KonsortSWD

Konsortium für die
Sozial-, Verhaltens-, Bildungs- und
Wirtschaftswissenschaften
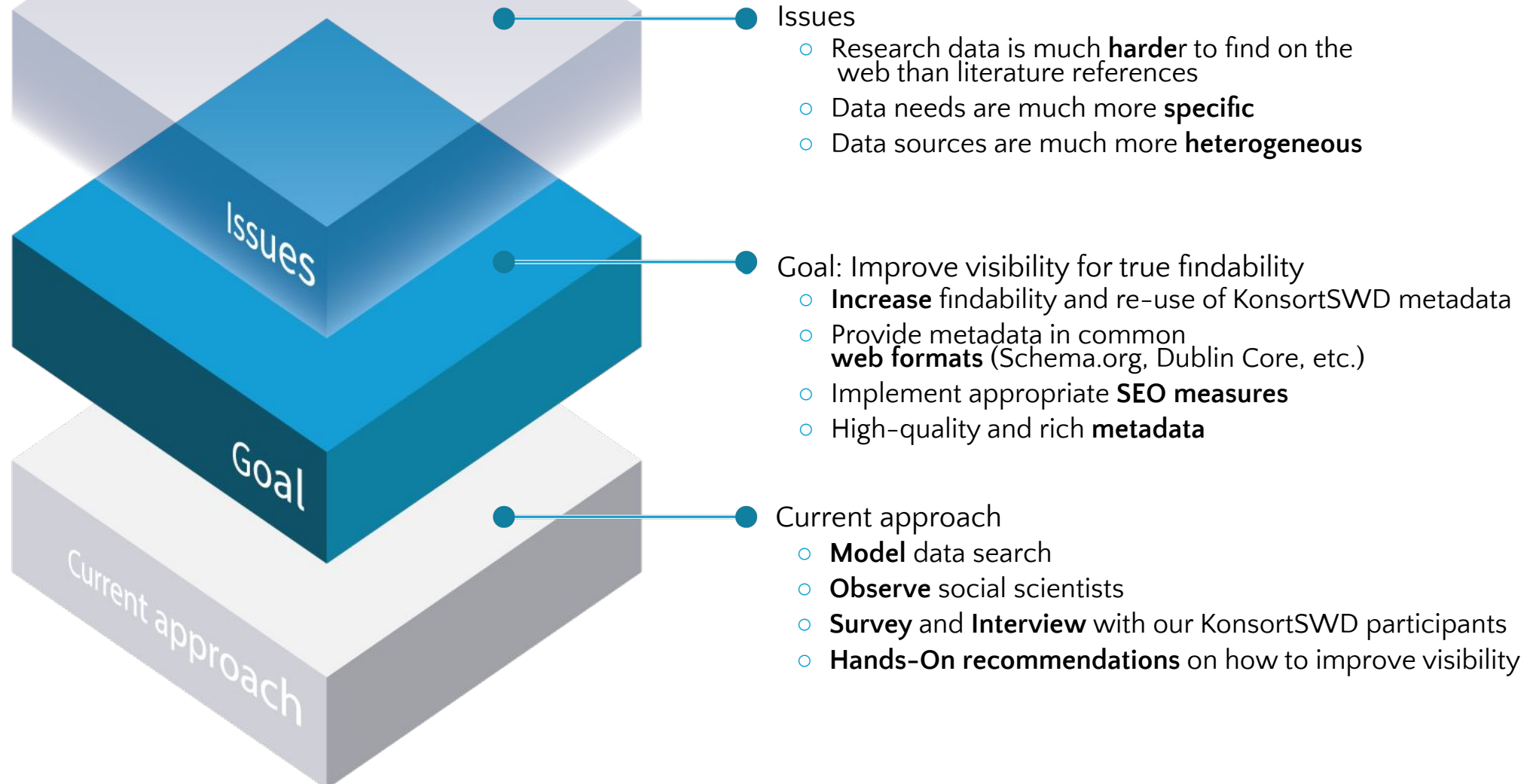
Gute Daten.
Bessere Forschung!

Datenangebote besser auffindbar machen
TA.5-M.2

Brigitte Mathiak (GESIS)
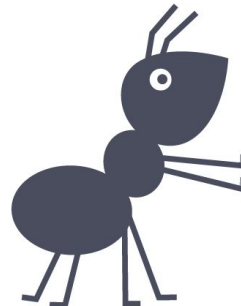
A brief introduction to findability

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

Goal: Improve visibility for true findability
- ○ **Increase** findability and re-use of KonsortSWD metadata
- ○ Provide metadata in common **web formats** (Schema.org, Dublin Core, etc.)
- ○ Implement appropriate **SEO measures**
- ○ High-quality and rich **metadata**

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

Issues

Goal

Current approach

What is our discovery model?

**Data discovery is a complex process [1]**
- **75%** of researchers often rely on literature review
- **59%** of researchers rely on search engines
- **41%** use domain data repositories

We 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

Future work
- How to support data discovery within the data repository?
- How can we make sure relevant tools and resources are found as well?
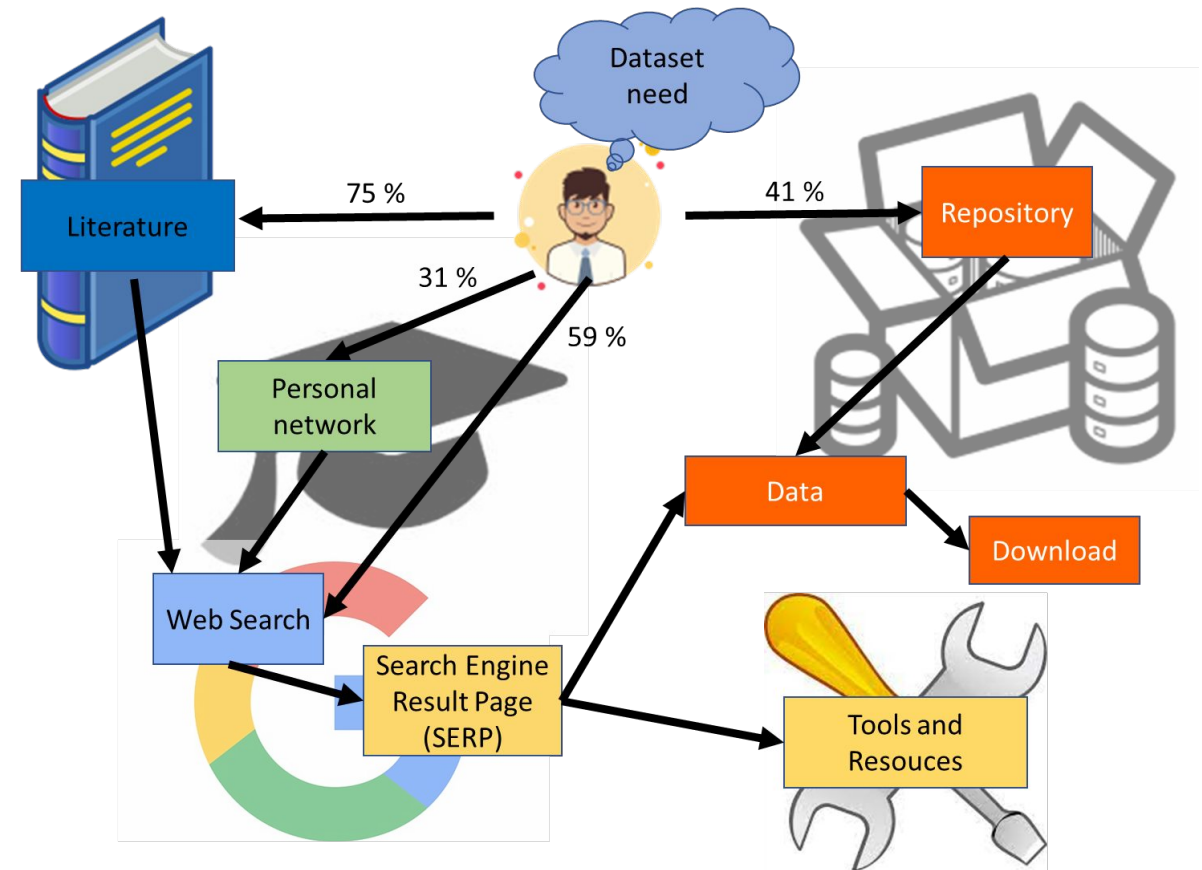
The percentages are taken from [1] and denote percentages of users who use this method "Often" in dataset search, rather than "Occasionally" or "Never".

[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

Observation study

"In the context of your research you need research data.
For today, you decide to start with the search for research data."
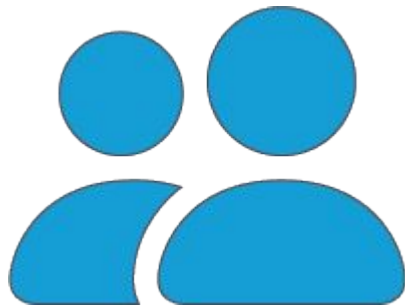
We observed **12** social scientists searching for research data in their natural office environment
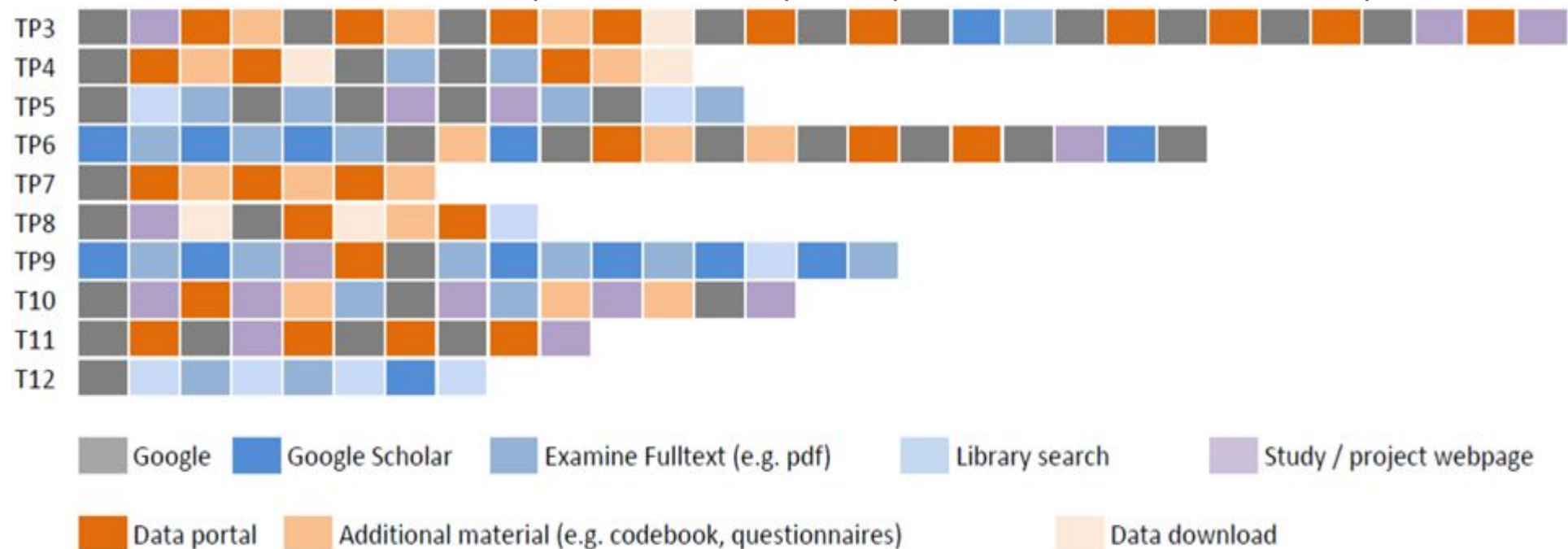
We recorded the screencasts and what they would explain about their process

Semi-structured interviews after the experiment

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.

The Data Findability Survey

Survey Design

- The survey ran from to
  Apr 15 – Apr 30 2021

- The survey was sent to **54** participants of KonsortSWD

- **42%** (n = 23) response rate

- **17** questions, organized in **6** sections

Initial results

- KonsortSWD is seen as potential support to their data findability efforts (**> 95%**)

- **88 %** of the respondents report a data repository/Website for their data

The Data Findability Survey: Results

Participants take great effort to improve findability

- Variety of **measures** are being taken, e.g. Open metadata, rich documentation and use of standards

- Participants are **highly interested** in improving data findability

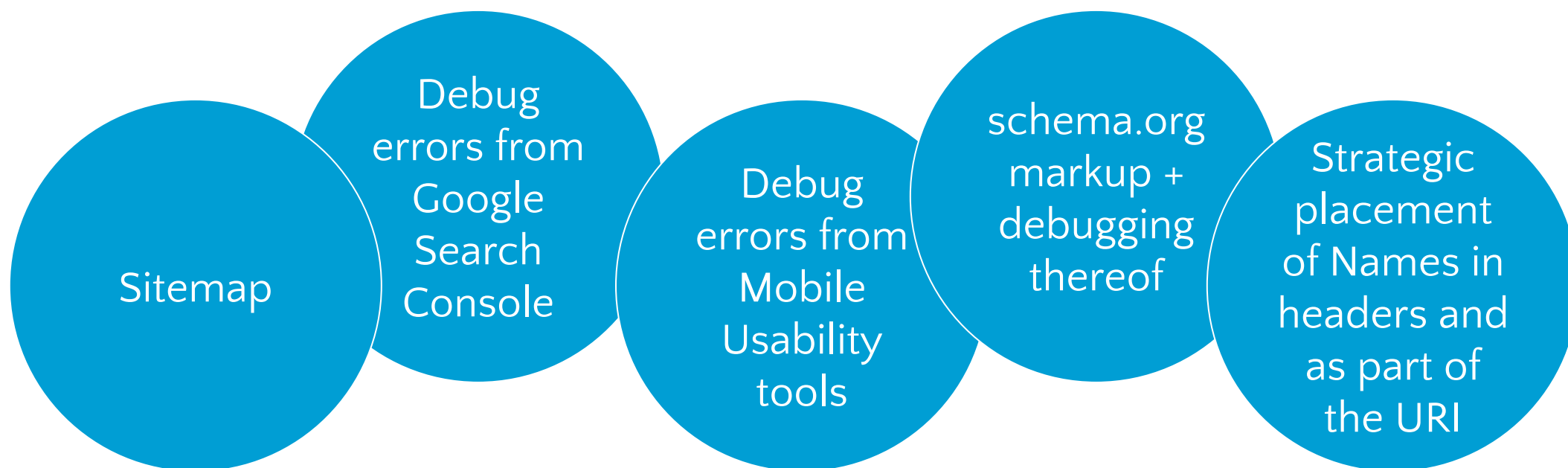- There is a **gap** in User tracking and other SEO practices that would improve web visibility

Data findability measures

Interested in improving data findability?
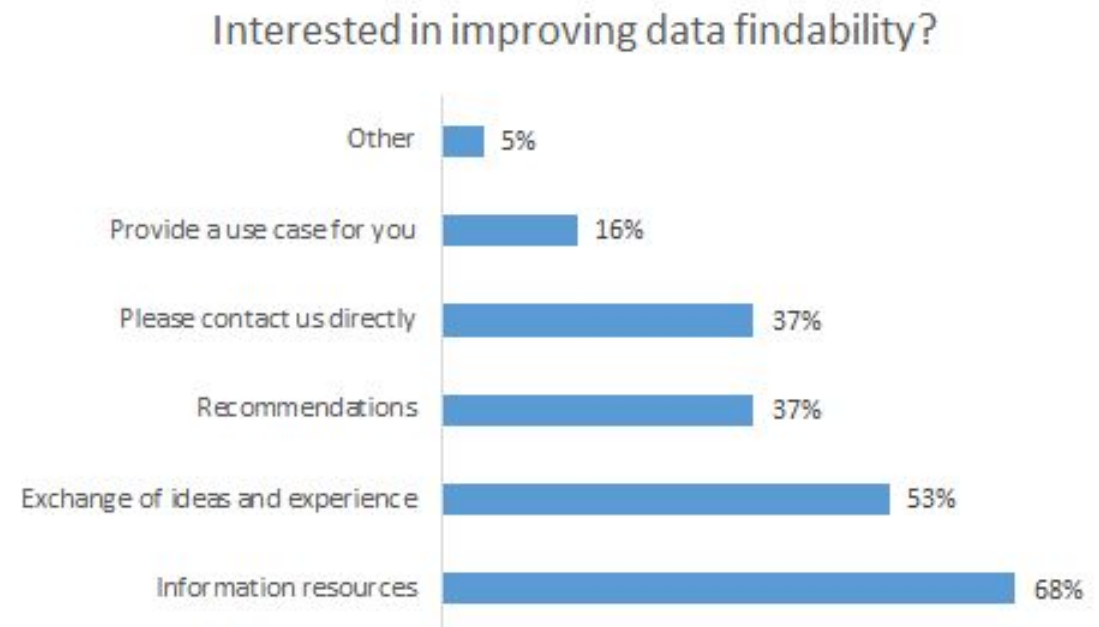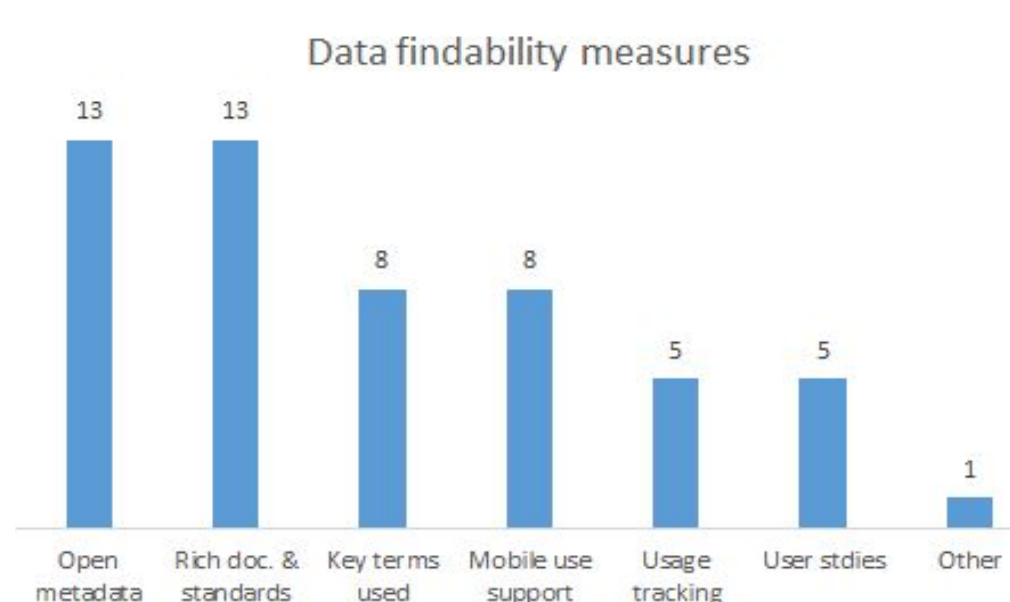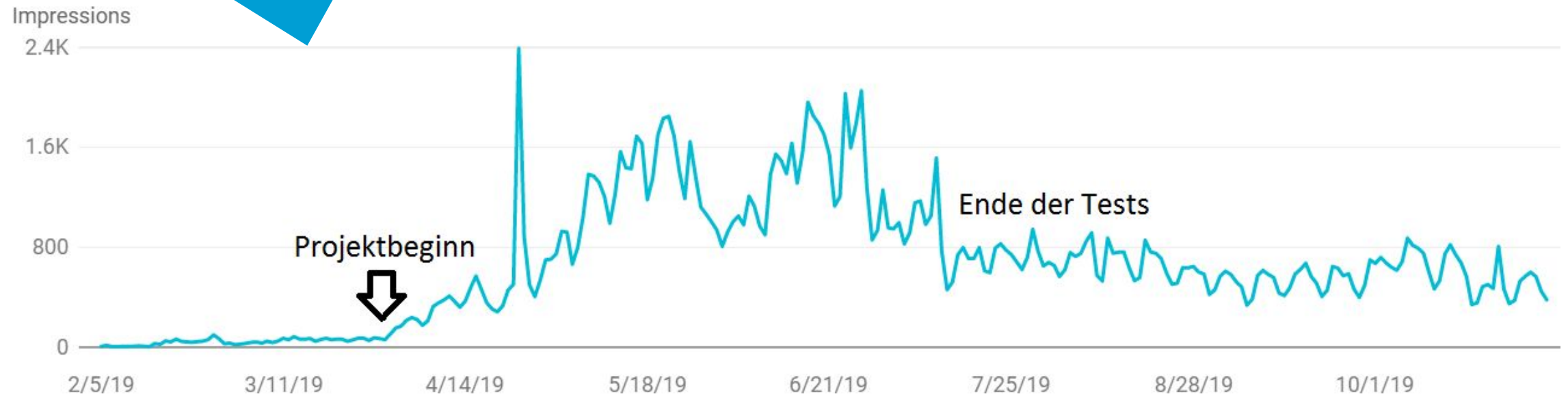
Impressions measures how often a web page was displayed by gesisDataSearch as a search result by Google on the respective day.
(Source: Google Search Console)

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

If you didn't check, if it worked, it probably **didn't.**

Log analysis – Which Keywords to use?

❑   Keywords you should use for dataset landing pages [1]

     ❑   "Dataset", Datensatz, Datenbank

     ❑   Specific to Social Science: Question, Questionaire, Skala (if applicable)

❑   Make sure you not only have the name, but also the acronym, as well as informal names in your metadata

❑   All relevant metadata for findability should be in title and description, as these are the only text fields to be indexed by everyone

published in [1] Younes, Yousef, Brigitte Mathiak, and Fidan Limani. 2021. "Keywords for Data Discovery." Open Science Fair Lightning Talk.

Tangent: Why should I bother?

If my metadata is in the relevant central indices, then it should be findable, right?

Large metadata collections for research data, such as DataCite, OpenAIRE, Google Dataset Search are not fully searchable through web search, currently

Surveys show that they are not well known among our target group [1,2,3]

For all research data, we tested, the original DOI landing page was ranked highest by far

This is in stark contrast to literature, where Google Scholar and other assorted metadata indices rank highest

The last slide... wrap up.
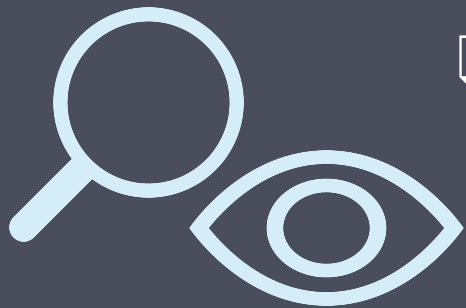
- ❏ A large proportion of users find data through web search
  - ❏ Therefore, it is a good idea to **improve web** visibility
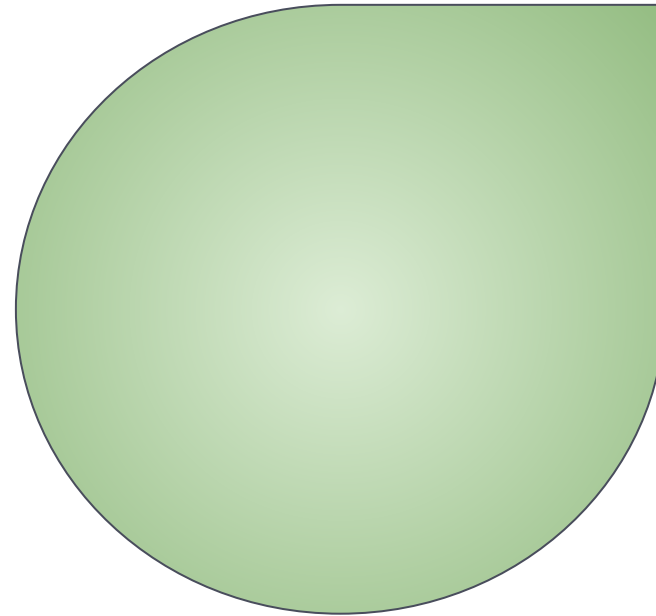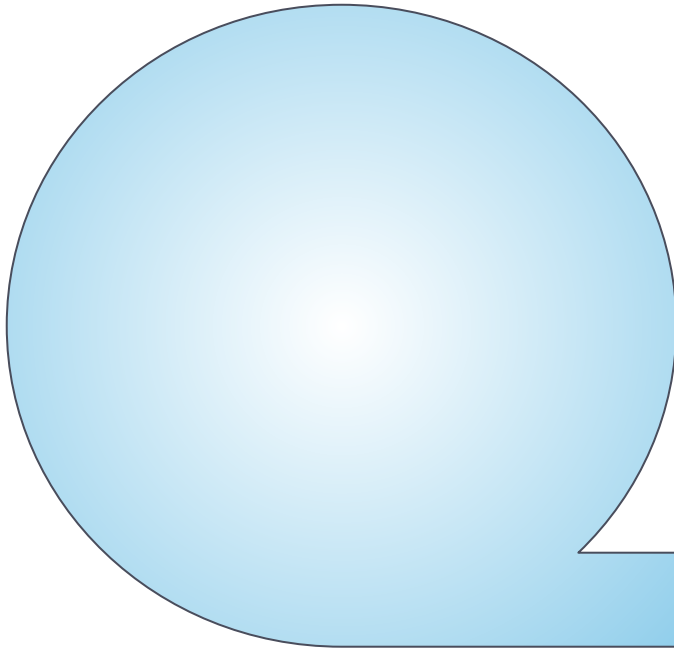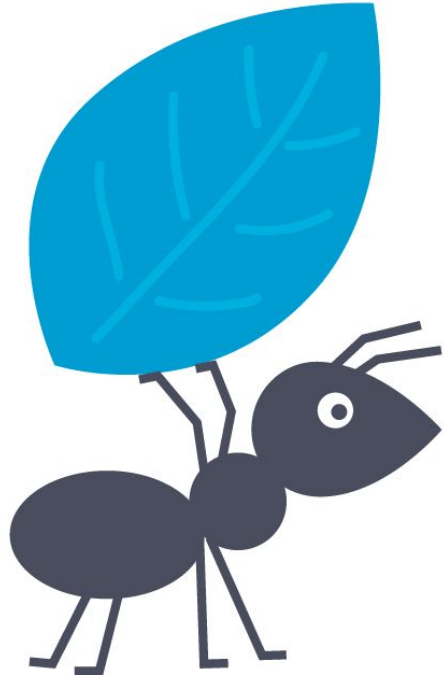 - ❏ KonsortSWD provides **guidelines** for that

- ❏ Other important factors are **literature** and word of **mouth**
  - ❏ Therefore, please make sure you are represented there

- ❏ We are looking for data donations to help us better understand web queries for research data.
  - ❏ If you have access to such data, please contact me! (brigitte.mathiak@gesis.org)

Discussion

Thank you!