gesis Leibniz Institute for the Social Sciences

Meet the Experts

Five Ways to Turn Your Dataset into Clickbait

Meet the Experts – GESIS online talks

Knowledge technologies for the Social Science: Access to Social Science Data and Services

Brigitte Mathiak, 15.2.2024

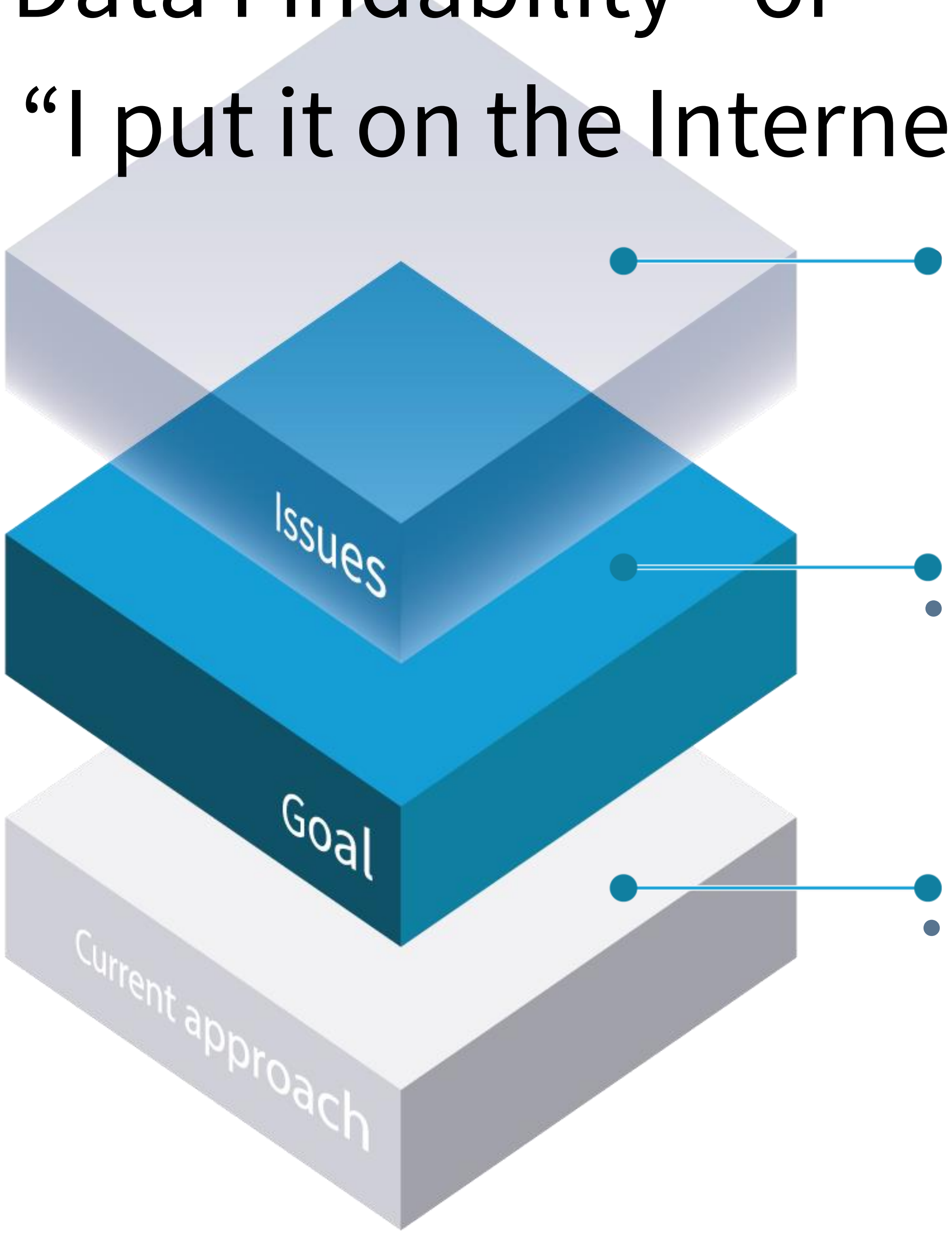
Leibniz Association

Speaker‘s introduction

Dr. Brigitte Mathiak

- Computer scientist
- I specialize in
  - Knowledge Discovery from Scientific Texts
  - Dataset Discovery
  - Research Infrastructure
- I am very active in the research data community
  - Research Data Alliance (co-chair of two groups)
  - Active in four consortia of the National German Research Data Infrastructure
  - Deputy spokesperson of the section (meta)data, terminologies and provenance
  - Co-chair of the GoFAIR Initiative on Data Discovery
- Projects
  - CodeInspector (Bring Social Science software into GESIS)
  - UnknownData (Find unknown datasets)
  - SmartER (Extraction of Author Affiliations)
- Contact: brigitte.mathiak@gesis.org

Data Findability - or “I put it on the Internet. What more do you want?”



- Issues
  - ○ Research data is much **harder** to find on the web than literature references
  - ○ Data needs are much more **specific**
  - ○ Data sources are much more **heterogeneous**
  - ○ Data sources are much more **distributed**

- Goal: Improve visibility for true findability
  - ○ **Increase** findability and re-use of metadata
  - ○ Provide metadata in common **web formats**
  - ○ PIDs for identifiability
  - ○ Implement appropriate **SEO measures**
  - ○ High-quality and rich **metadata**

- Current approach
  - ○ **Model** data discovery
  - ○ **Observe** social scientists
  - ○ **Survey** and **Interview** with our KonsortSWD datacenters
  - ○ **Hands-On recommendations** on how to improve visibility

Our model of the discovery process

Data discovery is complex [1]

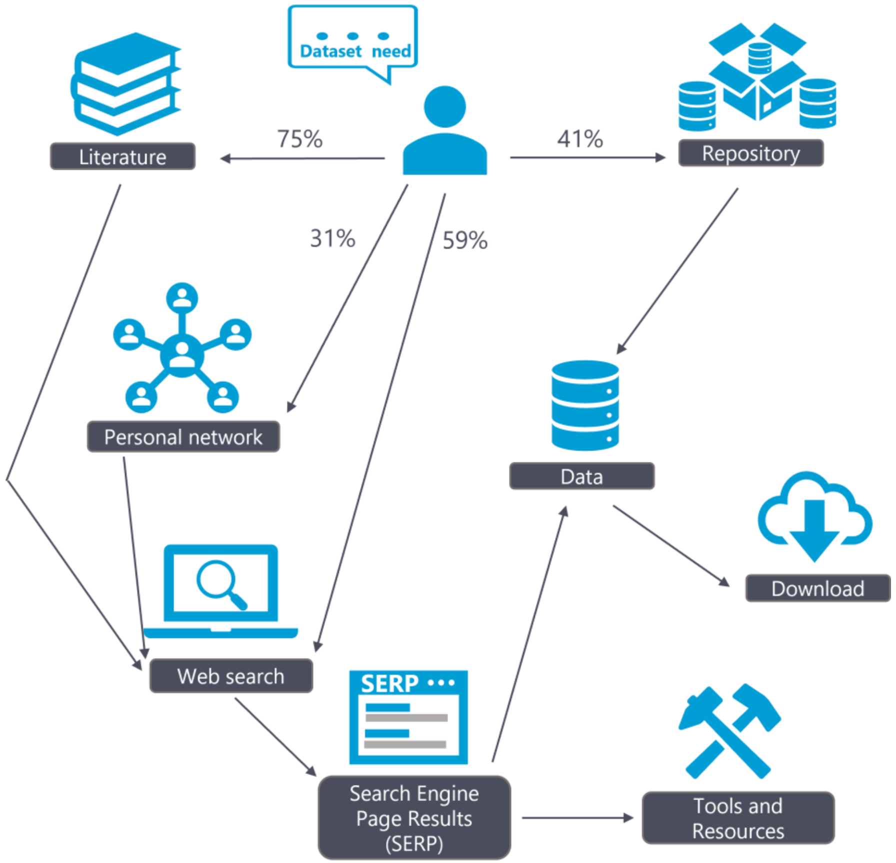
- **75%** of researchers often rely on literature review
- **59%** of researchers rely on search engines
- **41%** use domain data repositories

Focus on the lower path(s)

- If someone found the name of a dataset in the literature (<3% use links or DOI), how do they get to the download page?
- If someone types in a data-related query in a web search engine, will they find the relevant data

Single point of failure

- Dataset does not show up in web search
- Query terms do not match the description
- Description is not available the first place
- Users cannot find the relevant repository



The percentages are taken from [1] and denote percentages of users who use this method "Often" in dataset search, rather than "Occasionally" or "Never".

[1] 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

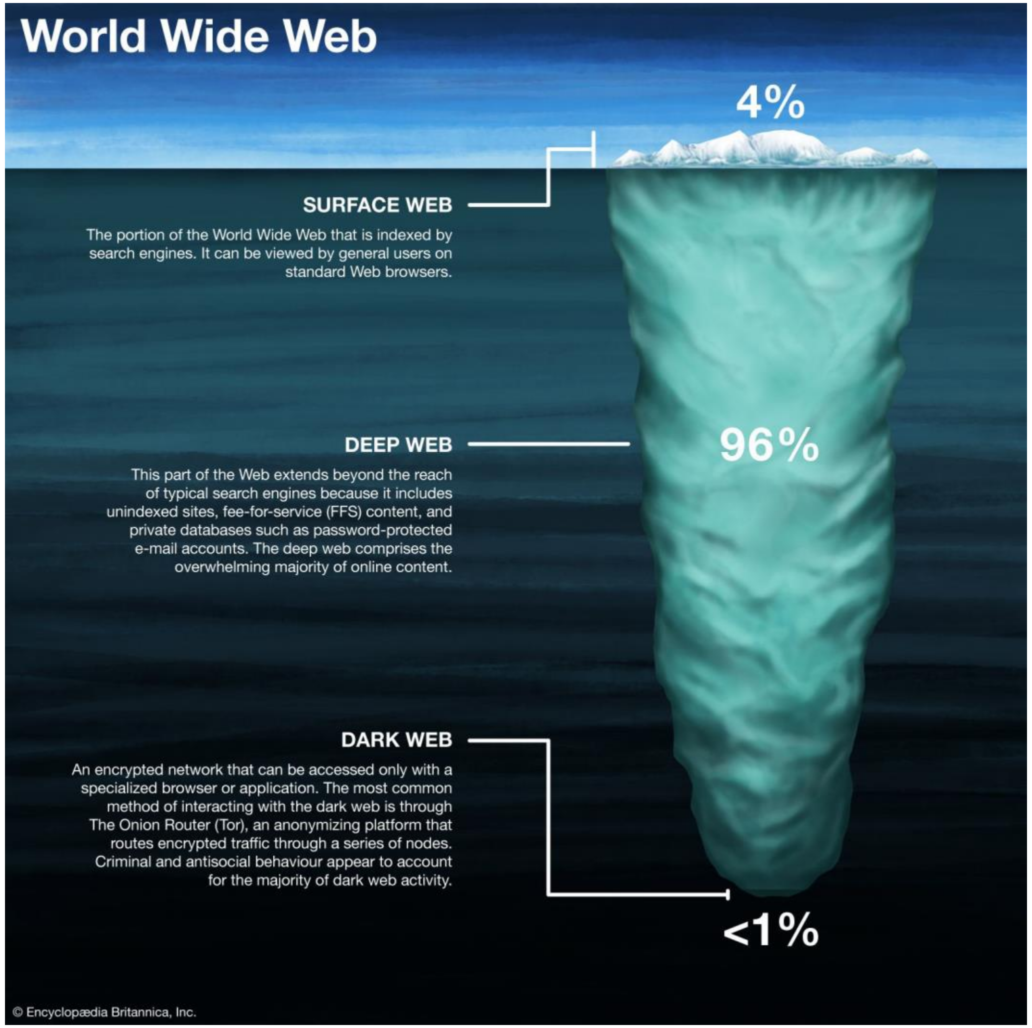
The crux of data findability

- Does the download link for your dataset appear, when you type in the name or acronym of the dataset into web search?

- This depends on the repository or website you choose for publication

Deep Web

- Most of the content on the internet is published, but is not actually available for search
- Unfortunately, this includes many sites concerning research data
- Repository owners need to invest considerable energy to keep their content floating



SURFACE WEB

The portion of the World Wide Web that is indexed by search engines. It can be viewed by general users on standard Web browsers.

DEEP WEB

This part of the Web extends beyond the reach of typical search engines because it includes unindexed sites, fee-for-service (FFS) content, and private databases such as password-protected e-mail accounts. The deep web comprises the overwhelming majority of online content.

DARK WEB

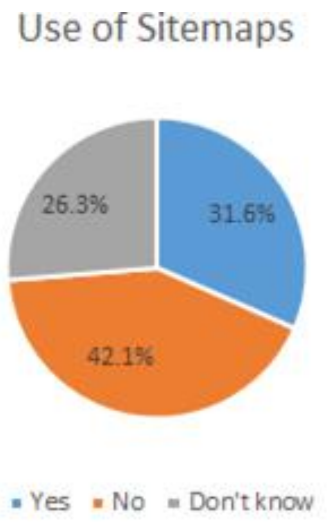
An encrypted network that can be accessed only with a specialized browser or application. The most common method of interacting with the dark web is through The Onion Router (Tor), an anonymizing platform that routes encrypted traffic through a series of nodes. Criminal and antisocial behaviour appear to account for the majority of dark web activity.

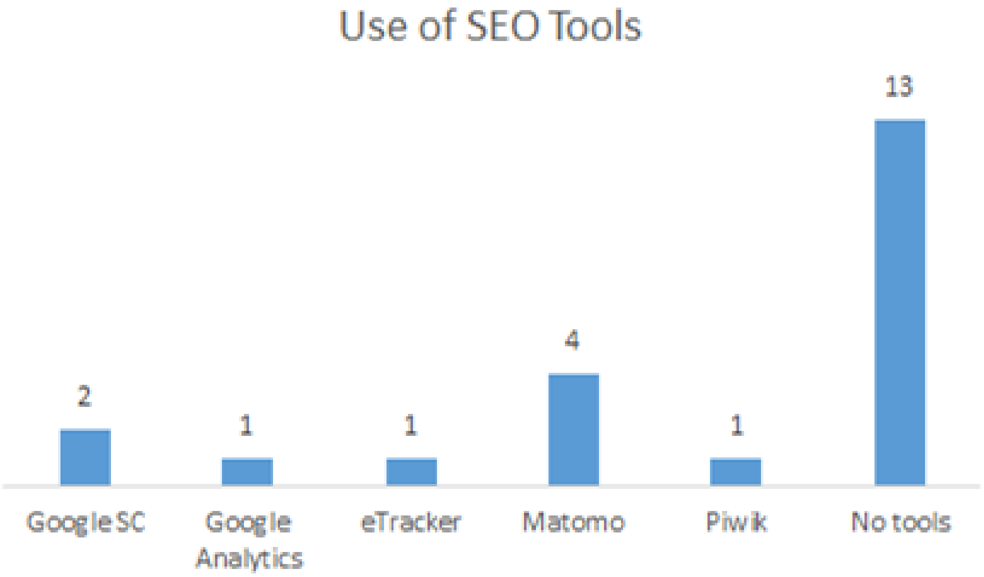
© Encyclopædia Britannica, Inc.

Survey on Social Science data centers

Available tools to guide SEO are underused

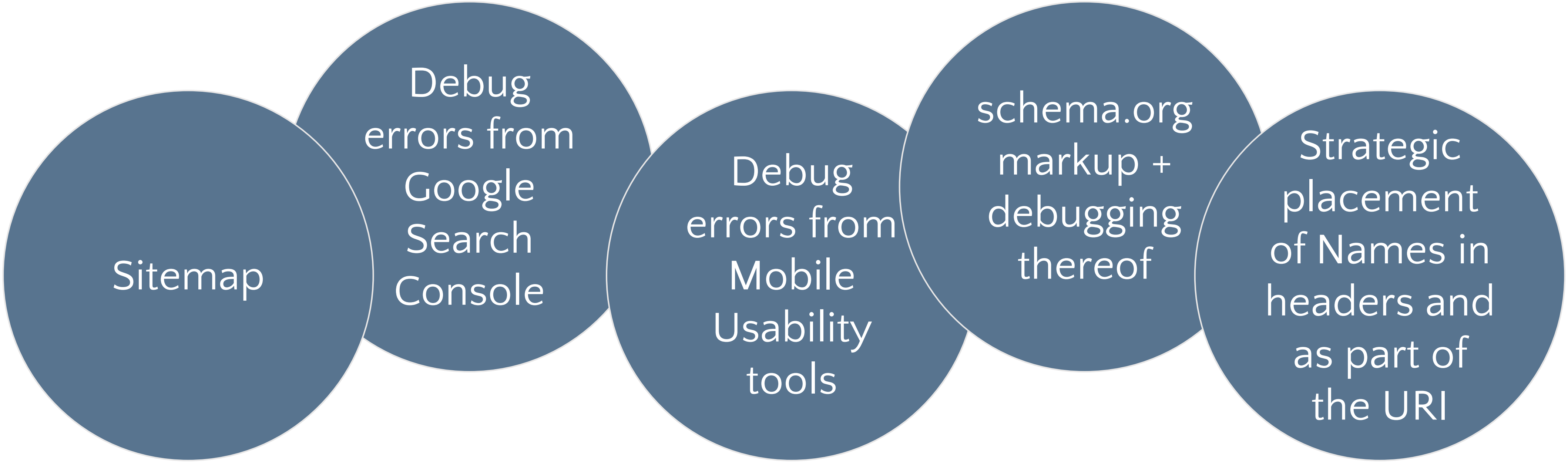
- **Sitemaps** are vital to communicate which content should be indexed
- Without proper tools to **monitor web traffic**, repositories cannot know what the problems are
- **Google Search Console** is free of charge and only displays data already collected, so there are no issues with privacy





What SEO measures did we take?

Most **important** measure is to use monitoring tools!



Sitemap

Debug errors from Google Search Console

Debug errors from Mobile Usability tools

schema.org markup + debugging thereof

Strategic placement of Names in headers and as part of the URI

And keeping with new trends which pop up every year…

Impact of SEO measures on the repository level

	2019	2020	2021	2022	2023
Impressions	(102.400)	513.000	1.163.824	1.930.956	5.340.000
Clicks	(6.840)	29.200	45.915	53.249	79.000

Total impressions (times that search results showed up on any result page on any query in the given year)
Total Clicks (times that someone clicked on the result in the given year)
Source: Google Search Console for search.gesis.org

- This is influenced by the other factors as well
- We also consult data center within the NFDI consortium KonsortSWD

Five Ways to turn your Dataset into Click Bait

- Deposit your data with a good repository
  - Look for repositories that are easy to find with web search
  - Make sure dataset specific subpages are findable
  - PIDs, ideally DOIs, are highly recommendable
  - Membership in national and international networks is also important
  - Having an additional project website for the data is also an option, if you want the flexibility

Observation study

*“In the context of your research you need research data.
For today, you decide to start with the search for research data.“*

We observed **12** social scientists searching for research data in their natural office environment

We recorded the screencasts and what they would explain about their process

Semi-structured interviews after the experiment

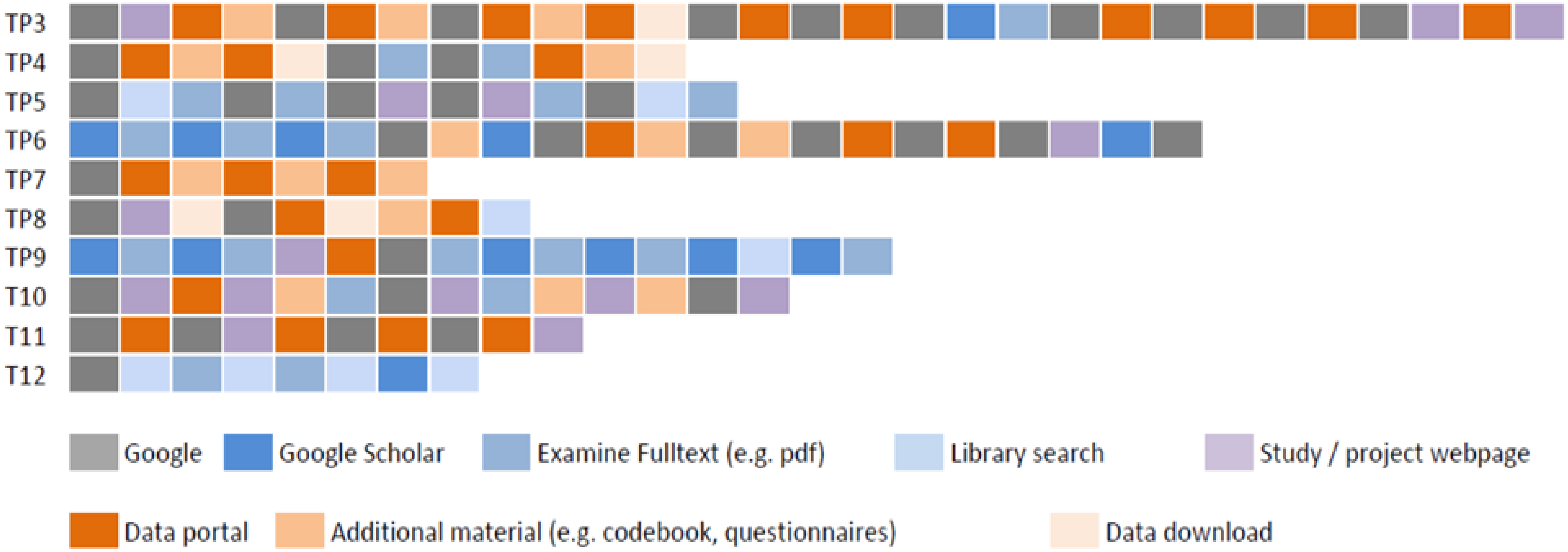
published in: [2] Krämer, T., Papenmeier, A., Carevic, Z., Kern, D., Mathiak, B. (2021). Data Seeking Behaviour in the Social Sciences. International Journal on Digital Libraries. https://doi.org/10.1007/s00799-021-00303-0

Findability

Meet the Experts

Interaction diagram

Visualization of interaction sequences of ten participants (P3–P12). Each box represents one interaction



- ❑ **83%** find their data through literature [2]
- ❑ **80%** in a survey among 1,458 social scientists [3]
- ❑ **75%** in a survey among 1,677 scientists [1]

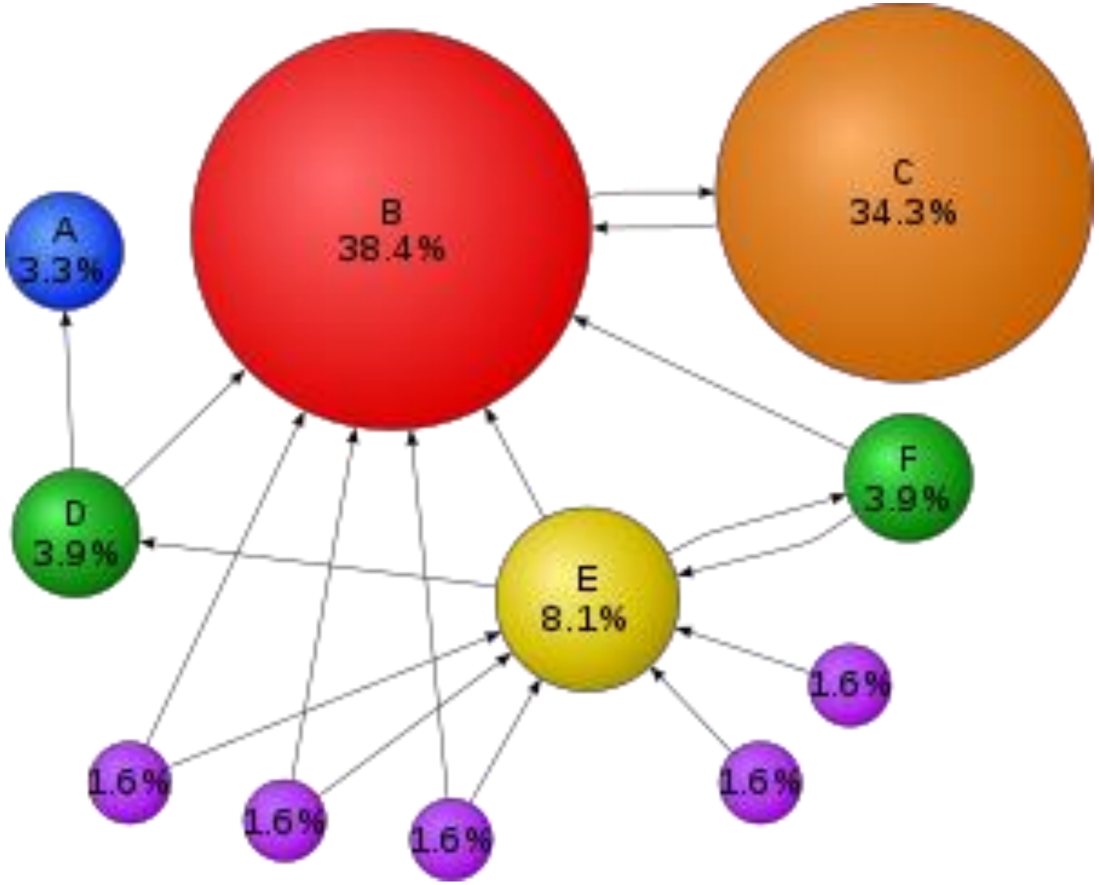
[3] Friedrich, Tanja. (2020). Looking for data: Information seeking behaviour of survey data users. 10.18452/22173.

Five Ways to turn your Dataset into Click Bait

- Deposit your data with a good repository, e.g. GESIS
- Write papers about your dataset
    - Make sure to use the DOI and the name and any name variants
    - Inspire others to write even more papers on your data

Ranking in Web search

PageRank algorithm



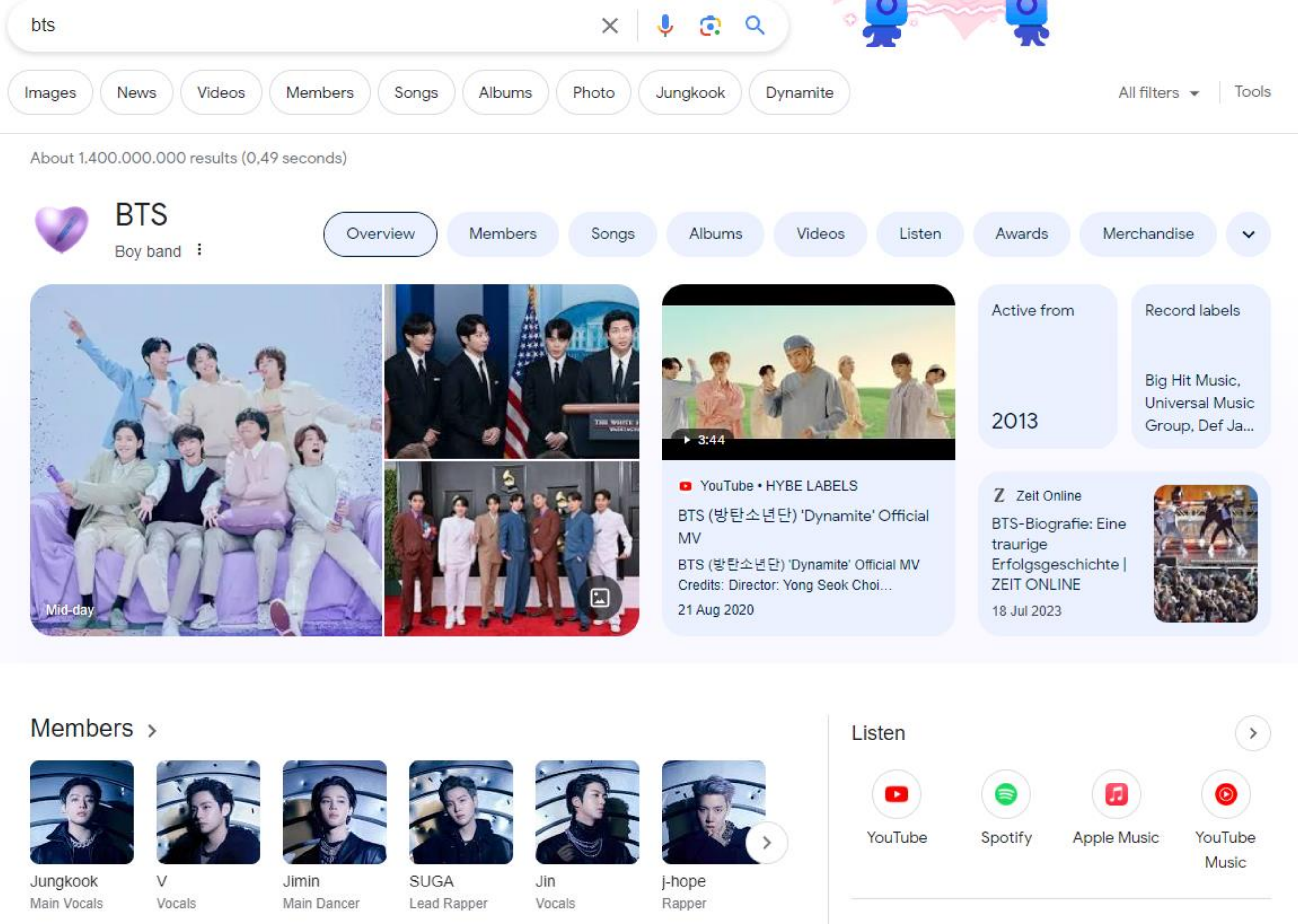
- Websites are more likely to rank high, when they are linked to from websites, which are ranked highly themselves
- Academic websites usually have a good starting rank
- As does Wikipedia

Five Ways to turn your Dataset into Click Bait

- Deposit your data with a good repository, e.g. GESIS
- Write papers about your dataset
- Make high class links to your dataset
    - Link to it from your personal homepage, e.g. CV, institution webpage
    - Use it in teaching
    - Consider writing a blog or Wikipedia article about it

Names and Acronyms

- Imagine you had a study on the Bundestag and you want to call it the Bundestagstudie (BTS)
- Let’s see what happens…



bts

Images | News | Videos | Members | Songs | Albums | Photo | Jungkook | Dynamite

All filters | Tools

About 1.400.000.000 results (0,49 seconds)

BTS

Boy band

Overview | Members | Songs | Albums | Videos | Listen | Awards | Merchandise

Mid-day

3:44

YouTube • HYBE LABELS

BTS (방탄소년단) 'Dynamite' Official MV

BTS (방탄소년단) 'Dynamite' Official MV Credits: Director: Yong Seok Choi...

21 Aug 2020

Active from

2013

Record labels

Big Hit Music, Universal Music Group, Def Ja...

Zeit Online

BTS-Biografie: Eine traurige Erfolgsgeschichte | ZEIT ONLINE

18 Jul 2023

Members

Jungkook	V	Jimin	SUGA	Jin	j-hope
Main Vocals	Vocals	Main Dancer	Lead Rapper	Vocals	Rapper

Listen

YouTube | Spotify | Apple Music | YouTube Music

Meet the Experts

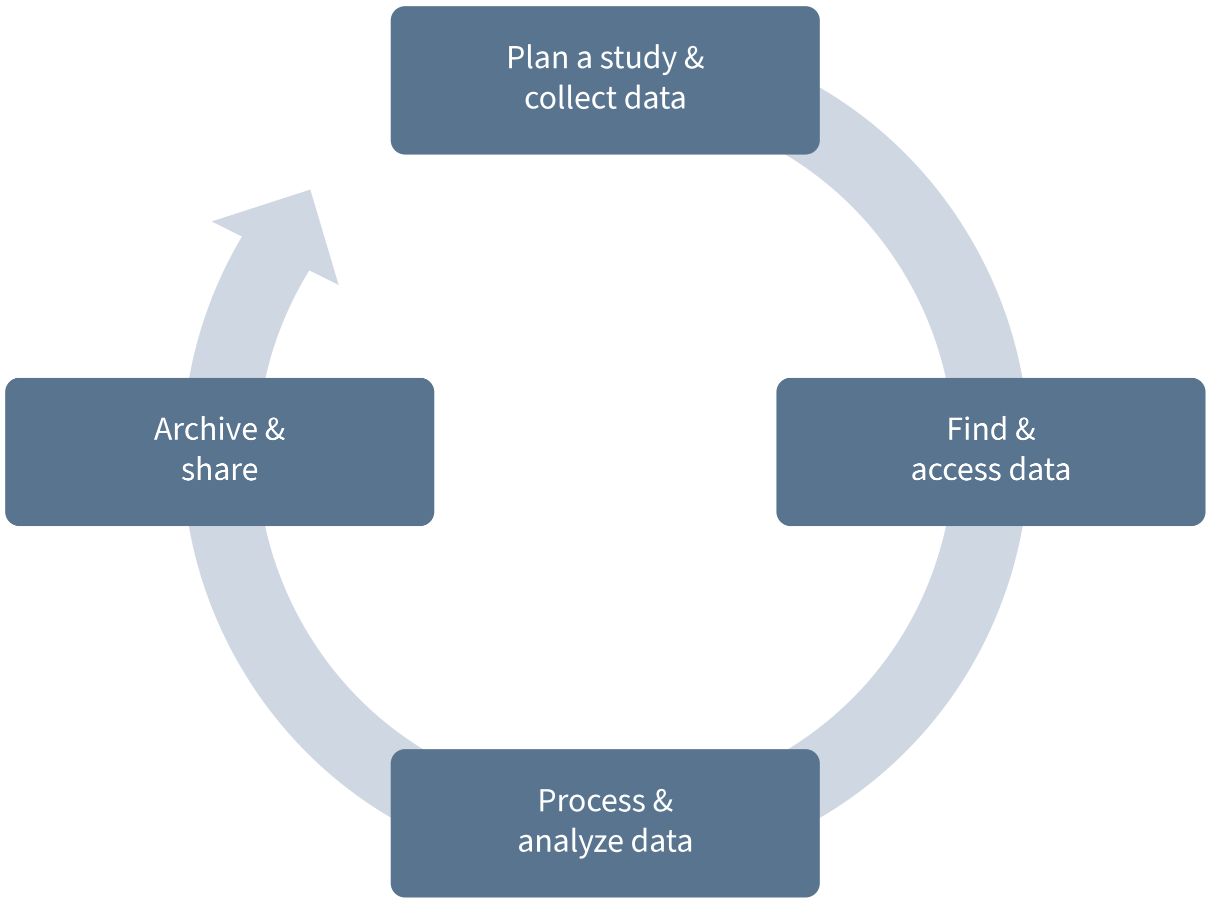
Names and Acronyms

- Imagine you had a study on the Bundestag and you want to call it the Bundestagstudie (BTS)
- That is not a good idea
- Likewise, please avoid common names and proper words, like family, age, etc.
- Make to sure Google both the name and the acronym beforehand

Five Ways to turn your Dataset into Click Bait

- Deposit your data with a good repository, e.g. GESIS
- Write papers about your dataset
- Make high class links to your dataset
- Google both the name and the acronym of your dataset, before you settle on a name

Research data cycle



- Finding and accessing data is not enough
- The research data cycle only comes into full swing when other scientists are able to re-use the data
- Data that cannot be re-used is not going to be cited

Five Ways to turn your Dataset into Click Bait

- Deposit your data with a good repository, e.g. GESIS
- Write papers about your dataset
- Make high class links to your dataset
- Google both the name and the acronym of your dataset, before you settle on a name
- Provide rich metadata and documentation, so others can re-use your data

The Future

- We are planning a study on the impact of ChatGPT on data retrieval
  - Both the actual impact on user behavior, but also how useful ChatGPT is to find data
- Structures are shifting and it is possible that soon metasearch engines will play a more important role for data retrieval

Expert contact & GESIS consulting

Contact: you can reach the speaker/s via e-mail:
Dr. Brigitte Mathiak

[brigitte.mathiak@gesis.org]

GESIS Consulting: GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website www.gesis.org for more detailed information on available services and terms.

Upcoming talks

- 15.02.2024: Five ways to turn your dataset into click bait
- 14.03.2024: Searching the social sciences with GESIS Search
- 11.04.2024: How knowledge graphs can help you to share research data and information
- 16.05.2024: Opportunities and challenges of Large Language Models for the social sciences
- 13.06.2024: Preserving and analysing large-scale Twitter data
- 11.07.2024: Introduction to scholarly information extraction

Please visit our meet-the-experts website:

https://www.gesis.org/en/services/sharing-knowledge/consulting-and-guidelines/meet-the-experts

Thank you for participating!