

# KonsortSWD Measure 5.2: Enhancing data findability.

## Milestones 4 and 5 report.

Fidan Limani<sup>1</sup>, Janete Saldanha Bach<sup>2</sup>, Brigitte Mathiak<sup>2</sup>

August 2023

### Abstract (in German)

Dieser Bericht beschreibt Empfehlungen zur Verbesserung der Auffindbarkeit von Daten für Forschungsdatenzentren (FDZ) innerhalb der KonsortSWD-Community. Er vergleicht die Ergebnisse der ersten (2021) und der zweiten (2022) Erhebungswelle unter den 44 Teilnehmer\*innen der 34 KonsortSWD-FDZs. Wir haben diese Ergebnisse ausgewertet, um die in der Community angewandten Praktiken der Datenauffindbarkeit zu verstehen. Die quantitative Methodik zur Erhebung der Datenauffindbarkeitspraktiken in den KonsortSWD-FDZ ist ein wesentlicher Bestandteil unserer Gesamtstrategie. Sie verfolgt eventuelle Veränderungen zwischen diesen beiden Erhebungen, indem sie die von den FDZ übernommenen oder verbesserten Tools und Dienste detailliert aufgelistet. Darauf folgen Abschnitte zur Diskussion dieser Ergebnisse und Testimonials der FDZ zu den von uns durchgeführten Beratungen. Darüber hinaus werden im Bericht die angewendeten Maßnahmen zur Datenauffindbarkeit und Monitoring-Tools beschrieben, insbesondere das SEO-Monitoring mit dem Sistrix-Tool, was uns ein besseres Verständnis für die Trends der Datenauffindbarkeit in der Community verschafft. Tools und Ressourcen für die Auffindbarkeit von Daten sind von entscheidender Bedeutung für die Nutzung der wertvollen Datensätze von KonsortSWD, die zwar zur Verfügung stehen, aber noch wenig genutzt werden. Um dies zu verbessern, werden Empfehlungen ausgesprochen, die ein Monitoring der Änderungen innerhalb der FDZ und die Verfolgung der implementierten Instrumente zur Messung der Auswirkungen der Datensichtbarkeit ermöglichen. Um eine schnelle Auffindbarkeit von Daten und eine einfache Zugänglichkeit zu ermöglichen, haben alle die FDZ das Potential, die Auffindbarkeit ihrer Daten zu verbessern. Von einfachen Techniken wie Sitemaps und SEO-Verbesserungen bis hin zu übergeordneten Diensten und Maßnahmen zur Auffindbarkeit hängt die Auffindbarkeit auch von Metadaten ab, die von den FDZ bereitgestellt werden. Nur so können Suchmaschinen ihre Suchergebnisse besser auswerten und einordnen. Wir stellen fest, dass mit der Einführung dieser Maßnahmen viele FDZ an den Rand ihrer technischen Möglichkeiten (Infrastruktur/Personal) kommen und skizzieren daher für die nächste Projektphase Maßnahmen vor, die keine technischen Ressourcen benötigen.

---

<sup>1</sup> ZBW – Leibniz Information Centre for Economics, Germany

<sup>2</sup> GESIS - Leibniz-Institute for the Social Sciences, Germany

## Abstract

This report describes the data findability improvement strategy recommendations for Research Data Centres (RDCs) across the KonsortSWD community. It compares results between the first (2021) and the second (2022) survey waves applied to the 44 participants within the 34 KonsortSWD RDCs. We assessed these results towards understanding the data findability practices adopted in these communities. The quantitative technique for surveying the data findability practices of the KonsortSWD repositories is an essential part of our methodology. It tracks any eventual changes between these two surveys, detailing the tools and services adopted or improved by RDCs, followed by the discussion and RDCs testimonials on the consultations we have provided. Moreover, the technique highlights the data findability practices and monitoring tools, describing the SEO monitoring with the Sistrix application, and providing a better understanding of the community's data findability trends. Data findability tools and resources are crucial to leveraging valuable datasets that are largely available but still under-exploited. As a result, it provides recommendations allowing change monitoring regarding the repository enhancements and tracking implemented instruments used to measure the impact of the data visibility. To enable quick data discoverability and easy accessibility, RDCs shall urgently improve their data findability practices. From simple techniques, such as sitemaps and SEO improvements, to higher-level discovery services and findability measures, RDCs shall also provide rich and structured metadata to enable search engines to harvest and rank their search results better. We find that doing so puts a strain on the technical resources available to many of the RDCs and we therefore outline more centralised measures to be undertaken in the next project phase.

### Keywords:

Research Data findability - impact measurement. Social Sciences Research data visibility. Search Engine Optimization (SEO). Metadata. Metadata standards.

---

<b>Project title</b>	Consortium for the Social, Behavioural, Educational, and Economic Sciences (KonsortSWD)
<b>Task Area 5</b>	Technical solutions
<b>Measure 2</b>	Enhancing data findability
Milestone 4 deliverables	Report
<b>Milestones 4 and 5</b>	[M4] Implementation of the strategy in selected KonsortSWD repositories, in particular by applying schema.org broadly, annotating KonsortSWD webpages with SEO-relevant keywords, implementing dedicated landing pages for digital KonsortSWD objects (Repository enhancements). [M5] Implementation of monitoring instruments to measure the impact of the visibility instruments (Tool).
<b>Authors</b>	Fidan Limani Janete Saldanha Bach Brigitte Mathiak
<b>Reviewed by</b>	Peter Mutschke (incl. textual contributions)
<b>Date</b>	27 December, 2022

---

Executive summary .....	5
<b>1. Introduction.....</b>	<b>7</b>
<b>2. Survey .....</b>	<b>7</b>
2.1 Survey design .....	7
2.3 Discussion.....	12
2.3.1 Testimonials.....	13
<b>3. SEO monitoring: Sistrix application.....</b>	<b>15</b>
3.1 Leibniz-Institute for Research and Information in Education – DIPF .....	16
3.2 Leibniz Institute for Economic Research - RWI.....	17
3.3 Leibniz Information Center for Economics - ZBW .....	18
3.4 German Center for Higher Education Research and Science Research - DZHW .....	18
3.5 Leibniz Institute for the German Language - IDS .....	19
3.6 Institute for Quality Development in Education - IQB .....	20
3.7 Summary of the Sistrix application .....	21
<b>4 Recommendations and future work .....</b>	<b>22</b>
4.1 Recommendations.....	22
4.2 Future work.....	23
<b>References .....</b>	<b>25</b>
<b>Appendix.....</b>	<b>26</b>
<b>Testimonials .....</b>	<b>31</b>
<i>DZHW</i> .....	31
<i>DIPF</i> .....	33

## Executive summary

This executive summary offers a comprehensive overview of the state of data findability practices within the KonsortSWD community. Drawing from ongoing and various feedback mechanisms (surveys, interviews, workshops, tool showcases, etc.), we have identified prevailing trends, challenges, and areas of opportunity for the community. The findings encompass the broader shift towards data findability, the role of SEO tools in facilitating such a shift, and an in-depth analysis of RDC domains through the Sistrix SEO tool.

We detected trends and changes in the KonsortSWD community based on our adopted methodology regarding how data findability practices have evolved. This summary provides stakeholders with critical insights to guide future strategies and interventions.

**Support Requirement for RDCs on Data Findability:** Survey results expressed that most respondents wanted support for Data Findability, consistent with the first two surveys. Most RDCs (81%) are keen on receiving support for data findability, although a slight decline in this interest was observed in the second survey.

**RDC Websites Dedicated to Data:** Another observed trend is the increase in RDCs' websites dedicated to data. The presence of such websites increased from 88% to 94% among organisations participating in both surveys. Conversely, the absence of such infrastructure decreased from 12% in the initial study to 6% in the subsequent one.

With a marked increase in dedicated websites and infrastructure for data, it is evident that organisations are recognizing the value of making data accessible and discoverable. However, the persistent demand for guidance indicates that there is much work to be done to offer tools, resources, and education on this topic.

**Sitemaps:** While awareness of sitemaps increased, the number of organisations not adopting them also rose. This is one of those cases that was hard to interpret based on the gathered feedback from the community.

**SEO Tools' adoption is increasing, highlighting their importance for data findability:** Their usage rose from 29% to 53%, whereas the cases of not using them declined from 42% to 21%. Many participants remain unfamiliar with these tools, while some institutions use multiple ones.

**RDC Presence in External Services is concentrated** in DataCite, FDZ Datensuche, and BASE as the top 3 services listing KonsortSWD RDCs data collections.

**Data Findability Measures:** Top measures include providing open metadata (69%), rich data documentation (47%), and linking to RDC sites (41%). Only some respondents (2 out of 17) felt data findability was not a priority.

**Resources allocated to Data Findability Improvement** still need to be higher; only 29% mentioned they had resources, while 41% of participants lacked dedicated resources. As an immediate solution, 18% incorporated such tasks within existing job roles.

**Collaboration on Data Findability:** Continuous collaboration and feedback have led to improvements in data findability strategies for several RDCs, and there is a general willingness from RDCs to collaborate on improving data findability. The main interests for such collaborations were information updates, exchanging ideas, and receiving advice.

**Testimonials from RDCs showed** that the impact of work is regularly monitored throughout project lifetimes, with consistent feedback from the KonsortSWD (RDC) community and highlighted the challenges and outcomes related to data findability. Their highlights include:

- **DZHW Testimonials show satisfaction** with the recommendations provided, leading to increased findability of their data collections. Data sets came with extensive metadata and were available through the RDC's search portal and other portals. Yet, they needed help improving data findability on popular search engines like Google and Bing
- **DIPF Testimonials evidence** their primary concern – that of enhancing their SEO strategy. They valued the opportunity to discuss with us for a second opinion on the SEO strategy they had in place and emphasised the need to resolve the disparity between data aggregator search engines and individual RDC websites.

**Monitoring RDC practices via an SEO tool** presents an evaluation of different RDCs using the Sistrix SEO monitoring tool to enhance data findability. Some of the important aspects to consider such a tool include:

- Sistrix was the chosen tool for monitoring due to its consistency in assessing changes resulting from data findability practices by RDCs;
- Instead of using RDC-specific domains for analysis, often higher-level institution domains were chosen to get a broader spectrum of SEO analysis, especially if the original domain did not generate significant metrics;
- Sistrix offers rich information for monitoring the SEO aspects of RDC domains, enhancing the understanding of a domain and potential areas of improvement;
- Low search volume domains result in minimal SEO insights. Updates in Sistrix features can make direct comparisons between analyses challenging;
- Despite the challenges, it is worth trying SEO tools for a powerful approach to data findability.

## 1. Introduction

Data findability tools and resources are crucial to leveraging valuable datasets that are largely available but still under-exploited. Research Data Centres (RDCs) shall urgently improve their data findability practices, from a minimal suite of techniques, such as sitemaps and SEO improvements, to higher-level discovery services and findability measures. Those practices enable their data to be quickly discoverable, easily found and understood, and accessed, allowing further analyses. In order to boost data findability, the related metadata plays an essential role since search engines harvest metadata standards and schemas, seeking rich and structured metadata to better rank the search results.

This report describes recommendations towards repositories for data findability improvement, considering two survey wave findings and feedback from the KonsortSWD RDCs. It highlights the data findability practices of the target community between the first and the second survey, allowing change monitoring regarding the repository enhancements and tracking implemented instruments used to measure the impact of the data visibility.

This report is organised as follows: section two describes the survey design and compares results between the first and second waves, detailing the tools and services adopted or improved by RDCs, followed by the discussion and some RDCs testimonials. Section three describes the SEO monitoring with the Sistrix application. Based on this, section four provides the recommendations, followed by further steps.

## 2. Survey

The quantitative technique for surveying the data findability practices of the KonsortSWD selected repositories is an important part of our methodology. As with the first survey, the goal is to elicit feedback from the broader community. Already present in the last report ([Limani et al., 2022](#)), we relied on the survey as our technique of choice for the second time. In this part we focus on the survey findings and its implications.

### 2.1 Survey design

After the interactions with the community - to understand and provide recommendations to them on improving their data findability practices - we wanted to track any eventual changes in between these two surveys. In order to track changes or trends in the community, we have kept the survey questionnaire<sup>3</sup> the same as the previous one. Thus, please refer to the Appendix for more questionnaire design details.

The survey ran from May 19 to June 30, 2022, and it was sent to 44 participants across the KonsortSWD community (consisting of 34 RDCs), with the possibility that more than one

---

<sup>3</sup> See the details in the Appendix.

participant per RDC can participate. The survey response rate was 47% (N = 21) and represents 50% (17 out of 34) of the RDCs, which we assessed as satisfactory as an initial step towards understanding the data findability practices adopted in these communities.

## 2.2 Survey results

One of the key motivations for having more than one survey on this topic was that of potentially detecting trends and changes in the KonsortSWD community based on the feedback, meetups, and the communication we have had since the first survey.

In this part, we analyse some of the results to highlight those that represent interesting trends from the second survey. To do so with a certain comparative relevance, in addition to the analysis of all respondents for this survey, we considered the respondents from the same organisations that also participated in the first survey. In this way, we could monitor any changes in data findability practices for these organisations.

The aspects we treat in this analysis include the (a) use of sitemaps and (b) SEO tools, (c) research data from KonsortSWD communities in discovery services, (d) adoption of data findability measures, and (e) resources designated for data findability improvement. Following is the discussion of the results.

Before we start with the selected aspects as was the case with the first survey, we wanted to know if and what type of help RDCs need when it comes to data findability. The majority of the respondents (81%) expressed the need for such a support, whereas a smaller group (19%) seemed comfortable with the topic. Although the number of participants that do not need support has slightly increased, this is still in line with the outcome from the first survey (see Fig. 1).

We observed a relative increase of RDCs' websites dedicated to data from survey 1 to survey 2; we also see this when comparing participants from the same institutions in both surveys. For the latter case, the presence of RDCs/websites grew from 88% to 94%, whereas the lack of such an infrastructure decreased from 12% to 6%.



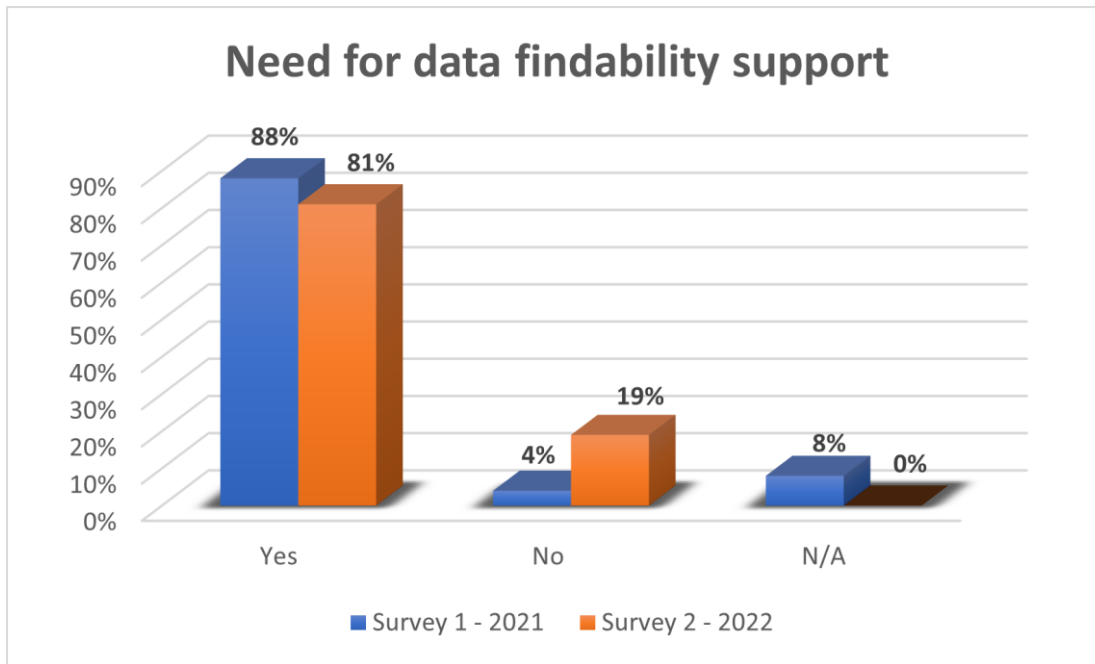


Figure 1. Need for data findability support.

**(a) Sitemaps** One thing we can infer from the adoption of sitemaps for the organisations that rely on websites to provide access to the data collections is that the number of participants not being aware of or being unsure of the use of sitemaps has decreased, whereas the number of those not adopting them has increased. Fig. 2 shows the difference in feedback between the two surveys we conducted.

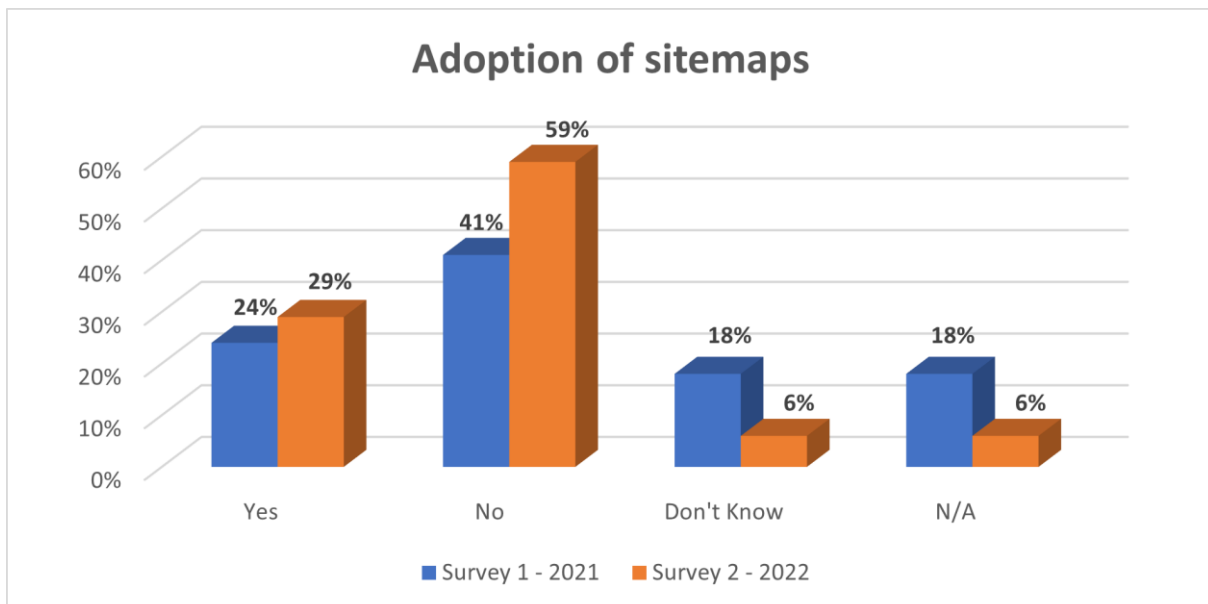


Figure 2. Adoption of sitemaps

**(b) SEO tools** These tools continue to be part of the RDC practices. Compared to the last survey, we see an increase in participants that use these tools (from 29% to 53%) (see Fig. 3); this is also reflected in the decrease in those that do not use such tools (from 42% to 21%). However, we still see feedback that indicates that participants are not familiar with such tools. Namely, there are those that do not know about the adoption of such tools at their RDC (12% in the first, 18% in the second survey), or that did not provide any answer to this question (18% in the first, and 6% in the second survey).

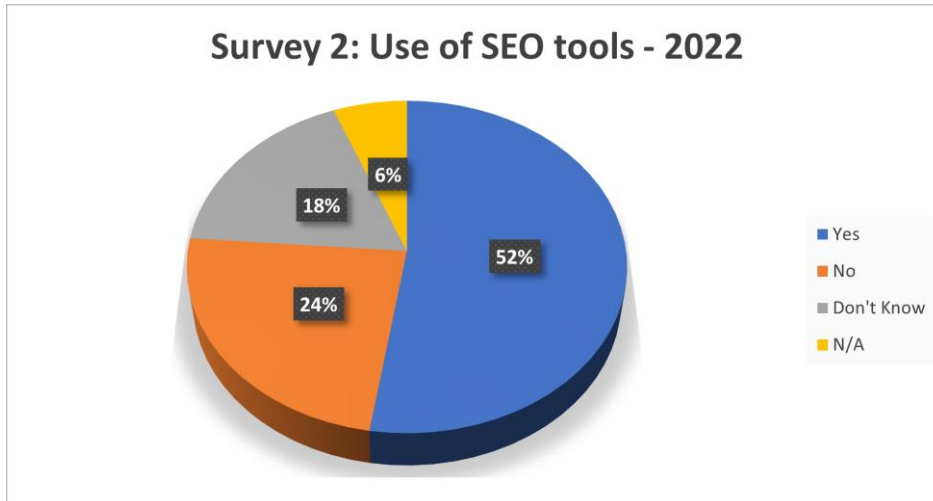


Figure 3. Use of SEO tools

It is important to note that some institutions reported using more than one SEO tool. For example, in the first survey, 5 participants report using 7 SEO tools, whereas in the second, 9 participants report 13 such tools (see Fig. 4). While the reasons for this can be manifold (different functionality needed not available in only one tool; usage of free tools alongside commercial ones, and so on), this is a promising finding, which ultimately should provide better SEO monitoring of a website.

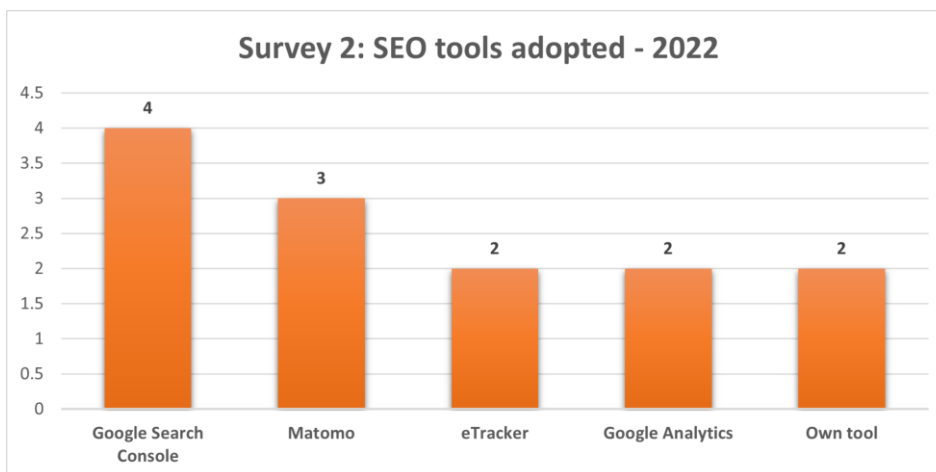


Figure 4. SEO tools adopted.

The information from the second survey about the type of sitemaps the participants adopt in their institutions, ranges from machine-readable directions only, to practices checked by Google Search Console, or relying on (the human readable) sitemaps to navigate a website. There are other approaches, such as limiting the information of the website hierarchy to the two top levels,

using the homepage to navigate to the RDC database, or using a menu in the header area to navigate to the data page.

**(c) Research data from KonsortSWD communities in discovery services** Researchers are not limited to use the RDC website or the website of an institution to access their data collections. Aggregators and other (data-specific) services also harvest, curate, and index such collections from RDCs. As such, in both surveys, we have been interested in seeing how visible/findable the data collections of the RDCs are as opposed to the same data present in other services that harvest it. While the focus was to test if they could improve their data findability practices (for cases where the dataset of an RDC is ranked below the same dataset harvested by another service, for example), we were also able to have a look in the most common services that also publish the collections of the KonsortSWD community. Fig. 5 lists these services; one can notice that DataCite<sup>4</sup>, FDZ Datensuche<sup>5</sup> and BASE<sup>6</sup> are the top 3 services that list the data collections of the KonsortSWD RDCs.

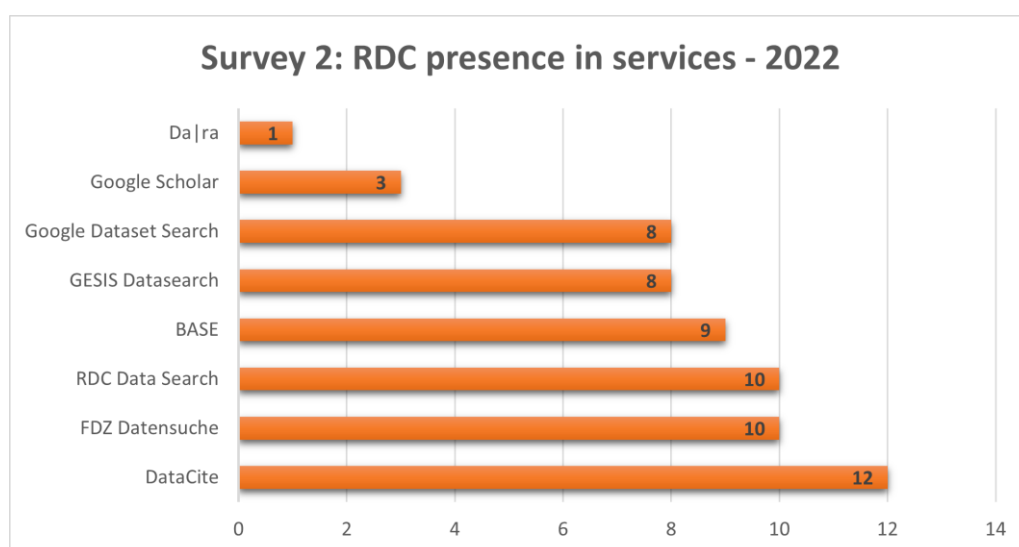


Figure 5. RDC presence in services

**(d) Data findability measures** We surveyed the measures that RDCs adopt to address or improve data findability for their data collections. The survey data show us that there is a variety of measures adopted, with the top 3 including “providing open metadata for their datasets” (available for search and harvest by other parties) for 69% of the participants; rich documentation of the data according to common standards in 47% of the cases; and encouraging/requesting from others to link to the RDC sites in 41% of the cases. Fig. 6 contains the rest of the measures different RDC participants reported in the survey. While the variety of the measures shows a certain dedication, we want to also point out that to some of the participants, data findability is currently not their (institution’s) priority (in 2 out of the 17 cases), including 3 participants that did not provide any feedback for this question.

<sup>4</sup> <https://search.datacite.org/>

<sup>5</sup> <https://www.fdz-bildung.de/datenarchiv.php>

<sup>6</sup> <https://www.base-search.net/>

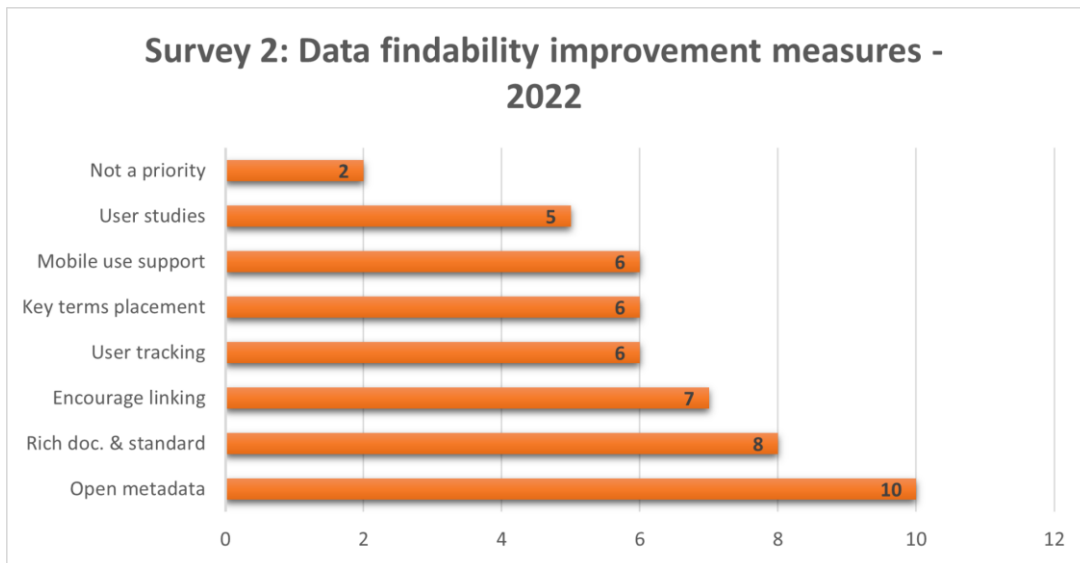


Figure 6. Data findability improvement measures 2022

**(e) Resources designated for data findability improvement** in the last part of the survey analysis, we focus on the allocated resources the RDCs have for their data findability, and their interest in collaborating with us – via the activities we plan in this measure – as means to address or improve their strategy.

As far as allocated resources go, the biggest group (41%) does not have resources dedicated to data findability, whereas 29% do. Instead of dedicated resources, 18% of the participants report that such tasks are carried out alongside the current job description of an employee. Finally, 12% of the participants did not provide an answer in this survey.

Lastly, there is a general willingness from participants from RDCs to collaborate with us on data findability aspects. There are several preferences to do so; the top 3 most popular ones include “being interested in getting information and updates about the topic”, “interested to exchange ideas and experiences”, and “advising on the steps individual RDCs can take”.

In a final note about the collaboration with RDCs, we would like to point out that we have been in continuous contact with the RDC representatives via different channels (interviews, meetups, email, survey, and so on), and this is shown in this analysis, too. One of the participants noted that, in addition to being interested in exchanging ideas and experiences, found the advice received from such discussions very useful.

## 2.3 Discussion

The majority of the RDCs are interested in getting support about data findability practices, although we do see a slight decreasing trend about it which can be challenging to interpret rightly given the gathered data via the survey.

SEO Tools remain an interesting part of the survey. The analysis shows that more and more participants/RDCs are adopting SEO tools as shown by the increased number of those who do and the decreased number of those who do not use SEO tools. There are still participants that are either not familiar with or do not know if such tools are in use for their RDC. We see this as room for improvement, of course, as SEO tools can play a significant role for data findability.

When it comes to services external to RDCs that harvest, index, or otherwise work with the collections in the RDCs, there are no changes since the first survey (DataCite, FDZ Datensuche and BASE remain as the top 3 services).

In addition to the willingness of participants across RDCs to continue their collaboration with us, we are starting to see cases where they are adopting the suggestions we are providing along the way and seeing improvements in their data findability strategies.

We would like to make a final note on the interpretation of the survey results. Namely, due to a number of reasons (staff changes at the institutions, their availability at the time of conducting the survey, technical changes at the RDCs or dataset websites, etc.), the feedback from participants from the same institution changes between the surveys. We believe that factors such as not having the same participants for all of our surveys, or different levels of knowledge for the different data findability aspects, and so on, to be a challenge to interpreting the survey results. For example, when it comes to the adoption of sitemaps, the awareness about them has increased with participants, but so has the number of those not adopting them, which conveys a conflicting message. Similarly, as pointed out in the beginning of this subsection, the decreasing trend for data findability support cannot be interpreted in a single way.

### *2.3.1 Testimonials*

Assessing the impact of our work in this measure is an important aspect that we monitor throughout the project lifetime. We continuously interact with and receive feedback from the KonsortSWD (RDC) community, such as via surveys, interviews, direct meetings, and so on. Testimonials are another such example that some RDCs chose as an alternative. In this way, they express their feedback about the data findability challenges they faced, as well as the results from adopting the recommendations we shared with them to increase data findability.

We received two testimonials from the following RDCs: German Center for Higher Education Research and Science Research (DZHW) and Leibniz Institute for Research and Information Education (DIPF). We next highlight some of the key points from them. Please note that the consultations differed in scope and were driven by the requirements and priorities of the institutions.

**DZHW** The team at the DZHW was satisfied with the recommendations we provided and was able to apply several of them and see an increased findability for their data collections.

The datasets provided by the DZHW are documented with rich metadata, accessible via the RDC's search portal<sup>7</sup>. Each dataset has a DOI that is registered via the dalra service<sup>8</sup>. The metadata

---

<sup>7</sup> <https://metadata.fdz.dzhw.eu>

<sup>8</sup> <https://www.da-ra.de/>

stored there are harvested by other portals, such as DataCite, BASE, VerbundDFB<sup>9</sup> and Gesis Search, which renders them available through these channels as well.

The challenge they were facing was the data findability of their collection in the more popular search engines – Google and Bing – that represent important “sources” of users searching for the datasets. The suggestions they adopted can be roughly classified as (1) adoption of popular metadata schemas, (2) change the practice around certain metadata elements, and (3) include additional information about the datasets.

- (1) Adoption of popular metadata schemas: Although all the datasets were available in a few other services in addition to the RDC, their visibility with the general search engines could be improved. In this point, they adopted the suggestion to adopt the two popular metadata standards – DublinCore<sup>10</sup> and Schema.org<sup>11</sup> – which have shown beneficial to address this challenge.
- (2) Inclusion of additional keywords as part of existing metadata: As pointed out during our meetings, when searching for datasets, users often include terms such as "dataset" and/or acronyms of specific studies present in the title. Since standard search engines such as Google primarily index only the titles of data packages, the suggestion was to add the term "dataset" and to mention the acronyms of studies in the title of data packages, if available.
- (3) Links between publications and data packages: The importance of such links stems from the fact that when searching for datasets, a large number of users usually do that as a result of reading a research paper that mentions that dataset. In such cases, any link that a dataset has to a (set of) publication increases the possibility for those users to find datasets, thus an important aspect to consider.  
Therefore, the DZHW team has put effort in improving links between datasets and publications by reflecting the requirement in their citation guidelines for external users, as well as by improving their internal processes to manage related publications.

You can find their testimonial in full in the Appendix section for testimonials.

**DIPF** The team at the DIPF hosts the Network for Research Data Education<sup>12</sup> (VerbundFDB) and the Research Data Centre for Education<sup>13</sup> (FDZ Bildung). The focus of the consultations with this partner was on their SEO strategy. Namely, they had already worked with an SEO consultant in the past and they had experience with it but wanted to get additional feedback to further their SEO goals with a team with a background on research data infrastructures.

We used the opportunity to discuss the SEO approach suggested to them, such as use of sitemaps, optimization of the information presented for a webpage that represents a datasets (for example, how to structure the `title` and `description` meta tags in the `head` element of an HTML page), explaining the role that each measure represents in the context of data findability.

The DIPF valued our feedback, as it allowed them to have a second opinion on the SEO strategy they had adopted so far. Moreover, they also stressed the importance of finding general policies that resolve the tension between data aggregator search engines and individual RDC websites.

---

<sup>9</sup> <https://www.forschungsdaten-bildung.de/>

<sup>10</sup> <https://www.dublincore.org/>

<sup>11</sup> <https://schema.org/>

<sup>12</sup> [www.forschungsdaten-bildung.de](http://www.forschungsdaten-bildung.de)

<sup>13</sup> [www.fdz-bildung.de](http://www.fdz-bildung.de)

This is especially important to the DIPF as it hosts such a data aggregator engine itself. The current state of affairs is that most of the traffic goes towards the landing pages of the individual RDCs, which makes the aggregators, in a way, superfluous. However, a shift in this landscape is highly likely, aspects that we have already discussed in Section 2.1, “Centralised or Decentralised approaches to SEO,” of our previous report (Limani et al. 2022, p. 44).

You can find their testimonial in full in the Appendix section for testimonials.

### 3. SEO monitoring: Sistrix application

SEO is an already established analysis technique in our work in this measure. Alongside techniques such as surveys or interviews, we have used them to identify issues with or growth opportunities for data findability for individual RDCs (cf. sections on this topic from Limani et al. 2022). Sistrix<sup>14</sup> is our tool of choice for such monitoring, which we have maintained throughout this measure. This provides a consistent approach and a way to monitor any changes resulting from changing data findability practices by the RDCs. Our SEO analysis includes the RDCs that we have monitored since the beginning of the project.

Sistrix allows one to measure various aspects of a domain; it is also possible to focus on other aspects, such as the host, directory, or URL. We have relied on a basic set of aspects to consider for the RDCs as we wanted to apply the same measurements across all the RDCs that discussed the topic of SEO with us.

- **Visibility Index (VI)** This is a Sistrix-specific measurement that, ultimately, provides a score for a given domain that reflects its visibility (the higher the score, the better the SEO standing for a domain), considering both the desktop and mobile cases. As such, it can be easily compared to different – even competing – domains to assess how one domain stands vis-a-vis another one. The VI calculations are out of the scope of this report<sup>15</sup>, but suffice it to say that it does this based on a rich representative set of keywords (and keyword searches), across domains from more than 30 countries, all based on the organic Google search results.
- **Keywords** This information category informs us on the role that keywords of a domain (could) play. The “organic” section provides the estimate of the average monthly traffic generated by the keywords. There are other sections within this group, such as keywords for which there are ads or keywords appearing in other features, such as news, etc.
- **Interesting rankings**<sup>16</sup> While there are potentially hundreds of keywords considered for the SEO analysis of a domain, certain keywords drive more traffic or are better positioned (by the number of clicks, for example).
- **Domain overview**<sup>17</sup> For this, we consider the following information: *Top countries*, *Top-10 rankings*, and *Number of URLs*.

We conducted several SEO analyses for the RDCs. Depending on the meetups or planned discussions, we have conducted a different set of Sistrix-based SEO analyses. We would like to

---

<sup>14</sup> <https://www.sistrix.de/>

<sup>15</sup> See more details at: <https://www.sistrix.com/support/sistrix-visibility-index-explanation-background-and-calculation/>

<sup>16</sup> See more details at: <https://de.sistrix.com/info/video/video/517085351>

<sup>17</sup> See more at <https://www.sistrix.com/tutorials/domain-overview/#Domain-Overview>

note that SEO tools are important in improving data findability. Such tools usually have different features, not always easily comparable across tools. This, however, is not important to our discussion of the data findability in this report, and that is why we try to capture the rationale for adopting any tool and specific features in the community.

Finally, a general remark about selecting the corresponding domains for the RDCs: SEO tools analyses depend on available data (search volume). This data comes from the traffic users generate when visiting these RDCs. In order to showcase the SEO tool capabilities, we often opted to use the higher-level domain of the institution the RDC was part of, instead of using the RDC's one domain. In this way, the RDC representatives could see the broader spectrum of analysis capabilities of the SEO tool in question (Sistrix in this case). Except for one of the cases in this report (see 3.6), we relied on higher domains for the RDCs presented in this section as, at the time of SEO analyses, the SEO tool was not able to generate many metrics for the corresponding RDC domains.

### **3.1 Leibniz-Institute for Research and Information in Education – DIPF**

For the DIPF, we conducted the SEO analyses during three different time periods: June 2021, February 2022, and December 2022. The VI score has increased during this period, with an exception for the first and the second part of this year. In any case, the VI shows that the visibility of this RDC is developing in the right direction, with a relatively significant improvement from February to December of this year.

Apart from the VI, the other aspects of the SEO analysis are relatively stable, that is, we do not see big changes between the different periods the SEO analysis were conducted. For example, we still see a similar set of keywords ranked as the most interesting ones in Sistrix, and the domain maintains the highest visibility in the German-speaking countries - Austria (AT), Switzerland (CH), and Germany (DE). The order was changed over these time periods (with Austria mainly leading as the country for which the domain had the highest visibility).



DIPF SEO measurements	Jun-21	Feb-22	Oct-22	Dec-22
Domain: <a href="https://fdz-bildung.de">fdz-bildung.de</a>				
Visibility index	0,0035	0,0056	0,0035	0,0230
Keywords	1.688 (Organic)	2.283 (Organic)	1.688 (Organic)	3.353 (Global)
Keywords (Germany) <sup>18</sup>	N/A	N/A	N/A	2.666
Organic	N/A	N/A	N/A	895
Interesting rankings	höchster bildungsabschluss (ranking: 3),	N/A	lernklima (ranking: 1)	N/A
	lernklima (ranking: 1),	N/A	kognitive aktivierung (ranking: 10)	N/A
	kognitive aktivierung (ranking: 10);	N/A	umsetzung englisch (ranking: 17)	N/A
Domain overview	AT, CH, DE	CH, AT, DE	AT, CH, DE	AT, CH, DE
Top-10 rankings	125	178	125	207
Number of URLs	538	619	538	824

Table 1. Four SEO points of measurement for the DIPF domain

### 3.2 Leibniz Institute for Economic Research - RWI

The SEO Analysis for RWI's domain shows a little fluctuation between the two data points - that of June 2021 and the recent one in December 2022. The overall VI score shows a relative decline, and the interesting rankings show different keywords for these two data points. Moreover, the domain overview conveys a difference between the number of keywords ranked on the first page of the SERP results, and that of URLs for which there is a ranking in the SERP results (see the listing below). The visibility of this domain remains present mainly in the German-speaking countries, which is Germany, Austria, and Switzerland.

<sup>18</sup> The keywords information has changed since the last check we did with the RDCs and now shows the number of keywords that can be measured both across (global) and in a country of choice (Germany, in our case). In this report, we list the former as researchers outside of Germany also access these RDCs.

RWI SEO measurements	Jun-21	Dec-22
Domain: <a href="https://www.rwi-essen.de">rwi-essen.de</a>		
Visibility index	0,131	0,0257
Keywords (Global)	19.958	4.168
Keywords (Germany)	-	3.634
Interesting rankings	astrazeneca wirksamkeit (ranking: 3)	rwi (ranking: 1)
	hässliche menschen (ranking: 1)	rwi essen (ranking: 1)
	hässlich menschen (ranking: 5)	unstatistik des monats (ranking: 2)
Domain overview	DE, CH, AT	DE, AT, CH
Top-10 rankings	1.247	258
Number of URLs	3.456	606

Table 2. Two SEO points of measurement for RWI's domain

### 3.3 Leibniz Information Center for Economics - ZBW

The ZBW is not a typical RDC from the KonsortSWD community in that it contains a relatively small dataset collection (in collaboration with a partner) - the JDA portal<sup>19</sup> - supporting journal submissions in the economics domain. The SEO analysis was unable to calculate the VI (this value was 0.000), possibly due to the low traffic to this document. It is worth noting that the VI index also considers the search volume for a given term or keyword as part of its calculation approach. Thus, the JDA does not have a measurable VI index we can compare or discuss in this part.

Nonetheless, since we had a meeting with the JDA team and discussed their practices that affect the data findability outcomes, we included this portal in the list of RDCs for which we conducted this SEO exploration.

### 3.4 German Center for Higher Education Research and Science Research - DZHW

We conducted two SEO analyses for the DZHW. The VI score for these two shows a decrease for the most recent analysis. In any case, we have to note that at times the VI score can change even in the span of a day, let alone a few months.

<sup>19</sup> <https://journaldata.zbw.eu/>

Here we have once more the different Sistrix feature updates: for the first analysis shows the number of keywords rankings from the organic Google SERP (no ads, etc.), and that from other features (news, etc.). In the second case, we have the number of different keyword rankings measured across all countries vs those measured only for a specific country (Germany, in this case).

For the first analysis, we see that there is information about the domain overview, whereas this is missing in the second one due to the low search volume. This could have also impacted the VI scores between these two analyses.

DZHW SEO measurements	Jun-21	Dec-22
Domain: <a href="https://dzhw.eu">dzhw.eu</a>		
Visibility index	0,0842	0,036
Keywords (Global) (Organic)	6.400	6.267
Keywords (Global) (Other)	5	N/A
Keywords (Germany)	N/A	5.042
Interesting rankings	N/A	brief (ranking: 11)
	N/A	dzhw (ranking: 1)
	N/A	otmane azeroual (ranking: 1)
Domain overview	CH, DE, AT	N/A
Top-10 rankings	449	N/A
Number of URLs	1.903	N/A

Table 3. Two SEO points of measurement for the DZHW domain

### 3.5 Leibniz Institute for the German Language - IDS

As mentioned earlier in the section, if Sistrix is not able to provide a rich set of results for a domain, it will often suggest choosing a higher domain, with the chance that it will have more search volume, thus more opportunities for analysis. Just as a comparison for this case, from the analysis we did in June, the VI for the subdomain was 0,0114, whereas that for the parent domain was 1,308.

While the VI score changes for these two points of measurement, we can see a difference in the number of keywords rankings measured. For the first experiment in June, we see that the Sistrix feature measures the organic vs other categories for the keywords, whereas in the second experiment in December, it measures the keyword rankings measured across and in

Germany. Moreover, the number of keywords and URLs for which there was a ranking on the first page of the search results are relatively comparable for both cases.

IDS SEO measurements	Jun-21	Dec-22
Domain: <a href="https://ids-mannheim.de">ids-mannheim.de</a>		
Visibility index	1,308	1,073
Keywords (Global) (Organic)	56.631	93.744
Keywords (Global) (Other)	10	N/A
Keywords (Germany)	N/A	71.172
Interesting rankings	N/A	wort mit j (ranking: 1)
	N/A	fakultativ (ranking: 5)
	N/A	worte mit k (ranking: 1)
Domain overview	CH, AT, DE	CH, DE, AT
Top-10 rankings	9.588	10.295
Number of URLs	6.580	6.624

Table 4. Two SEO points of measurement for the IDS domain

### 3.6 Institute for Quality Development in Education - IQB

For this last Sistrix exploration, we again see comparable VI scores for the two dates when we conducted these analyses. Depending on the (RDC) domain, we were able to include different information from the SEO experiment. As you can notice from the following listings, there was not much information about the “domain overview” section. On the other hand, from the “interesting rankings” section of the SEO report that Sistrix generates, we notice how the set of (top 3) keywords have not changed much.

IQB SEO measurements	Jun-21	Dec-22
Domain: <a href="http://iqb.hu-berlin.de">iqb.hu-berlin.de</a>		
Visibility index	0,1902	0,1683
Keywords (Global) (Organic)	28.285	33.546
Keywords (Global) (Other)	7	N/A
Keywords (Germany)	N/A	27.357
Interesting rankings	iqb (ranking: 1)	iqb (ranking: 1)
	iqb aufgabenpool (ranking: 1)	iqb aufgabenpool (ranking: 1)
	roadmovie albert ... (ranking: 1)	vera (ranking: 4)

Table 5. Two SEO points of measurement for the IQB domain

### 3.7 Summary of the Sistrix application

There are benefits, but also a few limitations to SEO monitoring as represented in this section. Sistrix offers us a way to easily monitor certain SEO aspects of RDC domains, without much involvement on the RDCs. The analysis it provides is rich in information and can direct our understanding about a domain and try to identify potential improvement. We need to point out that this depends on the domain itself; as seen with few of the cases, if the search volume is low, the analysis for that domain correlates, i.e., we get minimal information from an SEO perspective.

On another note, there can be updates in how Sistrix reports the results that render the comparison challenging. For example, if a feature changed between two SEO analyses for a domain, a simple comparison of the VI of their results will not suffice. One such example is the “organic,” “ads,” and “other” categories lately changed to “global” and “country-specific” when describing the keywords of an SEO analysis.

In any case, we invite the reader to try out an SEO tool – and Sistrix is but one of the possibilities – to examine different aspects of a domain. While such a task always includes SEO tool-specific terminology and concepts, it is worth trying it as a straightforward – and powerful – approach to data findability.

User feedback remains the final limitation for this section. Often, we can observe the changes in the visibility of a domain, and, while we can make informed assumptions about them, user feedback could provide the context to confirm such assumptions.

In conclusion, we do not detect any substantial changes in terms of data findability practices as shown from the SEO analyses we conducted for a few of the KonsortSWD partners.

## 4 Recommendations and future work

The findability of research data remains a challenge. In order to locate the datasets that best suit their needs, researchers spend an excessive amount of time seeking and gathering data. Numerous valuable datasets are still not fully utilised; although they are available, it is challenging to find them (see the “Executive Summary” section for the main findings). From multimodal analysis (surveys, feedback mechanisms, and informal conversations and direct engagements), an interconnected perspective emerged: the limited metadata and lack of IT expertise resources, along with the decentralised method of the RDC, hinder widespread improvements in findability. While there is a growing awareness and adoption of data findability practices, there remains room for improvement. The feedback suggests a need for consistent education, resources, and collaboration within the community. To effectively address these challenges and needs, we have compiled a list of targeted recommendations.

### 4.1 Recommendations

Creating and maintaining interoperable, publicly accessible data catalogues with cutting-edge discovery tools is necessary to promote the responsible sharing and use of data. Some basic features should be implemented to increase data visibility, such as using elementary search engine optimization (SEO) techniques. After all, metrics are indexed by search engines, and SEO-based improvement plays a crucial role in this sense. In addition, many users begin their search by querying search engines like Google, Google Dataset Search, Bing, Yahoo!, etc. instead of exploring a specialised data source. Yet, because of poor search engine optimization, specialised data are not visible. Adopting structured metadata that search engines can crawl is a vital feature of SEO in favour of data visibility. Nevertheless, many research data centres need to gain the knowledge, tools, and incentives necessary to document their data assets properly.

High-quality metadata creation involves expertise, resources, and financial support. Frequently, one or more of these three requirements still need to be fully satisfied. However, data frequently have insufficient and inadequate metadata. Because of this, many datasets are still incompletely documented and thus challenging to discover, find, and use. To meet these goals, extensive metadata must be included with the data, including technical and contextual information that data holders and/or curators may produce manually or programmatically. Research Data Centres are financed through funding agencies, universities, and public bodies not only to store data. Appropriate data findability remains a crucial task for RCDs as long-term preservation seeks to make data findable for an extended period. In this sense, funding budgets should consistently address not only discoverability and findability tools and automated solutions, but also qualified human resources, such as metadata specialists and information technology professionals, able to use quality software development methodologies upon current technologies.

RDCs need to support a specialised metadata team because metadata must be detailed, comprehensive, and structured according to the standards and schemas to serve the multiple goals of enhancing visibility, discoverability, and accessibility to achieve its purpose, which is data usability. Metadata standards and schemas are meant to promote and allow data retrieval. By adopting and applying those standards correctly, data users can identify interesting data since the preponderance of the content retrieved by search engines is provided by metadata indexing. Accurate descriptions can lead to the user's decision-making on data usage, potentially avoiding misunderstanding and misuse of data.

Data providers also benefit from the in-depth, enriched, and precise metadata, such as multiple datasets integration, evaluating the data's quality, tracking data used and measuring its impact. It brings advantages such as interoperability, transparency, and accountability, culminating in credibility for the data holders, their services, and products. The more data is visible and findable, the more its utilisation demand increases, advancing directly on the RDC's relevance in a given domain. The efficient cataloguing should be at the core of RDCs' interest, fostering the FAIR principles through a metadata-driven research ecosystem.

In our last report ([Limani et al., 2022](#)), we identified some of the areas in which improvements could enhance the findability. For example, the demonstration of SEO tools and their application to few of the community RDCs identified aspects that could improve the findability of their data collections. This included suggestions of different engagement levels, such as the inclusion of sitemaps, mapping of dataset pages with Schema.org, or using the (free) Google Search Console, to tasks that require conducting user studies, adopting certain metadata practices to describe the dataset pages, to optimising these pages for mobile users. You can find more details about these tasks in sections 2.2 and 2.3, as well as the report recommendations of the report. With the help of our partners, we tested these measures, and they were quite successful. Despite this, the overall visibility has remained flat. Our main result has to be that knowing what to do, does not make it happen.

We conducted interviews with almost a dozen RDCs and had informal talks with many more. Those talks revealed that the main crux is that technical personal resources are scarce. Due to the distributed and decentralised nature of the way research data is currently stored, it is hard to make any fundamental changes with a single impulse. Outreach from projects such as ours on Data findability, for the most part gets adopted/considered by the larger institutions and those with more funding, which are also institutions that already are well-positioned. Conversely, smaller RDCs do not have enough (technical) resources to connect and implement the recommendations from such a project. As a result, the overall findability of data is only minimally improved.

## 4.2 Future work

These unveiled problems pushed us to rethink our approach. For the next year, we plan a workshop on data findability, not only targeting KonsortSWD, but coming together with experts from all NFDI consortia to hear about their approaches. We are also looking at a landscape that is currently changing. Metadata harvesting for research data is given a lot of attention by the NFDI, European and international levels. Knowledge graphs and services are being built on top of that. So far, these efforts have yet to reach the critical mass to become the predominant way users interact with research data. However, going by the relevant activities (see for example [Stocker et al., 2023](#) for a dedicated working group on the adoption of Knowledge Graphs across NFDI projects), it seems inevitable that eventually they will. Our new efforts will therefore go in the direction of future-proofing the metadata.

Most of the RDCs are currently registering their data at the data registration agency dalra. However, they often use a minimal set of metadata and the mapping to the schema of the registration agency is often incomplete or even error ridden. GESIS and other larger institutions are no exception to this. While this made no difference in the past, we see a rising number of activities re-using this metadata, including harvesting metadata in different ways, such as through

JSON-LD schema.org. Only providing the minimum metadata necessary, while the easiest way to do it, will eventually place social science data at a disadvantage, if not remedied.

The new approach is, therefore, to leverage the metadata and try to ideally elevate it. This includes, supporting the data providers with direct feedback of what is lacking in terms of information, warning them about the consequences of “bad (meta)data.” A final part includes providing feedback on the machine-readability of the landing pages and the metadata provided there.

This approach is inspired by activities such as those in the context of FAIRSFAR<sup>20</sup>, and the tools they are offering. Moreover, we plan to build on activities such as this to keep up to date with international developments. Finally, ongoing, and established activities, such as the survey, consultations with RDCs and monitoring via SISTRIX, will continue as we enter the last year of the measure.

---

<sup>20</sup> FAIRsFAIR “Fostering FAIR Data Practices In Europe” see <https://www.fairsfair.eu/>.



## References

Fidan Limani, Yousef Younes, Janete Saldanha Bach, Valentina Hiseni, Peter Mutschke, & Brigitte Mathiak. (2022). KonsortSWD Measure 5.2: Enhancing data findability Milestones 1, 2, and 3 report (1.1). Zenodo. <https://doi.org/10.5281/zenodo.7224672>.

Stocker, Markus, Rossenova, Lozana, Shigapov, Renat, Betancort, Noemi, Dietze, Stefan, Murphy, Bridget, Bölling, Christian, Schubotz, Moritz, & Koepler, Oliver. (2023). Knowledge Graphs – Working Group Charter (NFDI section–metadata) (1.1). Zenodo. <https://doi.org/10.5281/zenodo.7515324>

## Appendix Questionnaire

### General information

1. Name of your institution
2. If it applies: Name of the RDC
3. Is it possible for you to be supported by KonsortSWD in improving the findability of your data?
4. Does your RDC have a data repository or a website for specific data sets?

### Research data websites

5. Please insert a link to the page(s) here
6. In your assessment: Is the data set best described / documented on your website or is there a more complete description / documentation elsewhere (e.g., at da | ra)?

### Technical information

7. What kind of meta information does your site provide?
8. Do you use tools to analyse (and/or optimise) the use of your website?
9. Does your website provide a sitemap?
10. If yes, what kind of sitemap is it?

### State of Search Engine Optimization (SEO)

11. Please give an example of a popular or, from your point of view, particularly important data record on your side.
12. Please enter the data set just mentioned on Google. If the first 10 results are not relevant, please add "data set" or "download". How is the result?
13. Are you satisfied with the results of this little exercise?
14. To the best of your knowledge, at which of the following services is the above record listed?

### Existing measures

15. Is there a person in your RDC/institute who is responsible for improving the findability and / or are there resources for this?
16. Please tick which of the following options describe measures that your RDC/your institution undertakes to improve the findability of research data.

### Interested in improved findability

17. Are you interested in working with us to improve the findability of the research data of your RDC? How?

### Questionnaire description

The questionnaire contains both open- and closed-ended questions. In order to provide complete information about it, in this section we provide all the information regarding the questionnaire design.

### General information

In this section, the goal is to collect information about the participant, such as their associating institution, RDC, or a data repository, as well as the possible improvement of the data findability of their collection. The questions 1 and 2 are free-text entries, whereas for questions 3 and 4 participants could choose “yes” or “no” for an answer.

**Q1 Name of your institution**

A required, open-ended question to capture the information about the KonsortSWD partner the survey participant is associated with.

**Q2 If it applies: Name of the RDC**

Open ended question that aims to capture the information about a specific RDC. This information could also be used to indicate an RDC domain that we could explore in parts of the methodologies, such as during the interviews, SEO analysis, and so on. This is an optional question.

**Q3 Is it possible for you to be supported by KonsortSWD in improving the findability of your data?**

An open-ended question that aims to capture the importance of the “Data findability” aspects and this measure to the participants.

**Q4 Does your RDC have a data repository or a website for specific data sets?**

This question is required; the options to answer this (closed) question include:

- Yes, we have a research data repository or a website dedicated to a specific dataset.
- No, we don't
- We choose to opt out of this measure

**Research data websites**

With the questions in this section, we try to understand URLs of data of interest to the participants, as well as their assessment of how documented a certain dataset is.

**Q5 Please insert a link to the page(s) here**

An optional, open-ended question, meant to collect the URL links to the website that the participant uses to work with data sets (search for them, etc.).

**Q6 In your assessment: Is the data set best described/documented on your website or is there a more complete description/documentation elsewhere (e.g. at dalra)?**

A closed-ended, optional question that tries to capture the role a given website that hosts datasets plays. The options include:

- Users should come to us first for all datasets we host.
- We have both cases.
- We mainly provide additional material. Users should come to us, but maybe not first.
- I don't know.

**Technical information**

This section is dedicated to collect the more technical information of the data websites/repositories that the RDCs use.

**Q7 What kind of meta information does your site provide?**

This is an optional, open-ended question that focuses on the metadata elements the different communities use when describing their datasets. For this question, the user can choose more than one option when answering. The options include:

- Schema.org
- Dublin Core
- <title>
- <meta name="description" ...
- Other <meta>
- Other

**Q8 Do you use tools to analyse (and/or optimise) the use of your website?**

This is similar to the previous question, with the available options that include:

- Google Search Console
- Google Analytics
- eTracker
- Something we build ourselves
- None
- Other

**Q9 Does your website provide a sitemap?**

This is a closed-ended, optional question, with the following (mutually exclusive) options (the user can only select one value):

- Yes
- No
- I don't know

**Q10 If yes, what kind of sitemap is it?**

This question tries to capture more information about the sitemap in use through the following options (the user can choose more than 1):

- Our sitemap is verified through Google Search Console.
- Our sitemap is a complete list of all our relevant web pages, including those that show research data.
- Users can navigate through our site via the sitemap.
- Other

**State of Search Engine Optimization (SEO)**

Through questions 11 through 14, we try to understand the data findability currently offered by data search services. To a certain extent, this will provide us with the basic expectation about the data findability of the participant.

**Q11 Please give an example of a popular or, from your point of view, particularly important data record on your side.**

An optional, open-ended question, where the participant provides a dataset name s/he is familiar with.

**Q12 Please enter the data set just mentioned on Google. If the first 10 results are not relevant, please add "data set" or "download". How is the result?**

The options for this optional question are a few, and the user can select all that apply, including:

- Our website is in the Top 3.
- Our website is in the Top 10.

- The entry for our website looks weird.
- Our highest-ranked entry is not the best to have on top.
- All websites in the Top 3 pertain to that dataset.
- Our website contains a link to at least one of the websites in the Top 3 that does not belong to us.
- At least one of the websites in the Top 3 that does not belong to us has an (easy to find) link to us.
- I see a Wikipedia entry.
- I see an Infobox on the right hand side.
- I see ads.
- If I were looking for that dataset, I would be satisfied with this result list.
- Knowing the dataset, I feel like there is at least one important resource missing in the result list.
- Other

**Q13 Are you satisfied with the results of this little exercise?**

An optional, multiple choice question with the following options:

- Yes
- No
- Other (For this option, the user can provide an additional explanation)

**Q14 To the best of your knowledge, at which of the services is the above record listed?**

For this multiple choice question, also optional, the user has the following options to choose from (s/he can choose more than 1 option):

- BASE: <https://www.base-search.net/>
- Google Scholar: <https://scholar.google.com/>
- Google Dataset Search: <https://datasetsearch.research.google.com/>
- DataCite: <https://search.datacite.org/>
- RatSWD: <https://www.ratswd.de/forschungsdaten/suche>
- GESIS Datasearch: <https://datasearch.gesis.org/>
- Other (If selected, the user can provide another service not mentioned above)

### Existing measures

The results of the “Data findability” measure are meant to be adopted by RDCs that do not necessarily have dedicated resources and, as a result, practices to address these challenges. To assess the current resources and needs to support the data findability efforts at participants’ organisations/RDCs, we have planned two questions for this section of the questionnaire.

**Q15 Is there a person in your RDC/institute who is responsible for improving the findability and/or are there resources for this?**

Multiple choice, optional question, for which the user can choose between (thus, mutually exclusive) the following options:

- Yes
- Employees are encouraged to do what they can on top of their normal responsibilities.
- No

**Q16 Please tick which of the following options describe measures that your RDC/institution undertakes to improve the findability of research data.**

Via this multiple choice, optional question, the user can select all the options that apply, including adding new ones if they see it fit. Thus, the options include:

- Our metadata is open and can be harvested by others.
- We make sure our metadata is rich and obeys the relevant standards.
- We monitor our user engagement, e.g. download statistics.
- We encourage others to link to us.
- We make sure that important keywords are placed prominently on the webpages.
- We conduct user studies to improve our usability.
- Our web pages are optimised for mobile access.
- Other

#### Interested in improving findability

In the final section of the questionnaire, we wanted to identify the participants interested in collaborating with us in assessing and then improving the data findability for their RDCs.

#### **Q17 Are you interested in working with us to improve the findability of the research data of your RDC? How?**

This is a multiple choice, optional question, where more than 1 answer is allowed. The list of options includes:

- We are interested in sharing our ideas and experiences with you.
- We are interested in being a use case partner.
- We want a consultation on what we can do to improve ourselves.
- Please send us some material on how to improve discoverability.
- Please contact us directly.
- We are not interested at this point. We may contact you later.
- Other

## Testimonials

### DZHW

DZHW GmbH | Lange Laube 12 | 30159 Hanover | Germany



Lange Laube 12  
30159 Hanover  
Germany  
 Postfach 29 20 | 30029 Hanover  
Germany  
 phone +49 511 450670-0  
fax +49 511 450670-960  
 [www.dzhw.eu](http://www.dzhw.eu)  
 [twitter.com/dzhw\\_info](https://twitter.com/dzhw_info)  
 [www.linkedin.com/company/dzhw-gmbh](https://www.linkedin.com/company/dzhw-gmbh)

#### Report on findability improvements of the FDZ-DZHW search portal as a result of the consultation by KonsortSWD Task Area 5 Measure 2.

*Andreas Daniel (Deputy Head of FDZ-DZHW and Product Owner of [metadata.fdz.dzhw.eu](https://metadata.fdz.dzhw.eu))*

The Research Data Centre for Higher Education Research and Science Studies (FDZ-DZHW) archives quantitative and qualitative data from the field of higher education research and science studies and makes them available for secondary use. The FDZ-DZHW is based at the German Centre for Higher Education Research and Science Studies (DZHW). The data packages provided by the FDZ-DZHW are documented with rich metadata and accessible via a search portal (<https://metadata.fdz.dzhw.eu>). Each published data package has a DOI that is registered via the da|ra service. The metadata stored at da|ra are harvested by various other portals such as DataCite, BASE, and Gesis Search, so the data packages can be found there as well.

We welcomed the initiative by Measure 2 of KonsortSWD Task Area 5 to provide advice on the current state of our system and how to improve our findability. Their analysis has revealed that although the metadata is well distributed in portals of the scientific community, the general findability of our data via search engines such as Google or Bing was not ideal. Accordingly, several improvements were suggested by the experts, which have now been implemented:

1. With Dublin Core and Schema.org, we have implemented two of the most important metadata standards to improve findability.
2. The research by Brigitte Mathiak and colleagues shows that potential data users often search for terms such as "dataset" and/or acronyms of specific studies. At the same time, standard search engines such as Google primarily index only the titles of data packages. Consequently, we added the term "dataset" to "data package" in the title (screenshot 1) and started to mention the acronyms of studies in the title of data packages, if available (screenshot 2).
3. The experts also emphasized the importance of the links between publications and data packages (see screenshot 3), as potential data users often come across data by reading publications. We have therefore increased our efforts to improve these links (e.g., by introducing a new citation guideline and improving our internal processes for managing related publications).

Hanover, 24.10.2022

#### Contact:

Dr. Andreas Daniel  
phone +49 511 450670-402  
email: [daniel@dzhw.eu](mailto:daniel@dzhw.eu)

Page 1 of 2

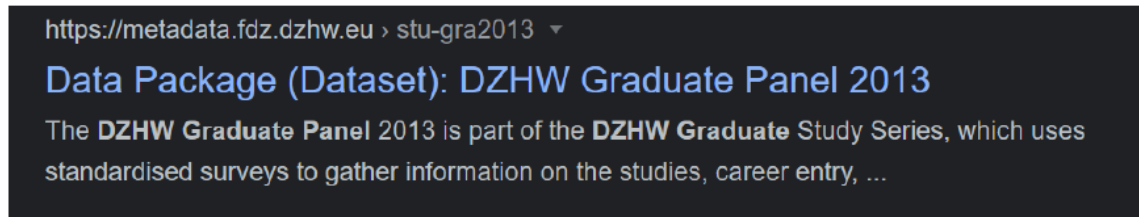
German Centre for Higher Education Research and Science Studies (DZHW)

Chairman of the Supervisory Board: Ministerialdirigent Peter Greisler  
Scientific Director: Prof. Dr. Monika Jungbauer-Gans  
Administrative Director: Dr. habil. Thorsten Kowalke

Bankdetails:  
Commerzbank AG Hannover  
IBAN: DE60 2504 0066 0308 4084 00  
BIC: COBADE33xxx

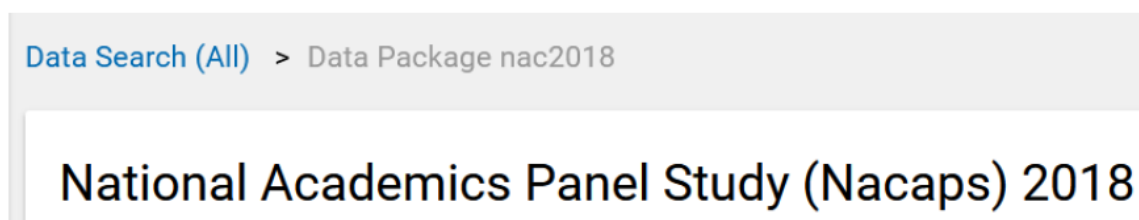
TAX ID No. 25/206/21502  
Registration Court:  
Amtsgericht Hannover | HRB 210251  
VAT No. DE291239300

We would like to thank Brigitte Mathiak and her team of TA5M2 for their targeted support. The discussions with them and their excellent advice have been very helpful for us and have led to significant improvements in our search portal. We look forward to further cooperation in the near future.



https://metadata.fdz.dzhw.eu > stu-gra2013 ▾  
**Data Package (Dataset): DZHW Graduate Panel 2013**  
The **DZHW Graduate Panel 2013** is part of the **DZHW Graduate Study Series**, which uses standardised surveys to gather information on the studies, career entry, ...

Screenshot 1 Term "Dataset" added to Data Package (here as Google search result)



Data Search (All) > Data Package nac2018  
**National Academics Panel Study (Nacaps) 2018**

Screenshot 2 Added study acronyms (here "Nacaps") to titles of data packages



**Related Objects** ↑

Surveys Instruments Questions Data Sets Variables **Publications (9)** Concepts

Q Search (Publications) Search

Sort by Relevance ▾ Items per page 10 ▾ 1 - 9 of 9 < >

**Übergang vom Studium in den Beruf - Wie gestaltet sich der Berufseinstieg von Sozialwi...**  
Lindner, Kristina (2022)  
Lindner, K. (2022). Übergang vom Studium in den Beruf - Wie gestaltet sich der Berufseinstieg von SozialwissenschaftlerInnen und welche spezifischen Bildungsinvestitionen stellen die Grundlage für eine erfolgreiche Karriere-Perspektive dar? [Masterthesis]...

**The role of sex segregation in gender wage gap among university graduates in Germany**  
Ransmayr, Juliane; Weichselbaumer, Doris (2022)  
Ransmayr, J. & Weichselbaumer, D. (2022). The role of sex segregation in gender wage gap among university graduates in Germany. Institute for the Study of Labor, Bonn.

Screenshot 3 Related publications in the search portal of the FDZ-DZHW



DIPF

## Kollegialer Austausch und Beratung zu Search Engine Optimization im Rahmen von KonsortSWD

Im Rahmen des Projekts KonsortSWD nahm der Arbeitsbereich Forschungsdaten Bildung am DIPF, in dem der Verbund Forschungsdaten Bildung (VerbundFDB, [www.forschungsdaten-bildung.de](http://www.forschungsdaten-bildung.de)) und das Forschungsdatenzentrum Bildung (FDZ Bildung, [www.fdz-bildung.de](http://www.fdz-bildung.de)) angesiedelt sind, das Beratungsangebot zu Fragen der Suchmaschinenoptimierung (SEO) durch ein Expertenteam von GESIS wahr. In diesem Kontext fanden, aufbauend auf einer vorab durchgeführten Befragung, zwei Beratungsgespräche statt.

Zielsetzung der Gespräche war es aus Sicht des DIPF, die bisherigen eigenen Aktivitäten im Bereich SEO mit unabhängigen Experten zu validieren, um so eine zusätzliche Rückkoppelung bzw. Ansatzpunkte zu weiteren Optimierungen zu erhalten – und dies von Experten mit starkem Bezug zu Forschungsdateninfrastrukturen.

Schon seit ca. Mai 2017 hat die Abteilung Informationszentrum Bildung am DIPF ihre Aktivitäten im Bereich SEO für die verschiedenen Online-Portale (Deutscher Bildungsserver, Fachportal Pädagogik, Forschungsdaten Bildung) intensiviert und arbeitet dabei mit der auf SEO spezialisierten Agentur get:traction zusammen. Entscheidend für die Einbindung externer Experten war die Tatsache, dass intern keine entsprechenden personellen Ressourcen für das Thema in ausreichendem Maß vorhanden sind und zudem mit einem erhöhten finanziellen Aufwand zu rechnen ist, diese aufzubauen und anschließend dauerhaft das Know-How auf dem State-of-the-Art vorzuhalten. Jedoch sind auch intern stets Ressourcen notwendig, um entsprechend inhaltlich Verantwortliche mit einzubinden und die entsprechenden IT-Anpassungen implementieren zu können.

Der Einstieg lief über ein befristetes Projekt, bei dem nach einer Analyse durch die Agentur verschiedene Maßnahmen für die Seite [www.forschungsdaten-bildung.de](http://www.forschungsdaten-bildung.de) konzipiert und umgesetzt wurden. Diese haben sich relativ zeitnah in einer positiven Entwicklung hinsichtlich der Auffindbarkeit der Inhalte über Google-Suchen niedergeschlagen. Maßnahmen waren vor allem:

- Optimierung der Informationsarchitektur der Website (z.B. stärkere Strukturierung einzelner Pages mit konsistenter Verwendung von Überschriften),
- Optimierung der Angaben im HEAD der einzelnen Seiten (z.B. spezifische Angaben in Title- und Description-Meta-Tag),
- Anwendung von schema.org und semantische Auszeichnung von Inhalten,
- Implementierung eines optimierten URL-Konzepts,
- Optimierung der Indexierung (z.B. Bereitstellung von Sitemaps, Vermeidung von duplicate Content und Definition von canonical links),
- Optimierung in Bezug auf Brand-Anfragen (v.a. Verbund Forschungsdaten Bildung, VerbundFDB) und
- redaktionelle Anpassungen bei Seiten zur Unterstützung des Forschungsdatenmanagements oder auch Suchergebnissen, um Datenbankinhalten (Infos zu Studien und Forschungsdatenbeständen) zugänglich für die Indexierung zu machen, um damit die Auffindbarkeit durch und Sichtbarkeit bei Google zu erhöhen.

Nach dieser initialen Phase läuft im Rahmen der Zusammenarbeit mit der Agentur ein systematisches Monitoring, um auf Fehlentwicklungen reagieren zu können, kombiniert mit einem regelmäßigen Reporting, bei dem Daten der Google Search Console zugrunde gelegt werden (v.a. Beobachtung von Impressions und Klicks, CTR und Position) sowie eine punktuelle Beratung bei aufkommenden Fragen.

Im Austausch mit den SEO-Experten von KonsortSWD wurden die oben genannten Maßnahmen erläutert,

kollegial diskutiert und insgesamt als zielführend eingeschätzt. Die Expertise der Kolleg\*innen aus KonsortSWD hat geholfen, die durchgeführten SEO-Maßnahmen noch einmal mit einem erweiterten multiperspektivischen Blick zu betrachten und zu bewerten.

Darüber hinaus war ein wichtiges Thema, das in den Gesprächen andiskutiert wurde, das Nebeneinander und Zusammenspiel des Meta-Angebots des VerbundFDB, bei dem Datenbestände FDZ-übergreifend sichtbar gemacht werden und spezifische Sichten auf Datenbestände bei den FDZ selbst. Hier konnte die Frage nicht abschließend beantwortet werden, wie SEO für die unterschiedlichen Seiten sinnvoll und für die Nutzenden zielführend angegangen werden kann, so dass es als sinnvoll angesehen wird hierzu weiterhin im Austausch zu bleiben.

**Contact:**

**Dr. Brigitte Mathiak**

GESIS – Leibniz Institute for the Social Sciences  
Knowledge Technologies for the Social Sciences – KTS  
Unter Sachsenhausen 6–8, D–50667 Cologne

[www.gesis.org](http://www.gesis.org)

[brigitte.mathiak@gesis.org](mailto:brigitte.mathiak@gesis.org)

Tel.: +49 221 47694-510

KonsortSWD is funded within the framework of the NFDI by the German Research Foundation (DFG) – project number: 442494171.



Diese Veröffentlichung ist unter der Creative-Commons-Lizenz (CC BY 4.0) lizenziert:

<https://creativecommons.org/licenses/by/4.0/>

doi: 10.5281/zenodo.8289917