How do we search for data?

Towards User-Driven Dataset Descriptions

Emilia Kacprzak[1,2], Laura Koesten[1,2], **Luis-Daniel Ibáñez[1]**, Elena Simperl[1]

[1] University of Southampton

[2] The Open Data Institute

Understand how users search for data

-->

Better design of descriptions and data search engines

Analysis

Qualitative:

Interviews with 20 data professionals

Goal:

- Insight on the dataset search process
- Identify user requirements

Quantitative:

Query log analysis from 4 Open Data portals from 3 countries (2M queries).

Goal:

Characterize queries currently issued to Data Portals. Compare to other verticals.

Qualitative analysis - Quotes

*“Documentation is most frustrating, **there's often data without documentation** and fishing for this information is the hardest bit.”*

*“It's very difficult first when you download new data, to have a quick idea of what the data represents, a **quick summary of the data**.”*

*“I'm looking at [..] the coverage of the data, so does it cover the **geographic area** I'm interested in? Or the **time period** that I'm interested in? And does it do that to the level of detail I need?”*

*“Helping people to understand what's in the data is incredibly useful and also what has been excluded from the data ….**what were the cleaning choices**? **What constitutes a valid record**?”*

Qualitative - User needs

- Filtering according to:
  - location, provenance, format, licence, time frame and date, publishing date, location of publication and data schema
- Details on how the original data was collected.
- Summary comprised of statistics and representative samples as a preview
- Historical evolution of data (when applicable)

Quantitative - User profile

- mostly desktop devices (85%)
- mostly active during weekdays and working hours
- 41% use Chrome, 31% use Internet Explorer
- 68% of queries came from Web Search Engines

Quantitative - User profile

- mostly desktop devices (85%)
- mostly active during weekdays and working hours
- 41% use Chrome, 31% use Internet Explorer
- 68% of queries came from Web Search Engines

Suggests data search is a work-related activity
Reliance on general-purpose search engines

Query characteristics

- Mostly **keyword queries** (less than 1% question queries)
- **short** (in average 2 words per query)
- 1 word queries represent **50%** of all queries

Query characteristics

- Mostly **keyword queries** (less than 1% question queries)
- **short** (in average 2 words per query)
- 1 word queries represent **50%** of all queries

Suggests exploratory search.

Query Characteristics - specifications

- File format or words such as “data” or “datasets”: 6%
- Location (data about a place): 5.6%
- time frame (data of certain year or month): 7.3%

Query Characteristics - specifications

- File format or words such as “data” or “datasets”: 6%
- Location (data about a place): 5.6%
- time frame (data of certain year or month): 7.3%

No dominant further query specification

Future Work

Analyse click and download patterns to estimate search effectivity

Further analysis of keyword queries: Do they refer to categories? to entities? to specific datasets?