IS YOUR DATA FINDABLE ON THE WEB?

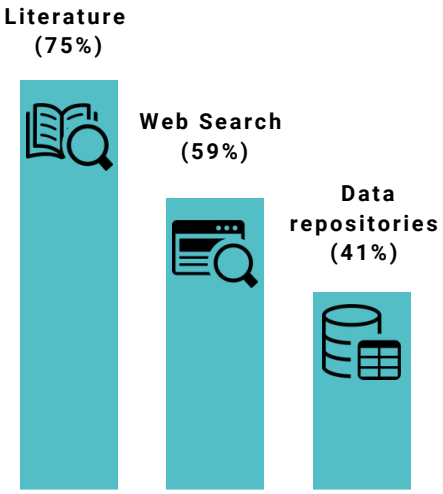
ENHANCING DATA FINDABILITY

HOW SCIENTISTS AND REPOSITORIES CAN IMPROVE THEIR DATA VISIBILITY

DATA DISCOVERY PROCESS

People use web search to find data

- Less than 6% of articles use DOIs to cite data [1];
- The majority of users will use web search also to find the data from the literature;
- As such it is very important for the findability of datasets to be well represented in Web Search.



[1] Mayo, Vision, & Hull (2016) Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. doi:10.5061/dryad.8q931

[2] Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020, April 30). Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review, 2. doi:10.1162/99608f92. e38165eb

It is crucial for findability that datasets are well represented in Web search

We explain how!

INCREASING DATA FINDABILITY FOR SCIENTISTS

- Write papers about the data
- Speak about it at conferences
- Encourage others to link to your data

INCREASING DATA FINDABILITY FOR REPOSITORIES

ALLOW YOUR DATA TO BE HARVESTED

Research (Meta)Data on the Web

- Each dataset should have exactly one dedicated webpage, directly accessible to all users;
- The title is the most important bit of information on this page, so it should be large, on the top and part of the URL;
- Include other known names, such as acronyms, in the title or description of the dataset;
- Mention certain keywords explicitly:
  - What is it? (a *dataset*)
  - What you can do with it ? (*download*, look at the *questionnaire*).

Adhere to (Web) Metadata Standards

- Schema.org has a type “Dataset” which you can use to describe the data and which is used widely for harvesting;
- Not all fields are harvested, though. We have seen title, description, identifier, license, provider and/or author, datePublished, spatialCoverage, temporalCoverage;
- Dublin Core is harvested indiscriminately. Use title, type, creator, identifier, date, description.

MONITORING APPROACHES

Google Search Console | Tools for Verification | User Tracking | Sitemaps

TOOLS & MONITORING

- Use Google Search Console to find out what works about your website and what doesn't;
- Use tools such as validator.schema.org and search.google.com/test/rich-results for embedded metadata. If you didn't check, assume it doesn't work;
- Conduct user tracking - with tools such as Matomo - to track different usage aspects and performance of a website and can improve the experience for visitors;
- Use sitemaps, which allows for search engine indexing.

OUR EXPERIENCE

Most effective measures are very easy to implement.

Frequently asked legal question:

Can I use Google Search Console?

Google Search Console allows website owners to access anonymized data that Google has already collected, so it is not relevant for the GDPR and privacy in the relationship between the website owner and the user of the site. The issue for active user tracking like Google Analytics or Matomo is more complex; please contact your privacy protection counsel.

KonsortSWD Task Area 5, Measure 2 , NFDI funding number 442494171
Authors: Fidan Limani, Valentina Hiseni, Janete Saldanha Bach, Brigitte Mathiak
Contact: brigitte.mathiak@gesis.org
DOI: 10.5281/zenodo.6760242

nfdi | KonsortSWD | gesis Leibniz Institute for the Social Sciences | ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics