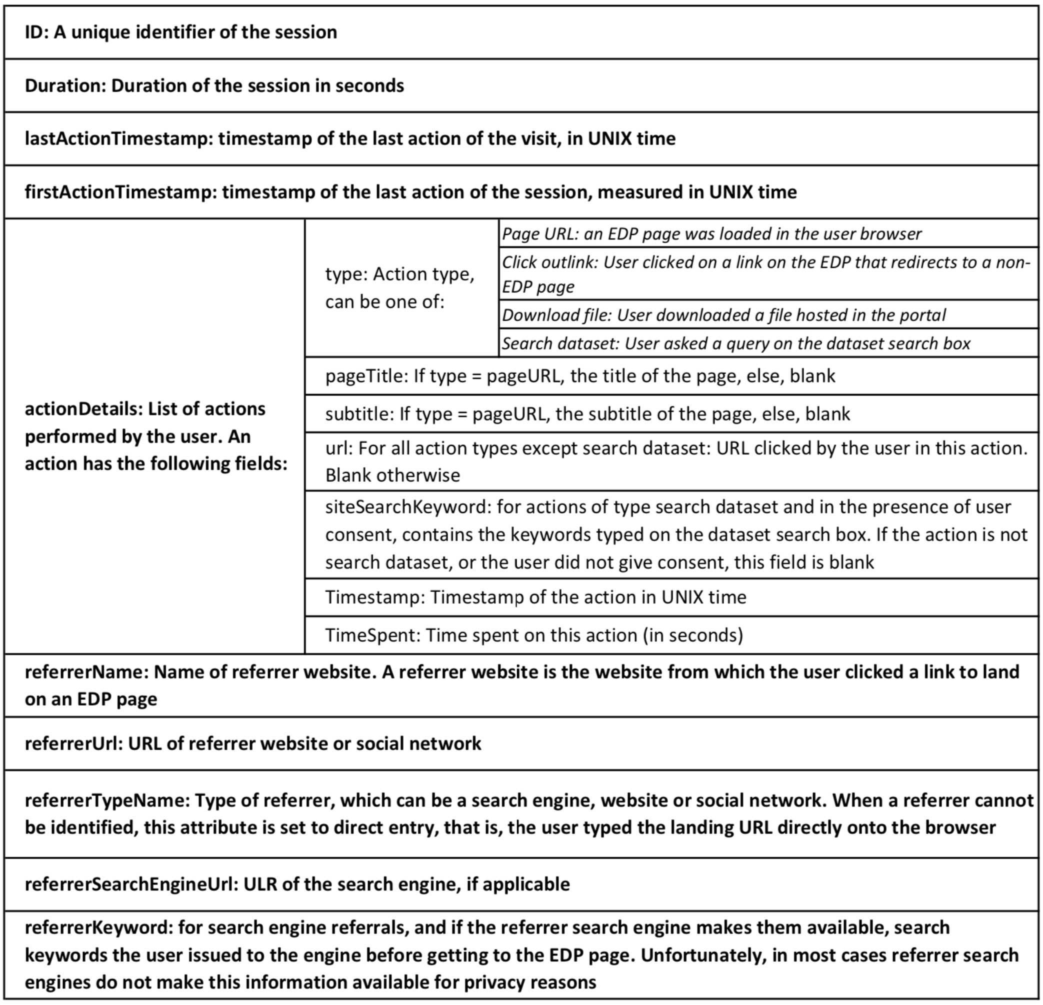
The European Data Portal, in contrast to previously analysed national data portal which use Google Analytics, uses the Matomo Web Analytics suite to log the actions of users of the portal for each of their visits. A description of the capabilities of Matomo Web Analytics is available on this link[1]. In the following we provide a summary of those data attributes which feature in our analysis:

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

1. Session timeout (in minutes): This parameter refers to how long a web analytics package should wait after the last recorded action to consider the session finished. If a user returns to the portal within this time, their subsequent actions will be recorded as part of the same session; otherwise, the activities will be recorded as a new session. In EDP this value set as 30 minutes – this means that if no activity has been recorded for 30 minutes, any subsequent activity will be considered to belong to a new session. Setting the right value for this parameter should be the subject of future studies to better reflect the realities of dataset search and allow for more accurate follow-up session analytics.

2. Exclusion of bots: Bots are automated agents that crawl websites. As we are interested in understanding the behaviour of people on the portal, visits from bots

[1] https://matomo.org/faq/general/faq_18254/



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

FIGURE 7.3: Summary of data attributes in session log

should be excluded. 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 make it to the portal. Periodic analysis of this dataset for internal EDP reporting found no indication of skewing due to bot traffic.

3. Time spent measurement: Matomo's EDP kept a default configuration that does not allow the measurement of time spent on the last page of a session. This means that the available duration (in seconds) of a session is a lower bound of the real time spent by the user. We consider this a limitation of this study.

As the analysed platform changed over time we decided to report our results for three major releases of the portal which were rolled out since March 2018. The collected sessions were split according to those timeframes into three disjoint sets. Table 7.1 summarises the characteristics, date range and number of sessions in each category.

Below we describe their timeframes and main changes which might have affected the dataset search analysis:

- EDPv1: From the 1st April 2018 to the 2nd April 2019 EDP's native dataset search engine was based on CKAN.
- EDPv2: From the 2nd April 2019 to the 6th March 2020, the dataset section and search engine migrated to a solution developed in-house by the EDP team. The URL scheme of the dataset section was changed, leading to a period where web search engines had to re-index those pages.
- EDPv3: From the 7th March 2020 onwards, improvements were introduced on the dataset search engine following the evolution of the DCAT-AP standard. This more or less coincided with the introduction of lockdowns in many European countries due to COVID-19 pandemic, which, as the logs show, affected traffic on the portal.

Portal version	Description	Date range (time in GMT time zone)	No. of sessions
v1	Dataset section and search engine based on CKAN	01/04/2018 00:00 – 01/04/2019 23:59	430.815
v2	Dataset section and search engine based on EDP's code base. Change of URL scheme on dataset section.	02/04/2019 00:00 – 06/03/2020 23:59	283.470
v3	Approximate start of COVID-19 outbreak in Europe. COVID-19 section on EDP Improvements on dataset indexation following DCAT-AP evolution	07/03/2020 00:00 – 30/06/2020 23:59	130.058

Table 7.1: EDP portal versions differences and split highlight.

We than further classified sessions according to the activities carried out by users. The classification looks as follows:

1. We call a session a *dataset search session* if one of the following applies:
   (a) The session includes actions related to entering a query into the search box one or more times.
   (b) The session includes actions related to filtering results with a facet one or more times.
2. We call a session a *dataset page session* if that session includes a visit to at least one dataset page (see Figure 7.2) and if the session is not a dataset search session. In other words, these are the session where the user landed straight on a dataset page without using the search box of the EDP. This happens, for instance, if the

user was referred to that dataset page via an external website or by a web search engine.

3. We call a session a *homepage bounce* when the user visits the EDP front page or the dataset search main page, but then does not follow up to any other parts of the portals and does not trigger any dataset searches.

4. We call a session a *section session* when the user visits at least one of the other main sections of the EDP site, including: News, Events & Highlights; Training & E-Learning; Reports Studies; and COVID-19. The COVID-19 section is also divided in sub-sections similar to the ones of the main portal, including a dataset section featuring datasets manually curated by the EDP's editorial team. We consider the sessions that contain a visit to a page in the COVID-19 dataset section separately (we refer to that sub-corpus COVID-Data in our analysis, see Table 7.2), and those that visit one of the other COVID-19 sections, but do not visit any COVID-19 dataset section page (which we refer to as COVID-Other, see Table 7.2).

7.3 Sessions Analysis

In this section we analyse the search and interaction logs to answer the three sets of research questions introduced earlier in each subsection.

7.3.1 Dataset search in context of the EDP

To understand how is the dataset search section of the EDP used, how it is influenced by differences in portal versions and how it compares to to other versions of the portal we first analyse the split of sessions according to their type as introduced in the Methodology section. Figure 7.4 compares the percentage of sessions that visit each section, for each portal version. Absolute numbers are detailed in Table 7.2.

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

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

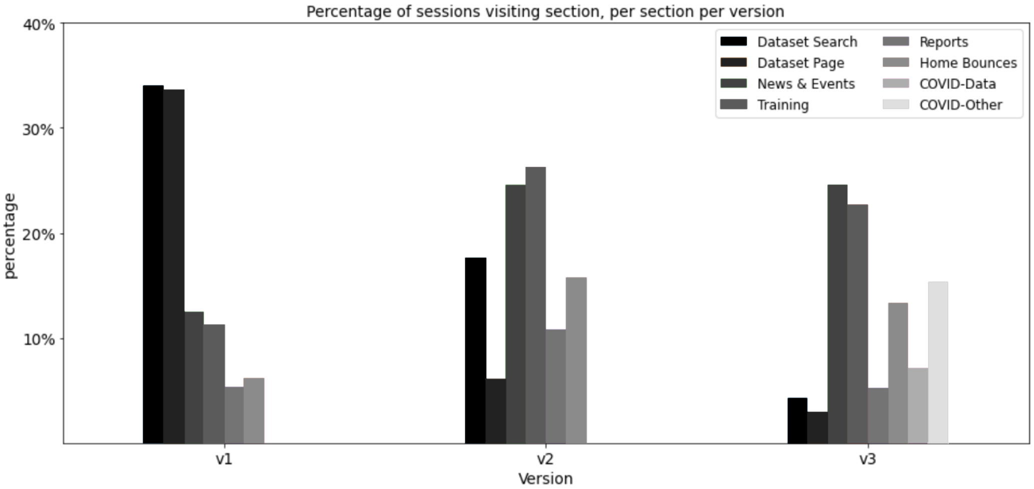


Figure 7.4: Percentage of sessions that visit each portal section

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

Table 7.2: Number of sessions per category per version of the EDP

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

Next, we analysed the number of cross-section sessions – that is sessions where the user visits multiple sections of the EDP site, we counted for each section the number of sessions that only visit that section and divided it by the number of sessions that visit that section or others. Results are shown in Figure 7.5. We observe that across all versions of the portal, more than 80% of visits to the dataset section do not crossover to other sections. There is a similar trend among the other sections except for the *Reports* section. This suggest that users are not looking for additional resources, which might allowing better understanding of a datasets once it is found. Potential interface changes might encourage further look though the portal resources.

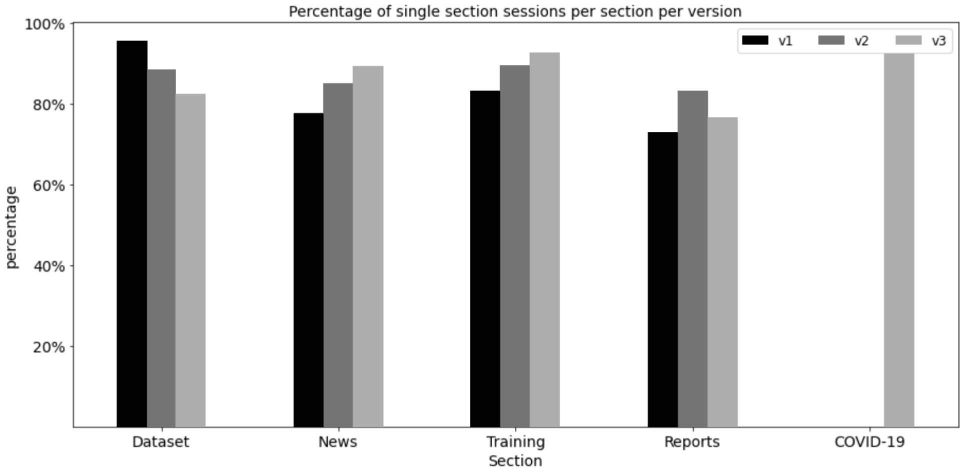


FIGURE 7.5: Percentage of single-section sessions

We took a closer look to the crossovers between the dataset section and the others, as shown on Table 7.3. Crossovers between sections in v1 was minimal, but increased for v2 and v3, in particular with the news, events & highlights sections. We believe this effect comes from the decrease in the number of visits to datasets (because of the change in URLs, for instance) combined with an increased use of highlights to announce new datasets. This showcases the added value of auxiliary content such as news, stories etc. that facilitate reuse.

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

TABLE 7.3: Crossover between the dataset section and other sections

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

A large majority of users visit only one section of the portal at a time. For the dataset sites (search, dataset descriptions) the number of crossovers with other sections was less than 5% in v1, but increased up to 18% in v3, mostly due to more links to and from news, events & highlights section. The trend is opposite for the news and training sections. In summary, there are significant variations between different release versions of the portal. Having gained a high-level understanding of the search and interaction sessions across the three versions of the EDP, we now investigate the search behaviour in terms of search strategies.

7.3.2 Dataset search strategies

In this section we analyse how people search for datasets which include the analysis of search strategies and used facets filters. This subsection expands on the queries analysis conducted in the previous chapter of this work.

In order to find a suitable dataset or datasets user might employ one of the the three different strategies to search on the EDP. For instance, they could: (a) type keywords into the search box; (b) apply facets to filter datasets; and (c) a combination of the above.

7.3.2.1 Analysed Logs

In order to answer our research questions we conducted additional filtering over the session log as not all search and interaction logs are relevant to understand search for datasets. To restrict the analysis only to the relevant sessions, we removed sessions where a user has visited the portal with aim different than dataset search (e.g. access news, learning materials etc.) In this section, we explain how we put together the subset of logs that are relevant to dataset search sessions.

Our starting point is the dataset search sessions dataset. As explained earlier, it is made of sessions that include at least a keyword query or the application of a filter via one or more facets (see Figure 7.2). Furthermore, we distinguish in our analysis between two scenarios for users to reach EDP's dataset search pages:

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

eu/data/datasets?keywords=karte). We refer to these as external dataset search sessions. Those are further split into :

(a) External search sessions that do not include any further queries or use of faceted search. In other words, the user has issued a query outside of the EDP and clicked a link that sent them to EDP pages. We call this set of sessions in our corpus external landing without further search.

(b) External sessions that include further use of keywords or facets past arrival on an EDP page. In the case, the user landed on some EDP content following an external link, then went ahead and searched on the EDP using native tools. We call this external landing with further search.

Table 7.4 shows the number of sessions of each of the datasets described above. In our analysis we cover both cases of internal search sessions types: search initiated on the portal and search initiated via external tools but continued internally. For the sake of simplicity, we refer to both these categories as internal dataset search for the remainder of this section.

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

Table 7.4: Breakdown of dataset search sessions by starting point and search continuation

7.3.2.2 Use of search box and facets

Earlier in this section we noted that users can follow different strategies to look for the data they need, including queries, facetted search or both. Figure 7.6 shows the percentage of sessions per search strategy for the three versions of the EDP. Using only facets to search is significantly more popular than the other two throughout the evolution of the portal.

Next, we calculated the share of sessions that use each of the available facets at least once. Figures 7.7 and 7.8 shows the results for sessions that used only facets and those relying on keywords and facets.

For those cases where the users relied only on faceted search, we note a slightly more even distribution of facet types in EDPv1 than EDPv2 and EDPv3. V1 was also the corpus for which the majority of recorded user activity was dedicated to search. In EDPv1,

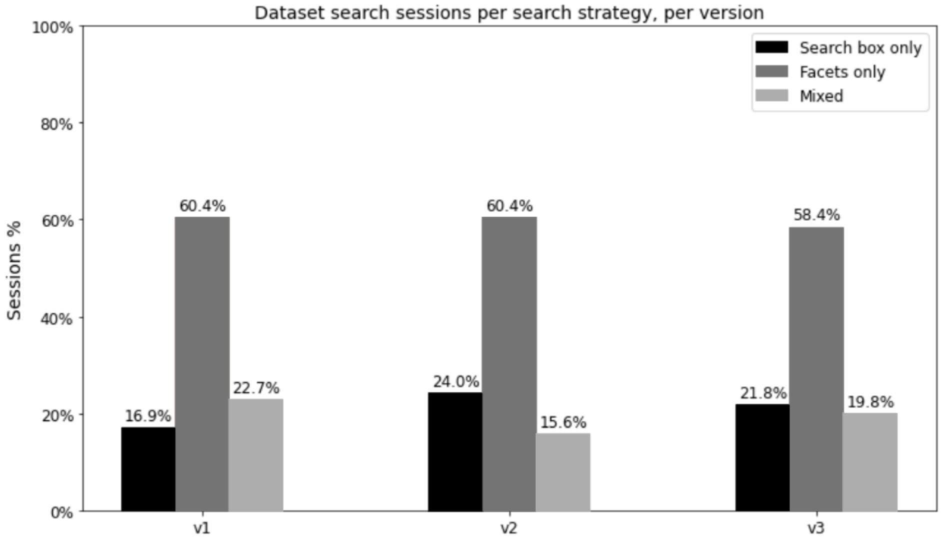


Figure 7.6: Dataset search sessions per search strategy, per version

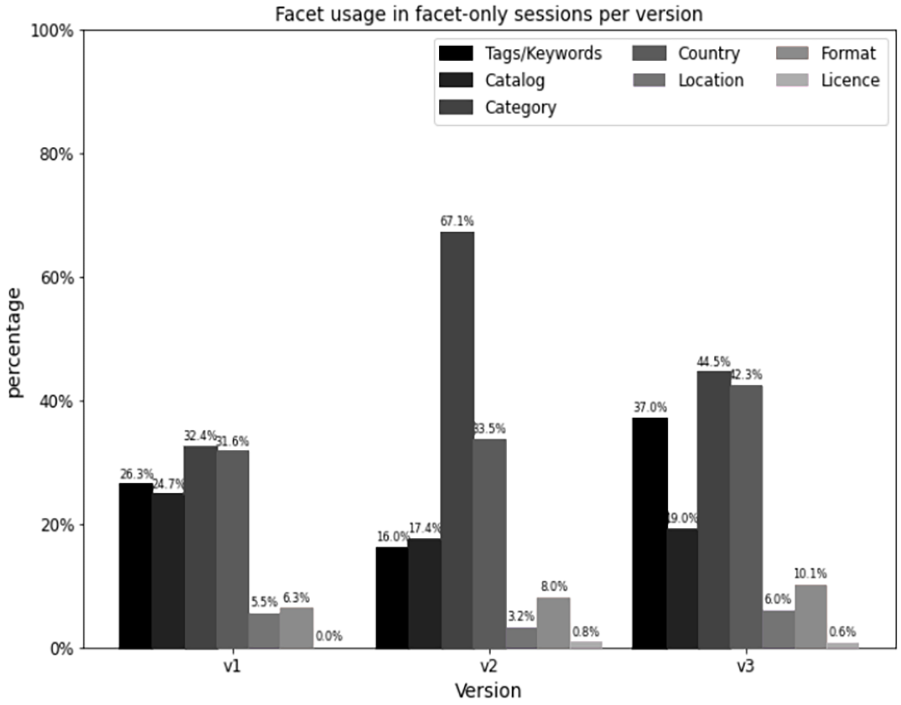


Figure 7.7: Use of facets per version for sessions that used facets

people used tags/keywords, catalogue, category and country extensively. For EDPv2 the category facet is by far the most popular; we hypothesize that this is a function of the homepage re-design, which includes a panel where users can start exploring the datasets along categories. For v3, the difference between category and the others does not stand out as much. Overall, we observe that facets related to the format or license of the dataset, as well as its geolocation are under-used. This is in stark contrast to some of the insights we gained from analysing search logs in previous chapter of this

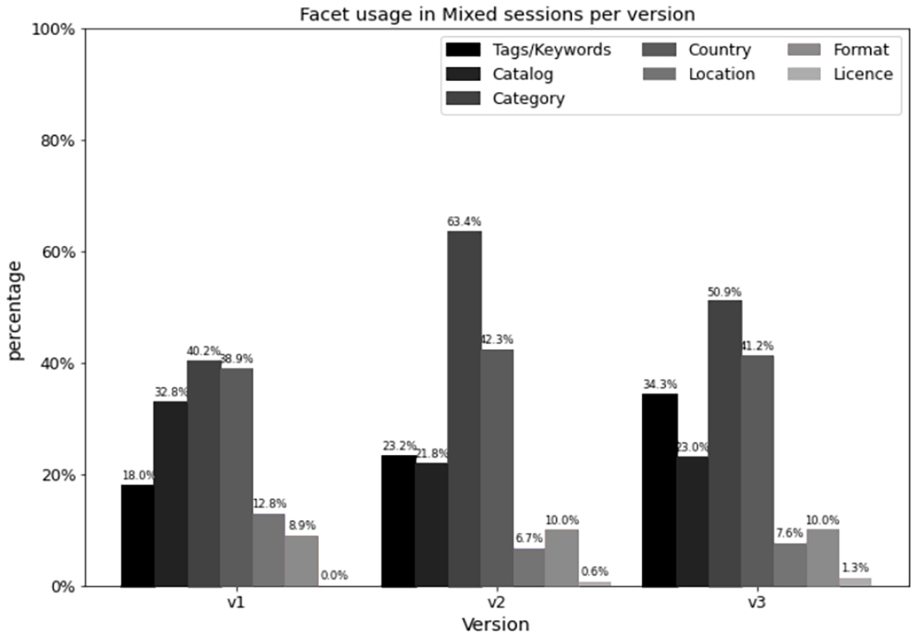


Figure 7.8: Use of facets per version for sessions that used search box and facets

work or from interviews with data practitioners about their search strategies (Koesten et al., 2017), which suggested that these are three key attributes that decide if a dataset will be eventually reused.

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

Finally, we calculated the most popular facet combinations for the two types of sessions from Figures 7.7 and 7.8. This helps us understand the type of information needs people have and the attributes of the data that matter to them while looking for data. Table 7.5 shows the combinations found in at least 5% of the sessions. Across all versions, the most common combination involves countries and categories. The popularity of this combination greatly increases from v2 onwards, possibly in relation to the new design of the site which promoted dataset exploration by category. We also note that during v1, there was a lesser use of combined facets, despite a higher share of search sessions overall.

In summary, the use of only facet filters is more common than the two other strategies combined. Country and category are the most popular, and also the most popular combination, followed by combinations of each of these two with the keywords facet.

Version	Only facets	Query & facets
EDPv1	(Country, Category) → 5.4%	(Country, Category) → 9.9%
		(Catalog, Category) → 8.3%
		(Country, Location) → 8.0%
		(Catalog, Country) → 7.6%
		(Tags, Country) → 6.6%
		(Country, Format) → 5.6%
EDPv2	(Country, Category) → 17.42%	(Country, Category) → 23.1%
	(Country, Keywords) → 6.8%	(Country, Keywords) → 10.7%
	(Catalog, Category) → 6.2%	(Catalog, Category) → 10.3%
	(Catalog, Country) → 6.2%	(Category, Keywords) → 10.0%
	(Category, Keywords) →5.5%	(Catalog, Country) → 9.0%
		(Country, Category, Keywords) → 6.8%
		(Country, Format) → 6.6%
		(Category, Format) → 6.2%
		(Catalog, Keywords) → 6.1%
		(Catalog, Category, Country) → 5.7%
EDPv3	(Country, Category) → 17.4%	(Country, Category) → 18.6%
	(Country, Keywords) → 9.8%	(Country, Keywords) → 11.7%
	(Catalog, Country) → 7.9%	(Category, Keywords) → 9.9%
	(Catalog, Category) → 6.5%	(Catalog, Country) → 9.9%
	(Country, Format) → 6.38%	(Catalog, Category) → 8.4%
	(Category, Keywords) → 6.0%	(Country, Format) → 6.0%
		(Country, Category, Keywords) → 5.2%
		(Catalog, Category, Country) → 5.1%
		(Catalog, Keywords) → 5.1%
		(Category, Format) → 5.0%

Table 7.5: Percentage of sessions with more than one facet per portal version. We only show combinations that reached more than 5%. Pairs are not ordered.

Format and licence are the least popular. Sessions that use the dataset search box use more combinations of facets than facet-only sessions.

7.3.3 Success in dataset search

We finally explore the users' success when they search for datasets. We assume that users who visit the dataset section of the portal have information needs of the form "Find a dataset(s) that is(are) relevant to my criteria". It is best practice to ask users who visited an online site to provide feedback on the purpose of their visit and whether they found what they were looking for. In the absence of such explicit user feedback, an alternative is to look for explicit actions (or lack thereof) that signal the success (or failure) in satisfying the information need of interest. As EDP did not collect explicit user feedback, we assume that sessions that include download or go-to-source activities were successful.

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

With respect to unsuccessful sessions (failures), we consider two scenarios: 1. "only SERP": The user only looks at Search REsult Pages and does not click on any of the result dataset pages. 2. "Dataset Page View" (DPV): The user clicks on at least one dataset page shown to them SERPs but does not download anything. Figure 7.9 shows for internal search sessions the percentage of successful, "Only SERP" and "DPV" sessions per version. According to our classification, 37% of relevant sessions were successful on EDPv1. That share dropped to 22% on EDPv2 and rebounded on EDPv3 to 40%. For EDPv1 and EDPv3 the proportion of "only SERP" and "DPV" failures is approximately similar, with a spike on "Only SERP" failures for v2. To verify if the spike is due to a temporal effect, we computed the distribution of "Only SERP" failures per month for versions v1 and v2, but we did not find any significant change. This suggests that the dataset search engine introduced in EDPv2 is less effective than in EDPv1, but the changes introduced in EDPv3 improved the previous situation.

Figure 7.10 shows the same data for external search sessions. We note a similar trend as observed for internal search, but with lower rates: EDPv1 and EDPv3 have around 25% of successful searches, but there is a drop in EDPv2 to less than 10%, combined with a spike on the number of "Only SERP" failures.

As we did with internal queries, we went on analyse the distribution of "Only SERP" failures per month for all versions to discover any temporary effects. This time we did find a larger number of failures on the first three months of v2. We analysed the titles and URLs of the landing pages of these sessions and found a high rate of "404 missing page" titles, suggesting that hits were linked to the wrong resources. Very likely this is a direct consequence of the change in the dataset section URL scheme introduced in EDPv2: many pages referred by web search engines that led to a failed search were facets. In particular, 78% were 'tags' facets (e.g. `http://www.europeandataportal.`

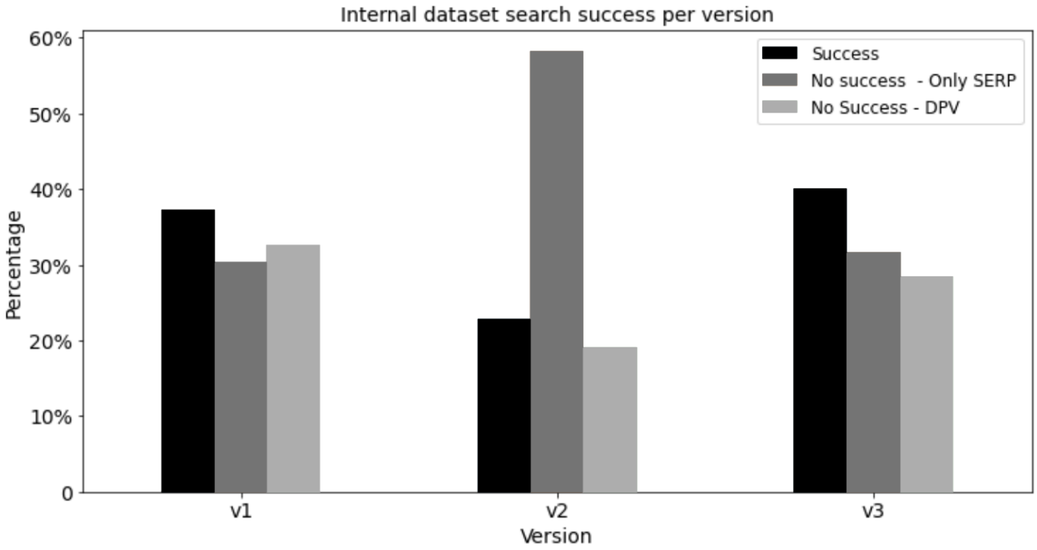


Figure 7.9: Internal dataset search success per EDP version

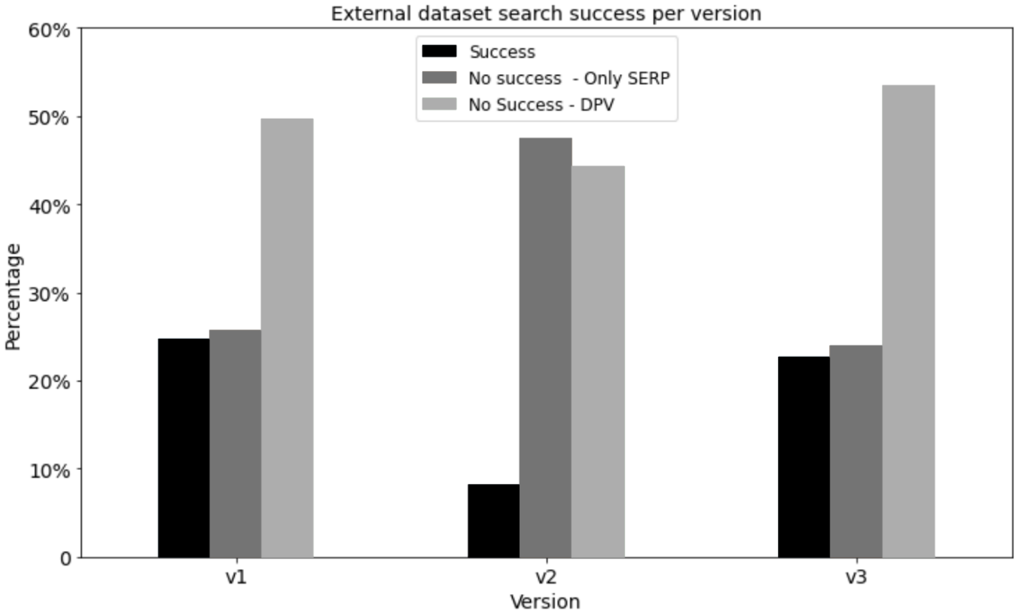


Figure 7.10: External dataset search success per EDP version

eu/data/en/dataset?tags=lidar). The tags facet was replaced by the keywords facet on EDPv2, but no redirection rules were made. By contrast, redirections to dataset pages and main categories were set up; this meant that the number of DPV failed search remained constant from EDPv1 to EDPv2. After July 2019, the number of "Only SERP" failures decreased to negligible levels; we believe this was the time it took web search engines to completely remove the old pages from their indexes. This explains why the success rate in EDPv3 is similar to EDPv1.

Figure 7.11 compares the success rate between internal and external searches. We observe that for EDPv1 and v3, internal searches are around 50% more successful than external ones. In EDPv2 the difference is more than 100%, however, this is due to the 404 pages issue already described.

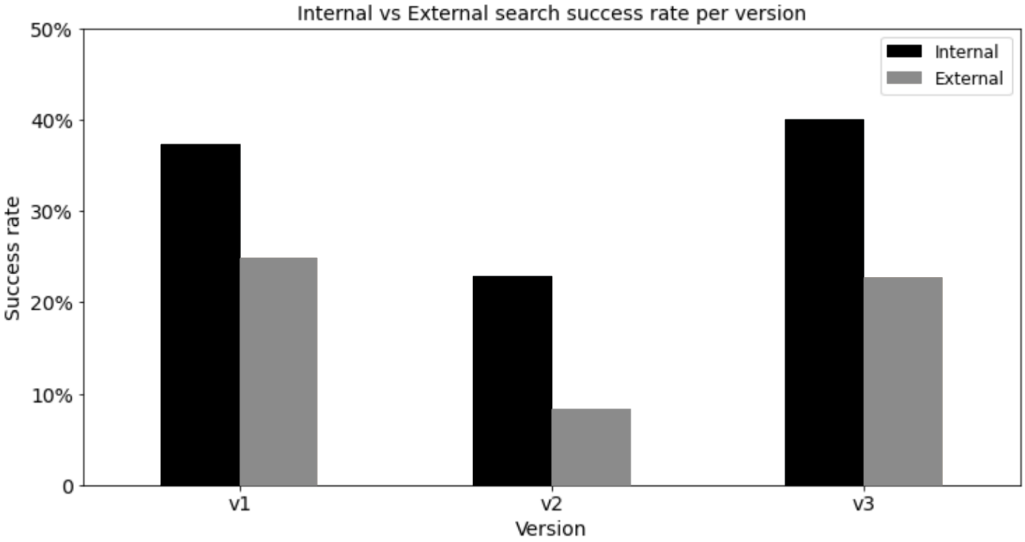


Figure 7.11: Success rate comparison between internal and external search, per EDP version

Figure 7.12 shows a breakdown of successful and failed searches for each of the search strategies including: use of only search box, use of only facets and, a mix of both strategies. Each bar chart corresponds to one version of the EDP: v1 top, v2 middle, v3 bottom. We notice that for EDPv1 and EDPv3 a mixed strategy led to more successful searchers than using just queries or facets alone. In EDPv2, the mixed strategy works better as well, but this is because of the very large number of only SERP failures for the other two strategies.

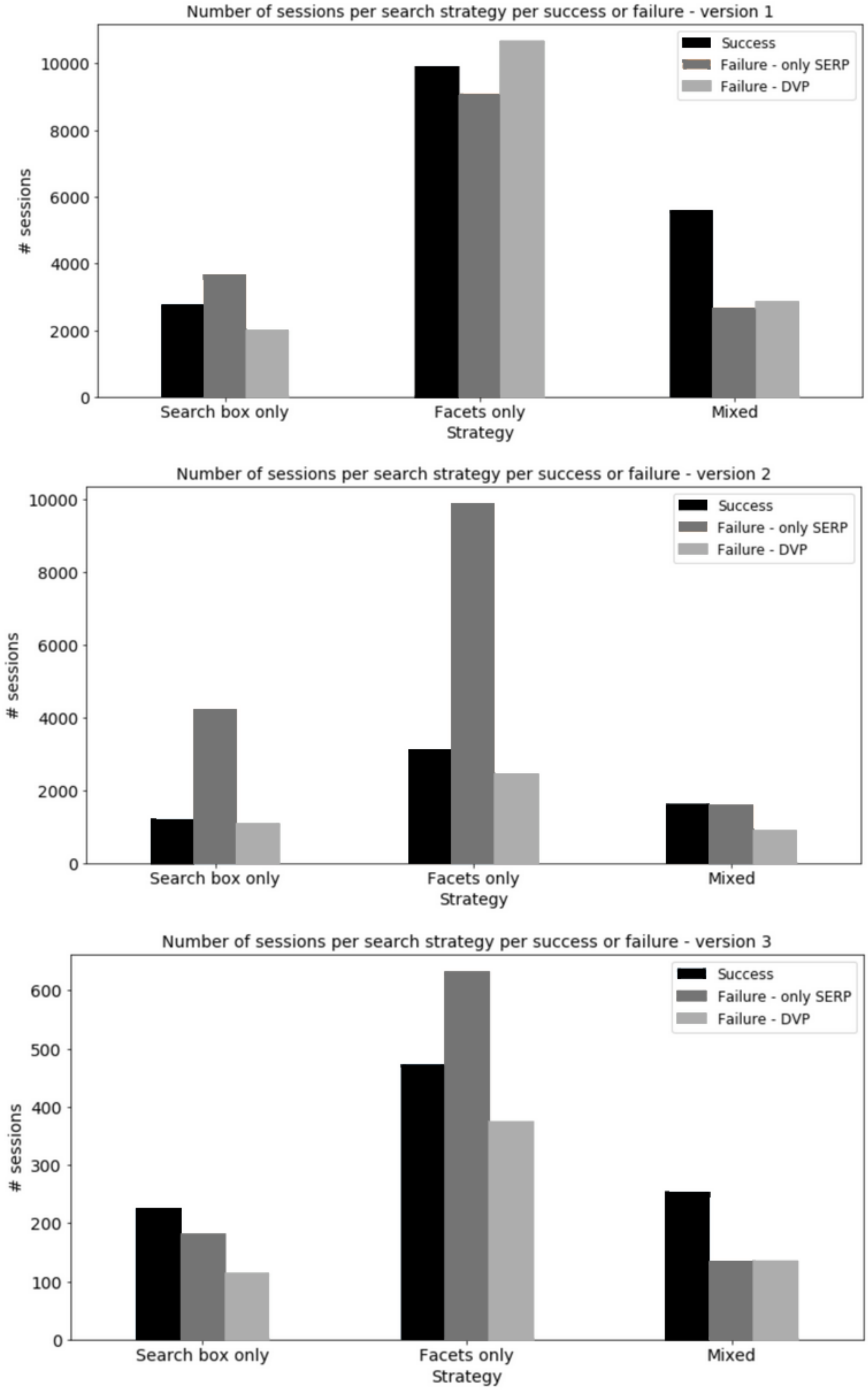


FIGURE 7.12: Success rate comparison between internal and external search, per EDP version

Finally, we employ the *alphabet* technique known from gene analysis to further analyse the session log and understand strategies leading users to the success in search (Fox et al., 2005; Hassan et al., 2010). Using introduced earlier session split into successful (sessions containing download or an outlink) and unsuccessful sessions (sessions without download or an outlink) we code each users action into a corresponding alphabet letters. Mappings used by us are summarised in Table 7.6. Data prepared in the following manner allowed as to analyse frequent patterns of both sets of data.

Alphabet	User Action
SQ	Issuing search query
SN	Search next page action
SR	Return to the search page
PD	Dataset page visit
PG	General dataset page visit
PH	Portal homepage visit
PF	Facetted search page visit
PV	Visualisation/distribution/resource page visit
PO	Other page visit
DO	Download or Outlink visit action

Table 7.6: Mapping of users actions to alphabet codes

Session Start. In our first piece of analysis we check what was the starting action of a successful and unsuccessful sessions. Table 7.7 shows the results split per portal version.

Action	Success			Nosuccess		
	v1	v2	v3	v1	v2	v3
SQ	2.01	6.36	8.49	3.69	7.72	12.66
PD	20.12	4.48	19.97	9.71	2.47	9.90
PG	1.82	7.16	4.83	1.24	4.47	3.43
PH	17.54	39.11	32.61	14.92	34.73	27.57
PF	51.93	23.81	12.4	67.38	43.90	31.43
PV	0.02	0.12	1.08	0.004	0.01	0.04
PO	5.92	16.98	19.47	3.05	6.7	14.98
DO	0.65	1.99	1.16	0	0	0

Table 7.7: First action in successful and unsuccessful sessions

Naturally, no unsuccessful sessions started with the download link or download link ('DO' action). As we can see a significantly larger proportion of successful sessions starts on a dataset page ('PD' action) which leads us to the hypothesis of such sessions starting based on the recommendation (the link to the portal was shared with a person). In order to verify this hypothesis we check the starting point for all sessions (Table 7.8) and sessions starting on dataset page in Table 7.9. In presented tables we see the split between different origins of a user who accessed the EDP portal.

Refferer	Success			Nosuccess		
	v1	v2	v3	v1	v2	v3
Search Engines	72.37	40.82	44.93	72.78	44.88	43.15
Direct Entry	18.0	39.19	43.93	18.28	36.68	47.30
Websites	9.16	18.29	9.23	8.01	15.58	6.65
Social Networks	0.46	1.69	1.83	0.93	2.86	2.9
Campaigns	0.01	0.01	0.08	0	0	0

Table 7.8: Percentage of sessions split by referrer type

Refferer	Success			Nosuccess		
	v1	v2	v3	v1	v2	v3
Search Engines	74.17	43.12	62.08	79.91	55.86	70.76
Direct Entry	15.42	39.68	27.5	12.76	29.49	19.13
Websites	10.10	16.14	9.17	7.27	14.06	8.66
Social Networks	0.30	0.79	1.25	0.06	0.39	1.44
Campaigns	0	0.26	0	0	0.2	0

Table 7.9: Percentage of sessions split by referrer type for which first action was landing on dataset page

We can see a slightly larger percentage of success sessions in comparison to unsuccessful sessions both in terms of 'direct entries' and 'social network' when session started on a dataset page. The reverse trend was observed for sessions referred from the 'search engine'. Presented differences were much less visible in overall session percentage for each referrer type.

We noticed that larger percentage of successful sessions started on the portals' homepage ('PH' action in Table 7.7). We believe that this is due to people deliberately coming to the portal with the intention of finding information and presumably knowing that the information they are looking for is there. Further investigation of referrers of sessions starting on portals' homepage uncovered that for both successful and unsuccessful sessions close to one-third of sessions were refereed from 'websites', 'direct entry' and ‘search engines’ for V1 and V2. For V3 two-thirds were ‘direct entry’, ‘search engines’ were 26-27% and ‘websites’ fall to roughly 6.5% for both successful and unsuccessful sessions.

Session Characteristics. In our next piece of analysis we wanted to understand the overall characteristics of successful sessions and what differentiated them from those which were unsuccessful.

Session type	v1	v2	v3
Success	15.96	16.79	17.40
Nosuccess	3.89	4.36	6.82

TABLE 7.10: Average number of actions within session

Table 7.10 shows the average session length for successful and unsuccessful sessions. The number of actions for successful queries is roughly three times larger than the number of actions in unsuccessful queries. In order to better understand the reasoning behind such large difference we decided to look at the distributions of of sessions lengths in both cases – which can be seen on Figure 7.13 and Figure 7.14.

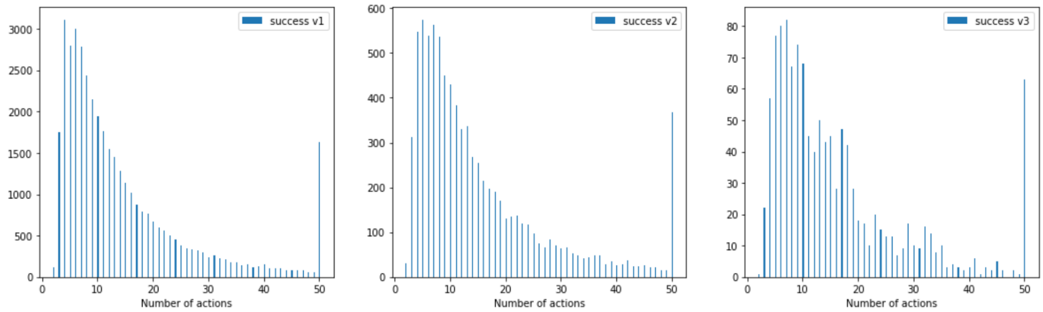


FIGURE 7.13: Success number of actions distribution, per EDP version

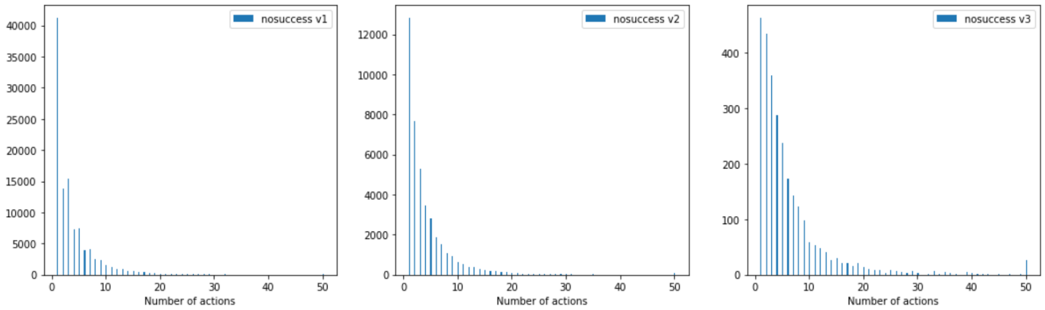


FIGURE 7.14: Nosuccess number of actions distribution, per EDP version

Presented figures show that indeed unsuccessful sessions are shorter with their peaks much more concentrated whereas successful sessions are overall longer. Extended length of many successful sessions could be a sign of different things: users having complex tasks requiring multiple datasets but equally well they could be the sigh of struggling in their search process.

In next step we conduct the analysis of average number of queries issued and facets actions in sessions. Table 7.11 confirms our previous findings as overall facets are much more popular in comparison to issuing a search query. This further strengthens our case from Chapter 6 stating that users use current search functionalities more as a filtering tool, not as a precise tool aimed at finding a specific dataset.

Session type	v1	v2	v3
SQ – Success	0.92	1.03	1.09
SQ – Nosuccess	0.38	0.43	0.57
PF – Success	4.24	4.18	3.66
PF – Nosuccess	2.01	1.76	2.23

Table 7.11: Average number of query issues (SQ) and facet searches (PF) within session

Finally, we look at the average time spent on results list after issuing a query or selecting a facet. We excluded from the analysis the actions for which it was not possible to determine the time spent due to the limitations explained earlier in Section 7.2.1. Table 7.12 shows average time spent on the result page for each type of search and for both of them combined. In most cases users spent more time after issuing a search query in comparison to faceted search. The exception was portals' v1 for which users after faceted search took longer time to look through the web page shown. We hypothesise that overall shorter time spend for V1 is connected with the change of the search functionality (from CKAN to custom made). Default CKAN setup presents 10 results per result page whereas the new functionality shows 15 results, causing the users to spend longer time, looking through results.

Search Type	Success			Nosuccess		
	v1	v2	v3	v1	v2	v3
SQ and PF	37.18	38.30	33.35	45.67	57.17	41.69
PF	37.63	36.7	30.93	46.88	55.65	41.48
SQ	35.16	44.72	41.5	40.1	63.41	42.59

Table 7.12: Average time in second spent on SERP page, per search type, per EDP version

Query Characteristics. Finally, we decided to investigate queries reformulations based on session being successful or unsuccessful. For each session we extracted SQ search query pairs (e.g. 'london' to 'london population' within one session). Next, we cleaned the queries of all special signs and we categorised each pair to one of the reformulations types. Result of this categorisation can be seen in Table 7.13. We categorised a query as a 'new query' if it was not other query type and the Lvenshtein distance is larger than 2 for one word queries and larger than 3 for queries longer than one word. If the Levenshtein distance was less or equal than above we categorised it as a 'spelling' reformulation. No significant differences in the distribution of reformulation types in successful and unsuccessful queries were visible.

Reformulation Type	Success			Nosuccess		
	v1	v2	v3	v1	v2	v3
New Query	65.96	68.52	63.31	62.02	63.15	63.39
Word/s Added	14.39	15.43	16.39	15.15	16.8	18.75
Word/s Removed	7.83	5.62	9.02	8.96	6.13	5.89
Spelling	10.7	8.74	9.92	12.66	11	9.82

Table 7.13: Percentage of reformulation types, per EDP version

7.4 Discussion

Below we present a discussion of our analysis in relating to our research questions. We discuss the search practices of portal users such as the keyword and faceted search usage, accessing datasets with use of different means such as general web search engines, and patterns in search through time and the changes on the versions of the European Data Portal.

7.4.1 Search practices of portal users

Our analysis shows the importance of filters when searching for data on the EDP platforms. 60% of dataset search users rely exclusively on facets, without using keywords and between 15 and 20% mix keywords and facets while searching.

In addition to previous findings on location-based queries, location-based facets were also prominent in our analysis. One reason may be that the portals indexed by the EDP have a broad geographic spread and include data resources from different administrative levels. Our current analysis supports specifically the importance of location-based filtering and the results suggest the recommendation of even more fine-grained filters of location. That would allow users to search for local datasets at different levels of granularity, for instance not just for countries but also for counties, cities, boroughs, etc. We know that the wrong granularity in terms of both location and time can easily result in the data not being usable for a task – as was indicated in our analysis of data requests in Chapter 5 and is further supported by Koesten et al. (2017). The user could perhaps download the relevant datasets, but end up not using it as the data is not aggregated at the right level and changing that, if possible at all, is costly, especially without appropriate technical skills and tools.

In contrast with the results of Chapter 6, we noticed a lesser prevalence of temporal information. This might be the case due to the fact that no filtering based on timeframes is possible on the EDP and the majority of portal-based search is done via facets. However, it would be interesting to explore the usefulness of time-based facets in respect to user needs in future qualitative work, as our other pieces of analysis suggests it is a core dimension people considered when looking for or selecting data to use.

Our findings also show that a large portion of filtering was done using categories of datasets. As this applies mostly to EDPv2, we attribute the popularity of this facet to changes in UI design. By showing a category panel on the landing page of the EDP dataset section, users are primed to explore the collection via this facet. It is possible to switch to a “search datasets by term” setting, however it is clearly not prioritised in the UI. This is a design choice that directly affects user interaction in dataset search on the portal and would therefore be an interesting starting point for user research, both in terms of validating the general search strategy through categories but also regarding potential subcategories to make filtering actions more precise. The results suggest that the success rate of a search seems to be higher if both keyword search and filtering strategies are applied, which has implications for how to support this interaction through the UI.

7.4.2 Search through web search engines

Our analysis confirmed previous findings on web search engines being the main tool for dataset search, next to human recommendations. The results show that the majority of users arrive at the dataset section through web search engines (more than 60% in EDPv1). They also document how important the link between the EDP and the search engine is, which led to a drastic decrease of traffic after EDPv2. Google’s general Web search engine refers 94% of the external visits to the dataset section, while Google dataset search only accounts for 4% percent of them.

The other factor to consider when users land on the dataset pages directly is the importance of the dataset preview page and the contextual information shown there. In previous chapters we shown the importance of some pieces of information which allow the user landing on the page to evaluate its usefulness for their task.

We found that users accessing the portal via web search engines tended to be less successful in their search activity. This might be because those who actively use the internal search functionality are a more informed user group who are likely more familiar with the EDP context and have matching expectations rather than landing on the site by chance. At the very least, this hints at different user groups, potentially with different intents and information needs.

7.4.3 Curated datasets with context

An interesting change of pattern could be noted in EDPv3, which contains COVID-19 datasets. We observed a decreased use of the dataset section with an increased use of the COVID section. 46% of external queries include terms such as ‘Covid’ or ‘Corona’ during that time frame. The EDP provided specific COVID-19 related pages, including a collection of key datasets and editorial pieces on the pandemic. The success of this section of the EDP sites means that the COVID-19 datasets are likely to become more popular and enable easier reuse by people, which in turn allows to improve discoverability of specific topics and in effect to improve discoverability of the underlying dataset. Such approach to a specific topic of interest brings to mind UK’s Office for National Statistics which also drifts away from a classical paradigm of search and which we hypothesised earlier is more trustworthy to the user due to such approach.

Other way to improve discoverability of the datasets in general web search engines is to improve the metadata available on the dataset pages in order to aid external search engines – which are currently used to access portals and get to the datasets. The intuition of what kind of information would be beneficial to be included in order to aid the indexing and in effect the discovery process was discussed already in terms of data request analysis and search query analysis, presented in previous chapters of this work, however, more detailed user studies of people interacting with search functionalities and with the data itself are needed.

7.4.4 Dataset Search as a Vertical

Finally, based on our findings from this chapter and Chapters 4, 5, 6 we would like to formalise our understanding of dataset search and what are it’s characteristics as a search vertical. Data, in order to be useful for an information need, must meet certain criteria. We believe that prior literature together with the findings of this work suggest that dataset search has unique characteristics which result in requirements that current dataset search functionalities do not fulfil. This suggests a large space for future research to improve current, and develop new, approaches for dataset search.

We believe that, in a retrieval scenario, datasets are complex to understand due to the ability to transform or analyse data, but also due to the different formats and structures that data can be stored in Koesten et al. (2017). Key findings include the importance of boundaries in dataset search for different information types, especially geospatial and temporal information, as well as the granularity of available data. One aspect is that this information can be expressed in different ways, sometimes very specific and sometimes very vague – therefore, descriptions would need to index ranges as well as exact values, and search interfaces would need to be flexible enough to enable fuzzier queries and potentially through more advanced filtering capabilities allowing to understand what data is available. As shown in this work some of the data portals already enable filtering datasets by geospatial coordinates inside a user-defined box. Another example is the U.S. national open data platform which includes the map preview of the geographical coverage of some of their datasets. UK's office for national statistics allows filtering time series data by custom periods of time, leveraging the fact that the underlying data is already in a time series format and it has a manageable size. However, further research is needed to extract the spatio-temporal and topical characteristics of datasets for their addition to metadata descriptions, in particular for the case of Big datasets that might contain different entities and granularities.

Many requests specified a data type and format, which underlines the complexity of data search in contrast to searching for documents. Majority of portals currently cover filtering through different dataset types, however it does not support known from the literature complex search activities. Many tasks with data involve comparing, contrasting or combining data with other data. This is reflected in the requests which often refer to other data, or specify that more than one dataset is needed for the task and that datasets need to be comparable (in terms of format, identifiers, etc.). Our findings suggest that often the successful retrieval of a dataset does not fulfil a users information need even if the retrieved dataset was fit for use, but can only be seen as a step towards it. Therefore functionalities supporting recommendations, or links to other datasets have the potential to be very valuable.

It is important to note that in terms of both geospatial and temporal information there could be more than one dataset equally relevant to a single information need. For example, when a user requests data from the "last 20 years", this could be returned as a number of equally relevant datasets, whether as one dataset covering the whole period, or as an individual dataset per year. Such requests could be fulfilled by automatically presenting an aggregation of the relevant datasets and, particularly if one of these datasets is not available, showing the timespan covered by the returned datasets.

7.5 Summary and Conclusions

In this chapter we analysed the session log of the European Data Portal which is a complex centralised resource for the data originating from all around Europe. Along its other functionalities it provides its users with the search functionality composed of query and facet based search. As users search for data using different methods: queries, facets or mix of both, the presented analysis expanded on research of queries and allowed to observe users through they search process and interaction with the system. We found that a majority of users search for datasets using only the facet filters, in particular, country and category. A possible further direction is to study the impact of the UX design, to rule out the users not using this facet is due to the portal design. Taking into account finding of this and previous chapters one of the main directions for UX research could be encouragement of querying based on available data with approaches for keywords suggestions, query reformulations or expanding facets with more tailored solutions (facets granularity wise, introduction of topic specific facets etc.). As a result of the findings regarding the fact that the metadata available for both internal search functionalities of data portals and external search functionalities – general web search engines is sparse, hence the extensive use of filtering options on a data portal along short search queries, in the next chapter we investigated the potential for the automatic metadata generation. Automatically generated metadata could aid each search related functionality: could add context for better indexing of the resources, could help user in understanding and assessment of the data for their task. This could be equally beneficial for internal search functionalities, allowing more complex filtering options. At last it could enable interlinking of related datasets which could lead to building similar or related datasets recommendation functionality.

Chapter 8

Semantic Labelling of Numerical Columns

In previous chapters we analysed how users of open data portals search for datasets and found that the following metadata fields could improve the most the searchability of datasets: the geospatial and temporal information about the dataset, and an overall understanding of what the dataset is about. In this chapter we look at the problem of dataset content understanding, specifically at the numerical columns present in the dataset. Numerical information, in form of columns, in reported to be the most popular column type (Mitlöhner et al., 2016). For geospatial and temporal metadata generation, some approaches have already being proposed (e.g. Neumaier and Polleres (2019)). This brings us to our third research question posed in this work [RQ3] *How to automate the generation of metadata? In this work we focus on one of the most under-investigated areas: semantic labelling of numerical columns.* Metadata generation overall is a costly process. Manual addition of metadata and creation of descriptions of data is a cumbersome process that not all data providers are effective in doing in their publishing process. The task is becoming even more troublesome when considering legacy data, published in the past with different metadata standards. This motivates the research and development of methods for automatically generating the metadata of a datasets one of which is to lift them to their semantic representation in the Linked Data world to add this additional, missing layer of information. While there are approaches on how to lift structured data to semantic web formats, most work to date focuses on textual fields rather than numerical data. This problem is referred to as a sub-domain of assigning semantic labels to columns task - assigning semantic labels to numerical columns. In this section we first look into previous work done in this research area and propose our own two level (row and column based) approach to add semantic meaning to numerical values in tables, called NUMER. We evaluate our approach using a benchmark (NumDB) generated for the purpose of this work. Some of the results presented in this chapter appeared in (Kacprzak et al., 2018a).

8.1 Semantic Labelling

Integrating tables in the Web of Data is useful for enriching structured knowledge bases (KB), improving search over data, or to enable question-answering systems to use larger corpora of information. Numerical values raise different challenges than entity matching from strings, and by numerical columns being the most popular column type (Mitlöhner et al., 2016), leading to the scientific problem of *semantic labelling of numerical columns*: given a table with numerical values where each row contains values that are assumed to refer to an entity, and a reference knowledge base, for each column, identify the property in the knowledge base that most likely describes it. Despite numerical columns being the most popular column type in open governmental datasets, existing approaches focus mostly on mapping textual data (Limaye et al., 2010; Mulwad et al., 2010; Syed et al., 2010; Venetis et al., 2011; Wang et al., 2012; Taheriyan et al., 2014). This is also reflected in the benchmarks available for the problem. For instance, one of the most commonly used benchmarks, T2D (Ritze et al., 2015; Ermilov and Ngomo, 2016), contains only 12 of 1748 tables with numerical columns disambiguated to DBpedia properties, partially because only very few columns from Open Data match do DBpedia.

In order to better understand challenges connected with this research problem we first conduct an analysis and replication of results of existing approaches to solve semantic labelling of numerical values presented by Pham et al. (2016) and Neumaier et al. (2016). Both derive from the Ramnandan et al. (2015) approach in which statistical tests are used to compare the values in the table with values of properties in the reference knowledge base. Properties whose values are statistically similar to column values are candidate semantic labels for the column. Neumaier et al. (2016) developed an Multi-level Semantic Labelling approach (henceforth MSL) in which they compare the distance between columns (seen as bags of numerical values) and nodes (where each node is a bag of numerical values) in a hierarchical background knowledge graph built upon DBpedia numerical data[1]. Pham et al. (2016) approach, called the Domain-independent Semantic Labeller (hereafter DSL), consider a list of similarity metrics for both textual and numerical values. For numerical values, they use a customised Jaccard similarity that works with ranges rather than sets of values. Following Ramnandan et al. (2015), they also use the Kolmogorov-Smirnov test to check distribution similarity between two sets of numerical values. More detailed description of the above approaches could be found in the Chapter 2.

With the insights gained from the analysis of replicability of those approaches we propose our own approach. Inspired by Venetis et al. (2011), who reported increased accuracy if a main (subject) column was identified, we introduce NUMER – an approach which uses the context of numerical columns to assign semantic labels to them. We leverage existing approaches for identifying the subject column of a table by matching textual

[1] DBpedia is described in Chapter 3

columns to entities in a knowledge base. We propose using the subject column of the table to pick potential labels which are then matched against the numerical column. Each cell in a subject column is disambiguated to a concept (entity) in the target KB. The numerical values associated with the subject columns are subsequently examined following a composite approach: (i.) a *column level analysis*, which looks at their distribution in a column; (ii.) a *row level analysis*, which compares each of them to the values associated to the disambiguated entities in the target KB. As a result, we generate a ranked list of properties for each numerical column, based on numerical properties of subject column numerical properties. By selecting a table-specific set of possible semantic labels based on the subject column we were able to narrow down the possible values in a KB to those that are likely related to the context of the table. The preselected semantic labels are than ranked according to their fit to data in a column. This reduces memory requirements (as only data related to those values needs to be processed), and may make it more suitable in cases where KB are large, diverse, or rapidly changing such as DBpedia or Wikidata.

To evaluate our approach, we created the benchmark NumDB (Piscopo and Kacprzak, 2018). This consists of tables with numerical values constructed from types and numerical properties from DBpedia. NumDB introduces two dimensions of benchmarking: first, it includes deviations in the values drawn from DBpedia to test the sensitivity of approaches to values that are not exactly the same as the ones in the target KB. Second, it considers versions of the same table with different number of entities, to test the accuracy of approaches when facing smaller versus larger tables. Our evaluation suggests that our approach, which includes both row and column levels of analysis, outperforms the state-of-the-art in terms of sensitivity to value deviation and effectiveness on smaller tables. NUMER shows itself more adaptable for use in a real world scenario in terms of time and memory consumption, as it does not require to generate the background knowledge necessary for approach proposed by Neumaier et al. (2016). NUMER shows itself more scalable for use in a real world scenario in terms of time and memory consumption, as it does not require to generate the background knowledge necessary for approach proposed by Neumaier et al. (2016), however, they pose similar limitation in terms of the usage of the same background knowledge.

8.2 Use case - Running example

In this section we introduce the general *Semantic Labelling* problem with the help of a use case. Semantic labelling is defined as follows: given a set of bags of values V (the columns) and a target domain ontology D, return a ranked list of mappings of bags of values to terms (properties or classes) in D. Note that in the general sense, values can be numerical or not, however, the focus of our work is on numerical values. Knowledge bases that use D are also used as input, to compare the values in them to the values in the columns. Figure 8.1 shows the general workflow of assigning semantic labels to bags of values.

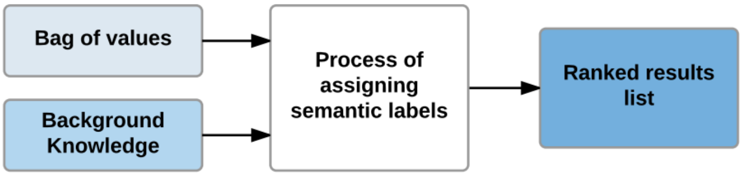


FIGURE 8.1: General workflow for assigning semantic labels.

Table 8.1 shows a dataset example containing information about the employees of a company. It contains 5 text column headers, `Name`, `Birth Place`, `Birth Date`, `Weight`, and `Height`. Assuming DBpedia's ontology and knowledge base as target. A candidate mapping from column header to ontology is shown in Table 8.2.

Name	Birth Place	Birth Date	Weight	Height
Joe Doe	London	01-01-1990	91kg	1.92
John Smith	Paris	03-12-1946	56.5	1.65
Martin Olaf	Warsaw	22-03-2000	63	1.80
Andrew Young	Madrid	09-12-1987	77kg	1.75
Matt Willis	Berlin	13-05-1973	98.2kg	1.89

TABLE 8.1: Example data source with different data types in each column.

Attribute	Ontology term
Name	*dbo:name*
Birth Place	*dbo:birthPlace*
Birth Date	*dbo:birthDate*
Weight	*dbo:weight*
Height	*dbo:height*

TABLE 8.2: Mapping of Table 1 to the DBpedia ontology. Ontology term originates from http://dbpedia.org/ontology/

8.3 Datasets

In this section, we describe the datasets used for reproducing the approaches. For our experiments, we use the same datasets as the ones used by MSL and DSL.

8.3.1 DSL dataset

DSL is trained and evaluated on four different datasets, provided by the authors, and covering different domains: **City** (Ramnandan et al., 2015), **Weather** (Ambite et al., 2009), **Museum** (Taheriyan et al., 2014) and **Soccer** data. Each of those originates in different sources. We did not alter the original datasets.

8.3.2 MSL dataset

The dataset used by MSL as Background Knowledge Graph was constructed from DBpedia by selecting 50 of the the most frequently used numerical DBpedia properties, excluding internal properties and not directly in the root path of the *http://dbpedia.org/ontology/* prefix. From this, 20% of the data per property was randomly selected as test data for the evaluation. In order to generate meaningful bags of numerical values, the authors propose to build the knowledge graph without applying the constraints that were used while building the background knowledge graph (i.e., child nodes can have overlapping subjects). Based on this test, the authors select random nodes from this knowledge graph, to be used as bags of values. The background knowledge graph was generated with the remaining 80% of the data. In the original paper, the dataset was created based on the 46 properties with numerical domains mentioned by them.

8.4 Evaluation metrics

The approaches we are focusing on in this chapter evaluate the accuracy with use of two different metrics.

Metric	Definition
MRR - mean reciprocal rank	Statistical measure for evaluation of approaches that are generating results list, ranked by probability of correctness score.
% of correct labels among the top-k ranked results	The accuracy score is calculated by checking top-k results list for the appearance of a correct semantic label. If the correct semantic label is among top-k labels in a ranked result list the accuracy is assigned to 1 in other case the accuracy score is 0.

Table 8.3: Evaluation metrics introduced in DSL and MSL

DSL is evaluated with use of Mean Reciprocal Rank (MRR) as an evaluation metric. However, it is important to note that the output of the approach is not simply a ranked result list, but a list of lists where multiple labels can be assigned the same confidence probability score. The authors argue that some bags of values could be assigned with multiple correct semantic labels (e.g from our running example the correct semantic labels for the *birthPlace* are both`dbo:city` and `dbo:capital`). The fact that multiple labels can have the same confidence probability could lead to better MRR scores since there is no penalty for giving the same probability of the correct label to many wrong labels. Also, the definition of MRR given by Craswell (2009) states that is an appropriate measure for *known item search* which is not the case in the semantic labelling problem. This description implies that the MRR metric might not be the most suitable.

8.5 Replicability

In this section we report on the replication of both DSL (Domain-Independent Semantic Labeler) and MSL (Multi-level Semantic Labelling of Numerical Values) and compare them against the results reported by the authors in their studies. We try to recreate, as much as possible, the original environment to generate our results. The following was performed in order to understand the state of the art approaches within the research area of assigning semantic labels to numerical column and to identify the core problems in this research area as discussed in Section 8.6 and further understand the scenarios for such approaches (see Section 8.7).

DSL - The metric used by DSL to evaluate their approach is the mean reciprocal rank (MRR). The methodology for their evaluation is the following: Given a dataset A that consists of n sources $\{s_1, s_2, ...s_n\}$; choose m labelled sources that $m < n$. For each source s_i in dataset perform semantic labelling with use of m labelled sources from s_{i+1} to s_{m+i+1}. The authors analysed two classifiers: Logistic Regression and Random Forests. Both of the classifiers achieve similar labelling accuracy, although Logistic Regression is reported to require less training and labelling time. In our experiment we evaluated the final set up of the system, using Logistic Regression.

The code for the DSL was published on GitHub[2]. We repeated the experiment reported in the paper for all four datasets: `city`, `weather`, `museum` and `soccer`. The experiment was run over each datasets with 1 to 5 labelled data sources. The exception is the Weather dataset was only tested with up to 3 labelled data sources since it has only 4 sources.

Table 8.4 presents the results for `city`, `museum` and `soccer` as a classifier training dataset along with our results. The `weather` dataset was not used as a classifier training dataset because it does not provide a sufficient number of feature vectors. The experiments were run with 1 to 5 labelled data sources for the datasets `city`, `museum`, and `soccer` and for 1 to 3 labelled data sources for the `weather` dataset (which has a total of 4 sources). The left column of each table shows the results reported by the authors of the study while the right column displays our results for the same set up.

[2] https://github.com/minhptx/iswc-2016-semantic-labeling

The second part of the original work is to evaluate the approach with the T2D Gold Standard. We could not perform this experiment due to unavailability of the preprocessed dataset and information about the set-up that was used in this experiment. We tried to perform a similar experiment by using the complete DBpediaAsTable[3] dataset, as well as its subset[4]. Unfortunately, all our attempts failed with exceptions during preprocessing step.

DS	TS	Number of labelled sources									
		1		2		3		4		5	
		orig.	repli.	orig.	repli.	orig.	repli.	orig.	repli.	orig.	repli.
S	S	0.913	0.915	0.932	0.925	0.932	0.938	0.941	0.943	0.945	0.941
	C	0.912	0.901	0.927	0.903	0.928	0.919	0.941	0.923	0.944	0.928
	M	0.914	0.916	0.928	0.932	0.930	0.943	0.939	0.955	0.944	0.959
M	S	0.471	0.463	0.665	0.644	0.719	0.707	0.755	0.757	0.790	0.767
	C	0.463	0.459	0.652	0.635	0.709	0.704	0.752	0.757	0.792	0.768
	M	0.472	0.456	0.659	0.637	0.706	0.704	0.713	0.744	0.730	0.746
C	S	0.913	0.915	0.932	0.925	0.932	0.938	0.941	0.943	0.945	0.941
	C	0.912	0.901	0.927	0.903	0.928	0.919	0.941	0.923	0.944	0.928
	M	0.914	0.916	0.928	0.932	0.930	0.943	0.939	0.955	0.944	0.959
W	S	0.899	0.845	0.951	0.909	0.977	0.909	-	-	-	-
	C	0.899	0.880	0.951	0.928	0.977	0.966	-	-	-	-
	M	0.902	0.902	0.955	0.955	0.977	0.977	-	-	-	-

Table 8.4: Comparison of DSL MRR scores for each dataset. (S: soccer, M: museum, C: city, W: weather, DS: dataset, TS: training set, orig.: original study results, repli.: replication study results)

As shown in Table 8.4, results of our replication study are almost matching the results obtained by Pham et al. (2016). However, we do not know the reason that caused the small differences between results reported by Pham et al. (2016) and our results. Initially our results differ significantly for the one reported in a paper. After contacting one of the authors we obtained certain parameters for training the model. The model performed better when it was trained on a combination of 1, 5 and 9 labelled domain data (of each training dataset) at the same time. With this training parameters, we were able to replicate the approach.

MSL - Neumaier et al. (2016) performed an evaluation in two different settings: in a controlled environment by splitting the background knowledge into training and testing, and over Open Data files. Since the latter was used to gain first insights and are evaluated mainly in an exploratory way, we replicate only the first experimental setting.

[3] http://wiki.dbpedia.org/DBpediaAsTables

[4] http://webdatacommons.org/webtables/goldstandard.html

To construct the background knowledge, DBpedia 3.9 dump was used, as described in Section 8.3. Two different aggregation methods are taken into account: majority vote and aggregated distance. Neumaier et al. (2016) build two knowledge graphs: an evaluation graph and a testing graph. Details on the generation of test nodes are shown in the Section 8.3. For this experiment the authors selected a maximum of 20% of leaf nodes per property, resulting in 33657 test nodes.

They propose to generate those results on five different levels:

- property level – aggregation of the neighbours by their properties e.g., "height" or "capacity" (ignoring any type or p-o context)
- type level – aggregation of the neighbours by their exact type
- root type level – aggregation of the neighbours by their root type (e.g. "Person" or "Location")
- all types level – aggregation of the neighbours by each of their types
- p-o level – aggregation of the neighbours by each of their p-o nodes

The source code of the approach was published on GitHub on request. A prototype source code used for the generation of results for evaluation was shared with us later and was said to be published openly in the future. The experiments performed in the paper were conducted on a machine with 30GB of RAM memory (with no further specifications made). For our test we have used a machine with 96GB of RAM and the process of performing MSL calculations was killed by the server due to *out of memory* issue. We repeated it on a machine with 256GB of RAM memory, the difference results of the evaluation study reported by the authors and our replication are shown in Table 8.5.

Top-k		Prop		Type		All-types		Root-type		P-o level	
neigh.	agg.	maj	avg	maj	avg	maj	avg	maj	avg	maj	avg
25	1	58.0	40.1	56.7	44.1	56.7	44.1	60.5	64.8	23.5	21.4
	5	91.7	81.7	90.2	80.6	90.2	80.6	93.6	93.6	62.9	45.8
	10	98.5	98.5	93.6	93.6	93.6	93.6	93.6	93.6	71.9	64.6
50	1	50.3	31.6	58.3	42.0	58.3	42.0	60.3	65.6	23.5	21.4
	5	80.8	93.3	91.5	78.8	91.5	78.8	95.9	95.9	62.9	45.8
	10	99.1	99.1	95.9	95.8	95.9	95.8	95.9	95.9	71.9	64.6

TABLE 8.5: Accuracy achieved in our replication study (maj: majority vote, avg: average distance, agg.: aggregated results)

We could not achieve any results for MSL with the reported machine specifications. However the results obtained with a machine with more memory are comparable to those originally reported. Results with 1 and 5 aggregated results are almost always consistently worse than in the original, but almost equal at $k = 10$, except for the p-o level, that has a difference of around 10%.

8.6 Discussion of existing systems

In the first part of this work we present a discussion and performed a replication of two systems targeting the problem of assigning semantic labels to data sources. The two approaches we are analysing are different, but both of them present solutions on the under-explored area of labelling of bags of numerical values.

Numerical values: While analysing CSV files we identified potential problems that approaches to label numerical values could look for improving upon. We could identify four general categories:

- **Precision and deviation.** Numerical values are subject to a variance which can be caused by rounding of values or difference in set-up of analysis (e.g. population of London might differ depending of the source of information). Therefore, comparison of this information with a specific knowledge base is subject to greater noise than textual information from the get-go.
- **Format as an indication of type.** Numerical values can have different precisions, which could indicate specific subsets of labels to be considered first in the process of assigning semantic labels to a bag of values. Similar conclusions could be drawn from the presence of negative numbers, which eliminate a vast number of properties that can not contain negative numbers. At last, we noticed that there might be cases in which analysing ranges of values might not be the best suited technique (e.g. when the same value is repeated through the whole column).
- **Formatting differences.** Numerical columns can contain a mixture of numerical and textual values on both cell level (e.g numbers with string indicators of type km or m) and column level (e.g. a column containing both numerical and textual). Furthermore, in some cases the same data was presented in different formats (e.g. temporal data), or even as ranges of values.
- **Table structure.** Some tables, and this is the case specifically for numerical columns, have an additional last row presenting the sum of all the values in a column. Others include additional headers in the table, to add structure for the reader, but also making the task of automatically understanding the table more difficult.

Benchmarks: We identified the need for larger and more specialised (e.g. type specific) benchmarks. T2D, while widely used, has issues: Not every column is aligned to DBpedia properties. Even if they were aligned, we could identify multiple properties with a numerical range, that we could not access on either the most recent DBpedia dump nor the DBpedia SPARQL endpoint. It seems they are outdated and need updating. That emphasises the need for a gold standard, that can be updated and accessed.

Lessons Learned: Through the presented experiments we summarise our thoughts in the following main points: (I) Different types of values should be treated differently; (II) precision or negative numbers could be an indication for possible semantic labels; (III) Connecting information from textual and numerical analysis - header and the rest of the file can have positive impact on informing the analysis of numerical values; (IV) There are structural challenges as: presence of sub-headers in a table, last row presenting the sum of values of a column; (V) Correlations between numerical columns in the same table.

Conclusion For MSL, we were able to reproduce most of the results that were reported in the original studies, with some variations resulting from the selection of random bags of values at each run of the evaluation, and with the caveat that a machine with 4 times more memory than reported was required. For DSL we reproduced the evaluation of the approach with four different datasets with minor differences between our and the reported results.

In the next section we describe a new approach targeting the problem of assigning semantic labels to numerical columns building on previous work showcased above.

8.7 Approach

Our approach comprises four stages, described below. We assume the availability of a tool that enables the match of textual cells to entities in the target KB, and of a tool that allows the identification of a subject column.

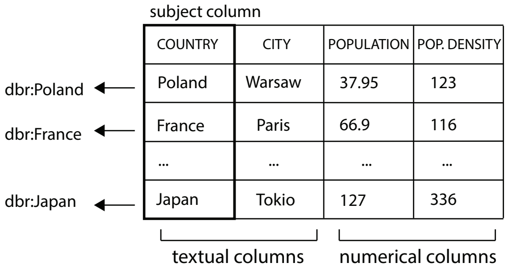
Definitions: We define a **table** T as a collection of related data on a specific topic. A table consists of m rows and n columns represented by a $m \times n$ matrix. Each row in a table has the same structure and can be seen as a single record of related data. Columns in a table are of specific type depending on their content; possible column types are: **Numerical and Textual Columns.** A numerical column is a column where more than 50% of cells contain at least one digit. We chose this definition to not rule out cells that contain units of measure (*e.g.*, 2 Kilometers) or dates. A column that is not numerical is considered textual. One column per table is a **Subject Column**. That is, a textual column which represents the main subject of the table and connects the other columns semantically through binary relations (Venetis et al., 2011; Wang et al., 2012; Ermilov and Ngomo, 2016). Those connections are represented through properties from a KB. The process of determination of subject columns is detailed in Section 2.7.

Scenarios. In this work we aim to disambiguate columns with numerical vales with use of the information from subject column in the table. We distinguish a list of scenarios that can be encountered when solving this problem. We look at the scenarios when the subject column is known and if the selection of subject column was not successful with existing approaches:

1. When the information on which textual column in the table is a subject column is known:
   (a) *Full match of numerical properties values* in the KB with numerical values in a table. This scenario is the most trivial and could be solved with basic matching techniques.
   (b) *Numerical values in the numerical column could deviate from the values in the KB.* Some values could be more distinct than others, some could be missing in the KB entirely. Our approach includes mechanisms to make the influence of the following problems negligible: the *Column Level Analysis* (presented in Section 8.7) compares the distribution of values in the numerical column against that of values from numerical properties, which helps when values are missing. We also take into account numerical properties connected to entities of each type that were recognised in the subject column, which helps minimise the influence of partially correct disambiguation of the cells in the subject column.
   (c) *Subject column cells can be disambiguated to a range of types* which could indicate a lack of consistency within the table or incorrect disambiguation of the cell value. Analysing tables per row allows us to compensate for the latter and detect the property the values of which are the closest to the values in the KB (Section 8.7 *Row Level Analysis*).
   (d) *Numerical columns can be properties of types different from the types of entities found in the subject column.* They can be properties of other textual columns (that is not the subject column) or not connected to any other column (their meaning could be identified from the context but not from any textual column present in the dataset). This scenario is out of the scope of this work and for such cases alternative approaches could be used (e.g. Neumaier et al. (2016)) which do not rely on additional information provided with the numerical column.
   (e) A property describing a numerical column could not be represented in the KB, in which case the approach will fail as the correct result of the disambiguation process is impossible to achieve. This would also be the case in case of values in which case unit conversion would be necessary in order to match the values.
2. Approaches for the detection of a subject column could fail, in which case the scenario is similar to the one described in 1.(d).

From these scenarios we can see that the knowledge of the subject column can be used as a basis for improving the accuracy and efficiency of labelling of numerical columns.

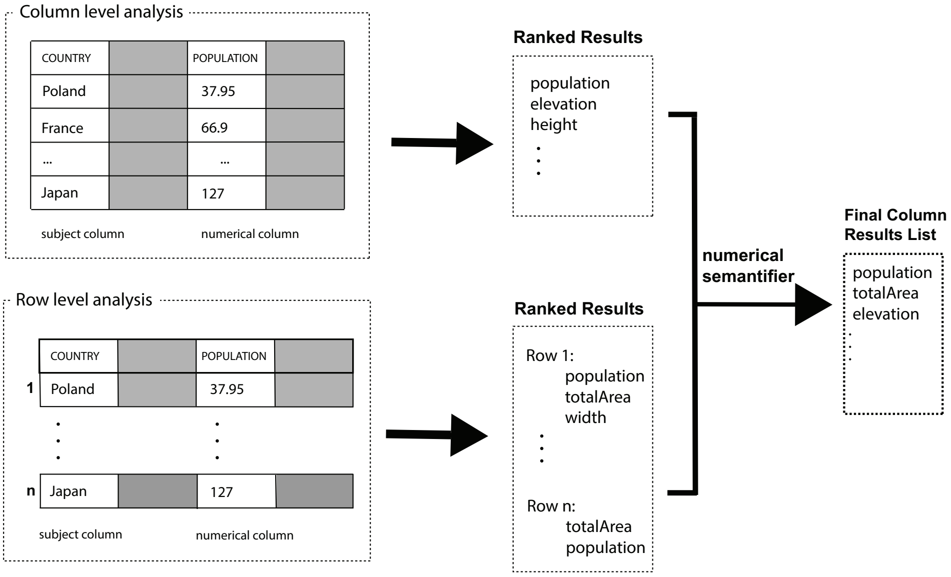
Preprocessing. We preprocess the input table as follows: (1) Partition columns into *numerical* and *textual* columns. Following the definition in Section 8.7, we define a numerical column as a column that has $\geq 50\%$ numerical values. (2) From the subset of textual columns, select one subject column. This may be done following any the approaches described in Section 2.7, we select most-left textual column in a table. (3) Match each cell in the subject column to an entity in the target KB. (4) In numerical columns, we strip out cells containing non-numerical characters (*e.g.*, "2Km"), leaving only numerical values (*e.g.*, "2").



COUNTRY	CITY	POPULATION	POP. DENSITY
Poland	Warsaw	37.95	123
France	Paris	66.9	116
...	...	...	...
Japan	Tokio	127	336

FIGURE 8.2: Information resulting from performing preprocessing steps a table.

Figure 8.2 shows the output of preprocessing our running example. Columns *Country* and *City* were classified as textual, *Country* was identified as the subject column. All values in the subject column were disambiguated to a DBpedia entity. Columns *Population* and *Population Density* were classified as numerical.



COUNTRY		POPULATION	
Poland		37.95	
France		66.9	
...		...	
Japan		127	

	COUNTRY		POPULATION	
1	Poland		37.95	
n	Japan		127	

FIGURE 8.3: An overview of the analysis stages in the semantification process.

Column Level Analysis. Similarly to Neumaier et al. (2016) and Pham et al. (2016), we compare the distribution of values in numerical columns with bags of values from the target KB. However, instead of comparing to all bags of values in the target KB, we consider only the properties that have a semantic relation with the types of the entities identified by the subject column, hence reducing both number of comparisons and memory requirements. From the entities identified in the subject column in the preprocessing stage, we query the target KB for the list of all types associated to them. A sample list of types could be: [*dbo:CapitalCity, dbo:Country, dbo:PopulatedPlace*]. Next, for each type, we generate a list of all its instances in the target KB. Then, for each entity, we select properties of *rdf:type owl:DatatypeProperty* associated to it. For example, Poland, *dbo:PopulatedPlace* has the properties *population, area, existsFrom*, and *dbo:Country* has *population, area, populationDensity.* For each property, we select its associated values and compare them to those in the numerical columns using the two-sample Kolmogorov–Smirnov test (Ramnandan et al., 2015), as shown in Equation 8.1.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{8.1}$$

The output of the comparison is a list of properties for each numerical column, ordered by the probability given by the KS test. The sample output for a *population* column could be as follows: `[(populationTotal, 0.98), (populationDensity, 0.44), (elevation, 0.14), (area, 0.08)]`.

Row Level Analysis. Comparing value distributions does not necessarily result in meaningful matches. The size of a numerical column in a dataset and popularity of a specific property in the KB influences the accuracy of the results when comparing the distributions. To improve accuracy, we also analyse numerical values based on the context provided by the row they are in.

In this level of analysis, we perform the following steps for each row in the table: (1) From the entity disambiguated in the subject column at preprocessing stage, we query all properties of `rdf:type owl:DatatypeProperty` associated to it, together with their values. (2) Next, we compute the relative difference (Equation 8.2) between the value of the cell in the numerical column and the value of each of the properties collected in step 1. The intuition is that the property with the smallest relative difference is the right match for the value.

$$\text{Rel diff}(val_1, val_2) = \left| \frac{val_1 - val_2}{max(|val_1|, |val_2|)} \right| \tag{8.2}$$

In our running example (Figure 8.3), for the cell *Poland* 37.95 in the column *Population* we compute the relative difference with each of the values of the 33 numerical properties and the value in numerical column (here `37.95`). Producing a ranking of candidate labels per row ordered by decreasing relative difference.

(3) Finally, we generate a final ranking of candidate labels from the rankings generated in step 2. This is done by selecting all properties that appear in any of the lists. For each unique property we also assign its best relative difference value, intuitively giving more importance to a property that was able to exactly match one row. In case of a tie, we break it by computing the average position between all intermediate lists.

Numerical Semantifier In the last step, we create the final ranking by combining the outputs from the row level analysis (relative distance and average position) and the column level analysis (probability). Concerning the column level results list, a higher score represents a higher rank. As regards the row level analysis, a lower score means a higher rank. In order to merge the two analyses we order all outputs based on the closeness of the predictions to the highest ranking, independently of whether these represent outputs of row or column analysis (distance and probability). In case of a tie we prioritize the row level analysis labels over the column level analysis, as we identified in our evaluation row level analysis performs better than column level analysis on average.

Our final results list consists of predictions of semantic labels for a numerical column with their confidence score. We call the overall approach *NUMerical SemantifiER* (NUMER).

8.8 Evaluation

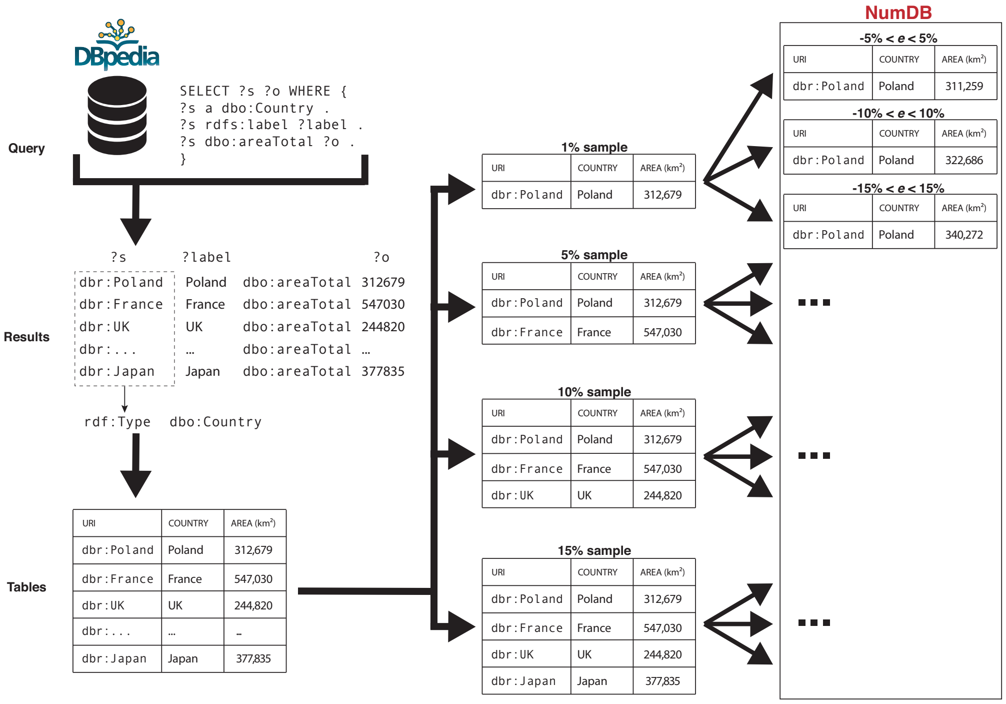
We evaluated our approach under two aspects: resource consumption (**Ev1.**) and accuracy in matching the correct property (**Ev2.**). For both, we compared against the approach developed by Neumaier et al. (2016) as a baseline – which we refer to as (MSL). We generated a new benchmark for the purpose of our evaluation, which is described in the following section. All code used and results of the evaluation are available in a Github repository[5].

8.8.1 Benchmark

To evaluate our approach we needed a set of tables containing at least one textual subject column (*i.e.*, the column to which the values in the other columns refer) and one numerical column. We generated a benchmark by extracting tables from DBpedia. Each table has three columns: the first is the subject textual column; the second contains the DBpedia URIs corresponding to the entities listed in the first column (ignored by our algorithm); and the third numerical column.

We extracted the tables according to the following process: first, we took a set of properties that could be mapped to numerical columns, namely those identified in (Neumaier

[5] https://github.com/chabrowa/semantification



```
SELECT ?s ?o WHERE {
?s a dbo:Country .
?s rdfs:label ?label .
?s dbo:areaTotal ?o .
}
```

```
?s            ?label                    ?o
dbr:Poland    Poland   dbo:areaTotal  312679
dbr:France    France   dbo:areaTotal  547030
dbr:UK        UK       dbo:areaTotal  244820
dbr:...       ...      dbo:areaTotal  ...
dbr:Japan     Japan    dbo:areaTotal  377835
```

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	312,679
dbr:France	France	547,030
dbr:UK	UK	244,820
dbr:...	...	...
dbr:Japan	Japan	377,835

1% sample

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	312,679

5% sample

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	312,679
dbr:France	France	547,030

10% sample

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	312,679
dbr:France	France	547,030
dbr:UK	UK	244,820

15% sample

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	312,679
dbr:France	France	547,030
dbr:UK	UK	244,820
dbr:Japan	Japan	377,835

NumDB

-5% < *e* < 5%

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	311,259

-10% < *e* < 10%

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	322,686

-15% < *e* < 15%

URI	COUNTRY	AREA (km²)
dbr:Poland	Poland	340,272

FIGURE 8.4: Tables in the benchmark were created by extracting data from DBpedia and transposing it into tables. The figure represents the whole pipeline.

et al., 2016). These included the 46 most popular numerical DBpedia properties[6]. Second, we extracted the type of information (*i.e.*, the classes) of all subject entities for each property p_i in the property set and the number of entities of each type. For each property, we left out all classes whose entities comprised less than 0.1% of all the property subjects, in order to exclude possible erroneous triples, and selected a number of random classes above this threshold, 10 when available, less otherwise. Subsequently, for each subset of classes we took all triples $p_i(i, o)$ for the corresponding property, where $type(i, C_j)$ for C_j in the class subset. Labels were collected for all entities and properties. All these steps were performed by querying the live DBpedia endpoint[7]. We transposed all the resulting triples into tables (see Figure 8.4). This produced a total of 389 tables, which we used to generate our benchmark. We created tables with different levels of sampling and introduced varying degrees of errors, with respect to the data subsequently used to disambiguate them, to also allow conclusions around the robustness of our approach in respect to inaccuracies in the data. We used four different sample sizes (*Very Small*:1% of all entities; *Small*:5%; *Medium*:10%; *Large*:15%). A statistical description of each sample can be seen in Table 8.6. For each sample size we generated three additional tables, to which a degree of error e of $-5\% < e < 5\%$, $-10\% < e < 10\%$, or $-15\% < e < 15\%$ to each value v was introduced. The total number of tables created was 3952, 247 for each combination of sample size and error degree. The dataset is available at (Piscopo and Kacprzak, 2018).

[6] 50 most popular properties excluding those linking to DBpedia internal ids.
[7] https://dbpedia.org/sparql

Set	Statistics					MSL			NUMER		
	#rows	Median	Avg	S.dev.	Δ	Total	Avg	Δ	Total	Avg	Δ
V.S	11,456	79.5	137.69	127.76	-	555	2.256	-	2168	8.815	-
S	56,604	390	682.75	633.08	3.94	630	2.561	0.45	3147	12.793	0.14
M	113,054	808.5	1366.58	1265.55	1.00	816	3.317	0.14	3564	14.486	0.30
L	169,484	1255.5	2069.16	1899.65	0.50	936	3.915	0.11	3973	16.152	0.18

Table 8.6: Set statistics and processing time for NUMER and MSL (V.S - very small, S - small; M - medium; L - large set; Avg - average; S.dev - standard deviation).

8.8.2 Evaluation Results

For both resource consumption and accuracy we compared the performance of NUMER and MSL for each table size and degree of error in the NumDB benchmark. The resource consumption evaluation (**Ev1.**) included processing time and memory consumption. The accuracy evaluation (**Ev2.**) examined the percentage of correctly disambiguated columns, examining both the top 1 and the top 3 semantic labels on the ranked results list. Moreover, we generated scores for each level of analysis (*i.e.*, row and column) separately and compared it against the overall score, in order to gain a better overview of the influence of different levels of analysis on the results. Finally, we applied ANOVA to test for statistical significance between table sizes and between various degrees of error. We believe that this range of experiments was able to provide a better picture of the performance of our approach and to detect directions for further research.

Experiment Setup We deployed a SPARQL endpoint for DBpedia v.2016-4 using Virtuoso and AWS services[8]. We run the evaluation on a virtual machine with 6 cores and 66 GB of memory running Ubuntu Linux. MSL was evaluated using code provided by authors on an associated Github account[9].

Resource consumption We tested the overall performance of NUMER and MSL by measuring the **processing time** and **RAM** consumption to assign semantic labels to NumDB datasets without deviation. It is important to notice that both approaches differ significantly in their implementation. MSL requires to build a background knowledge, which in our experiment environment took 01:02:22 and 16.48GB of memory. Keeping a large amount of data in the memory allowed the MSL approach to analyse the tables with an average of 3 seconds per file. However, in the current set-up we used, following Neumaier et al. (2016)'s evaluation, only 46 DBpedia properties. The resources required to build the background knowledge will grow with the number of properties used. As all the necessary information is selected based on the subject column, NUMER does not require prior set-up, resulting in a significantly lower memory consumption. However, requesting all of the information from the DBpedia endpoint at run time resulted in an average processing time of 13 seconds per table. Table 8.6 the processing times per set.

[8] https://aws.amazon.com/marketplace/pp/B012DSCFEK

[9] https://github.com/sebneu/number_labelling

Set	top k	Row Level				Column Level			
		V.S	S	M	L	V.S	S	M	L
0% dev	1	75.61	77.24	73.98	77.64	28.46	34.96	36.18	34.15
	3	93.50	95.93	93.50	95.53	40.65	55.28	56.10	54.88
5% dev	1	75.20	77.24	76.83	78.46	23.58	30.89	30.89	28.05
	3	93.50	95.53	92.68	95.53	35.77	50.00	52.03	48.78
10% dev	1	75.61	73.98	71.14	77.64	22.76	28.05	28.05	28.86
	3	93.09	95.12	93.09	93.90	37.40	48.37	48.78	47.56
15% dev	1	75.20	73.58	69.11	76.83	23.58	26.02	26.83	26.02
	3	93.09	93.90	92.28	94.31	36.59	46.75	47.15	45.12

Table 8.7: Percentage of correctly assigned labels within top 1 and top 3 results in a results list for NUMER approach split by level of analysis (V.S - very small set, S - small set; M - medium set; L- large set).

NUMER – Levels of analysis. The row level analysis achieved better scores compared to the column level (Table 8.7). On the other hand, the combination of both levels (Table 8.8) was often more accurate of to the best performing scores of each level of analysis alone. We found varying levels of accuracy, the **row level analysis** performed consistently well compared to the column level analysis. The difference between accuracy scores by table size was not statistically significant, in contrast to a comparison by error deviation. The performance of the **column level analysis** differed significantly by table size but not by error deviation. The column level analysis used the KS test to assign semantic labels to bags of numerical values. The lower levels of accuracy of the column level analysis suggest a higher dependency on the deviation of the numerical values in a specific column than the row level analysis. Concerning NUMER, which integrated row and column level analysis, it was able to assign semantic labels with a higher degree of precision than the two approaches it is based on, selecting the correct semantic label in over 80% of cases regardless of sampling size or error rate.

Comparative evaluation. We compared the performance of NUMER and MSL for all tables sizes and error degrees within the NumDB benchmark. NUMER consistently outperforms the latter, across all the dimensions in which the datasets change (Table 8.9). NUMER was not affected by variations in table sizes, whereas it was sensible to different degrees of error in the data. Accuracy for top 1 results drops 10.5% on average when introducing any error in the original data from DBpedia. On the other hand, MSL's performance significantly decreased according to both table size and error degree. Overall, the behaviour of MSL appears to be similar to that of our column level analysis, to which it had similar, yet higher, scores. Nevertheless, whereas MSL's performance rises as table size increase, column level analysis' scores are roughly constant for *small*, *medium*, and *large* table sizes, dropping only for *very small* tables.

Set	top k	NUMER			
		V.S	S	M	L
0%	1	90.65	93.50	91.87	93.09
	3	93.50	96.34	93.90	95.93
5%	1	80.49	84.15	78.86	81.30
	3	93.50	95.53	93.09	95.93
10%	1	78.05	78.86	75.61	80.89
	3	93.09	98.37	93.50	94.31
15%	1	78.46	76.02	73.98	78.05
	3	93.09	94.72	93.90	94.72

Table 8.8: Percentage of correctly assigned labels within top 1 and top 3 results in a results list for NUMER.

Set	top k	MSL – average distance				MSL – majority vote			
		V.S	S	M	L	V.S	S	M	L
0%	1	40.65	34.96	53.66	40.24	55.28	40.24	58.13	40.65
	3	60.16	62.20	73.98	73.98	78.86	76.42	77.24	75.20
5%	1	39.43	30.89	48.37	33.74	50.41	34.15	49.59	32.52
	3	53.25	52.44	64.63	63.01	66.67	66.26	65.45	64.23
10%	1	39.02	30.89	47.56	30.08	47.15	31.30	48.78	30.49
	3	52.85	52.44	62.20	59.35	63.01	60.57	62.60	60.98
15%	1	38.21	30.89	42.28	28.86	45.12	27.64	43.90	28.86
	3	52.85	52.85	58.94	56.91	59.76	57.32	59.76	59.35

Table 8.9: Percentage of correctly assigned labels within top 1 and top 3 results in a results list for MSL. (V.S - very small set, S - small set; M - medium set; L - large set).

8.9 Discussion and Limitations

The experiments used to evaluate NUMER enabled us to gain a number of insights about its performance, which indicate directions for future research. NUMER was highly accurate in predicting semantic labels for numerical columns, outperforming the state of the art. MSL, the approach used as a baseline, achieves better scores over the column level analysis aspect of NUMER; however, comparing the combination or row and column level analysis, NUMER outperforms MSL consistently. In most cases, the row level analysis is responsible for most of the accuracy of the whole approach. Only when there is no deviation the integration of the column level analysis yields a significant increase in accuracy.

The results in Table 8.8 show a large difference in terms of performance between top 1 and the top 3 results. Additional scoring factors could be introduced based on other columns or additional textual information available together with the table besides the subject column, in order to improve the top 1 result. The correct semantic labels could be listed after the top 3 (*e.g.*, as a 4 semantic label in a result list), to provide users with a set of potentially valid semantic properties from which they could choose the correct one. This type of interaction may be applied to several contexts, *e.g.*, aiding the process of metadata generation, when generating a dataset summary, or to create dataset to train a more sophisticated machine learning model to assign properties.

When comparing both approaches according to their time and memory consumption, NUMER requires longer time ($13s$) to analyse a single NumDB table than MSL ($\sim 3s$). However, it does not need to generate the background knowledge which, in the case of MSL, carries a cost in memory consumption and initialization time. We believe that this makes NUMER more suitable for use in a personal usecase scenario where one might not want to wait for the background knowledge generation, but might not be of importance when implementing for a large corpora tables (e.g. on a open data platform).

Neumaier et al. (2016) deliberately excluded any additional textual information in MSL. Conversely, NUMER requires textual information in the table to detect potential correct semantic properties. This makes our approach more dependent on textual content in the data: the lack of a subject column or multiple subject columns would likely have a negative impact on the results. A possible solution to that could be to combine NUMER and MSL depending on the presence of the subject column in the table. Moreover, we used textual information in the tables only to disambiguate subject columns to DBpedia entity types. In the future, methods to extract further semantic information from text should be explored, *e.g.*, finding relations between the extracted entities, in order to better understand how different elements in a table relate to each other which could further inform the task of assigning semantic labels to numerical values.

Limitations. As with most approaches there are some limitations connected to this approach. First, a subject column might not be present, or several columns may be considered as subjects. Those scenarios present an additional layer of complexity which would require approaches that are independent of a subject column or other, more tailored solutions. Second, we evaluated our approach by using a set of synthetic tables extracted from DBpedia. Although we processed our tables to test our approach under different conditions, an evaluation in a real world scenario, *i.e.*, with tables found on web pages, should be carried out in the future to provide more solid indications about the applicability of NUMER. Third, there is still work to be done before the presented work could be used in the context of the Open Data datasets. Current content of Open Data datasets might not appear on ant Knowledge Base. This is an additional difficulty when working with Open Data and further development is necessary to allow lifting of such data into the linked data format.

8.10 Conclusion

We presented NUMER – an approach to derive semantic representations of numerical values in tables. Approaches to add semantic meaning to numerical values are particularly valuable, as these represent the most popular column type in open governmental datasets (Mitlöhner et al., 2016). We applied a column level analysis – based on the types of entities found in the subject column of the table and the related values in a KB – matched to the column values in the table. We further applied a row level analysis in which we matched the individual values in a row to the corresponding entity in the KB and approximate the closest numerical values linked to this specific entity. This enabled us to create a table-specific ranked list of potential semantic labels for numerical columns. Automatically inferring the meaning of numerical values found in tables has the potential to significantly improve the discovery of structured data as it can add context to otherwise obscure values. We evaluated our approach using a benchmark (NumDB), created by us, and investigate the influence of the number of rows (percentage of entities of specific type in the KB) and the influence of (intentionally introduced) deviation in the data. We can see that both levels of analysis have a positive influence on the final score in our approach, outperforming the state-of-the-art under the given conditions.

Existing benchmarks have shown not to be useful when the focus of evaluation in the task of assigning semantic labels is mainly on numerical columns. For instance, in T2D (Ritze et al., 2015; Ermilov and Ngomo, 2016), contains only a few tables contain numerical columns disambiguated to DBpedia properties. This indicates a need for new reliable benchmarks to test approaches such as MSL and NUMER, preferably in a real world scenario, without the bias of automatically generated tables. NumDB, although automatically generated, can be seen as a step in that direction. Benchmark involving datasets originating from the open data platforms along knowledge bases targeted at the type of information is which is the most commonly published by the publishers would benefited the field greatly. Disambiguation of columns as a publishing step could be troublesome, however, after initial background knowledge development the publishers could be supported with suggestions on the meaning of the dataset columns with use of approaches such as NUMER or MSL. We believe that providing additional information on the menaing of the numerical (and textual) columns can, for instance, support search over tables on the web and make numerical columns discoverable even if their meaning is not explicitly available. We further see the potential of our approach to be used in recommendation systems for datasets by finding similar or semantically connected tables (Goel et al., 2012).

Chapter 9

Conclusion and Future Work

In this thesis we looked into the problem of dataset search and how to improve the searchability of datasets. We analysed existing state of dataset search and applied quantitative and qualitative methods on search and request logs from five open data portals to identify useful features in the process of dataset search – both from the system and user perspective – which led us to determine the characteristics that make dataset search different from Web search. We then proceeded to look into one of the dimensions of dataset search, in particular, the prevalence of numerical values which lead us to the the scientific problem of semantic labelling of columns with numerical values. After analysing existing approaches targeted at this problem we proposed our own approach in which in addition to analysis of numerical column we analyse the subject column of the dataset. The content of this chapter summarises the contributions made through the presented work and highlights future work directions for improving searchability of datasets.

9.1 Summary

The overall research question of the presented thesis is *how can we improve searchability of datasets that are published on the web?* We have further split this question into three sub-questions:

- [RQ1] Understanding Dataset Search; How users search for data? What search strategies they use?
- [RQ2] What information should be included in the metadata to improve the dataset search process?
- [RQ3] How to automate the generation of metadata? In this work we focus on one of the most under-investigated areas: semantic labelling of numerical columns.

In order to answer *RQ*1 and *RQ*2 we conducted an qualitative and quantitative analysis of users interactions on five open data portals: queries they issue to find data, requests they submit to get access to datasets fitting their specific needs and overall sequence of actions (sessions) when searching for data. Our main findings can be summarised as follows:

We hypothesised that dataset search is a working-hours related activity. We found that most queries issued directly on the portals (i.e., the internal queries) were related to datasets in the area of business and economy. By contrast, external queries were topically more diverse, with topics such as society and towns and cities appearing regularly. We also noticed differences in the ratio of question queries - a larger percentage of external queries included question queries.

Dataset queries are overall short. Carevic et al. (2020) also found (comparing dataset search to publication search) that on average, the length of a dataset search query is shorter which is in line with our findings. Interview study with data users in 2017, shown that many users do not expect that the search functionality will be able to provide relevant data for longer and more specific queries and therefore issue short queries (Koesten et al., 2017). This is also in line with our finding from crowd-generated queries, where crowdworkers issued much longer queries when taken out of the data portal environment. As this was not a real information retrieval scenario this data can be biased towards even longer queries, and as often happens the truth could lie in the middle.

Data search queries on data portals are different to those issues on general web search. There is a difference in topics, length and structure between dataset queries issued directly to data portals and dataset queries issued to web search engines. For instance, a larger percentage of external queries included question queries, which might be due to the increasing ability of large search engines to support natural language type queries.

Common properties to describe datasets are temporal and geospatial coverage, with varying levels of granularity. Queries including some indication of time were almost five times more frequent than in web search (Nunes et al., 2008), suggesting that datasets have a stronger relationship to time than documents. This can include the time frame the data represents (data about a particular year) or the creation time of a dataset (the time the data was collected and published, or the frequency of updates). This findings were supported by analysis of data requests were users were describing datasets with use of restrictions about location, temporal information, specific data type and/or specific granularity (e.g., year/month/day). One of the popular metadata standards of open data platforms – DCAT already includes properties for temporal and geospatial description of datasets, and our findings suggest that providing fine-grained descriptions of these properties could improve search experience.

Our finding regarding queries and data requests suggest that users have dataset specific selection criteria. This is in line with Gregory et al. (2019) findings where authors looking specifically at dataset search amongst researchers in a large scale survey have found that for almost 90% data collection conditions and methodology was important or extremely important in their decisions, which was also considered the key fact to establish trust in the data. This was followed by information about data processing and handling as well as topical relevance. The ease of accessing data was also considered very important. Most of these results are mirrored in a mixed-methods study by Koesten et al. (2020) looking at selection criteria for datasets, different aspects of relevance, quality and usability.

In terms of analysis of full sessions we found that users when searching for data use in majority facets in opposition to query search. Sessions mostly originate from general web search engines, and successful session more often start on directly on a dataset page and are more likely to be longer in terms of action count but shorter in terms of time spent on portals' search result list, which might be a result of struggling with finding the relevant dataset. Above finding suggest that improvements in the area of information indexable by external search engines might be overall beneficial to search outcome. In terms of internal search functionalities more advanced, tailored functionalities should be explored, such as dynamically adapting list of facets. Finally, dedicated section within data portals (such as the one regarding the COVID-19 pandemic) can enhance users understanding and encourage the reuse of datasets on the specific topic. Such direction for data presentation should be further explored. The presented results were achieved with use of data originating from open data portals and with other types of data portals such as corporate or scientific data portals, a generalisability study is necessary. We, however, believe that our findings are applicable to open government data portals of all sizes and the proposed methodology for analysis of logs of data portals is applicable in the task of discovering the relevant metadata fields.

Looking at our last research question ($RQ3$) we conclude from our previous analysis that approaches supporting the provision of context through semi-automated or automated metadata creation could be beneficial to the whole process of dataset search. Aside more difficult structure of datasets in comparison to documents with natural language, which can already cause users difficulties in understanding them and drawing conclusions, majority of it's content are numerical values, which adds another layer of difficulty. This issue is present for both understanding datasets by users and by search algorithms that rank their relevance to user search query. As per usual, adding more information to the metadata describing a specific resource (dataset) can aid data discovery and reuse, as suggested by our analysis and other dataset search literature. In order to aid the metadata generation process we looked into the problem of adding semantic information to datasets, focusing on a problem of generating links between classes in Knowledge bases to columns with numerical values. We identified a number of potential problems and future research directions when labelling numerical columns: precision and deviation in values, format of the values and non machine-readable structure of the table. We then proposed the NUMER approach focusing on the first of the issues, using the context of subject column in the semantic labelling process. We evaluated our approach with use of NumDB benchmark which was automatically generated with use of DBpedia for this purpose. We compared our results with *Multi-level semantic labelling of numerical values* approach proposed by Neumaier et al. (2016). We achieved better results under our test conditions, however, out approach uses subject column, which is not necessary for *Multi-level semantic labelling of numerical values* approach. In the real world a combination of both approaches might be necessary in order to provide full-on experience. This types of approaches can benefit the dataset search process in various ways, not only through providing context to the meaning of data but also through allowing to build a tailored experience through more advanced search interfaces (e.g. tailored facets) or recommendation systems for datasets or search queries.

Presented approach is only in one of the range of research topics targeted at additional metadata generation. This ranges from aiding publishers in their publishing efforts by clearer guidelines on what metadata and in what form could be useful to a potential end user of the dataset, manual metadata generation by voluntaries (to the extent possible since they are not the publisher of the data and might not have the same information on it), linking datasets, columns or data points to the Linked Data, automatic topic detection to the datasets to cluster similar datasets. All of the above could be beneficial to the search functionality which could be tailored towards those kinds of information in terms of search functionality itself (indexing such information) but also supporting users through adding more tailored facets, advanced search boxes or query extension and reformulation suggestions. Furthermore, the proposed qualitative and quantitative methodology for analysis of the search functionalities on data portals could be utilised in different setting to understand dataset search in a broader than through the Open Data Portals perspective as it was done in this work.

9.2 Future Work

As the problem of aiding users and systems in the process of dataset retrieval is still relatively unexplored (in comparison to different search verticals) there are multiple research directions which could be undertaken, following on the research presented in this thesis.

Understanding users of datasets on their dataset retrieval journeys if far from being fully explored and understood. We believe that further studies of users' interactions with search tools are necessary in order to gain a fuller picture of this search vertical. We think the analysis of domains specific portals (e.g. portals with academic datasets) and understanding of users interactions and expectations at various stages of the retrieval process is needed. Understanding of how users interact with the result page and how they judge the relevance of each result list item or how they interact and evaluate the information on the dataset page after selecting a result from result list is as important as development of under the hood functionalities judging the datasets relevance based on users input.

Development of dedicated search functionalities taking into account users' needs are necessary. Recently, Google introduced the Dataset Search engine which is a great initial tool, however, better understanding of how users search is needed in order to propose more tailored dataset ranking algorithms (Brickley et al., 2019). Tailored search interfaces also need to be considered to support users needs and those functionalities. Our findings from crowd queries show a potential for a design space which will encourage users to ask more complex queries. This, however, needs to go in line with search functionalities developments, which will be able to handle such queries.

Search functionalities will need more reliable and tailored metadata in order to perform relevance judgements. Further approaches to support or fully automate metadata generation would be beneficial also to the publishers – as a means to ease the task of metadata creation. They could be additionally useful to a potential user during for example the evaluation of the dataset relevance to their task. Some of the potential research directions in this area is description of datasets generation, evaluation of spatial and temporal coverage of the datasets based on their content or gaining additional insights based on linking datasets with other datasets through lifting them to the semantic web. Along the development of new approaches lifting the structured datasets into the semantic web we need to focus on development of ontologies covering concepts present in datasets (e.g. the ones on Open Data platforms). This would allow to benefit from the wast amounts of research already done with use of the most popular knowledge bases to be utilised to the open data with use of such open data targeted knowledge bases.

Improved metadata and employment of information gathered in search logs can be beneficial not only to the search scenario but also to other functionalities known from different search verticals such as different items recommendation based on their similarity or complementarity to already selected datasets. Analogously, functionalities such as query recommendations, reformulation or next query keyword suggestions could be proposed.

With developments in different search verticals and developments of machine learning based approaches pushing the boundaries of the state-of-the-art in many spaces, we believe that dataset retrieval space needs creation of large corpora serving as a benchmark as well as a training set for the new approaches. New benchmark creation is one of the key aspects in various research problems. Such benchmarks could benefit machine learning based relevance ranking functionalities to further exploring user needs. Corpora of a sufficient size and information would allow evaluation of different approaches and could potentially serve as a base of a new TREC conference track allowing for the problem of dataset retrieval to gain more traction and kick-start rapid development.

We hope that the work presented in this thesis allows further broadening our understanding of dataset search which will allow creation of advanced tools to find and understand the data, benefiting us all with more informed decision-making.

Bibliography

Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *The VLDB Journal*, 24(4):557–581, 2015.

Marco D. Adelfio and Hanan Samet. Schema extraction for tabular data on the web. *Proc. VLDB Endow.*, 6(6):421–432, April 2013. ISSN 2150-8097.

Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 3–10, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7.

Qingyao Ai, Susan T. Dumais, Nick Craswell, and Dan Liebling. Characterizing email search using large-scale behavioral logs and surveys. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1511–1520, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0.

Ahmad Alobaid, Emilia Kacprzak, and Oscar Corcho. Typology-based semantic labeling of numeric tabular data. *Semantic Web*, (Preprint):1–16, 2019.

José Ambite, Sirish Darbha, Aman Goel, Craig Knoblock, Kristina Lerman, Rahul Parundekar, and Thomas Russ. Automatically constructing semantic web services from online sources. *The Semantic Web-ISWC 2009*, 2009.

Daniel P. Ames, Jeffery S. Horsburgh, Yang Cao, Jirí Kadlec, Timothy L. Whiteaker, and David Valentine. Hydrodesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling and Software*, 37:146–156, 2012.

B Saravana Balaji, S Balakrishnan, K Venkatachalam, and V Jeyakrishnan. Automated query classification based web service similarity technique using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 12(6):6169–6180, 2021.

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In

SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004, pages 321–328, 2004.

Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, 1992.

Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 8–14, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-434-8.

Wondwossen Mulualem Beyene. Metadata and universal access in digital library environments. *Library Hi Tech*, 2017.

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Tabel: Entity linking in web tables. In *Proceedings of the 14th International Semantic Web Conference ISWC, 2015*, pages 425–441, 2015.

Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pages 1365–1375, 2019.

Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833, 2012.

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002. ISSN 0163-5840.

Michael J. Cafarella, Alon Halevy, and Jayant Madhavan. Structured data on the web. *Commun. ACM*, 54(2):72–79, 2011. ISSN 0001-0782.

Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: Exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1): 538–549, August 2008. ISSN 2150-8097.

Zeljko Carevic, Dwaipayan Roy, and Philipp Mayr. Characteristics of dataset retrieval sessions: Experiences from a real-life digital library. *arXiv preprint arXiv:2006.02770*, 2020.

Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. Dataset search: a survey. *The VLDB Journal*, 29(1):251–272, 2020.

Xiaoling Chen, Anupama E Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, Yueling Li, Nansu Zong, Min Jiang, et al. Datamed–an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association*, 25(3):300–308, 2018.

Malcolm Clark, Yunhyong Kim, Udo Kruschwitz, Dawei Song, Dyaa Albakour, Stephen Dignum, Ulises Cerviño Beresi, Maria Fasli, and Anne De Roeck. Automatically structuring domain knowledge from text: An overview of current research. *Information Processing and Management*, 48(3):552–568, 2012.

Nick Craswell. Mean reciprocal rank. In *Encyclopaedia of Database Systems*. Springer, 2009.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines - Information Retrieval in Practice*. Pearson Education, 2009. ISBN 978-0-13-136489-9.

Ranjeet Devarakonda, Giriprakash Palanisamy, Bruce E. Wilson, and James M. Green. Mercury: reusable metadata management, data discovery and access system. *Earth Science Informatics*, 3(1):87–94, 2010. ISSN 1865-0481.

Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In *International Semantic Web Conference*, pages 260–277. Springer, 2017.

Ivan Ermilov, Sören Auer, and Claus Stadler. User-driven semantic mapping of tabular data. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 105–112, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1972-0.

Ivan Ermilov and Axel-Cyrille Ngonga Ngomo. TAIPAN: Automatic Property Mapping for Tabular Data. In *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management*, volume 10024, 2016. ISBN 978-3-319-49003-8.

Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *European Semantic Web Conference*, pages 519–534. Springer, 2014.

Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2), 2005.

Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the First International Workshop on Location and the Web*, LOCWEB '08, pages 49–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-160-6.

Aman Goel, Craig A Knoblock, and Kristina Lerman. Exploiting Structure within Data for Accurate Labeling Using Conditional Random Fields. In *Proceedings of the 14th International Conference on Artificial Intelligence (ICAI)*, 2012.

Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. Google fusion tables: Web-centered data management and collaboration. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1061–1066, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0032-2.

Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. Lost or found? discovering data needed for research. *Harvard Data Science Review*, 2019.

Nidhi Grover and Ritika Wason. Comparative analysis of pagerank and hits algorithms. In *International Journal of Engineering Research and Technology*, volume 1. ESRSA Publications, 2012.

Tobias Grubenmann, Abraham Bernstein, Dmitry Moor, and Sven Seuken. Financing the web of data with delayed-answer auctions. In *Proceedings of the 2018 World Wide Web Conference*, pages 1033–1042, 2018.

R. V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, January 2016. ISSN 0001-0782.

Ido Guy, Sigalit Ur, Inbal Ronen, Sara Weber, and Tolga Oral. Best faces forward: a large-scale study of people search in the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 221–230, 2010.

Luis-Daniel Ibáñez, Laura Koesten, Emilia Kacprzak, and Elena Simperl. Characterising dataset search on the european data portal: an analysis of search logs, 2020.

Bernard J Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3), 2006.

Bernard J. Jansen and Amanda Spink. An analysis of web searching by european alltheweb.com users. *Information Processing and Management*, 41(2):361–381, 2005. ISSN 0306-4573.

Bernard J. Jansen and Amanda Spink. How are we searching the world wide web?: A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006. ISSN 0306-4573.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000. ISSN 0306-4573.

Daxin Jiang, Jian Pei, and Hang Li. Mining search and browse logs for web search: A survey. *ACM Transactions on Intelligent Systems and Technology*, 4(4):57:1–57:37, 2013. ISSN 2157-6904.

Steve Jones, Sally Jo Cunningham, Rodger McNab, and Stefan Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152–169, 2000. ISSN 1432-5012.

Emilia Kacprzak, José M Giménez-García, Alessandro Piscopo, Laura Koesten, Luis-Daniel Ibáñez, Jeni Tennison, and Elena Simperl. Making sense of numerical data-semantic labelling of web tables. In *European Knowledge Acquisition Workshop*, pages 163–178. Springer, 2018a.

Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, and Elena Simperl. Characterising dataset search—an analysis of search logs and data requests. *Journal of Web Semantics*, 55:37–55, 2019.

Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. Characterising dataset search queries. In *Companion Proceedings of the The Web Conference 2018*, pages 1485–1488, 2018b.

Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. *A Query Log Analysis of Dataset Search*, pages 429–436. Springer International Publishing, Cham, 2017. ISBN 978-3-319-60131-1.

Navjot Kaur and Himanshu Aggarwal. Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement. *International Journal of Information Technology*, 10(1):99–110, Mar 2018. ISSN 2511-2112.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009. ISSN 1554-0669.

Dagmar Kern and Brigitte Mathiak. Are there any differences in data set retrieval compared to well-known literature retrieval? In *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings*, pages 197–208, 2015.

Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The web as a graph: measurements, models, and methods. *Computing and combinatorics*, pages 1–17, 1999.

Craig A. Knoblock, Pedro Szekely, José Luis Ambite, Aman Goel, Shubham Gupta, Kristina Lerman, Maria Muslea, Mohsen Taheriyan, and Parag Mallick. Semi-automatically mapping structured sources into the semantic web. In *The Semantic Web: Research and Applications, ESWC 2012*, pages 375–390, 2012.

Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies*, 135:102367, 2020.

Laura M. Koesten, Emilia Kacprzak, Tennison Jenifer, and Elena Simperl. 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, New York, NY, USA, 2017. ACM.

Sven R. Kunze and Soren Auer. Dataset Retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing*, sep 2013. ISBN 978-0-7695-5119-7.

Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. *ACM Transactions on Information Systems*, 19(3):242–262, 2001.

Steve Lavalle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 52(2), 2011. ISSN 15329194.

Oliver Lehmberg and Christian Bizer. Web table column categorisation and profiling. In *Proceedings of the 19th International Workshop on Web and Databases*, WebDB '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343107.

Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016. ISBN 9781450341448.

Xin Li, Bing Liu, and Philip S. Yu. *Time Sensitive Ranking with Application to Publication Search*, pages 187–209. Springer, New York, 2010. ISBN 978-1-4419-6515-8.

Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. *PVLDB*, 3(1):1338–1347, 2010.

Meiyu Lu, Srinivas Bangalore, Graham Cormode, Marios Hadjieleftheriou, and Divesh Srivastava. A Dataset Search Engine for the Research Document Corpus. In *2012 IEEE 28th International Conference on Data Engineering*, apr 2012. ISBN 978-0-7695-4747-3.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Johann Mitlöhner, Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Characteristics of open data csv files. In *Open and Big Data (OBD)*. IEEE, 2016.

Varish Mulwad, Tim Finin, Zareen Syed, and Anupam Joshi. Using linked data to interpret tables. In *Proceedings of the First International Workshop on Consuming Linked Data*, volume 665 of *CEUR Workshop Proceedings*, 2010.

Felix Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2013.

Sebastian Neumaier and Axel Polleres. Enabling spatio-temporal search in open data. *Journal of Web Semantics*, 55:21–36, 2019.

Sebastian Neumaier, Jürgen Umbrich, Josiane Xavier Parreira, and Axel Polleres. Multi-level Semantic Labelling of Numerical Values. In *Proceedings of the 15th International Semantic Web Conference*, 2016.

Sérgio Nunes, Cristina Ribeiro, and Gabriel David. Use of temporal expressions in web search. In *European Conference on Information Retrieval*, pages 580–584. Springer, 2008.

Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. Struggling and success in web search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1551–1560, 2015.

Adan Ortiz-Cordova, Yanwu Yang, and Bernard J. Jansen. External to internal search: Associating searching on search engines with searching on sites. *Information Processing Management*, 51(5):718–736, 2015.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

Minh Pham, Suresh Alse, Craig A Knoblock, and Pedro Szekely. Semantic labeling: a domain-independent approach. In *International Semantic Web Conference*. Springer, 2016.

Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, April 2002. ISSN 0001-0782.

Alessandro Piscopo and Emilia Kacprzak. Numdb, 2018.

Erhard Rahm and Philip A Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

S. K. Ramnandan, Amol Mittal, Craig A. Knoblock, and Pedro A. Szekely. Assigning semantic labels to data sources. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC Proceedings*, pages 403–417, 2015.

Dominique Ritze, Oliver Lehmberg, and Christian Bizer. Matching HTML Tables to DBpedia. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pages 10:1–10:6, 2015.

Colin Robson and Kieran McCartan. *Real world research*. John Wiley & Sons, 2016.

Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Jeffrey S Grethe, Hua Xu, Ian M Fore, Jared Lyle, Anupama E Gururaj, Xiaoling Chen, et al. Dats, the data tag suite to enable discoverability of datasets. *Scientific data*, 4:170059, 2017.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Sunny Sharma and Vijay Rana. Web search personalization using semantic similarity measure. In Pradeep Kumar Singh, Arpan Kumar Kar, Yashwant Singh, Maheshkumar H. Kolekar, and Sudeep Tanwar, editors, *Proceedings of ICRIC 2019*, pages 273–288, Cham, 2020. Springer International Publishing. ISBN 978-3-030-29407-6.

Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 525–534, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. In *ACm SIGIR Forum*, volume 33. ACM, 1999.

Ayush Singhal, Ravindra Kasturi, Vidyashankar Sivakumar, and Jaideep Srivastava. Leveraging Web Intelligence for Finding Interesting Research Datasets. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, volume 1, 2013. ISBN 978-0-7695-5145-6.

Amanda Spink, Seda Ozmutlu, Huseyin C Ozmutlu, and Bernard J Jansen. U.s. versus european web searching trends. In *ACM Sigir Forum*, volume 36. ACM, 2002.

Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American society for information science and technology*, 52(3), 2001.

Zareen Syed, Tim Finin, Varish Mulwad, and Anupam Joshi. Exploiting a web of semantic data for interpreting tables. In *Proceedings of the 2nd Web Science Conference, 2010*, 2010.

Mona Taghavi, Ahmed Patel, Nikita Schmidt, Christopher Wills, and Yiqi Tew. An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces*, 34(1), 2012.

Mohsen Taheriyan, Craig A. Knoblock, Pedro Szekely, and José Luis Ambite. A scalable approach to learn semantic models of structured sources. In *Proceedings of the International Conference on Semantic Computing*, 2014.

David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment and evolution of open data portals. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 404–411, Aug 2015.

Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. Recovering semantics of tables on the web. *Proc. VLDB Endow.*, 4(9):528–538, June 2011. ISSN 2150-8097.

Stefaan Verhulst and Andrew Young. Open data impact when demand and supply meet. Technical Report March, GOVLAB, 2016.

A. Walker, B. Pham, and A. Maeder. A Bayesian framework for automated dataset retrieval in geographic information systems. In *10th International Multimedia Modelling Conference, 2004. Proceedings.*, 2004. ISBN 0-7695-2084-7.

Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Qili Zhu. Understanding tables on the web. In *Conceptual Modeling - 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings*, pages 141–155, 2012.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. Evidence aggregation for answer re-ranking in open-domain question answering.(2018). In *Proceedings of the 6th International Conference on Learning Representation, Vancouver, Canada, 2018 April 30-May*, volume 3, pages 1–14, 2018.

Wouter Weerkamp, Richard Berendsen, Bogomil Kovachev, Edgar Meij, Krisztian Balog, and Maarten de Rijke. People searching for people: Analysis of a people search engine log. SIGIR '11, New York, NY, USA, 2011. ISBN 978-1-4503-0757-4.

Ryen W White, Matthew Richardson, and Wen-tau Yih. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 135–136, 2015.

Ryen W. White, Sheng Wang, Apurv Pant, Rave Harpaz, Pushpraj Shukla, Walter Sun, William DuMouchel, and Eric Horvitz. Early identification of adverse drug reactions from search log data. *Journal of Biomedical Informatics*, 59:42 – 48, 2016. ISSN 1532-0464.

Dominik Wienand and Heiko Paulheim. Detecting incorrect numerical data in dbpedia. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen

Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges*. Springer International Publishing, 2014. ISBN 978-3-319-07443-6.

Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. Query log analysis of an electronic health record search engine. In *AMIA annual symposium proceedings*, volume 2011, 2011.

Philip S. Yu, Xin Li, and Bing Liu. Adding the temporal dimension to search " a case study in publication search. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 543–549, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2415-X.

Shuo Zhang and Krisztian Balog. Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(2): 1–35, 2020.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. Tfidf, LSI and multi-word in information retrieval and text categorization. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12-15 October 2008*, pages 108–113, 2008.

Ying Zhang, Bernard J Jansen, and Amanda Spink. Time series analysis of a web search engine transaction log. *Information Processing & Management*, 45(2), 2009.

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M. Jose. Which vertical search engines are relevant? In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1557–1568, 2013.